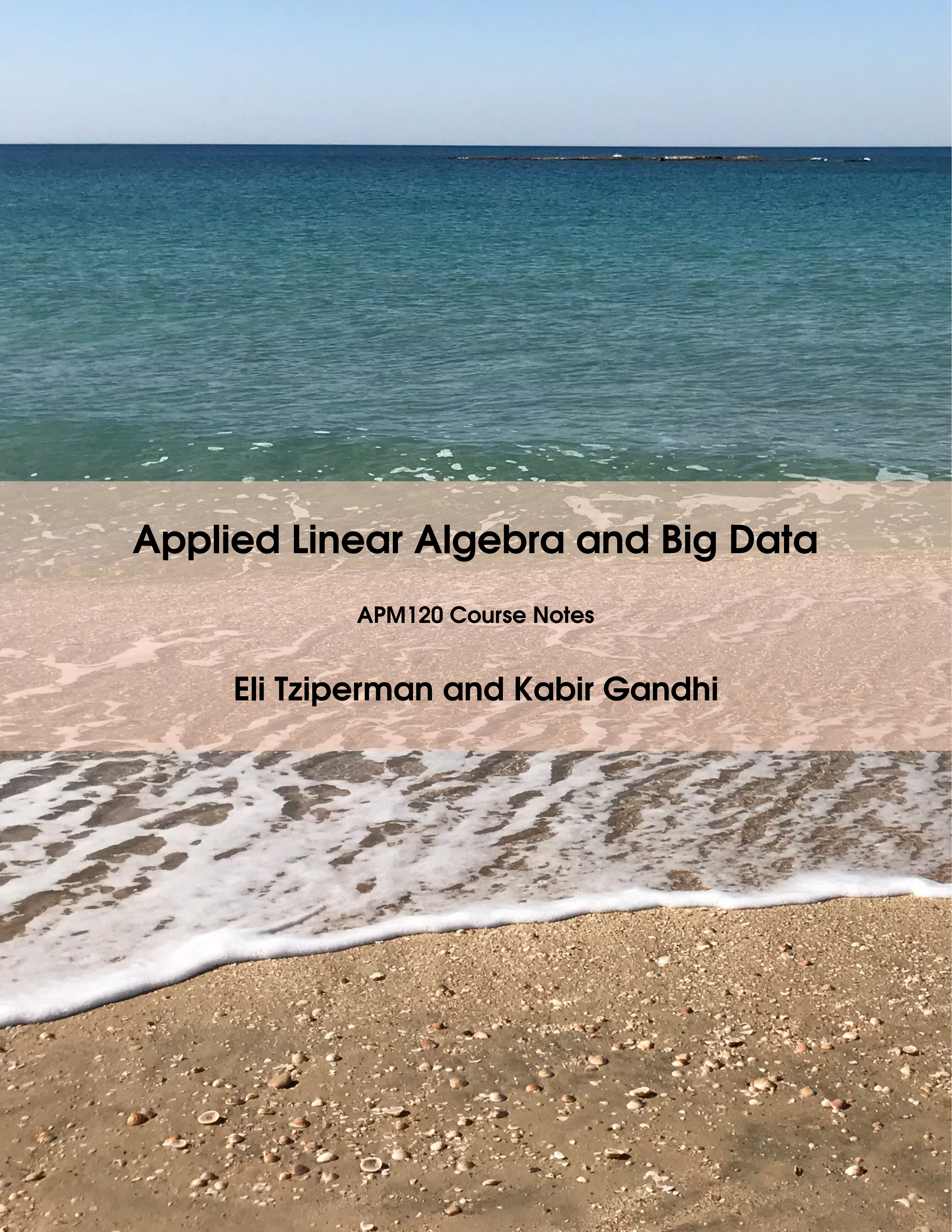
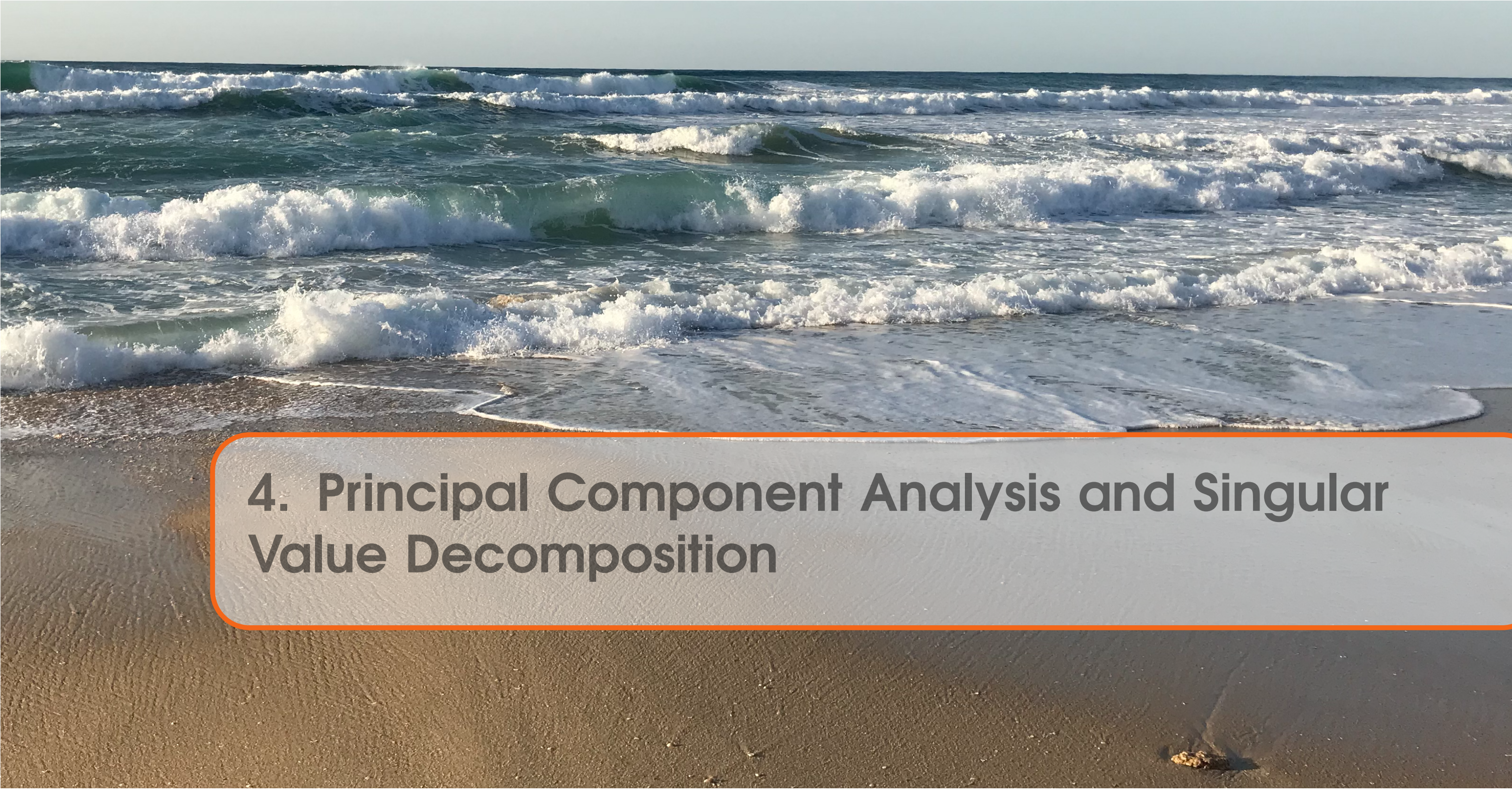


Applied Linear Algebra and Big Data

APM120 Course Notes

Eli Tziperman and Kabir Gandhi



4. Principal Component Analysis and Singular Value Decomposition

4.1 Principal Component Analysis (PCA) from the covariance matrix

4.1.1 Motivation

Principle Component Analysis, also known as “Factor Analysis” or “Empirical Orthogonal Functions”, is a powerful yet simple technique for uncovering relationships within data, with applications in finance, weather and climate, social sciences and more. Some interesting applications include,

Quantitative finance: Consider a vector containing stock prices, where the vector is given daily, for several years, and can therefore be represented as a data matrix whose columns are the daily data. We wish to find out which stock prices vary together, and which stocks vary opposite to, or independently of, one another. PCA allows us to call upon the covariance matrix to more clearly understand these kinds of relationships. See data demo PCA_small_data_example_using_covariance.m/py.

Weather and climate systems: See the animation of sea surface temperature (SST) during El Niño events in slides 1-4. More generally, we can use PCA to look for trends and correlations between different weather variables.

Artificial 2d data example: example_2d_PCA.m/py, shows animation of data. In this demo we are looking for an “efficient representation” without knowing how data were actually constructed.

4.1.2 Derivation

Consider N vectors of size $M \times 1$, $\mathbf{f}_n = f_{mn}$. Each column vector could represent, for example, data at a given time, and the different components of $\mathbf{f}_n$ could correspond to, for example, either data – such as temperature – at different locations, or, say, prices of stocks of different companies at this time. Let $\mathsf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_N) = f_{mn}$ be the $M \times N$ data matrix containing the entire data set. While we are using a terminology that is based on the subscript representing time, this is not necessarily the case. For example, the elements of $\mathbf{f}_n$ could be the number of high school students taking courses in social sciences, biology, literature, math and physics (that is, $M = 5$), and the subscript n could represent different schools.

The first step is to remove the mean from each row of the data matrix, defining a mean-less primed data matrix,

$$f'_{mn} = f_{mn} - \frac{1}{N}\sum_{i=1}^{N} f_{mi} \tag{4.1}$$

and from now on we are going to drop the prime and assume the mean is zero.

Our objective is to find M orthogonal vectors $\mathbf{u}_j$ of dimension $M \times 1$ (these are the principal components) that best describe the variability in the data. By this we mean that we look for $\mathbf{u}_1$ of magnitude one such that $\sum_{n=1}^{N}(\mathbf{u}_1 \cdot \mathbf{f}_n)^2$ is maximal. Maximizing this projection means that $\mathbf{u}_1$ is similar, up to a minus sign, to the typical patterns of the data vectors $\mathbf{f}_n$. Next, we look for $\mathbf{u}_2$ of magnitude one such that $\sum_{n=1}^{N}(\mathbf{u}_2 \cdot \mathbf{f}_n)^2$ is maximal again, and is also orthogonal to the first $\mathbf{u}_1$, i.e. $\mathbf{u}_1 \cdot \mathbf{u}_2 = 0$. Then, in general, we wish $\mathbf{u}_j$ to be of magnitude one and to maximize $\sum_{n=1}^{N}(\mathbf{u}_j \cdot \mathbf{f}_n)^2$ while being orthogonal to all previous vectors. Define a matrix U whose columns are the vectors $\mathbf{u}_j$, and use the data matrix notation to rewrite the sum to be maximized as,

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_j \cdot \mathbf{f}_n)^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_j^T \mathbf{f}_n)^2 &= \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{m=1}^{M} U_{mj}F_{mn}\right)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{m=1}^{M} U_{mj}F_{mn}\right)\left(\sum_{k=1}^{M} U_{kj}F_{kn}\right) \\
&= \sum_{m=1}^{M}\sum_{k=1}^{M} U_{mj}\left(\frac{1}{N}\sum_{n=1}^{N} F_{mn}F_{kn}\right)U_{kj} \\
&= \mathbf{u}_j^T\left(\frac{1}{N}FF^T\right)\mathbf{u}_j.
\end{aligned}$$

The matrix that appears here,

$$\mathsf{C}_{M\times M} \equiv \frac{1}{N}\mathsf{F}_{M\times N}\mathsf{F}^T_{N\times M}$$

is the covariance matrix, whose element $c_{ij} = \frac{1}{N}\sum_{n=1}^{N} f_{in}f_{jn}$ is the time averaged product (covariance) of data elements i and j.

We next show that the covariance matrix is positive semi-definite,

$$\begin{aligned}
\mathbf{x}^T\mathsf{C}\,\mathbf{x} = \sum_{ij} x_i C_{ij} x_j &= \frac{1}{N}\sum_{ij} x_i\left(\sum_n F_{in}F_{jn}\right)x_j \\
&= \frac{1}{N}\sum_n\left(\sum_i x_i F_{in}\right)\left(\sum_j F_{jn}x_j\right) \\
&= \frac{1}{N}\sum_n\left(\sum_i x_i F_{in}\right)^2 \geq 0.
\end{aligned}$$

and this implies the eigenvalues are non-negative.

The constrained optimization problem of maximizing $\sum_{n=1}^{N}(\mathbf{u}_j \cdot \mathbf{f}_n)^2$ and requiring $\mathbf{u}_j$ to be of unit magnitude is solved using Lagrange multipliers by maximizing,

$$\mathbf{u}_j^T \mathsf{C} \mathbf{u}_j + \lambda(1 - \mathbf{u}_j^T \mathbf{u}_j).$$

By requiring the derivative with respect to the elements of $\mathbf{u}_j$ to vanish, we find that the optimal vectors are eigenvectors of the covariance matrix,

$$\mathsf{C}\mathbf{u}_j = \lambda \mathbf{u}_j.$$

Because the covariance matrix is symmetric, the eigenvectors $\mathbf{u}_j$ are orthogonal, as we required above. The projection of the eigenvector $\mathbf{u}_j$ on the data – the quantity being maximized – is equal to the corresponding eigenvalue,

$$\mathbf{u}_j^T \mathsf{C} \mathbf{u}_j = \mathbf{u}_j^T \lambda_j \mathbf{u}_j = \lambda_j,$$

and therefore the maximal projection occurs for the eigenvectors corresponding to the largest eigenvalues. This also provides some intuition for the discussion of the variance explained by each PC in section 4.1.4 below.

We can project the data at a given time, $\mathbf{f}_n$, on a principal component $\mathbf{u}_j$, to obtain the amplitude for this principal component at that time,

$$t_{jn} = \mathbf{f}_n \cdot \mathbf{u}_j.$$

This corresponds to a time series $\{t_{j1}, \ldots, t_{jN}\}$ which represents the amplitude of the principal component $\mathbf{u}_j$ as function of time. Time series are more generally referred to as the principal component “expansion coefficients” or “scores”, in particular in applications where the different data vectors do not represent different times as mentioned above. There are M such time series, and we can put them as the rows of an $M \times N$ matrix as $\mathsf{T} = (t_{jn})$. Remembering that we defined the M principal components to be the columns of the $M \times M$ matrix $\mathsf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_M)$, we can write the equation used to derive the time series matrix as

$$\mathsf{T} = \mathsf{U}^T \mathsf{F}.$$

Multiplying by the orthogonal matrix of principal components U, which is also the inverse of U^T, we find,

$$\mathsf{F}_{M \times N} = \mathsf{U}_{M \times M} \mathsf{T}_{M \times N}.$$

This equation corresponds to an expansion of the data in terms of the eigenmodes (principal components). Writing it for one time,

$$\mathbf{f}_n = \sum_{j=1}^{M} \mathbf{u}_j t_{jn}.$$

Remember that the PCs, calculated as eigenvectors, are defined up to a minus sign. This last expression also shows that if we multiply one of the principal components by a minus sign, the corresponding time series will also be multiplied by a minus sign. Thus the interpretation of the time series should take into account the sign of the PC.

If some of the PCs are judged not important, for example, if we decide that they represent only noise in the data, we may therefore wish to reconstruct the data using only $k < M$ PCs. Then,

$$\begin{aligned}
\mathsf{F}_{\text{reconstructed}} &= \sum_{j=1}^{k} \mathbf{u}_j t_{jn} \\
&= \mathsf{U}_{M\times k}\mathsf{T}_{k\times N} \\
&= \mathtt{U(:,1:k)*T(1:k,:)}
\end{aligned}$$

4.1.3 Example of PCA

Consider the following data set that represents the deviations from the long-term mean of the CO_2 concentration in the atmosphere (first line), the global mean surface temperature (second) and the area of summer sea ice in the Arctic (third) over 9 consecutive years,

$$\mathsf{F} = \begin{pmatrix} 0 & -1.75 & 0.25 & -1.0 & 0.5 & -0.25 & 0.75 & 0.5 & 1.0 \\ -3.0 & -0.5 & -1.75 & 0.25 & -0.5 & 1.0 & 0.75 & 1.75 & 2.0 \\ 0 & 1.75 & -0.25 & 1.0 & -0.5 & 0.25 & -0.75 & -0.5 & -1.0 \end{pmatrix},$$

where each column represents a year.

The PCA steps are as follows,

1. Remove the mean from each row of the data matrix as in (4.1).
2. Calculate the covariance matrix by calculating $\mathsf{F}\mathsf{F}^T/N$, where $N = 9$ is the number of years.
3. Calculate the eigenvectors of the covariance matrix and normalize them; these are the principle components. Place the eigenvectors as columns of a matrix U, where each column is a principal component of the data matrix.
4. Calculate the time series, $\mathsf{T} = \mathsf{U}^T\mathsf{F}$, where F is the original dataset and U is the matrix of eigenvectors. If the data matrix F has dimensions $M \times N$, the time series T should be the result of multiplying matrices that are $M \times M$ and $M \times N$ and so should also be of dimensions $M \times N$.

The covariance matrix is,

$$\mathsf{F}\mathsf{F}^T/N = \mathsf{C} = \begin{pmatrix} 0.6944 & 0.3472 & -0.6944 \\ 0.3472 & 2.361 & -0.3472 \\ -0.6944 & -0.3472 & 0.6944 \end{pmatrix}.$$

We see that the diagonal terms are the variance of the three variables, that is, the average of $(CO_2)^2$, of the temperature squared and of the ice area squared (because each variable (row) in the data matrix already has a zero mean). The off-diagonal terms are the covariance of the different variables. The values indicate that the CO_2 and global temperature vary in the same direction, hence positively correlated (entry c_{12}), while they are both negatively correlated with sea ice (c_{13}, c_{23}). That is, CO_2 and temperature tend to increase/decrease together, while sea ice vary in the opposite direction.

The eigenvalues and eigenvectors (PCs) of the covariance matrix are calculated, and sorted by the value of the eigenvalues from large to small,

$$\mathsf{D} = \begin{pmatrix} 2.57 & 0 & 0 \\ 0 & 1.18 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathsf{U} = \begin{pmatrix} 0.272 & 0.653 & 0.707 \\ 0.923 & -0.385 & 0 \\ -0.272 & -0.653 & 0.707 \end{pmatrix}.$$

The PCs $\mathbf{u}_i$ are orthogonal vectors whose structure best represents the nine data vectors. In other words, $\mathbf{u}_1$ is a vector that is as parallel as possible to the data vectors. $\mathbf{u}_2$ is as parallel as possible to the data vectors while also being perpendicular to $\mathbf{u}_1$. Because the data space is three-dimensional, the last eigenvector $\mathbf{u}_3$ is actually completely determined by the requirement that it is perpendicular to the first two PCs. The principal components corresponding to the largest eigenvalues represent the main modes in the data.

The eigenvector corresponding to the largest eigenvalue is the first PC
$\mathbf{u}_1 = (0.272, 0.923, -0.272)^T$. From a climate point of view, this PC represents a variability mode in which when the CO_2 and global temperature both increase, the sea ice decreases, and vice versa.

Furthermore, we may be interested in calculating the time series, T, for this data, which we can find by multiplying the transpose of the matrix of principle components, U, by the original data set, F. We calculate the time series to be,

$$\mathsf{T} = \begin{pmatrix} -2.77 & -1.41 & -1.48 & -0.314 & -0.189 & 0.787 & 1.1 & 1.89 & 2.39 \\ 1.16 & -2.09 & 1.0 & -1.4 & 0.845 & -0.711 & 0.69 & -0.0212 & 0.535 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The time series of the first and most dominant PC, $\mathbf{u}_1$, shows a continuous increase over time. Given that, from our interpretation of the principle components, CO_2 concentration and temperature vary in the opposite direction of sea ice coverage, we can conclude from the time series that temperature and CO_2 concentration increase with time, while sea ice coverage decreases. The time series of the second mode, shows alternating signs, indicating short-term variability superimposed on the slower climate change represented by the largest mode. The PC corresponding to the zero eigenvalue, is positive for CO_2 and sea ice and zero for temperature. It explains zero percent of the variance (which will be discussed further in the next section), presumably because even during climate variability in this data set, the CO_2 and sea ice never vary together while leaving the temperature unchanged. Thus, this principal component does not represent a physical mode.

Two eigenvalues dominate since one is zero (yielding an empty bottom row in the time series), and as a result the data can effectively be reconstructed using only two PCs. Let the reconstructed data be $\mathsf{F}_{recontructed}$, we can calculate $\mathsf{F}_{recontructed} =$ `U(:,1:2)*T(1:2,:)`. Indeed, since one of the eigenvalues is zero (rather than, say, much smaller than the other two eigenvalues), the reconstructed data set is identical to the original.

4.1.4 Fraction of variance explained by PC modes

We found so far that the eigenvectors of the covariance matrix are the principal components that best describe the variability in the data, and that the projection of these modes on the data is given by the time series. We consider further the role of the eigenvalues of the covariance matrix. First, introduce the concept of “total variance” using a simple example. If the data vectors are given by $\mathbf{f}_n = (x_n, y_n, z_n)^T$, using the expression for the variance of x_n, $var(x) = \sum_{n=1}^{N} x_n^2/N$, and similarly for y_n and z_n, we define the total variance of these data as $var(x_n) + var(y_n) + var(z_n)$. In terms of the data matrix, the variance of the ith variable in the data vectors (the ith row of F) is,

$$\frac{1}{N}\sum_{n=1}^{N}(f_{in})^2,$$

and the total variance is the sum over all of these,

$$\text{total variance} = \sum_{i=1}^{M}\left(\frac{1}{N}\sum_{n=1}^{N}(f_{in})^2\right). \tag{4.2}$$

Next, use the fact that this total variance is exactly the trace (sum of diagonal elements) of the covariance matrix, which is also equal to the sum of eigenvalues, $\text{trace}(\mathsf{C}) = \sum_{j=1}^{M}\lambda_j$. As a result, the fraction of the variance explained by the ith PC, $\mathbf{u}_i$ is,

$$\frac{\lambda_i}{\sum_{j=1}^{M}\lambda_j}. \tag{4.3}$$

We can consider another way to understand the issue of "fraction of total variance explained". Let $\mathbf{u}_1$ be the first PC, corresponding to the largest eigenvalue of C. Given also the corresponding time series, we can create a partial reconstruction of the data using this first PC only,

$$\mathsf{F}^{PC1} = \mathbf{u}_1\mathbf{t}_1 = \mathtt{U(:,1)*T(1,:)} \quad (= M\times N).$$

Now calculate the total variance of this reconstructed data using (4.2). The ratio of this total variance to that calculated using the full rather than reconstructed data, is the fraction of the total variance explained by the first PC, and should be equal to the first eigenvalue divided by the sum of all eigenvalues, as derived above.

4.1.5 PCA from covariance matrix in Matlab

```
%% calculate covariance matrix:
C=F*F'/N;
%% calculate PCs (matrix U) and eigenvalues:
[U,D]=eig(C);
%% calculate time expansion coefficients:
T=U'*F;
%% fraction of variance explained by each PC:
trace_C=trace(C);
fraction_variance=diag(D)/trace_C;
%% reconstruct data using only k PCs (assuming they are sorted!):
F_reconstructed=U(:,1:k)*T(1:k,:);
```

4.1.6 Examples and additional issues

Examples, activities and demos,

1. Complete the demo of PCA analysis of time series of four stocks PCA_small_data_example_using_covariance.m/py
2. Consider the idealized 1d example, demonstrating the meaning of the covariance matrix, and the calculation of PCs for both $\cos(kx)\cos(\omega t)$ and for random data: one_dim_covariance_matrix_and_PCA_tutorial.m/py
3. Idealized 2d example: which shows the calculated PC structure and time series: example_2_PCA.m/py
4. Application: El Nino. PCA_Equatorial_SST_example_using_covariance.m/py
5. Global SST: 1st PC=global warming, 2nd=ENSO: Compo and Sardeshmukh (2010), Fig 10, p 1967.