Lecture 7. Multivariate Normal distribution

Definition 1.

Let random vector $\vec{\xi} = (\xi_1, \dots, \xi_m)$ has $\vec{a} = E\,\vec{\xi}$, E – expectation and Σ – symmetric matrix, positive, non degenerate. It is said that the $\vec{\xi}$ has Normal distribution N ($\vec{a}$, Σ), if density function is equal to

$$f_{\vec{\xi}}(\vec{x}) = \frac{1}{(\sqrt{2\pi})^m \sqrt{|\det \Sigma|}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{a})^T \Sigma^{-1} (\vec{x} - \vec{a})\right\} \qquad (1)$$

for any $\vec{x} \in \mathrm{R}^m$.

<u>Statement:</u>

Covariance matrix of random vector $\vec{\xi}$, having density function (1) equal to Σ.

<u>Proof:</u> Let introduce vector $\vec{\eta}$, having multivariate standard Normal distribution and density function:

$$f_{\vec{\eta}}(\vec{x}) = \frac{1}{(\sqrt{2\pi})^m} \exp\left\{-\frac{1}{2}\,\vec{x}^T\,\vec{x}\right\}.$$

Matrix Σ – symmetric, positive, non degenerate, so according to Lemma2 of previous lecture $B = \sqrt{\Sigma}$ exists.

Let $\vec{\xi} = B\,\vec{\eta} + \vec{a}$. Construct $\vec{\xi}$ distribution:

$$f_{\vec{\xi}}(\vec{x}) = |\det B|^{-1}\, f_{\vec{\eta}}(B^{-1}(\vec{x} - \vec{a})) =$$

$$= \frac{1}{(\sqrt{2\pi})^m \sqrt{|\det \Sigma|}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{a})^T (\mathrm{B}^{-1})^{\mathrm{T}} \mathrm{B}^{-1} (\vec{x} - \vec{a})\right\} \qquad (2)$$

Here $(\mathrm{B}^{-1})^{\mathrm{T}}\mathrm{B}^{-1} = \mathrm{symm} = \mathrm{B}^{-1}\,\mathrm{B}^{-1} = \Sigma^{-1}$; we will show that $\vec{\xi}$ has distribution (2).

Expectation of $\vec{\xi}$ is equal $\vec{a}$, because $\mathrm{E}(B\,\vec{\eta} + \vec{a}) = B\ \mathrm{E}\ \vec{\eta} + \vec{a} = \vec{a}$. Find covariance matrix for $\vec{\xi}$:

$D\vec{\xi} = E(B\,\vec{\eta}\,(B\,\vec{\eta})^{T}) = B\,E\left(\vec{\eta}\,\vec{\eta}^{T}\right)B^{T} =$ [based on $E\left(\vec{\eta}\,\vec{\eta}^{T}\right) = D\vec{\eta} = \boldsymbol{E}$ – identity matrix] $= B\;B^{T} = \Sigma$.

Theorem 1.

1. Let random vector $\vec{\eta}$ consists of independent normal random variables; B – non degenerate matrix, then vector $\vec{\xi} = B\,\vec{\eta} + \vec{a}$ has multivariate Normal distribution $N(\vec{a}, \Sigma)$ with vector of average $\vec{a}$ and covariance matrix $\Sigma = B\;B^{T}$.

2. Let random vector $\vec{\xi}$ has Normal distribution $N(\vec{a}, \Sigma)$, Σ –positive matrix, then $\vec{\eta} = B^{-1}(\vec{\xi} - \vec{a})$, consists of consists of independent standard normal distributed random variables.

Remarks:

a) For normal distributed random vector $\in R^{m}$ with dependent coordinates having the density of joint distribution, the transformation by some non-degenerate matrix to vector with independent, normal distributed coordinates exists.

b) (inverse to (a)) standard normal distributed random vector can be transform to vector with given multivariate normal distribution.

Based on theorem 2 of **lecture 4-10**: multiplying by orthogonal matrix doesn't change joint distribution of standard normal vector coordinates.

Lemma. If vector $\vec{\xi}$ has covariance matrix $\Sigma = D\vec{\xi}$, then covariance matrix of B $\vec{\eta} = B\vec{\xi}$ is equal to $D\vec{\eta} = B\;\Sigma\,B^{T}$. But for *good matrix (symm., positive)* magic transformation exists.

$$Q\;\Sigma\,Q^{T} = D = diag(\sigma_1^2, \dots, \sigma_m^2) \qquad (3)$$

Q – orthogonal matrix, so for any $\vec{\xi}$ – normal

$$\vec{\eta} = Q\vec{\xi} \qquad (4)$$

transformation normal distributed vector with dependent coordinates to vector with independent coordinates (not necessarily identically distributed).

Theorem 2.

Let random vector $\vec{\xi}$ has multivariate Normal distribution N $(\vec{a}, \Sigma)$ with density function (2). Coordinates of this vector are independent if and only if they are correlated, that is when the covariance matrix is the diagonal.

Proof: Only in the case of diagonal matrix Σ with $\Sigma(i,i) = \sigma_i^2 = D\xi_i$ quadratic form $(\vec{x} - \vec{a})^T \Sigma^{-1} (\vec{x} - \vec{a})$ turns sum of squares

$$\sum\nolimits_{i,j} (x_i - a_i)\, \Sigma^{-1}(i,j)\big(x_j - a_j\big) = \sum\nolimits_i (x_i - a_i)^2 / \sigma_i^2$$

and density function decomposes into product of coordinates densities.

Remark: Uncorrelated random values may be dependent and even functionally dependent if their joint distribution doesn't have density (2).

Theorem 3. *Multivariate central limit theorem (MCL)*

Let $\vec{\xi}^{(1)}, \vec{\xi}^{(2)}, \ldots$ – a sequence of independent and identically distribute random vectors, each of them has average $\mathrm{E}\,\vec{\xi}^{(1)} = \vec{a}$ and non-degenerate covariance matrix Σ. Denote $S_n = \vec{\xi}^{(1)} + \ldots + \vec{\xi}^{(n)}$ – vector of partial sum. If n → ∞ then a weak convergence of vector distributions take place

$$\vec{\eta}^{(n)} = \frac{S_n - n\vec{a}}{\sqrt{n}} \Rightarrow \vec{\eta}, \qquad \vec{\eta} \in N(\vec{0}, \Sigma)$$

Consequence:

In terms of *MCL* theorem takes place convergence

$$\left\| \vec{\eta}^{(n)} \right\|^2 \Rightarrow \|\vec{\eta}\|^2$$

Distance and metrics

Definition 2. Function $d: X \times X \to R$ is called distance if it satisfies the following conditions:

a) nonnegative: $d(x,y) \geq 0$, for any $x,y \in X$

b) identity: $d(x,y) = 0$, if and only if $x=y$

c) symmetry: $d(x,y) = d(y,x)$, for any $x,y \in X$

d) inequality of the triangles: for any three points $x,y,z \in X$

$d(x,y) \leq d(x,z) + d(z,y)$

Definition 3.

Euclidean distance between two points $x=(x_1, \dots, x_p)^T$ and $y=(y_1, \dots, x_p)^T$ – is function of form

$$d_E(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(x-y)^T (x-y)}$$

$$d_E(x,0) = \|x\|_2 = [\text{– Euclidean norm of } x] = \sqrt{x^T x}$$

Definition 4. *Normalized Euclidean distance .*

$$\mathrm{d}(x,y) = d_E(u,v) = \sqrt{\frac{(x_1 - y_1)^2}{{s_1}^2} + \cdots + \frac{(x_p - y_p)^2}{{s_p}^2}} = \sqrt{(x-y)^T \mathrm{D}^{-1} (x-y)}$$

D – diag$({s_1}^2, {s_2}^2, \dots, {s_p}^2)$ – diagonal matrix; s_i – scaling factor

$$\|x\| = d(x,0) = d_E(u,0) = \sqrt{(x_1/s_1)^2 + \cdots + (x_p/s_p)^2} = \sqrt{x^T \mathrm{D}^{-1} x}.$$

The normalized distance is used also for characteristics, measured in different units or greatly different values.

Definition 5.

The statistics distance or Mahalanobis distance between two points $x=(x_1, \dots, x_p)^T$ and $y=(y_1, \dots, x_p)^T$ is called function of the form

$$d_M(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$

$$d_M(x,0) = \|x\|_s = \sqrt{x^T S^{-1} x}$$

– norm of x, S – covariance matrix for N points;

$$S = \frac{1}{N-1} A^T A$$

here A has follow form:

$$A = \begin{pmatrix} (z_{11},\cdots,z_{1p})^T - (z_{01},\cdots,z_{0p})^T \\ \vdots \\ (z_{N1},\cdots,z_{Np})^T - (z_{01},\cdots,z_{0p})^T \end{pmatrix}$$

$(z_{01},\cdots,z_{0p})^T$ – fixed point, f.e. center of the class $(\bar{z}_1,\cdots,\bar{z}_p)^T$ and

$z_i = (z_{i1},\cdots,z_{ip})^T$, $z_i = (z_{j1},\cdots,z_{jp}) \in \mathrm{R}^p \quad for \quad \forall\, i,j$;

$$S_{ij} = \sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(z_{ki}-\bar{z}_i)(z_{kj}-\bar{z}_j),$$

$$d_M(x,\bar{x}) = \sqrt{(x-\bar{x})^T S^{-1}(x-\bar{x})}$$

Note, this distance usually use for classification or clusterization problems.