Stage 1 revision 11.13

\textbf{Principle Components Analysis}, knows as \textbf{PCA}, is a model that enable people to 删reduces data points in a high-dimensional vector space to that in a lower dimensional one像之前的那版一样改, and at the same time,删 while ensuring that the information they contain be is expressed to the greatest extent. Here, we are going to reduce a $10$-dimensional vector (the ten factors mentioned before) to a $n$-dimensional one.

Assume that the space where the original data points lies in locate is $V$, and the new space is $V\_{\text{PCA}}$. Generally, when applying the PCA method ‘the 。。。method’去掉只留‘PCA’, we are actually in fact looking for a new set of 3 \textbf{basis} and extracting the original vectors' \textbf{projection} on this $n$-dimensional hyperplane. The PCA process is a linear transformation, which means the process it can be written in the following form:

\begin{equation}

W = M\_{\text{PCA}} \cdot X

\end{equation}

where $X$ is a $10$ by $N\_{\text{time}}$ matrix ($N\_{\text{time}}$ stands for the number of samples) and $W$ is a $3$ by $N\_{\text{time}}$ matrix. To be specific, we select some vectors in $V$ one by one, which will later become the basis of $V\_\text{PCA}$. Later, with the basis picked 加具体描述：如何生成的basis?, we project the data points to $V\_\text{PCA}$, so that the new data points are加now $n$-dimensional and can be used in further prediction of CO2 of further use in the CO2 projection.

When selecting the $k$-th vector, we first ensure made sure that the $k$-th vector is \textbf{linear independent} with the previous $k-1$ ones. Under that condition, We Under such circumstances, we then pick the vector best representing information that are not carried by select the vector expressing the most comprehensive information that cannot be found in the first $k-1$ vectors, which specifically means to pick a vector that minimizes its \textbf{covariance} with the first $k-1$ vectors.

\vspace{12pt}

The covariance of two $n$-dimensional data points $\boldsymbol{a}$, $\boldsymbol{b}$ can be calculated with the following formula:

\begin{equation}

Cov(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{n-1} \sum\limits\_{i=1}^n

(\boldsymbol{a\_i} - \bar{\boldsymbol{a}}) (\boldsymbol{b\_i} - \bar{\boldsymbol{b}})

\end{equation}

where

\begin{equation}

\bar{\boldsymbol{a}} = \frac 1 n \sum\limits\_{i = 1}^n \boldsymbol{a\_i}

\end{equation}

Let Tthe covariance of more than two vectors is be represented by a symmetric matrix $\Sigma$, which is given by the formular below:

\begin{equation}

Cov(\boldsymbol{x\_1} \cdots \boldsymbol{x\_n}) =

\begin{bmatrix}

Cov(\boldsymbol{x\_1}, \boldsymbol{x\_1}) & \cdots & Cov(\boldsymbol{x\_1}, \boldsymbol{x\_n}) \\

\vdots & \ddots & \vdots \\

Cov(\boldsymbol{x\_n}, \boldsymbol{x\_1}) & \cdots & Cov(\boldsymbol{x\_n}, \boldsymbol{x\_n})

\end{bmatrix}

\end{equation}

The covariance of vectors, or data points, tells the how in which way and to what extent they are similar to each other. Thus, when upon adding the k-th basis vector to $V'$, we are able to compute the \textbf{eigenvector and eigenvalue} covariance among all $k$ vectors and . We then pick the vector that minimizes eigenvalue, which measures the revelance of the $k$ vectors. With $V'$ established the establishment of ???, we can derive $M\_\text{PCA}$ and then calculate $W$.