Functional Data Analysis (Lecture 4) - PCA

Zoltán Szabó

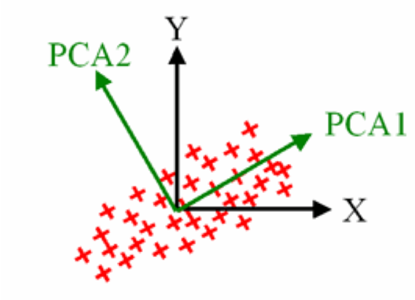
November 8, 2016

One-page summary

- Covered topics:
  - PEN_L-regularized least squares,
  - smoothing with constraints,
  - curve registration.

One-page summary

- Today:
  1. dimensionality reduction,
  2. principal component analysis (PCA; in $\mathbb{R}^d$ first):
     - see continuous registration.



PCA

PCA – intuition

- Given: a set of observations $X = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^D$.
- Goal: find the best d-dimensional subspace approximating X.
- $d \ll D$: compression (images, music, ...).

PCA formulation: $d = 1$

- We are looking for the best one-dimensional projection.

- $\mathbb{E}$:= empirical expectation (population: similarly).
- In other words: $\mathbb{E}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$.
- Assumption: $\mathbb{E}\mathbf{x} = \mathbf{0}$

PCA formulation: $d = 1$

- We are looking for the best one-dimensional projection.

- $\mathbb{E}$:= empirical expectation (population: similarly).
- In other words: $\mathbb{E}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$.
- Assumption: $\mathbb{E}\mathbf{x} = \mathbf{0}$
  - centering: $\mathbf{x} \to \mathbf{x} - \mathbb{E}\mathbf{x}$.

PCA: projection

- One-dimensional projection:
  - $\mathbf{w}$ with $\|\mathbf{w}\|_2 = 1$,
  - $\hat{\mathbf{x}} = \underbrace{\langle \mathbf{w}, \mathbf{x} \rangle}_{=\text{score}} \mathbf{w}$.
- Mean of the projection is zero:

$$\mathbf{0} \stackrel{?}{=} \mathbb{E}\hat{\mathbf{x}} = \mathbb{E}\left[\langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}\right]$$

PCA: projection

- One-dimensional projection:
  - $\mathbf{w}$ with $\|\mathbf{w}\|_2 = 1$,
  - $\hat{\mathbf{x}} = \underbrace{\langle \mathbf{w}, \mathbf{x} \rangle}_{=\text{score}} \mathbf{w}$.
- Mean of the projection is zero:

$$\mathbf{0} \stackrel{?}{=} \mathbb{E}\hat{\mathbf{x}} = \mathbb{E}\left[\langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}\right] = \Big\langle \mathbf{w}, \underbrace{\mathbb{E}\mathbf{x}}_{=\mathbf{0}} \Big\rangle \mathbf{w}.$$

PCA: MSE of the residual

Residual:

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathbf{x} - \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}\|_2^2$$

PCA: MSE of the residual

Residual:

$$\begin{aligned}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle \mathbf{w}\|_2^2 \\ &= \left(\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle \mathbf{w}\right)^T\left(\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle \mathbf{w}\right)\end{aligned}$$

PCA: MSE of the residual

Residual:

$$\begin{aligned}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\,\mathbf{w}\|_2^2\\ &= \left(\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\,\mathbf{w}\right)^T\left(\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\,\mathbf{w}\right)\\ &= \|\mathbf{x}\|_2^2-2\,\langle\mathbf{w},\mathbf{x}\rangle^2+\langle\mathbf{w},\mathbf{x}\rangle^2\underbrace{\|\mathbf{w}\|_2^2}_{=1}\end{aligned}$$

PCA: MSE of the residual

Residual:

$$
\begin{aligned}
\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x} - \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}\|_2^2 \\
&= (\mathbf{x} - \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w})^T (\mathbf{x} - \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}) \\
&= \|\mathbf{x}\|_2^2 - 2 \langle \mathbf{w}, \mathbf{x} \rangle^2 + \langle \mathbf{w}, \mathbf{x} \rangle^2 \underbrace{\|\mathbf{w}\|_2^2}_{=1} \\
&= \|\mathbf{x}\|_2^2 - \langle \mathbf{w}, \mathbf{x} \rangle^2 .
\end{aligned}
$$

PCA: MSE of the residual

Residual:

$$\begin{aligned}
\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w}\|_2^2\\
&= (\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w})^T(\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w})\\
&= \|\mathbf{x}\|_2^2 - 2\langle\mathbf{w},\mathbf{x}\rangle^2 + \langle\mathbf{w},\mathbf{x}\rangle^2\underbrace{\|\mathbf{w}\|_2^2}_{=1}\\
&= \|\mathbf{x}\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2.
\end{aligned}$$

MSE:

$$\min_{\mathbf{w}} \leftarrow \mathbb{E}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 = \mathbb{E}\left[\|\mathbf{x}\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2\right]$$

PCA: MSE of the residual

Residual:

$$\begin{aligned}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle\,\mathbf{w}\|_2^2\\ &= \left(\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle\,\mathbf{w}\right)^T\left(\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle\,\mathbf{w}\right)\\ &= \|\mathbf{x}\|_2^2 - 2\langle \mathbf{w},\mathbf{x}\rangle^2 + \langle \mathbf{w},\mathbf{x}\rangle^2\underbrace{\|\mathbf{w}\|_2^2}_{=1}\\ &= \|\mathbf{x}\|_2^2 - \langle \mathbf{w},\mathbf{x}\rangle^2.\end{aligned}$$

MSE:

$$\begin{aligned}\min_{\mathbf{w}} \leftarrow \mathbb{E}\,\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \mathbb{E}\left[\|\mathbf{x}\|_2^2 - \langle \mathbf{w},\mathbf{x}\rangle^2\right]\\ &= \underbrace{\mathbb{E}\,\|\mathbf{x}\|_2^2}_{\text{independent of } \mathbf{w}} - \mathbb{E}\,\langle \mathbf{w},\mathbf{x}\rangle^2.\end{aligned}$$

PCA: MSE of the residual

Residual:

$$
\begin{aligned}
\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \|\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w}\|_2^2 \\
&= (\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w})^T(\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w}) \\
&= \|\mathbf{x}\|_2^2 - 2\langle\mathbf{w},\mathbf{x}\rangle^2 + \langle\mathbf{w},\mathbf{x}\rangle^2\underbrace{\|\mathbf{w}\|_2^2}_{=1} \\
&= \|\mathbf{x}\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2.
\end{aligned}
$$

MSE:

$$
\begin{aligned}
\min_{\mathbf{w}} \leftarrow \mathbb{E}\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 &= \mathbb{E}\left[\|\mathbf{x}\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2\right] \\
&= \underbrace{\mathbb{E}\|\mathbf{x}\|_2^2}_{\text{independent of } \mathbf{w}} - \mathbb{E}\langle\mathbf{w},\mathbf{x}\rangle^2.
\end{aligned}
$$

$\Leftrightarrow$ Maximize the mean squared projection.

PCA: maximize variance

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + var(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 = \Big(\underbrace{\mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle}_{=0}\Big)^2 + var(\langle \mathbf{w}, \mathbf{x} \rangle).$$

PCA: maximize variance

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + var(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 = \Big(\underbrace{\mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle}_{=0}\Big)^2 + var(\langle \mathbf{w}, \mathbf{x} \rangle).$$

To sum up:

$$\text{Minimize MSE of the residual} : \mathbb{E}\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}} \Leftrightarrow$$

PCA: maximize variance

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + var(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 = \Big(\underbrace{\mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle}_{=0}\Big)^2 + var(\left\langle \mathbf{w}, \mathbf{x} \right\rangle).$$

To sum up:

$$\text{Minimize MSE of the residual} : \mathbb{E}\left\| \mathbf{x} - \hat{\mathbf{x}} \right\|_2^2 \to \min_{\mathbf{w}} \Leftrightarrow$$

$$\text{Maximize mean squared projection} : \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 \to \max_{\mathbf{w}} \Leftrightarrow$$

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + var(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\langle \mathbf{w}, \mathbf{x}\rangle^2 = \Big(\underbrace{\mathbb{E}\langle \mathbf{w}, \mathbf{x}\rangle}_{=0}\Big)^2 + var(\langle \mathbf{w}, \mathbf{x}\rangle).$$

To sum up:

$$\begin{aligned} \text{Minimize MSE of the residual} &: \mathbb{E}\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}} \Leftrightarrow \\ \text{Maximize mean squared projection} &: \mathbb{E}\langle \mathbf{w}, \mathbf{x}\rangle^2 \to \max_{\mathbf{w}} \Leftrightarrow \\ \text{Maximize variance of the projection} &: var(\langle \mathbf{w}, \mathbf{x}\rangle) \to \max_{\mathbf{w}}. \end{aligned}$$

PCA: max. variance → optimization

Using the bilinearity of covariance:

$$var(\langle \mathbf{w}, \mathbf{x} \rangle) = cov(\langle \mathbf{w}, \mathbf{x} \rangle, \langle \mathbf{w}, \mathbf{x} \rangle)$$

PCA: max. variance → optimization

Using the bilinearity of covariance:

$$\begin{aligned} var\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) &= cov\left(\langle \mathbf{w}, \mathbf{x}\rangle, \langle \mathbf{w}, \mathbf{x}\rangle\right) \\ &= cov\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) \end{aligned}$$

PCA: max. variance → optimization

Using the bilinearity of covariance:

$$\begin{aligned} var\left(\langle \mathbf{w}, \mathbf{x} \rangle\right) &= cov\left(\langle \mathbf{w}, \mathbf{x} \rangle, \langle \mathbf{w}, \mathbf{x} \rangle\right) \\ &= cov\left(\mathbf{w}^T \mathbf{x}, \mathbf{w}^T \mathbf{x}\right) \\ &= \mathbf{w}^T cov(\mathbf{x}) \mathbf{w} =: \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \to \max_{\|\mathbf{w}\|_2 = 1}. \end{aligned}$$

PCA: max. variance → optimization

Using the bilinearity of covariance:

$$\begin{aligned} var\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) &= cov\left(\langle \mathbf{w}, \mathbf{x}\rangle, \langle \mathbf{w}, \mathbf{x}\rangle\right) \\ &= cov\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) \\ &= \mathbf{w}^T cov(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}. \end{aligned}$$

Lagrange function, solving for 'derivatives $=0$':

.

PCA: max. variance → optimization

Using the bilinearity of covariance:

$$\begin{aligned} var\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) &= cov\left(\langle \mathbf{w}, \mathbf{x}\rangle, \langle \mathbf{w}, \mathbf{x}\rangle\right) \\ &= cov\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) \\ &= \mathbf{w}^T cov(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \rightarrow \max_{\|\mathbf{w}\|_2=1}. \end{aligned}$$

Lagrange function, solving for 'derivatives = 0':

$$L(\mathbf{w}, \lambda) = \underbrace{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

.

PCA: max. variance → optimization

Using the bilinearity of covariance:

$$\begin{aligned} var\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) &= cov\left(\langle \mathbf{w}, \mathbf{x}\rangle, \langle \mathbf{w}, \mathbf{x}\rangle\right) \\ &= cov\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) \\ &= \mathbf{w}^T cov(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \rightarrow \max_{\|\mathbf{w}\|_2=1}. \end{aligned}$$

Lagrange function, solving for 'derivatives = 0':

$$L(\mathbf{w}, \lambda) = \underbrace{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

$$0 = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda} = -(\mathbf{w}^T\mathbf{w} - 1),$$

.

PCA: max. variance $\rightarrow$ optimization

Using the bilinearity of covariance:

$$\begin{aligned} var\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) &= cov\left(\langle \mathbf{w}, \mathbf{x}\rangle, \langle \mathbf{w}, \mathbf{x}\rangle\right) \\ &= cov\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) \\ &= \mathbf{w}^T cov(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \rightarrow \max_{\|\mathbf{w}\|_2=1}. \end{aligned}$$

Lagrange function, solving for 'derivatives $= 0$':

$$L(\mathbf{w}, \lambda) = \underbrace{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

$$0 = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda} = -(\mathbf{w}^T\mathbf{w} - 1),$$

$$\mathbf{0} = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = 2\mathbf{\Sigma}\mathbf{w} - 2\lambda\mathbf{w}.$$

PCA: max. variance $\rightarrow$ optimization

To sum up: $\mathbf{\Sigma w} = \lambda \mathbf{w}$, $\|\mathbf{w}\|_2 = 1$. $\Rightarrow$ $\mathbf{w}^*$: eigenvector associated to $\lambda_{\text{max}}(\mathbf{\Sigma})$.

PCA: $d \geqslant 1$

PCA ($d \geqslant 1$): basis, approximation

- Goal: approximate with a d-dimensional subspace.
- ONB in the subspace: $\{\mathbf{w}_i\}_{i=1}^d$, $\langle \mathbf{w}_i, \mathbf{w}_j \rangle = \delta_{ij}$.
- Approximation ($\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]$):

$$\hat{\mathbf{x}} = \sum_{i=1}^{d} \underbrace{\langle \mathbf{w}_i, \mathbf{x} \rangle}_{=\text{scores}} \mathbf{w}_i = \mathbf{W}\mathbf{W}^T\mathbf{x}.$$

PCA ($d \geqslant 1$): residual

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2$$

PCA ($d \geqslant 1$): residual

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)$$

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)$$
$$= \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x}$$

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)$$
$$= \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x}$$
$$= \|\mathbf{x}\|_2^2 - \left\|\mathbf{W}^T\mathbf{x}\right\|_2^2,$$

PCA ($d \geqslant 1$): residual

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$
\begin{aligned}
\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 &= \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right) \\
&= \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x} \\
&= \|\mathbf{x}\|_2^2 - \left\|\mathbf{W}^T\mathbf{x}\right\|_2^2,
\end{aligned}
$$

$$
\mathbb{E}\,\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \mathbb{E}\,\|\mathbf{x}\|_2^2 - \mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2.
$$

PCA ($d \geqslant 1$): residual

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$
\begin{aligned}
\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 &= \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right) \\
&= \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x} \\
&= \|\mathbf{x}\|_2^2 - \left\|\mathbf{W}^T\mathbf{x}\right\|_2^2,
\end{aligned}
$$

$$
\mathbb{E}\,\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \mathbb{E}\,\|\mathbf{x}\|_2^2 - \mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2.
$$

Thus $\min_{\mathbf{W}} \mathbb{E}\,\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \Leftrightarrow \max_{\mathbf{W}} \mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2$.

Let $\mathbf{y} = \mathbf{W}^T \mathbf{x}$:

$$\mathbb{E} \|\mathbf{y}\|_2^2 - \|\mathbb{E}\mathbf{y}\|_2^2 = var(\mathbf{y})?$$

PCA ($d \geqslant 1$): mean squared projection $\rightarrow$ variance

Let $\mathbf{y} = \mathbf{W}^T\mathbf{x}$:

$$\mathbb{E}\,\|\mathbf{y}\|_2^2 - \|\mathbb{E}\mathbf{y}\|_2^2 = var(\mathbf{y})?$$
$$= \mathbb{E}\left[\sum_i y_i^2\right] - \sum_i (\mathbb{E}y_i)^2 = \sum_i var(y_i) \Rightarrow$$

PCA ($d \geqslant 1$): mean squared projection $\rightarrow$ variance

Let $\mathbf{y} = \mathbf{W}^T\mathbf{x}$:

$$\mathbb{E}\left\|\mathbf{y}\right\|_2^2 - \left\|\mathbb{E}\mathbf{y}\right\|_2^2 = var(\mathbf{y})?$$

$$= \mathbb{E}\left[\sum_i y_i^2\right] - \sum_i (\mathbb{E}y_i)^2 = \sum_i var(y_i) \Rightarrow$$

$$\mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2 - \left\|\underbrace{\mathbb{E}[\mathbf{W}^T\mathbf{x}]}_{=\mathbf{W}^T\mathbb{E}\mathbf{x}=\mathbf{0}}\right\|_2^2 = \sum_i var\left(\left(\mathbf{W}^T\mathbf{x}\right)_i\right) \rightarrow \max_{\mathbf{W}}.$$

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \boldsymbol{\Sigma} = cov(\mathbf{x}).$$

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \mathbf{\Sigma} = cov(\mathbf{x}).$$

- $\mathbf{\Sigma}$: symmetric, positive semi-definite $\Rightarrow \{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \boldsymbol{\Sigma} = cov(\mathbf{x}).$$

- $\boldsymbol{\Sigma}$: symmetric, positive semi-definite $\Rightarrow \{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.
- Variance decomposition: $cov(\mathbf{x}) = \sum_{i=1}^D \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \boldsymbol{\Sigma} = cov(\mathbf{x}).$$

- $\boldsymbol{\Sigma}$: symmetric, positive semi-definite $\Rightarrow$ $\{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.
- Variance decomposition: $cov(\mathbf{x}) = \sum_{i=1}^D \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.
- Energy preserved using d components: $\sum_{i=1}^d \lambda_i \Rightarrow$

$$R^2 = R^2(d) := \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i} \in [0,1].$$

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \mathbf{\Sigma} = cov(\mathbf{x}).$$

- $\mathbf{\Sigma}$: symmetric, positive semi-definite $\Rightarrow \{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.
- Variance decomposition: $cov(\mathbf{x}) = \sum_{i=1}^D \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.
- Energy preserved using d components: $\sum_{i=1}^d \lambda_i \Rightarrow$

$$R^2 = R^2(d) := \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i} \in [0, 1].$$

- In practice: choose d such that $R^2 \approx 0.8 - 0.9$.

PCA: alternative view/definition

Recursive formulation (deflation approach):

- Assume: we have $\mathbf{w}_0$.
- Ask for the first PC of the residual ($\mathbf{w}_1$).
- Iterate.

This leads to an equivalent definition.

Summary

- Dimensionality reduction.
- Approximation with a 'small' dimensional subspace.
- Top eigenvectors/-values of the observation covariance.

We covered the '$\mathbb{R}^d$ part' of Chapter 8 in [1], and [3].