Ensimag 2^{nd} year
MSIAM 1^{st} year

Statistical Analysis and Document Mining

Worksheet n°1

Exercise 1. Let U and V be two independent random variables with distribution uniform over $[0,1]$. Let $X = U + V$ and $Y = U - V$.

1. Compute the expectation and covariance matrix of $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$.

2. Prove that X and Y are uncorrelated but not independent.

Indications. Recall that the variance of the uniform distribution is $1/12$. Compute $E[X]$, $E[Y]$, $Var[X]$, $Var[Y]$ and $Cov(X,Y)$. The solution is $E[Z] = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $K_Z = \begin{pmatrix} 1/6 & 0 \\ 0 & 1/6 \end{pmatrix}$. X and Y are uncorrelated but not independent. It is possible because the vector is not Gaussian.

Exercise 2. Let X be a random vector in $\mathbb{R}^n$ and A be a deterministic $m \times n$ matrix.

1. Prove that $K_X = E\left[(X - E[X])(X - E[X])^T\right] = E[XX^T] - E[X]E[X]^T$.

2. Prove that $K_{AX} = AK_XA^T$.

3. Use 2. to derive again the result of exercise 1.

Indications. For 1 and 2, this is just a manipulation of vectors and matrices. For 3, write $Z = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$

Exercise 3. Let $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ be a Gaussian vector with mean $\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$.

1. Compute the density of this distribution.

2. Using $f_{Y|X=x}(y) = \dfrac{f_{(X,Y)}(x,y)}{f_X(x)}$, compute the distribution of Y given $X = x$.

3. What is the best prediction of Y given $X = x$?

Indications.

1. Use the general formula of the density of a Gaussian vector.
$$f_{(X,Y)}(x,y) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}(2x^2+y^2+2xy-8x-6y+10)\right]$$

2. $f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(y-(3-x)^2)\right]$ so the distribution of Y given $X=x$ is $\mathcal{N}(3-x,1)$.

3. The best prediction of Y given $X=x$ is $3-x$.

Exercise 4. Let X be a random variable in $L^2 = \{X; E[X^2] < +\infty\}$. By definition, the best approximation of X by a constant is the orthogonal projection of X on the space D of constant random variables. Prove that this best approximation is $E[X]$.

Indications. The orthogonal projection of X on D is the constant b such that $\|X-b\|^2 = \min_{a\in D}\|X-a\|^2$. $\|X-a\|^2 = Var[X]+(E[X]-a)^2$, so this norm is minimal for $a = E[X]$. The last formula can be written $\|X-a\|^2 = \|X-E[X]\|^2+\|E[X]-a\|^2$, which is the Pythagorean theorem applied to the triangle $(X, E[X], a)$.

Exercise 5. Let $\begin{pmatrix} X \\ Y \end{pmatrix}$ be a Gaussian vector in $\mathbb{R}^2$. Let $Z = Y - E[Y] - \frac{Cov(X,Y)}{Var[X]}[X-E[X]]$.

1. Compute $E[Z]$ and $Var[Z]$.
2. Prove that X and Z are independent.
3. Derive the distribution of Y given $X=x$.
4. Use 3. to derive again the result of exercise 3.

Indications.

1. $E[Z]=0$ and $Var[Z] = Var[Y] - \frac{Cov(X,Y)^2}{Var[X]}$.

2. $\begin{pmatrix} X \\ Z \end{pmatrix}$ is a linear transform of $\begin{pmatrix} X \\ Y \end{pmatrix}$, so it is also a Gaussian vector. $Cov(X,Z)=0$, so X and Z are independent.

3. The distribution of Y given $X=x$ can be derived from that of Z given $X=x$ by a translation. The distribution of Z given $X=x$ is the distribution of Z. Finally, the distribution of Y given $X=x$ is normal with mean $E[Y]+\frac{Cov(X,Y)}{Var[X]}[x-E[X]]$ and variance $Var[Y]-\frac{Cov(X,Y)^2}{Var[X]}$. For Gaussian vectors, the best prediction of Y given $X=x$ is affine.

4. In the case of exercise 3, the mean and variance are $3-x$ and 1.