Principal Component Analyis

Literature

Motivation

Getting an idea how `PCA` works and what kind of problems can be solved.

Background

1. A measurement is defined as a collection of `K` items.

2. In general there are `N` measurement

3. The `j-th` measurement has `K` measured items. These items are arranged into a column vector denoted $\mathbf{d}_j$.

$$\mathbf{d}_j = \begin{bmatrix} d_1[j] \\ \vdots \\ d_k[j] \\ \vdots \\ d_K[j] \end{bmatrix} = \begin{bmatrix} d_{1,\,j} \\ \vdots \\ d_{k,\,j} \\ \vdots \\ d_{K,\,j} \end{bmatrix}$$

$d_k[j] = d_{k\ j}$ denotes the `j-th` measurement / data of the `k-th` item.

---

centering the data set

The mean value of the data set is computed from all measurements $\mathbf{d}_j :\ j = 1, \ldots, N$ by taking the *element-wise* average of each measurement item. The mean value is defined as a column vector:

$$E(\mathbf{d}) = \begin{bmatrix} E(d_1) \\ \vdots \\ E(d_k) \\ \vdots \\ E(d_K) \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{j=1}^{N} d_{1,j} \\ \vdots \\ \frac{1}{N}\sum_{j=1}^{N} d_{k,j} \\ \cdots \\ \frac{1}{N}\sum_{j=1}^{N} d_{K,j} \end{bmatrix}$$

The **centered** data vector are denoted $\mathbf{x}_j :\ j = 1, \ldots, N$. They are computed from data vectors $\mathbf{d}_j$ by removing the mean value.

$$\mathbf{x}_j = \begin{bmatrix} d_1[j] - E(d_1) \\ \vdots \\ d_k[j] - E(d_k) \\ \vdots \\ d_K[j] - E(d_K) \end{bmatrix} = \begin{bmatrix} d_{1,\,j} - E(d_1) \\ \vdots \\ d_{k,\,j} - E(d_k) \\ \vdots \\ d_{K,\,j} - E(d_K) \end{bmatrix} = \begin{bmatrix} x_{1,\,j} \\ \vdots \\ x_{k,\,j} \\ \vdots \\ x_{K,\,j} \end{bmatrix}$$

Now we define a unit vector $\mathbf{w}: \ \in \mathrm{R}^K;\ ||\mathbf{w}|| = 1$. This vector shall be to determine the component of the `j-th` measurement in the direction of vector $\mathbf{w}$. This vector is denoted $\mathbf{p}_j$. It is the projection of vector $\mathbf{x}_j$ onto $\mathbf{w}$. The residual vector $\mathbf{r}_j$ is orthogonal to vector $\mathbf{p}_j$.

To summarise (projection vector & residual vector):

$$\mathbf{p}_j = \left(\mathbf{x}_j^T \cdot \mathbf{w}\right) \cdot \mathbf{w} \tag{1}$$

$$\tag{2}$$

$$\mathbf{r}_j = \mathbf{x}_j - \mathbf{p}_j \tag{3}$$

$$\mathbf{r}_j = \mathbf{x}_j - \left(\mathbf{x}_j^T \cdot \mathbf{w}\right) \cdot \mathbf{w} \tag{4}$$

The squared norm of the residual $\mathbf{r}_j$ is computed from:

$$||\mathbf{r}_j||^2 = \mathbf{r}_j^T \cdot \mathbf{r}_j = \left(\mathbf{x}_j - \left(\mathbf{x}_j^T \cdot \mathbf{w}\right) \cdot \mathbf{w}\right)^T \cdot \left(\mathbf{x}_j - \left(\mathbf{x}_j^T \cdot \mathbf{w}\right) \cdot \mathbf{w}\right) \tag{5}$$

$$= \mathbf{x}_j^T \cdot \mathbf{x}_j - \mathbf{x}_j^T \cdot \left(\mathbf{x}_j^T \cdot \mathbf{w}\right) \cdot \mathbf{w} - \left(\mathbf{x}_j^T \cdot \mathbf{w} \cdot \mathbf{w}\right)^T \cdot \mathbf{x}_j + \left(\mathbf{x}_j^T \cdot \mathbf{w} \cdot \mathbf{w}\right)^T \cdot \mathbf{x}_j^T \cdot \mathbf{w} \cdot \mathbf{w} \tag{6}$$

$$= \mathbf{x}_j^T \cdot \mathbf{x}_j - 2 \cdot \left(\mathbf{x}_j^T \cdot \mathbf{w}\right)^2 + \left(\mathbf{x}_j^T \cdot \mathbf{w}\right)^2 \tag{7}$$

$$= \mathbf{x}_j^T \cdot \mathbf{x}_j - \left(\mathbf{x}_j^T \cdot \mathbf{w}\right)^2 = ||\mathbf{x}_j||^2 - \left(\mathbf{x}_j^T \cdot \mathbf{w}\right)^2 \tag{8}$$

We have used the unit vector property of vector $\mathbf{w}$ namely $||\mathbf{w}|| = 1$.

adding up the squared residual

The mean squared error is the expectation of the squared residuals for all `N` measurements.

$$MSE(\mathbf{w}) = \frac{1}{N}\sum_{j=1}^{N} ||\mathbf{x}_j||^2 - \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{x}_j^T \cdot \mathbf{w}\right)^2$$

To make $MSE(\mathbf{w})$ as small as possible the term

$$V = \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{x}_j^T \cdot \mathbf{w}\right)^2$$

must be maximised by appropriate choice of vector $\mathbf{w}$. Re-writing V yields:

$$V = \frac{1}{N}\sum_{j=1}^{N} \mathbf{w}^T \cdot \underbrace{\left(\mathbf{x}_j \cdot \mathbf{x}_j^T\right)}_{\mathbf{X}_j} \cdot \mathbf{w} = \mathbf{w}^T \cdot \underbrace{(\frac{1}{N}\sum_{j=1}^{N} \mathbf{X}_j)}_{\mathbf{X}} \cdot \mathbf{w} = \mathbf{w}^T \cdot \mathbf{X}$$

$\mathbf{X}_j = \mathbf{x}_j \cdot \mathbf{x}_j^T \ : \ \in \mathrm{R}^{K \times K}$ is a square matrix computed as the outer product of centered data vector $\mathbf{x}_j$.

$\mathbf{X}_j$ has elements $x_{l,\,m}[j]$.

$$x_{l,\,m}[j] = x_l[j] \cdot x_m[j] = (d_l[j] - E(d_l)) \cdot (d_m[j] - E(d_m)) \tag{9}$$
$$x_{l,\,m}[j] = d_l[j] \cdot d_m[j] - d_l[j] \cdot E(d_m) - d_m[j] \cdot E(d_l) + E(d_l) \cdot E(d_m) \tag{10}$$

The sum of all matrices $\mathbf{X}_j$ is matrix $\mathbf{X}$:

$$\mathbf{X} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{X}_j$$

with elements $x_{l,\,m}$:

$$x_{l,\,m} = E(d_l \cdot d_m) - E(d_m) \cdot E(d_l) - E(d_m) \cdot E(d_l) + E(d_l) \cdot E(d_m) \tag{11}$$
$$\tag{12}$$
$$x_{l,\,m} = E(d_l \cdot d_m) - E(d_m) \cdot E(d_l) \tag{13}$$

For $l = m$ matrix elements $x_{l,\,l}$ are just the variance:

$$x_{l,\,l} = E(d_l \cdot d_l) - E(d_l)^2 = E(d_l^2) - E(d_l)^2$$

For $l \neq m$ matrix elements $x_{l,\,m}$ are the covariance.

Hence matrix $\mathbf{X}$ is also known as the covariance matrix.

summary

It has been demonstrated that by appropriate choice of projection vector $\mathbf{w}$ the maximisation of

$$V = \mathbf{w}^T \cdot \mathbf{X} \cdot \mathbf{w}$$

minimises the mean squared error $MSE(\mathbf{w})$.

The next step is to determine the optimum vector $\mathbf{w}$ with the constraint that $\mathbf{w}$ has unit length ($||\mathbf{w}|| = 1$).

Finding the vector minimising the MSE

The optimum vector must be computed with the unit length constraint.

Using the `Lagrange` multiplier method, the objective function $F(\mathbf{w},\lambda)$ may be expressed like this:

$$F(\mathbf{w},\lambda) = \mathbf{w}^T \cdot \mathbf{X} \cdot \mathbf{w} - \lambda \cdot (\mathbf{w}^T \cdot \mathbf{w} - 1)$$

ToDo

Get a solid understanding of constrained optimisation with `Lagrange` multipliers. (in this notebook I just copied the method without having understood how it works)

The solution vector is found by setting derivatives

$$\frac{\partial}{\partial \mathbf{w}} F(\mathbf{w},\lambda) = 0 \tag{14}$$
$$\frac{\partial}{\partial \lambda} F(\mathbf{w},\lambda) = 0 \tag{15}$$

$$\frac{\partial}{\partial \mathbf{w}} F(\mathbf{w},\lambda) = 2 \cdot \mathbf{X} \cdot \mathbf{w} - 2 \cdot \lambda \cdot \mathbf{w} = \mathbf{0}$$

$$\frac{\partial}{\partial \lambda} F(\mathbf{w},\lambda) = \mathbf{w}^T \cdot \mathbf{w} - 1$$

Leading to

$$\mathbf{X} \cdot \mathbf{w} = \lambda \cdot \mathbf{w} \tag{16}$$
$$\mathbf{w}^T \cdot \mathbf{w} = 1 \tag{17}$$

From the first equation we conclude that the optimum vector $\mathbf{w}$ has been found as an **eigenvector** of the covariance matrix.

The second equation just states the fact that $\mathbf{w}$ has unit length.

We have defined

$$V = \mathbf{w}^T \cdot \mathbf{X} \cdot \mathbf{w}$$

inserting the optimum vector yields:

$$V(\mathbf{w}) = \lambda \mathbf{w}^T \cdot \mathbf{w} = \lambda ||\mathbf{w}||^2 = \lambda$$

The MSE is minimised if we choose the eigenvector $\mathbf{w} = \mathbf{w}_1$ with the largest eigenvalue $\lambda = \lambda_1$.

For the residual vector we get:

$$\mathbf{r}_{j(1)} = \mathbf{x}_j - \left(\mathbf{x}_j^T \cdot \mathbf{w}_1\right) \cdot \mathbf{w}_1$$

and for the average of squared residuals:

$$\frac{1}{N}\sum_{j=1}^{N}||\mathbf{r}_{j(1)}||^2 = \frac{1}{N}\sum_{j=1}^{N}||\mathbf{x}_j||^2 - \mathbf{w}_1^T \cdot \mathbf{X} \cdot \mathbf{w}_1 \tag{18}$$

$$\frac{1}{N}\sum_{j=1}^{N}||\mathbf{r}_{j(1)}||^2 = \frac{1}{N}\sum_{j=1}^{N}||\mathbf{x}_j||^2 - \lambda_1 \tag{19}$$

Going one step further

We are going to project vector $\mathbf{r}_{j(1)}$ onto some other vector $\mathbf{w}_2$ and get a new residual vector $\mathbf{r}_{j(2)}$.

$$\mathbf{r}_{j(2)} = \mathbf{r}_{j(1)} - \left(\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2\right) \cdot \mathbf{w}_2$$

Then vector $\mathbf{w}_2$ shall be chosen such as to minimise the sum of the squared residuals:

$$\frac{1}{N}\sum_{j=1}^{N}||\mathbf{r}_{j(2)}||^2 = \frac{1}{N}\sum_{j=1}^{N}\mathbf{r}_{j(2)}^T \cdot \mathbf{r}_{j(2)} = \frac{1}{N}\sum_{j=1}^{N}(||\mathbf{r}_{j(1)}||^2 - \left(\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2\right)^2) = \frac{1}{N}\sum_{j=1}^{N}||\mathbf{r}_{j(1)}||^2$$

$$\frac{1}{N}\sum_{j=1}^{N}||\mathbf{r}_{j(2)}||^2 = (\frac{1}{N}\sum_{j=1}^{N}||\mathbf{x}_j||^2 - \lambda_1) - \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2\right)^2$$

Choose $\mathbf{w}_2$ to maximise

$$\frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2\right)^2$$

$$\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2 = \left(\mathbf{x}_j - \left(\mathbf{x}_j^T \cdot \mathbf{w}_1\right) \cdot \mathbf{w}_1\right)^T \cdot \mathbf{w}_2 \tag{20}$$

$$\tag{21}$$

$$\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2 = \mathbf{x}_j^T \cdot \mathbf{w}_2 - \mathbf{w}_1^T \cdot \left(\mathbf{x}_j^T \cdot \mathbf{w}_1\right) \cdot \mathbf{w}_2 \tag{22}$$

At this point we postulate the vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are **orthonormal**.

$$\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2 = \mathbf{x}_j^T \cdot \mathbf{w}_2$$

Hence

$$\frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{r}_{j(1)}^T \cdot \mathbf{w}_2\right)^2 = \frac{1}{N}\sum_{j=1}^{N}\mathbf{w}_2^T\left(\mathbf{x}_j\mathbf{x}_j^T\right) \cdot \mathbf{w}_2 = \mathbf{w}_2^T \cdot (\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T) \cdot \mathbf{w}_2 = \mathbf{w}_2^T \cdot \mathbf{X} \cdot \mathbf{w}_2$$

In [ ]: