Recitation 9. May 11

Focus: probability (discrete and continuous), random variables, principal component analysis (PCA)

A $\boxed{\textbf{random variable}}$ is a quantity X that takes values in $\mathbb{R}$. It can be either:

- **discrete**: X takes only countably many possible values x_i each with probability p_i
- **continuous**: X is associated to a probability distribution $p(x)$ (where $p : \mathbb{R} \to \mathbb{R}$ is a function).

The $\boxed{\textbf{mean}}$ (sometimes called "expected value") $E[X]$ of X is the quantity:

- $\boxed{\sum_i x_i p_i}$ if X is discrete
- $\boxed{\int_{-\infty}^{\infty} xp(x)\ dx}$ if X is continuous

The mean is linear: if X, Y are random variables and $a, b \in \mathbb{R}$, then $E[aX + bY] = aE[X] + bE[Y]$.

Given two random variables X, Y, their $\boxed{\textbf{covariance}}$ $\Sigma_{XY} = E[(X - E[X])(Y - E[Y])]$ is:

- $\boxed{\sum_{ij} p_{ij}(x_i - \mu)(y_j - \nu)}$ if X is discrete
- $\boxed{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu)(y - \nu)p(x, y)\ dxdy}$ if X is continuous

The covariance of X with itself is called the $\boxed{\textbf{variance}}$ Σ_{XX}.

Given n random variables $X_1, \ldots, X_n$, we may assemble them into a vector $\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, called a $\boxed{\textbf{random vector}}$.

The $\boxed{\textbf{covariance matrix}}$ of these random variables $X_1, \ldots, X_n$ is the matrix

$$K = \begin{bmatrix} \Sigma_{X_1X_1} & \cdots & \Sigma_{X_1X_n} \\ \vdots & \ddots & \vdots \\ \Sigma_{X_nX_1} & \cdots & \Sigma_{X_nX_n} \end{bmatrix} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T], \qquad \text{where} \quad \boldsymbol{\mu} = E[\boldsymbol{X}] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}$$

K is always positive semidefinite. It is positive definite unless a linear combination of $X_1, \ldots, X_n$ is constant.

$\boxed{\textbf{Principal component analysis (PCA)}}$ involves diagonalizing the covariance matrix:

$$\boxed{K = QDQ^T}$$

where Q is orthogonal and D is diagonal. This means that the random vector $\boldsymbol{Y} = Q^T\boldsymbol{X}$ has diagonal covariance matrix D, i.e. its entries are uncorrelated random variables (i.e. have covariance 0). In other words:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} q_{11} & \cdots & q_{n1} \\ \vdots & \ddots & \vdots \\ q_{1n} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \qquad \Rightarrow \qquad \Big\{ Y_i = q_{1i}X_1 + \cdots + q_{ni}X_n \Big\}_{i \in \{1, \ldots, n\}}$$

are linear combinations of $X_1, \ldots, X_n$ that are (by construction) uncorrelated. The individual variances of the random variables $Y_1, \ldots, Y_n$ are the diagonal entries of the diagonal matrix D.

1. Sample from the numbers 1 to 1000 with equal probabilities 1/1000, and look at the last digit of the sample, squared. This square can end with $X = 0, 1, 4, 5, 6$, or 9. What are the probabilities p_0, p_1, p_4, p_5, p_6 and p_9 that each of these digits occurs among the sample? Compute the mean and variance of X.

Solution:

2. Let A, H, and W denote random variables corresponding to the age, height, and weight of dogs at a local shelter, respectively. Suppose the random vector $\begin{bmatrix} A \\ H \\ W \end{bmatrix}$ takes two values, $\begin{bmatrix} 7 \\ 20 \\ 132 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ 24 \\ 120 \end{bmatrix}$ with probabilities p and $1-p$ respectively. Compute the covariance matrix of A, H, and W.

Solution:

3. Suppose now that the random variables A, H, W from above instead have the covariance matrix

$$K = \begin{bmatrix} 3 & -1 & 2 \\ -1 & 3 & -2 \\ 2 & -2 & 6 \end{bmatrix}.$$

Find three linear combinations of A, H, W which are pairwise uncorrelated random variables. What is the variance of each?

Solution:

4. Let X be a random variable, with mean μ and variance σ^2. Compute $E[X^2]$ in terms of μ and σ.

Solution: