The <u>covariance matrix</u> of the transformed representation is <u>diagonal</u>. This can be shown as follows.

The data matrix $\underline{X}$ (old basis) of the original representation will get transformed to $\underline{X}'$ in the new representation

$$\boxed{\underline{X}' = \underline{X}\ \underline{V}}$$

Here $\underline{V}$ is a matrix of eigenvectors as columns arranged from <u>left</u> <u>to right</u> such that their corresponding <u>eigenvalues</u> would form a <u>decreasing order</u>. $\lambda_1 \geq \lambda_2 \geq \ldots$

Let $\boxed{\underline{C}', \underline{\mu}'}$ be the covariance matrix and the mean vector in the new representation. $\underset{d\times 1}{\underline{\mu}'}$ is a $d \times 1$ vector

original

$$\underset{d\times d}{C} = \frac{X^T X}{m} - \underline{\mu}\,\underline{\mu}^T \qquad \boxed{\underline{\mu}' = \underline{V}^T \underline{\mu}}$$

$$\underline{X}' = \underline{X}\ \underline{V}$$

transformed

$$\underset{d\times d}{C'} = \frac{X'^T X'}{m} - \underline{\mu}'\,\underline{\mu}'^T$$

$$= \frac{(XV)^T (XV)}{m} - (\underline{V}^T\underline{\mu})(\underline{V}^T\underline{\mu})^T$$

$$= \frac{\underline{V}^T X^T X \underline{V}}{m} - \underline{V}^T\underline{\mu}\,\underline{\mu}^T\underline{V}$$

Subsituting $\boxed{\underline{X}^T\underline{X} = m\underline{C} + m\,\underline{\mu}\,\underline{\mu}^T}$ gives

$$C' = \frac{V^T(\cancel{m}\,\underline{C} + \cancel{m}\,\underline{\mu}\,\underline{\mu}^T)\underline{V}}{\cancel{m}} - \underline{V}^T\underline{\mu}\,\underline{\mu}^T\underline{V}$$

$$C' = \boxed{V^T C V} + \cancel{V^T \mu \mu^T V} - \cancel{V^T \mu \mu^T V}$$

$$\underset{d \times d}{C'} = \Lambda \qquad \because \boxed{C = V \Lambda V^T} \quad V^T C V = V^T V \Lambda V^T V = I \Lambda I = \Lambda$$

$$\therefore V^T C V = \Lambda$$

Thus, the transformed representation has a diagonal covariance matrix.

∴ the directions (i.e. the new basis vectors) that maximize the variance are also the directions that remove correlations.

In the new representation X', only the first $k < d$ columns will show a significant variation in values.

$$\underset{\text{data matrix}}{\underset{m\times d}{X'}} = m \overset{\xleftarrow{\quad d \text{ features} \quad}}{\left[\begin{array}{cccc|cc} x_1^1 & x_1^2 & x_1^3 & \cdots & \cdots & x_1^d \\ & \text{high variation} & & & \text{low variation} & \\ & & & & & \end{array}\right]}$$

$$\begin{array}{cccc} \lambda_1 & \lambda_2 & \cdots & \lambda_d \end{array} \qquad k \,\big|\, d-k \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

The remaining $d-k$ features will show less variation
$\Rightarrow$ they take values close to the mean of the feature.

∴ representing X' using just k features would preserve the maximum variance.

∴ only the first k columns of the transformed data matrix need to be retained. The remaining columns (features) can be discarded.

Dimensionality Reduction

Thus, PCA can be used to obtain a new representation of data in which the dimensionality is reduced.

So far we have discussed an interpretation of PCA in which the new low dimensional representation could preserve the maximum variance in the data. ① variance preserving interpretation

② In another interpretation of PCA, we seek a low dimensional representation such that when the data is projected back from the low dimensional subspace to the original space, the total squared distance between the original and the recovered vectors is minimal.

$$\begin{tikzcd} d \arrow[d, bend left, "W"] \\ n \arrow[u, bend left, "U"] \end{tikzcd}$$

$n<d$

That is, the reconstruction error is minimal.

Find $\underset{n\times d}{W}$ and $\underset{d\times n}{U}$ that $\displaystyle \min_{U,W} \sum_{i=1}^{m} \Big\| \underset{\text{original}}{x_i} - \underbrace{U W x_i}_{\text{recovered vectors}} \Big\|^2$

$n<d$

The proof is given in the text book.

It first shows that $\boxed{W = U^T}$

Then it shows that minimizing the reconstruction error is equivalent to maximizing $\text{trace}\Big(U^T \underbrace{\sum_i x_i x_i^T}_{A} U\Big)$

Then it shows that $\text{trace}\Big(U^T \sum_i x_i x_i^T U\Big) \leq \sum_{i=1}^{\cancel{m}\, n} \lambda_i$

where λ_i are the largest n eigenvalues of the eigenvalue eigenvector decomposition of matrix $\underbrace{A = \sum_i x_i x_i^T}$

The proof shows that setting the columns of matrix $\underline{U}$ as the n leading eigenvectors of A will ensure that

$$\text{trace}(\underline{U}^T A \underline{U}) = \sum_{i=1}^{n} \lambda_i$$

It can be showed that the reconstruction loss is the sum of discarded eigenvalues $\sum_{i=n+1}^{d} \lambda_i$. distortion

An efficient way to apply PCA when $d \gg m$

Consider vectors $\underline{x}_i$ which have been mean-centered.

The covariance matrix $\underline{C} = \underline{X}^T \underline{X}$ where $\underline{X}_{m \times d}$ = data matrix.

Clearly, $\underline{C}_{d \times d}$ is very large.

Instead, we formulate another matrix $\underline{B}_{m \times m} = \underline{X}\,\underline{X}^T$

$\underline{B}$ is $m \times m$ matrix.

The $(i,j)^{th}$ element of B is given as $\langle x_i, x_j \rangle$.

$\underline{B}$ is also called as the Gram matrix.

We obtain the eigen decomposition of $\underline{B}$.

If $\underline{u}$ is an eigenvector of $\underline{B}$ then $\underline{B}\,\underline{u} = \lambda \underline{u}$

$$\underline{X}\,\underline{X}^T \underline{u} = \lambda \underline{u}$$

$$X^T (X X^T) \underline{u} = \lambda \underline{X}^T \underline{u}$$

since $C \equiv \underline{X}^T \underline{X}$, we have