Miary zależności między zmiennymi

Mgr inż. Szymon Łukasik
szymonl@pk.edu.pl

Wprowadzenie

Celem niniejszego ćwiczenia jest zapoznanie Państwa z miarami zależności pomiędzy dwoma zmiennymi losowymi *X* i *Y*. Mogą one być również postrzegane jako miary zależności pomiędzy dwoma współrzędnymi X_1, X_2 zmiennej losowej wielowymiarowej. Zostaną omówione dwie podstawowe miary tego typu : wariancja i korelacja oraz geometryczna ich interpretacja.

Kowariancja

Wariancja może być postrzegana jako miara zmienności zmiennej losowej. Zdefiniujmy dla dwóch zmiennych *X, Y* miarę postaci:

$$\mathrm{cov}(X,Y) = E[(X - E[X]) \cdot (Y - E[Y])]$$

to tzw. kowariancja. Gdy jest ona większa od zera to można przypuszczać, że gdy zmienna *X* przyjmie wartość większą od swojej wartości oczekiwanej, to to samo stanie się ze zmienną *Y.* Jeżeli kowariancja jest mniejsza od zera to, gdy zmienna *X* przyjmie wartość większą od swojej wartości oczekiwanej, to zmienna *Y* przyjmie wartość mniejszą od *E[Y].* Proszę zwrócić uwagę, że $\mathrm{cov}(X,X) = \mathrm{var}(X)$ oraz na to, że miara ta jest symetryczna.

Współczynnik korelacji

Po dodatkowym unormowaniu kowariancji przez odchylenia standardowe tj.

$$\rho_{(X,Y)} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

otrzymujemy współczynnik zwany współczynnikiem korelacji.

Gdy jest on równy zero to można wnioskować o całkowitej liniowej niezależności dwóch zmiennych. Gdy 1 – zmienne są liniowo zależne tj. $Y = aX + b$ i $a > 0$. A gdy -1 to $a < 0$.

Współczynniki te często umieszcza się w macierzach – stąd macierz kowariancji i macierz korelacji:

$$COV = \Sigma = \begin{bmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{bmatrix}$$ i analogicznie:

$$R = \begin{bmatrix} 1 & \rho_{(X,Y)} \\ \rho_{(X,Y)} & 1 \end{bmatrix}$$

Macierze te są kwadratowe, a rozmiar zależy od ilości współrzędnych danej zmiennej losowej lub liczności zbioru zmiennych branych pod uwagę.

Estymatory dla kowariancji i współczynnika korelacji

Jak można estymować w/w charakterystyki mając próby $x_1, x_2, \ldots, x_m$ i $y_1, y_2, \ldots, y_m$?

$$\mathrm{cov}(X,Y) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{m-1}\sum_{i=1}^{m} x_i y_i - \frac{1}{m(m-1)}\sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i$$

i oczywiście:

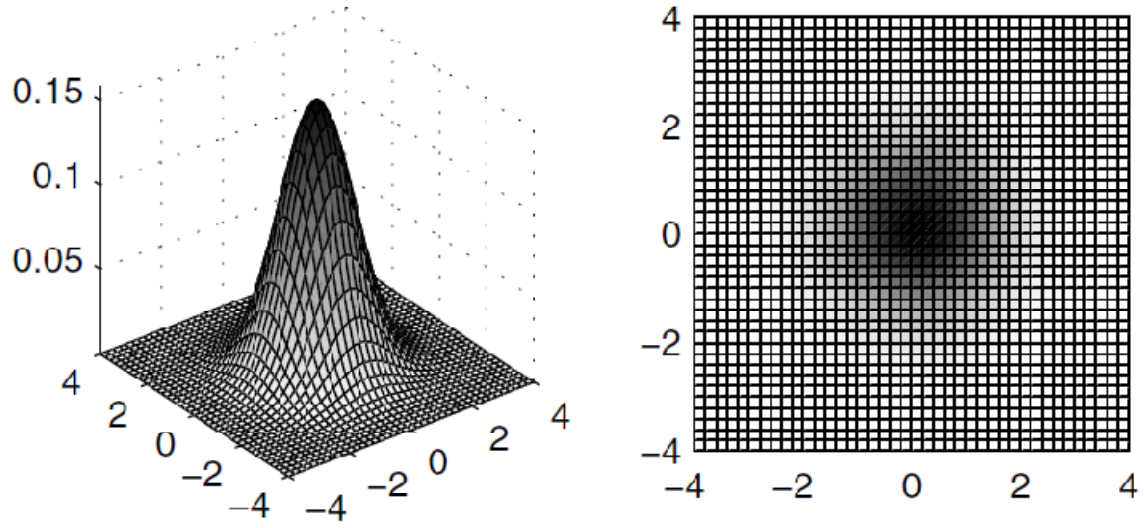
$$\hat{\rho}_{(X,Y)} = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

Rozkłady wielowymiarowe (na przykładzie rozkładu normalnego – ciekawostka, nie obowiązuje na kolokwium końcowym):

Niech x i μ stanowią wektory n-wymiarowe określające element próby losowej X oraz wartość oczekiwaną rozkładu (dla poszczególnych współrzędnych), natomiast Σ oznacza macierz kowariancji. Funkcja gęstości wielowymiarowego rozkładu normalnego przyjmuje wtedy postać:

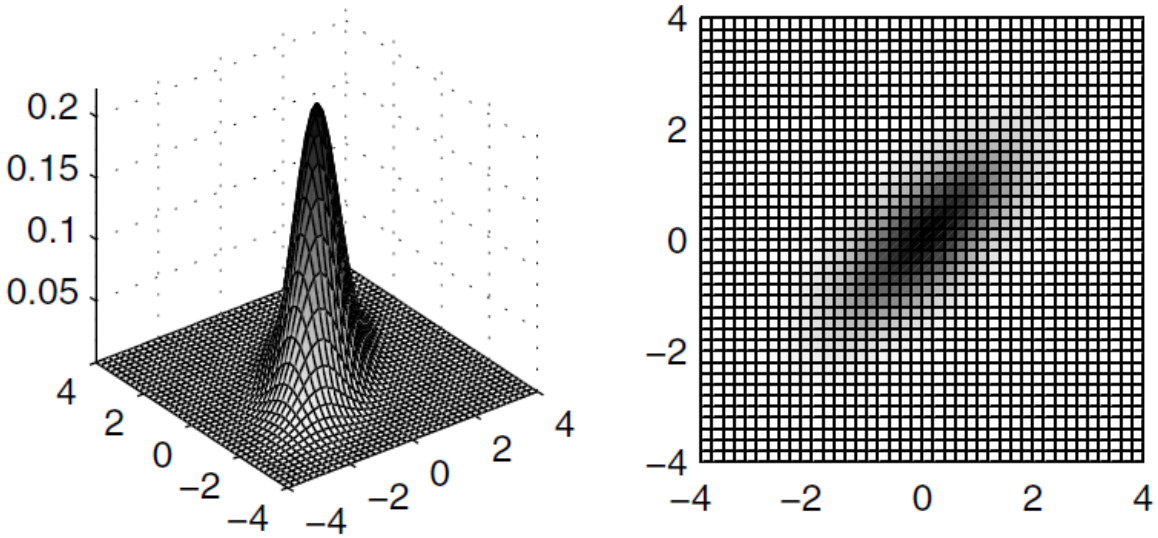
$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Jej ilustracja dla zerowego wektora wartości oczekiwanych oraz $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$:



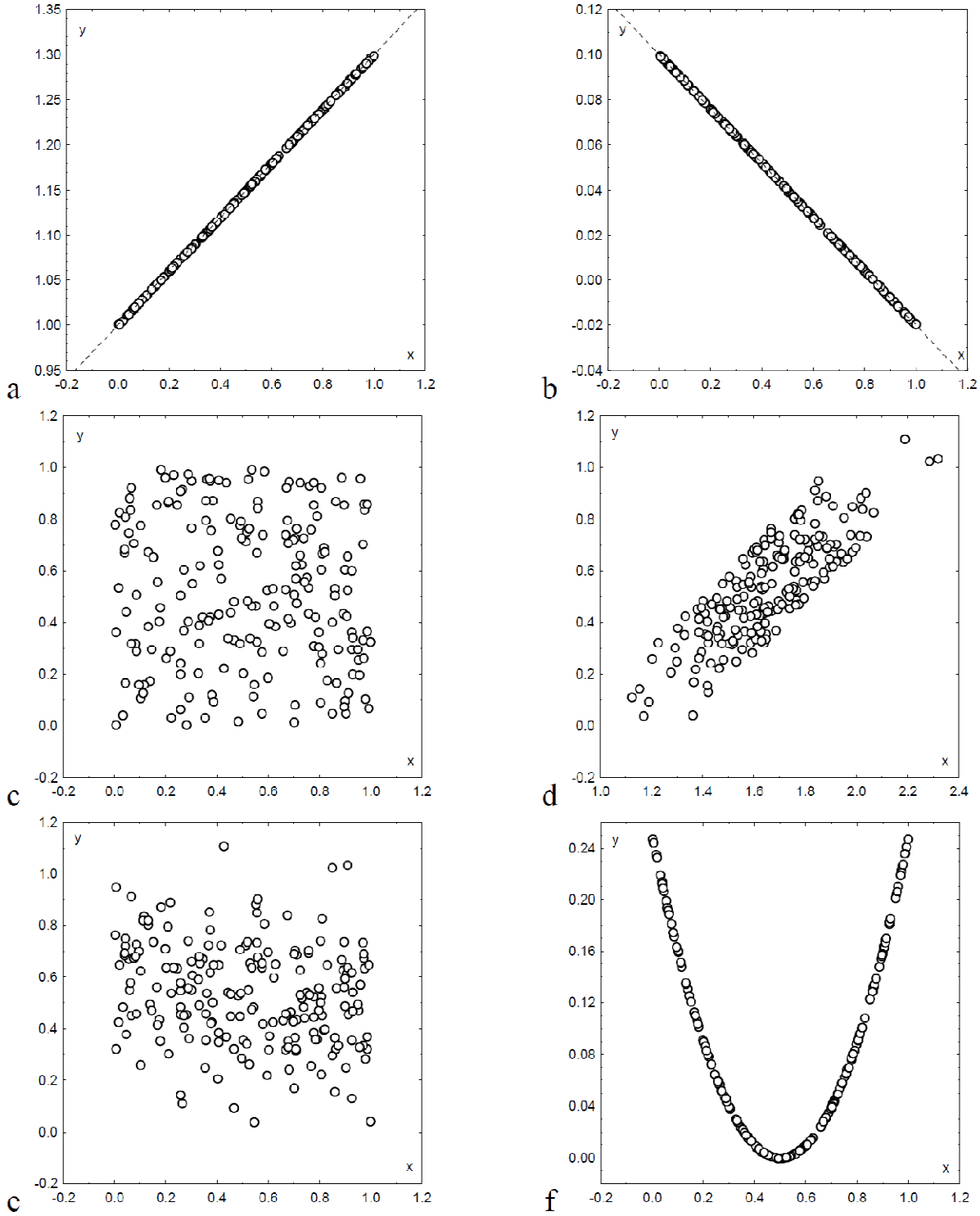
(źródło: Wendy L. Martinez and Angel R. Martinez, Computational Statistics Handbook with MATLAB, (Chapman & Hall/CRC, 2001).

A dla zerowego wektora wartości oczekiwanych oraz $\Sigma = \begin{bmatrix} 1 & 0{,}7 \\ 0{,}7 & 1 \end{bmatrix}$:



(źródło: Wendy L. Martinez and Angel R. Martinez, Computational Statistics Handbook with MATLAB, (Chapman & Hall/CRC, 2001).

Geometryczna interpretacja korelacji



Współczynniki korelacji: a) r = 1; b) r = –1; c) r = 0; d) r = 0.81; e) r = – 0.21; f) r = 0.04.
(źródło: Joaquim P. Marques de Sá, *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*, Springer, 2007).

Regresja liniowa

Zagadnienie regresji liniowej przy zmiennej X traktowanej jako zmienna niezależna (wyjaśniająca) i zmiennej Y traktowanej jako zależna (wyjaśniana), przy zastosowaniu

metody najmniejszych kwadratów, sprowadza się do określenia współczynników *a*, *b* takich że:

$$y_i = bx_i + a + \varepsilon_i, \quad i = 1, \dots, m$$

a suma kwadratów błędów ε_i:

$$s(a,b) = \sum_{i=1}^{m} {\varepsilon_i}^2 = \sum_{i=1}^{m} (y_i - bx_i - a)^2$$

była minimalna.

Współczynniki spełniające te założenia określa się następująco:

$$\hat{b} = \frac{cov(X,Y)}{var(X)}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Jeżeli zmienną Y opiszemy przy użyciu takiej krzywej regresji tj.:

$$\hat{y}_i = \hat{b}x_i + \hat{a}$$

To miarę dopasowania krzywej regresji do danych jest współczynnik determinacji (dopasowania) dany wzorem:

$$r^2 = \frac{\sum_{i=1}^{m} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}$$

i stanowiący kwadrat współczynnika korelacji.