3.8.1 Principal Component Analysis (PCA)

We begin by considering the problem of representing all of the vectors in a set of n d-dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by a single vector $\mathbf{x}_0$. To be more specific, suppose that we want to find a vector $\mathbf{x}_0$ such that the sum of the squared distances between $\mathbf{x}_0$ and the various $\mathbf{x}_k$ is as small as possible. We define the squared-error criterion function $J_0(\mathbf{x}_0)$ by

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{x}_k||^2, \tag{78}$$

and seek the value of $\mathbf{x}_0$ that minimizes J_0. It is simple to show that the solution to this problem is given by $\mathbf{x}_0 = \mathbf{m}$, where $\mathbf{m}$ is the sample mean,

$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k. \tag{79}$$

This can be easily verified by writing

$$\begin{aligned} J_0(\mathbf{x}_0) &= \sum_{k=1}^{n} ||(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_k - \mathbf{m})||^2 \\ &= \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2\sum_{k=1}^{n} (\mathbf{x}_0 - \mathbf{m})^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2(\mathbf{x}_0 - \mathbf{m})^t \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 + \underbrace{\sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2}_{independent\ of\ \mathbf{x}_0}. \end{aligned} \tag{80}$$

Since the second sum is independent of $\mathbf{x}_0$, this expression is obviously minimized by the choice $\mathbf{x}_0 = \mathbf{m}$.

The sample mean is a zero-dimensional representation of the data set. It is simple, but it does not reveal any of the variability in the data. We can obtain a more interesting, one-dimensional representation by projecting the data onto a line running through the sample mean. Let $\mathbf{e}$ be a unit vector in the direction of the line. Then the equation of the line can be written as

$$\mathbf{x} = \mathbf{m} + a\mathbf{e}, \tag{81}$$

where the scalar a (which takes on any real value) corresponds to the distance of any point $\mathbf{x}$ from the mean $\mathbf{m}$. If we represent $\mathbf{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$, we can find an "optimal" set of coefficients a_k by minimizing the squared-error criterion function

$$\begin{aligned} J_1(a_1, \ldots, a_n, \mathbf{e}) &= \sum_{k=1}^{n} ||(\mathbf{m} + a_k\mathbf{e}) - \mathbf{x}_k||^2 = \sum_{k=1}^{n} ||a_k\mathbf{e} - (\mathbf{x}_k - \mathbf{m})||^2 \\ &= \sum_{k=1}^{n} a_k^2 ||\mathbf{e}||^2 - 2\sum_{k=1}^{n} a_k \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2. \end{aligned} \tag{82}$$

Recognizing that $||\mathbf{e}|| = 1$, partially differentiating with respect to a_k, and setting the derivative to zero, we obtain

$$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m}). \tag{83}$$

Geometrically, this result merely says that we obtain a least-squares solution by projecting the vector $\mathbf{x}_k$ onto the line in the direction of $\mathbf{e}$ that passes through the sample mean.

SCATTER MATRIX

This brings us to the more interesting problem of finding the *best* direction $\mathbf{e}$ for the line. The solution to this problem involves the so-called *scatter matrix* $\mathbf{S}$ defined by

$$\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t. \tag{84}$$

The scatter matrix should look familiar—it is merely $n - 1$ times the sample covariance matrix. It arises here when we substitute a_k found in Eq. 83 into Eq. 82 to obtain

$$\begin{aligned} J_1(\mathbf{e}) &= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= -\sum_{k=1}^{n} [\mathbf{e}^t(\mathbf{x}_k - \mathbf{m})]^2 + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= -\sum_{k=1}^{n} \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t\mathbf{e} + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= -\mathbf{e}^t\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2. \end{aligned} \tag{85}$$

Clearly, the vector $\mathbf{e}$ that minimizes J_1 also maximizes $\mathbf{e}^t\mathbf{S}\mathbf{e}$. We use the method of Lagrange multipliers (described in Section A.3 of the Appendix) to maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to the constraint that $||\mathbf{e}|| = 1$. Letting λ be the undetermined multiplier, we differentiate

$$u = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^t\mathbf{e} - 1) \tag{86}$$

with respect to $\mathbf{e}$ to obtain

$$\frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e}. \tag{87}$$

Setting this gradient vector equal to zero, we see that $\mathbf{e}$ must be an eigenvector of the scatter matrix:

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}. \tag{88}$$

In particular, because $\mathbf{e}^t\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^t\mathbf{e} = \lambda$, it follows that to maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$, we want to select the eigenvector corresponding to the largest eigenvalue of the scatter matrix. In other words, to find the best one-dimensional projection of the data (best in the least-