### Dimensionality Reduction

**Vector Quantisation:**

If your data is clustered, you can approximate each input by its cluster assignment. E.g. given a probability $p_i$ that $x_i$ is in cluster $k$.

$$x_i \rightarrow \left[ \frac{p_i}{k} \right] \in \text{New } k\text{-dimensional representation}.$$ 

**Covariances:**

- **Random Variables** $X^A, X^B \sim P(X^A, X^B)$ with $\mu^A = E[X^A] = 0$, $\mu^B = E[X^B] = 0$
- **Variance**: $\text{Var}(X) = E[(X^A - \mu^A)^2] = E[X^2]$
- **Covariance**: $\text{Cov}(X^A, X^B) = E[(X^A - \mu^A)(X^B - \mu^B)] = E[X^AX^B]^\top$; $\text{COV}(X^A, X^B) = \text{VAR}(X)$

$E[XY] = \begin{cases} 
>0 & \text{positively correlated: } i) X^i > 0, X^j > 0 \quad (\text{and vice versa}) \\
=0 & \text{uncorrelated} \\
<0 & \text{negatively correlated: } i) X^i > 0, X^j < 0 \quad (\text{and vice versa}) 
\end{cases}$

**Covariance Matrix**: If $X \sim P$ is a vector $X = [X_1, X_2, \ldots, X_d]^\top$

Assume data $D = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$

$$\mu = E[X] = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\text{weak law of large numbers.}$$

$$\mu = E[X] = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\text{In Covariance Matrix of all r.v. in } X, \text{ i.e. } X_1, X_2, \ldots, X_d$$

$$C_{ij} = \text{Cov}(X_i, X_j)$$

$$C_{xx} = \text{VAR}(X)$$

**Match:**

\[
\begin{bmatrix}
1 & 2 \\
3 & 4
\end{bmatrix}
\]
**Principal Component Analysis**
(Pearson 1901)

Data \( \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d \) but are truly from a lower-dimensional subspace reed.

Idea: Find basis vectors for this subspace and project data onto it. \( \Rightarrow \) Leads to \( r \)-dimensional representation

**Step 1 of PCA:** Center data

\[ \mathbf{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \]
subtract mean: \( \mathbf{\bar{x}} \rightarrow \mathbf{x}_i - \mathbf{\bar{x}} \)

**Step 2:** Find first principal component

PCA finds the subspace that contains maximum variance.

**PCA:** Find \( \mathbf{u} \) s.t. after projection, \( \mathbf{u}^T \mathbf{x}_i \) variance is maximized.

\[
\max_{\mathbf{u} \neq \mathbf{0}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{u}^T \mathbf{x}_i)^2 = \max_{\mathbf{u} \neq \mathbf{0}} \mathbf{u}^T \mathbf{S} \mathbf{u} = \max_{\mathbf{u} \neq \mathbf{0}} \mathbf{u}^T \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} = \mathbf{u}^T \mathbf{\Sigma} \mathbf{u} = \lambda \mathbf{u}
\]

We only care about the direction.

Lagrangian:

\[
\max_{\mathbf{u}} \min_{\lambda \geq 0} \mathbf{u}^T \mathbf{\Sigma} \mathbf{u} - \lambda (\mathbf{u}^T \mathbf{u} - 1)
\]

Lagrangian:

\[
\max_{\mathbf{u}} \min_{\lambda \geq 0} \mathbf{u}^T \mathbf{\Sigma} \mathbf{u} - \lambda (\mathbf{u}^T \mathbf{u} - 1) = \mathbf{u}^T \mathbf{\Sigma} \mathbf{u} - \lambda = 0 \Rightarrow \mathbf{u}^T (\mathbf{\Sigma} - \lambda \mathbf{I}) \mathbf{u} = 0 \Rightarrow \mathbf{\Sigma} \mathbf{u} = \lambda \mathbf{u}
\]

\( \mathbf{u} \) is an eigenvector of \( \mathbf{\Sigma} \)

\( \mathbf{\Sigma} \) has \( d \) eigenvectors: \( \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d \) s.t. \( \mathbf{u}_i = \lambda_i \mathbf{u}_i \).

Sort eigenvectors such that \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \)

\( \mathbf{u}_1 \) is the first (aka leading) principal component: \( \Rightarrow \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d] \)

\( \mathbf{u}_2 \) is the second,

**PCA:**

\[
\mathbf{x}_i \rightarrow \mathbf{U}(\mathbf{x}_i - \mathbf{\bar{x}}) \]
Reconstruction:

PCA dimensionality reduction: \( \tilde{z}_i = U^T(\bar{x}_i - \mu) \)

PCA reconstruction: \( \hat{x}_i = U \tilde{z}_i + \mu \)

Quit: proof that if \( r=d \) the reconstruction is perfect (i.e. \( \hat{x}_i = x_i \)).

PCA de-correlates dimensions:

Correlation matrix of \( z_1, \ldots, z_m \):

\[
C = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^T = \frac{1}{n} \sum_{i=1}^{n} U^T x_i (U x_i)^T = \frac{1}{n} \sum_{i=1}^{n} U^T x_i x_i^T U = \frac{1}{n} U (\sum_{i=1}^{n} x_i x_i^T) U
\]

\[
= \frac{1}{n} U^T C U \Rightarrow [C_{ij}] = U^T C U \triangleq \lambda_i \text{; } i \neq j\text{; } C \text{ eigenvectors}
\]

\[
[C] = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \Rightarrow C = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}
\]

How to pick \( r \): \( \lambda_i \) is the variance within \( r \)th PC.

If you project onto \( r \) dimensions you lose \( \left( \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} \right) \) fraction of the total variance.

\[
\left( \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} \right) \text{ Denoising: Pick smallest } r \text{ such that } \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq 0.95
\]

Singular Value Decomposition: \( X = U S V^T \)

(for centered data) \( \uparrow \text{ principal eigenvalues of } C \)

\( S \) is the projected data

principal components