Statistical Analysis and Document Mining
Preliminaries 1

- Random vectors
- Gaussian vectors
- Probabilistic foundations of regression

Grenoble INP ensimag

Random vectors

Random vector in $\mathbb{R}^n$: $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = (X_1, \ldots, X_n)^T$, where the X_i's are real random variables.

Expectation : $E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$

Covariance matrix : $K_X = \begin{pmatrix} k_{11} & \ldots & k_{1n} \\ \vdots & & \vdots \\ k_{n1} & \ldots & k_{nn} \end{pmatrix}$

where $k_{ij} = Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$. $k_{ii} = Var(X_i)$.

Linear transformation of a random vector : $AX + b$, where A is a deterministic $m \times n$ matrix and b is a deterministic vector in $\mathbf{R}^m$.

- Expectation : $E[AX + b] = AE[X] + b$
- Covariance matrix : $K_{AX+b} = K_{AX} = AK_X A^T$

Quadratic form of a random vector : $X^T AX$, where A is a deterministic $n \times n$ matrix.

- Expectation : $E\left[X^T AX\right] = E[X]^T AE[X] + trace(AK_X)$

  where $trace(A) = \sum_{i=1}^{n} a_{ii}$, sum of the diagonal elements.

Distribution of X

- Discrete case : Probability

$$\forall x = (x_1, \ldots, x_n)^T \in \mathbf{R}^n,\ P(X = x) = P(X_1 = x_1, \ldots, X_n = x_n)$$

- Continuous case : Density $f_X(x) = f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n)$

$$\forall B \in \mathcal{B}(\mathbf{R}^n),\ P(X \in B) = \int_B f_X(x)dx$$

Marginal distribution :

$$f_{X_i}(x_i) = \int_{\mathbf{R}^{n-1}} f_X(x_1, \ldots, x_n)dx_1, \ldots dx_{i-1}dx_{i+1} \ldots dx_n$$

Independence

If the components X_i of X are independent, we have :

- Discrete case : $P(X = x) = \prod_{i=1}^{n} P(X_i = x_i)$.
- Continuous case : $f_X(x) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

If $X_1, \dots X_n$ are independent, for all functions $g_1, \dots g_n$,

$$E\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n} E\left[g_i(X_i)\right]$$

If X and Y are independent real random variables, $Cov(X, Y) = 0$. But the converse is false.
If the components of a random vector X are independent, the covariance matrix K_X is diagonal.

Gaussian vectors

A random vector $X = (X_1, \ldots, X_n)^T$ is a **Gaussian vector** iff every linear combination of its components $\sum_{i=1}^{n} a_i X_i$ has a normal distribution.

X is said to have a normal distribution $\mathcal{N}_n(\mu, \Sigma)$ or $\mathcal{N}(\mu, \Sigma)$, where $\mu = E[X] \in \mathbf{R}^n$ and $\Sigma = K_X \in \mathbf{R}^{n \times n}$.

The density of X is :

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

Properties of Gaussian vectors

- For Gaussian vectors, independence and uncorrelation (covariance zero) are equivalent.
- Each sub-vector of a Gaussian vector is still a Gaussian vector.
- If $X \sim \mathcal{N}(\mu, \Sigma)$, $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.
- $X \sim \mathcal{N}(\mu, \Sigma)$ iff there exists A such that $X = AZ + \mu$, where the components of Z are iid $\mathcal{N}(0, 1)$ and $\Sigma = AA^T$.
- If $X \sim \mathcal{N}_n(\mu, \sigma^2 Id)$, $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu_i)^2 = \frac{1}{\sigma^2}\|X - \mu\|^2$ has the χ^2_n distribution.

Spectral representation of Gaussian vectors

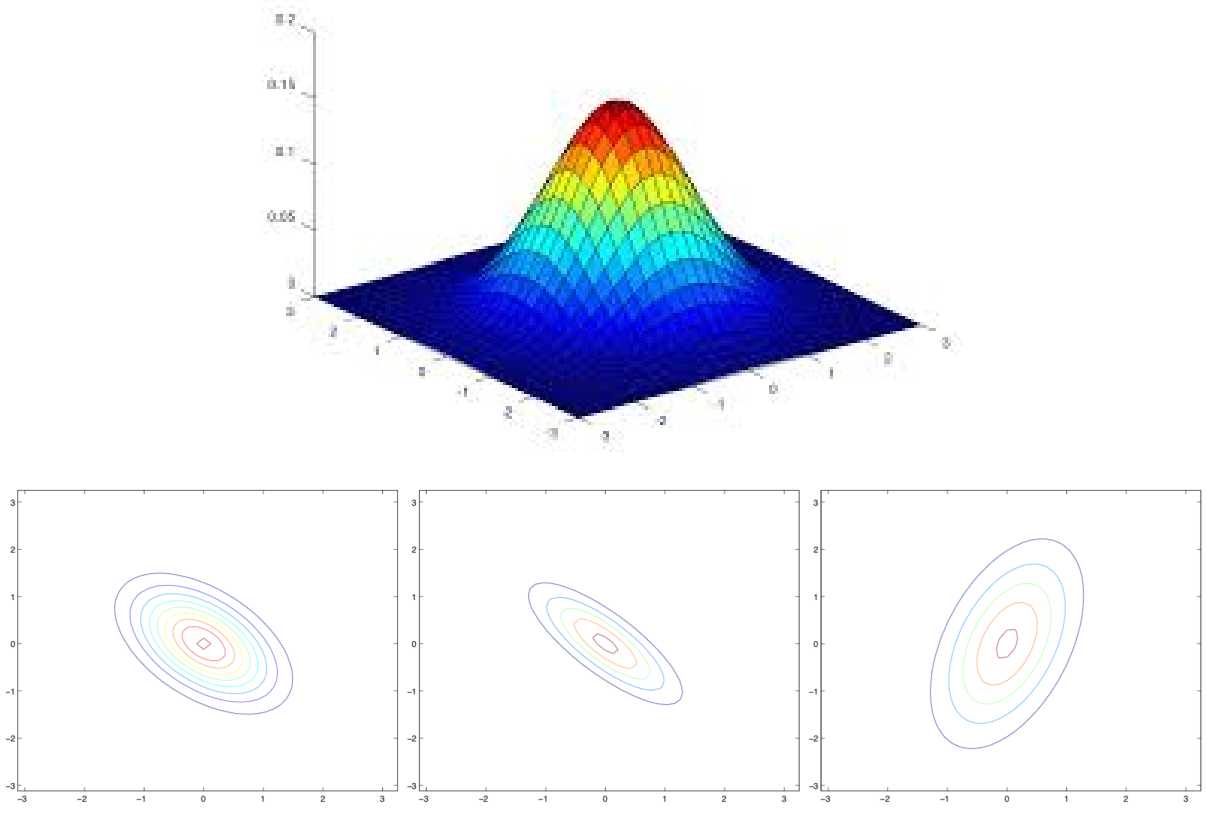
There exists an orthogonal matrix U ($UU^T = U^T U = Id$) such that $K_X = U\Lambda U^T$, where :

- Λ is the diagonal matrix made of the eigenvalues of K_X.
- The columns of U are the eigenvectors of K_X.

If $X \sim \mathcal{N}(\mu, \Sigma)$, $\Sigma = U\Lambda U^T$, then $Y = U^T(X - \mu) \sim \mathcal{N}(0, \Lambda)$. It means that any Gaussian vector can be linearly transformed in a Gaussian vector with centered and independent components.

The equidensity contours of a multivariate normal distribution are ellipsoids centered at the mean. The directions of the principal axes of the ellipsoids are given by the eigenvectors of Σ.

Example, bivariate normal distribution



Density and equidensity contours of a bivariate normal distribution

Probabilistic foundations of regression

X and Y : two real random variables.

If X and Y are dependent, it is logical to use the observed value x of X in order to predict the value of Y.

$\Longrightarrow$ we look for a function φ such that $\hat{Y} = \varphi(X)$ is "as close as possible" to Y.

Probabilistic foundations of regression

X and Y : two real random variables.

If X and Y are dependent, it is logical to use the observed value x of X in order to predict the value of Y.

$\Longrightarrow$ we look for a function φ such that $\hat{Y} = \varphi(X)$ is "as close as possible" to Y.

Criteria

- $E[Y - \varphi(X)] = 0$.
- $Var[Y - \varphi(X)]$ is minimal.

Best prediction of Y given X

X and Y are assumed to belong to $L^2 = \{X; E[X^2] < +\infty\}$.

Scalar product in L^2 : $< X, Y >= E[XY]$.
Associated norm : $\|X\|^2 = E[X^2]$. $Var[X] = \|X - E[X]\|^2$

The best approximation of a random variable X by a constant is the orthogonal projection of X on the space of constant random variables.

Best prediction of Y given X

X and Y are assumed to belong to $L^2 = \{X; E[X^2] < +\infty\}$.

Scalar product in L^2 : $< X, Y >= E[XY]$.
Associated norm : $\|X\|^2 = E[X^2]$. $Var[X] = \|X - E[X]\|^2$

The best approximation of a random variable X by a constant is the orthogonal projection of X on the space of constant random variables. This best approximation is $E[X]$.

Best prediction of Y given X

X and Y are assumed to belong to $L^2 = \{X; E[X^2] < +\infty\}$.

Scalar product in L^2 : $< X, Y >= E[XY]$.
Associated norm : $\|X\|^2 = E[X^2]$. $Var[X] = \|X - E[X]\|^2$

The best approximation of a random variable X by a constant is the orthogonal projection of X on the space of constant random variables.
This best approximation is $E[X]$.

Let $L^2_X = \{\varphi(X); \varphi : \mathbf{R} \to \mathbf{R}\}$.

The best prediction of Y given X is the orthogonal projection of Y on L^2_X.
This best prediction is $E[Y|X]$.

Useful results on the conditional expectation

- **Law of total expectation**

$$E[Y] = E\left[E[Y|X]\right]$$

- **Pythagorean theorem**

$$\|Y - E[Y]\|^2 = \|Y - E[Y|X]\|^2 + \|E[Y|X] - E[Y]\|^2$$

- **Law of total variance**

$$V[Y] = E\left[Var[Y|X]\right] + Var\left[E[Y|X]\right]$$

Best affine prediction of Y given X

Affine prediction : $\hat{Y} = \beta_1 X + \beta_0$.

Best affine prediction of Y given X

Affine prediction : $\hat{Y} = \beta_1 X + \beta_0$.

The best affine prediction of Y given X is

$$\hat{Y} = E[Y] + \frac{Cov(X, Y)}{Var(X)} [X - E[X]]$$