Chapter 2

Blind Signal Separation

Blind Signal Separation (BSS) is a now widely used technique for the identification and separation of mixed signals. It is called blind because usually we don't have much information about the forms in which these signals are actually mixed, although we usually might identify and interpret the different signals if they could be separated.

In this chapter we present two techniques that will be used in the present work: Principal Component Analysis, which is a way to decorrelate linearly a set of random variables and Independent Component Analysis, which is based in the assumption that the signals that where mixed are not only uncorrelated, but also independent. As we will see, both techniques have a wide range of application in time-series data, but the latter can be viewed as an extension of the former.

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a classical multivariate analysis method that has two basic ideas. The first one is to find a linear transformation $\mathbf{V}$ that transform a given zero-mean $N-$dimensional random vector[1] $\vec{X}$, which may have correlated random variables, into a new $N-$dimensional random vector $\mathbf{V}\vec{X} = \vec{Z}$ that has uncorrelated random variables. On the other hand, the second idea is to maximize the variance of each linear transformation of the elements of $\vec{X}$, i.e.,

[1]This asumption is made in order to simplify the notation. However, in practice we can always take a sample of this random vector and subtract the empirical mean in order to simplify the problem.

maximize the variance of each element of the random vector $\vec{Z}$, which are called the principal components. As we will see, the linear transformation that projects the random vector $\vec{X}$ to an uncorrelated random vector $\vec{Z}$ is not unique, and we'll search for the optimal one in the context of time series analysis.

2.1.1 A derivation of the principal components

The question now is: how do we find this transformation? The classic derivation (Jolliffe, 2002) uses the fact that we want to first maximize the variance of each linear combination of $\vec{X}$, i.e., maximize the variance of the principal components, obtaining the desired transformation componentwise. The idea is that the desired transformation matrix, $\mathbf{V}$, contains the coefficients of these linear transformations. Let's denote the elements of this matrix by $v_{i,j}$. Then, the i-th principal component (the i-th element of the vector $\vec{Z}$) is given by

$$Z_i = v_{1,i}X_1 + v_{2,i}X_2 + ... + v_{N,i}X_N = \vec{v}_i^T\vec{X},$$

where the elements of the vectors $\vec{v}_i$ are the coefficients of the linear combination of the elements of $\vec{X}$ corresponding to this i-th principal component. Our first task is to maximize the variance of the first principal component, $E\left[Z_1^2\right] = \vec{v}_1^T\mathbf{\Sigma}_X\vec{v}_1$. Note that, however, we need to constrain the values of the vectors $\vec{v}_i$ in order to do this. The constraint that we'll impose is $\vec{v}_i^T\vec{v}_i = 1$. In summary, the problem is stated as follows: maximize the function $f(\vec{v}_1) = \vec{v}_1^T\mathbf{\Sigma}_X\vec{v}_1$ given the constraint $g(\vec{v}_1) = \vec{v}_i^T\vec{v}_i = 1$. Using the method of Lagrange multipliers (and remembering that the covariance matrix is symmetric), we have

$$\vec{\nabla}_1 f = 2\mathbf{\Sigma}_X\vec{v}_1 = \lambda_1\vec{\nabla}_1 g = 2\lambda\vec{v}_1 \Longrightarrow \mathbf{\Sigma}_X\vec{v}_1 = \lambda\vec{v}_1,$$

where the operator $\vec{\nabla}_1$ represents the gradient with respect to the elements of $\vec{v}_1$. Here we see that $\vec{v}_1$ is an eigenvector of the covariance matrix of $\vec{X}$, where the corresponding eigenvalue is the Lagrange multiplier, λ_1. To obtain the second principal component, we repeat the maximization problem that we made for the first one but now we add one more constraint: we want Z_1 and Z_2 to be uncorrelated, i.e. $h(\vec{v}_1, \vec{v}_2) = \mathrm{Cov}(Z_1, Z_2) = E[Z_1^T Z_2] = \vec{v}_1^T\vec{v}_2\lambda_1 = 0$. In other words, the vectors of coefficients are orthogonal. Using again the method of Lagrange multipliers, but now with two constraints (normality of $\vec{v}_1$ and orthogonality between $\vec{v}_1$ and $\vec{v}_2$) and,

therefore, two multipliers λ_2 and λ_3 we have

$$\vec{\nabla}_2 f = 2\mathbf{\Sigma}_X \vec{v}_2 = \lambda_2 \vec{\nabla}_2 g + \lambda_3 \vec{\nabla}_2 h = 2\lambda_2 \vec{v}_2 + \lambda_3 \vec{v}_1.$$

Taking the dot product with respect to $\vec{v}_1^T$ from the left, we can see that $\lambda_3 = 0$. This implies that

$$\mathbf{\Sigma}_X \vec{v}_2 = \lambda_2 \vec{v}_2,$$

and, again, $\vec{v}_2$ is an eigenvector of $\mathbf{\Sigma}_X$, where λ_2 is the corresponding eigenvalue. We can repeat this process N times to find that the i-th coefficient vector is given by the i-th eigenvector of the covariance matrix $\mathbf{\Sigma}_X$. Because of this, one way to obtain the desired linear transformation $\mathbf{V}$ that takes the vector $\vec{X}$ and transforms it to the new random vector $\vec{Z}$ of uncorrelated random variables can be obtained by finding the eigenvectors of the covariance matrix of $\vec{X}$, and letting each eigenvector be a row of $\mathbf{V}$ (note that, because of this, $\mathbf{V}$ is orthogonal). This can be efficiently done for any sample covariance matrix via a Singular Value Decomposition (SVD) algorithm, which, for our case (real signals) will lead to the SVD decomposition $\mathbf{\Sigma}_X = \mathbf{EDE}^T$, where $\mathbf{E}$ is a matrix that contains the eigenvectors in the columns and $\mathbf{D}$ is a diagonal matrix with the corresponding eigenvalues. Note that the transformation that we found, $\mathbf{V} = \mathbf{E}^T$, gives the following covariance matrix for $\vec{Z}$:

$$E\left[\vec{Z}\vec{Z}^T\right] = \mathbf{E}^T E\left[\vec{X}\vec{X}^T\right]\mathbf{E} = \mathbf{E}^T\mathbf{EDE}^T\mathbf{E} = \mathbf{D}. \tag{2.1}$$

2.1.2 Interpretation of the Principal Components

Perhaps the most important feature of the principal components is that their respective eigenvalues give information about which principal component has the largest variance. The higher the eigenvalue, the higher the variance. This can be observed from equation (2.1), where for a given principal component, the variance is given by

$$E\left[Z_i Z_i^T\right] = E\left[\vec{v}_i^T \vec{X}\vec{X}^T \vec{v}_i\right] = \vec{v}_i^T \mathbf{\Sigma_X} \vec{v}_i = \lambda_i.$$

The importance of this is given because the direction of the vector of coefficients that define each principal component, $\vec{v}_i$, define directions of maximum dispersion.

Consider, for example, the random variables $X \sim N(0,1)$, $Z \sim N(0,1/2)$ and $Y = X + Z$. For the random vector $\vec{X} = (X,Y)^T$, it is straightforward to check

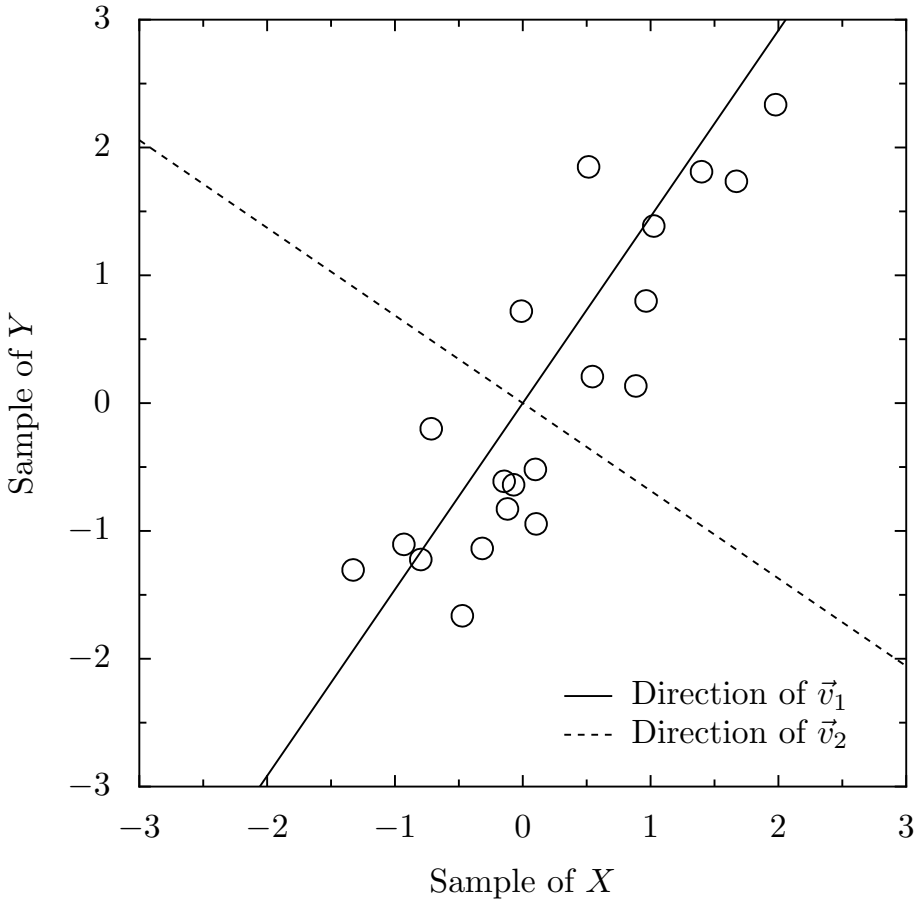

Figure 2.1: Samples of the random variables X and Y in the example.

that the eigenvalues of the covariance matrix are $\lambda_1 = 2.3$ and $\lambda_2 = 0.2$, where the corresponding eigenvectors are $\vec{v}_1 = (-0.6, -0.8)^T$ and $\vec{v}_2 = (-0.8, 0.6)^T$. 20 samples of the random variables X and Y where taken and the values obtained are shown in Figure 2.1. The direction of the eigenvectors is also plotted.

As we derived, the direction of $\vec{v}_1$ has maximum dispersion. However, note that the direction of the eigenvectors seem to form the "principal axes of the data", which is actually a property of the coefficient vectors. Jolliffe (2002) sumarizes a large number of properties and ways in which the principal components and their corresponding vectors of coefficients (the eigenvectors of the covariance matrix) can be interpreted. Perhaps the most important property is dimensionality reduction: the fact that the N principal components, Z_i, can be reduced to $q < N$ principal components in order to minimize the sum of the squared perpendicular distances of

the samples measured from this subspace (the sum of the squared perpendicular distance from the lines formed by the eigenvectors in Figure 2.1 to the samples). This is extremely useful when we are dealing with high dimensional random vectors: it means that we can apply our derived linear transformation to the random vector $\vec{X}$, obtain the principal components, select the ones that explain most of the variation on our data and analyse that sub-set of the data with minimum loss of information.

How can we interpret all this in the time series context? As we showed in Chapter 1, time series analysis is way more complex than just talking about random variables, because a process is a collection of random variables and, therefore, the probabilistic nature of it is changing with time. Unfortunately, PCA can't take into account this fact because we don't have enough information about the different random variables at each time index. This may seem rather dissapointing but, as will be shown, it is a good starting point in the analysis of time series.

The idea in applying PCA to indexed series is that, in practice, we may collapse a given process into a single random variable and at the same time see this random variable as a sum of other different random variables (which in the context of time series were also different processes). For example, the logarithm of the flux of a star as measured from an instrument may be thought as a random variable which is the sum of different atmospheric and instrumental effects (thinking that those effects multiplicatively modulate the emmited flux from the star). This makes sense if we think in the distributions of these random variables: different processes may produce different distributions when collapsed in a single random variable. To illustrate this concept, consider the measurement of the logarithm base 10 of the flux of a star as a function of time, $z_1(t)$, plotted in Figure 2.2. Here, the different values can be thought as being realizations of different random variables $Z_1(t)$. On the other hand, observing the frequency distribution of the star's flux, it can also be thought as realizations of a single random variable. Therefore, this frequency distribution can also be thought as our measurement of the distribution of the collapsed process, thinking of it as the measurement of the distribution of a single random variable.

A related subject of this kind of analysis is Functional Data Analysis, a concept that was introduced by Ramsay and Silverman (1997), which uses PCA in the context of continuous deterministic series, i.e., they assume that the data is a sample

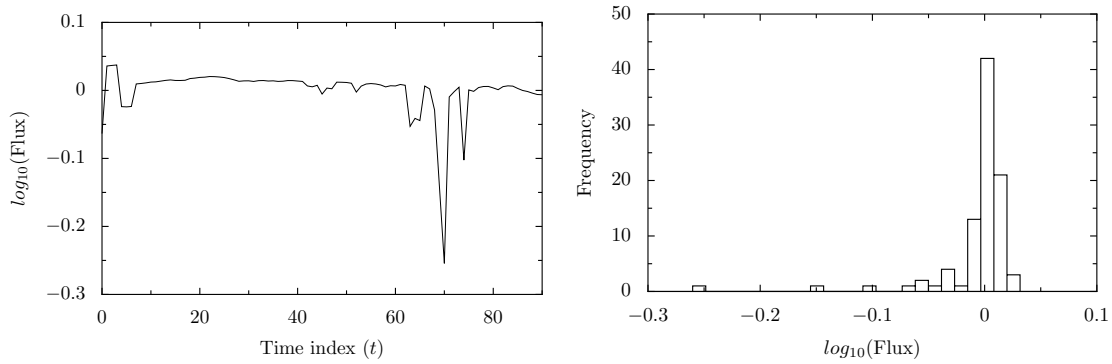

Figure 2.2: (Left) Light curve of a star (logarithm in base 10 of the measured stellar flux). (Right) Frequency distribution of the values of the logarithm's stellar flux.

of some continuous function of some index (e.g. time). In this context, they argue that if we take independent measures of some variables (e.g. curves of human growth, econometric time series, etc.) what PCA actually does is to decompose each observation into a suitable orthonormal basis (the coefficients) whose (uncorrelated) coefficients are the principal components. In this sense, the principal components show special features of the data, which has very good results in a wide variety of areas (Ramsay and Silverman, 2002).

To see how this applies to our collapsed processes, consider the random vector $\vec{X} = (X_1, X_2, ..., X_N)^T$, where this time the random variables may represent the different outcomes of N stochastic processes $X_i(t)$, $i = 1, 2, ..., N$, collapsed in them, e.g., the logarithm of the fluxes of different stars measured from an instrument (where now the PDF of each random variable has to be thought of as "the probability density of obtaning a given value for the flux"). Applying the linear transformation $\mathbf{V}$, recall that the i-th principal component is given by

$$Z_i = \sum_{j=1}^{N} v_{i,j} X_j.$$

Because the linear transformation $\mathbf{V}$ is orthonormal, $\mathbf{V}^{-1} = \mathbf{V}^T$ and the i-th random variable can be written as

$$X_i = \sum_{j=1}^{N} v_{j,i} Z_j.$$

This is almost what we where searching for! The above expresion can be thought as an expansion of the random variable X_i in terms of an orthonormal basis (the vectors of coefficients) and coefficients given by another (uncorrelated) random variable (the principal components). In the limit $N \to \infty$, this is known as the Karhunen-Loève theorem or expansion, which states that a random variable can be represented as an infinite linear combination of orthogonoal functions, whose coefficients are uncorrelated random variables. Because in our case we have a limited set of samples for each random variable, PCA may be seen as a truncated form of this expansion, which is known in the signal proccesing jargon as the Karhunen-Loève transform. Note that this transform is somewhat different from the usual transforms: here **the coefficients are the random variables** and the deterministic vectors contained in the linear transformation $\mathbf{V}$ are the functions.

In practice what we actually have are samples of the random variables, $x_i(t)$ (the different time series for each star), which are collected in order to create a data matrix, $\mathbf{X}$, where each row represents a different star and each column is a time index. Then, after subtracting the mean from each row we create the **sample covariance matrix** where the element (i, j) of that matrix is

$$\hat{\mathbf{\Sigma}}_X(i,j) = \frac{1}{M}\sum_{t=1}^{M} x_i(t)x_j(t),$$

where $x_i(t)$ is the i-th star's time series and M is the number of samples. Once we obtain the sample covariance matrix $\hat{\mathbf{\Sigma}}_X$, we obtain its eigenvalues and eigenvectors and obtain the linear transformation $\mathbf{V}$. Finally, we apply this transformation to our data matrix $\mathbf{X}$ to obtain

$$\mathbf{VX} = \mathbf{Z},$$

where the i-th row of the matrix $\mathbf{Z}$ is the corresponding time series for the i-th principal component. In summary, what this really means is that we can find projections of the samples where the resulting time series are uncorrelated from each other.

The interesting interpretation about PCA in the time series context is, then, that the principal components define a set of uncorrelated signals that best explain our data when properly weighted by the coefficient vector. Furthermore, the eigenvalues asociated with each principal component are a measure of "how important" is a given

principal component time series in order to explain our observed data. However, note that the obtained expansion is given in terms of uncorrelated random variables... is it possible to make an analogous expansion using **independent random variables** (which was what we wanted in the first place)? This question will be answered in the next section.

2.1.3 PCA and Whitening

We end this section with a discussion in a subject that will be of fundamental importance on the next section, which is the "whitening" of a random vector. As was stated in equation 2.1, given our linear transformation $\mathbf{V}$, the new random vector $\vec{Z}$ is uncorrelated but it is not white (i.e. their random variables have different variances). This is desirable because, as we'll see in future sections, it simplifies a lot of calculations and interpretations of the random variables. In this sense, a white random vector is "better" than an uncorrelated random vector.

It easy to show that a transformation that can project the initial random vector $\vec{X}$ to a white random vector $\vec{Z}$ is given by $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T$, because

$$E\left[\vec{Z}\vec{Z}^T\right] = \mathbf{D}^{-1/2}\mathbf{E}^T E\left[\vec{X}\vec{X}^T\right]\mathbf{E}\mathbf{D}^{-1/2} = \mathbf{I}.$$

This is called a **whitening transform**, for obvious reasons. It is interesting to note, however, that in fact any transformation of the form $\mathbf{V} = \mathbf{P}\mathbf{D}^{-1/2}\mathbf{E}^T$, where $\mathbf{P}$ is an orthogonal matrix will make a whitening transform. In order to simplify the notation, we'll use the transform $\mathbf{\Sigma}_X^{-1/2} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T$, which is widely used in the signal processing literature.

2.2 Independent Component Analysis

In the past section we saw that PCA is a powerful analysis tool, because it can decompose a given random variable into a series of uncorrelated random variables, properly weighted by an orthonormal basis. However, recall that, as we showed on Chapter 1, uncorrelated does not mean independent, so the components obtained with PCA may still be dependent of each other. With this in mind we posed the following question: if there exists a decomposition of a random variable in terms