Covariance Matrix Example

In this exercise you will get a hands-on experience with the key theoretical concepts behind Principal Component Analysis. You will work with the following 2-dimensional and 3-dimensional data sets:

- 2-dimensional: `zeroVarUnCov.dat`, `smallVarUnCov.dat`, `smallVarPosCov.dat`, `smallVarNegCov.dat`, `smallerVarUnCov.dat`, `smallerVarPosCov.dat`, `smallerVarNegCov.dat`
- 3-dimensional: `Var3dCov.dat`

Each `.dat` file provides a d-dimensional data set in the form of $d \times N$ matrix $\mathbf{D} = (\mathbf{x}^1, \ldots, \mathbf{x}^N)$ that stores $N = 500$ d-dimensional data, $d = 2, 3$, as columns. The data set is obtained by repeated independent draws from a vector random variable $\mathbf{X} = (X_1, \ldots, X_d)^T$. Using the data set we can calculate an estimation $\mathbf{C}$ of the covariance matrix $Cov[\mathbf{X}]$. In particular if all the random variables are centered, i.e. $\mathbb{E}[X_i] = 0$, $i = 1, \ldots, d$, the estimation can be given as

$$Cov[\mathbf{X}] \approx \mathbf{C} = \frac{1}{N}\mathbf{D}\mathbf{D}^T.$$

We simply call the estimation $\mathbf{C}$ “covariance matrix” below.

The eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ of $\mathbf{C}$ specify the new axes. We can project the data set $\mathbf{D}$ onto the new axes as follows:

$$\tilde{\mathbf{D}} = \mathbf{V}^T\mathbf{D}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ stores the eigenvectors of $\tilde{\mathbf{C}}$ as columns.

We know that the covariance matrix $\tilde{\mathbf{C}}$ of the points expressed in the new axes is diagonal (all covariances vanish),

$$\tilde{\mathbf{C}} = \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & \lambda_d \end{bmatrix}$$

and $\lambda_1, \ldots, \lambda_d$ are eigenvalues corresponding to the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$.

Let us confirm this fact by:

1. calculating $\tilde{\mathbf{C}}$ using the data set $\tilde{\mathbf{D}}$ in the new axes, in the same way as $\mathbf{C}$ in the original axes calculated by[1] $\mathbf{D}$.
2. projecting $\tilde{\mathbf{C}}$ back to the original axes, i.e. checking the equality

$$\mathbf{C} = \mathbf{V}\tilde{\mathbf{C}}\mathbf{V}^T.$$

Figure 1 demonstrates projection and back-projection on a 2-dimensional data set. It starts with a data set $\mathbf{D}$ with its non-diagonal covariance matrix $\mathbf{C}$, projects them onto the new axes, yielding a diagonal covariance matrix $\tilde{\mathbf{C}}$, and projects them back onto the original axes, yielding a matrix $\mathbf{V}\tilde{\mathbf{C}}\mathbf{V}^T$ that coincides with $\mathbf{C}$. The new axes, found by calculating eigenvectors of $\mathbf{C}$, are shown as red lines.

Play around with all 2-dimensional data sets, calculate the covariance matrices, obtain the eigenvectors, plot the new axes, project onto the new axes, verify that the eigenvector with maximal eigenvalue indeed captures

[1] It is worth asking whether the random variables are centered again in the new axes if they are centered in the original axes.

the data better than any of the original axes.

Figure 2 demonstrates the same procedure on a 3-dimensional data set. Data points are coloured either blue, red or black to help understanding. They spread on the triangle in Figure 2 with some noise that positions them slightly off the triangle plane.

Again, eigenvectors of $\mathbf{C}$ give the new axes $(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$ such that the triangle (containing most of the data variance) is on the $\tilde{X}_1\tilde{X}_2$-plane.

The table in Figure 2 shows 2-dimensional plots of original data $\mathbf{D}$ and projected data $\tilde{\mathbf{D}}$. We can observe that the triangle is not parallel with any of the X_1X_2-plane, X_2X_3-plane or X_1X_3-plane; and that the projection makes the triangle sit exactly on the $\tilde{X}_1\tilde{X}_2$-plane. Perform all these calculations yourself.

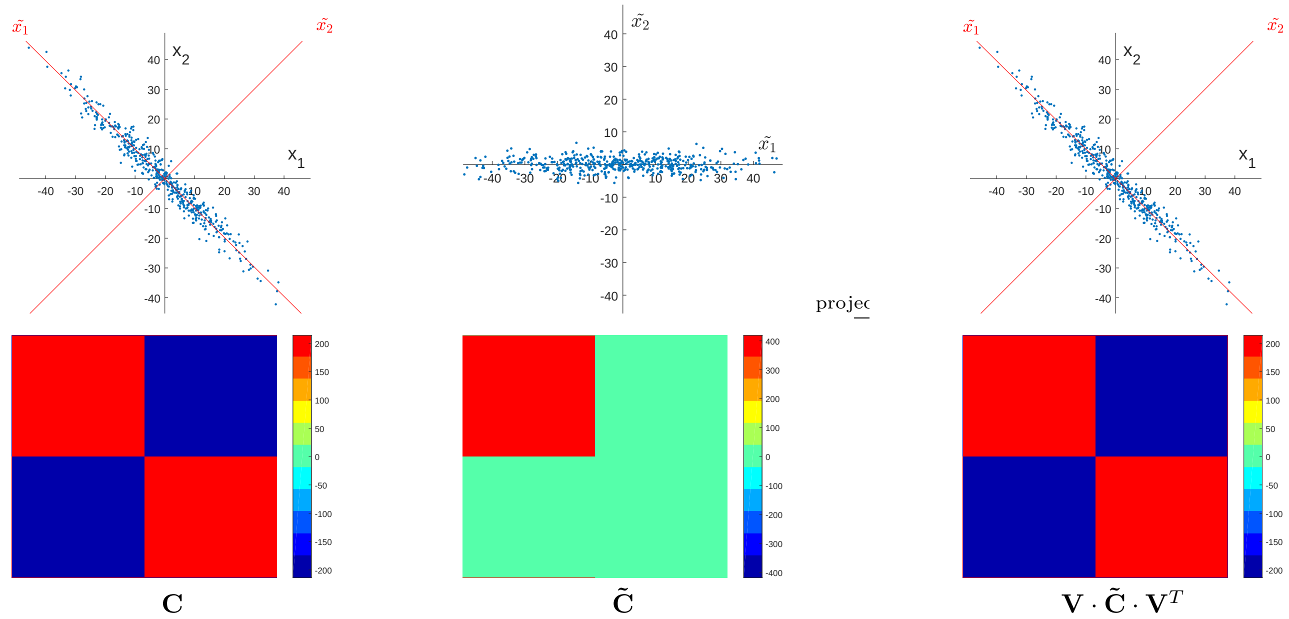

Figure 1: 2-dimensional demonstration

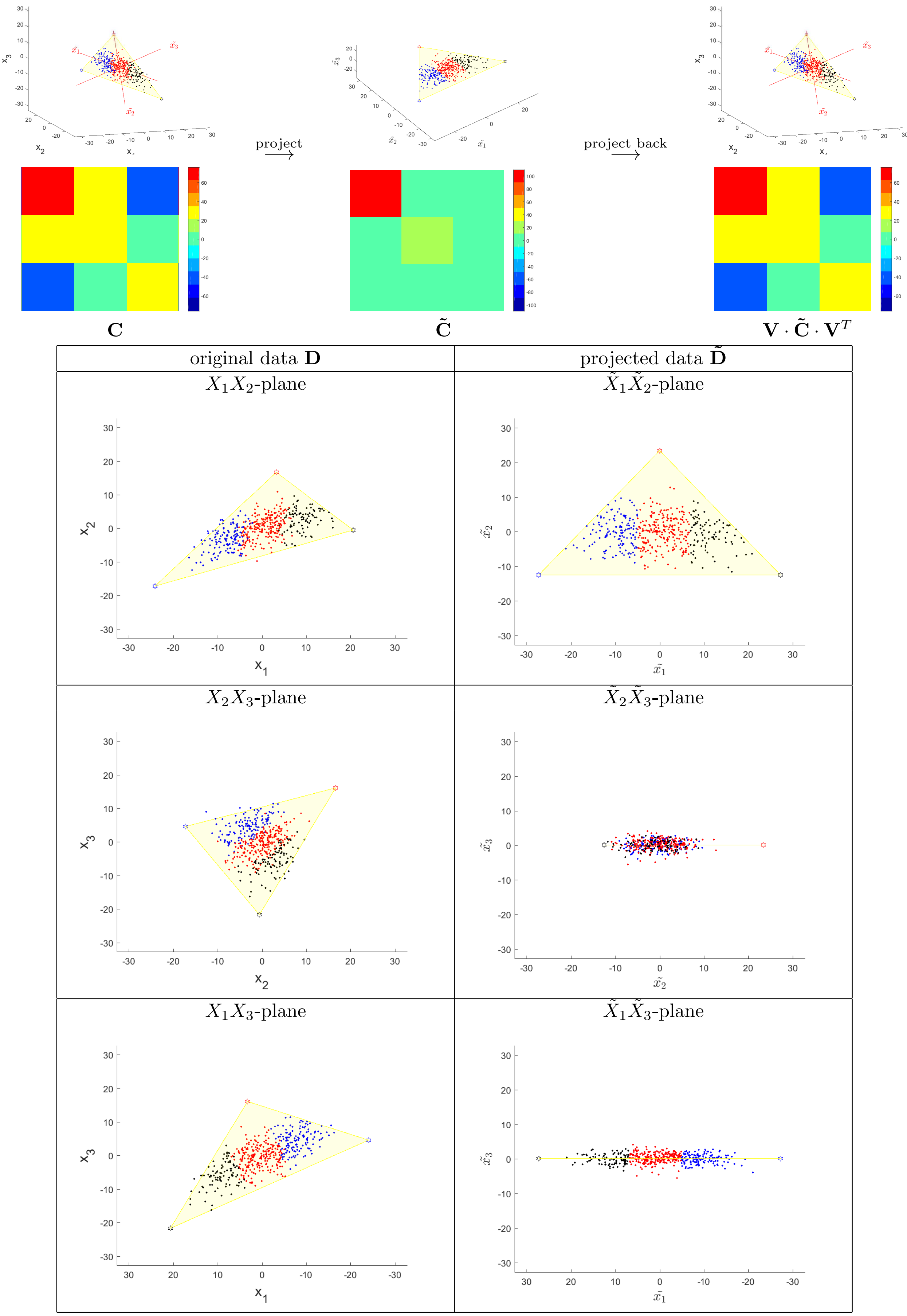

Figure 2: 3-dimensional demonstration