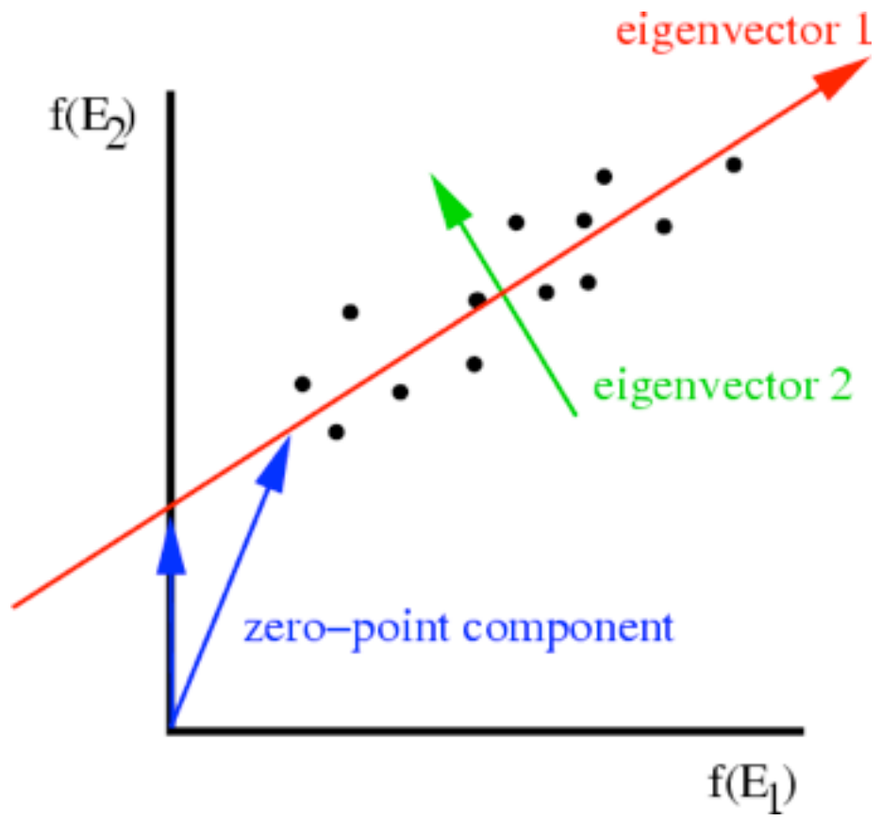
# Statystyczna analiza danych


- Anna Gambin
- Instytut Informatyki, Uniwersytet Warszawski

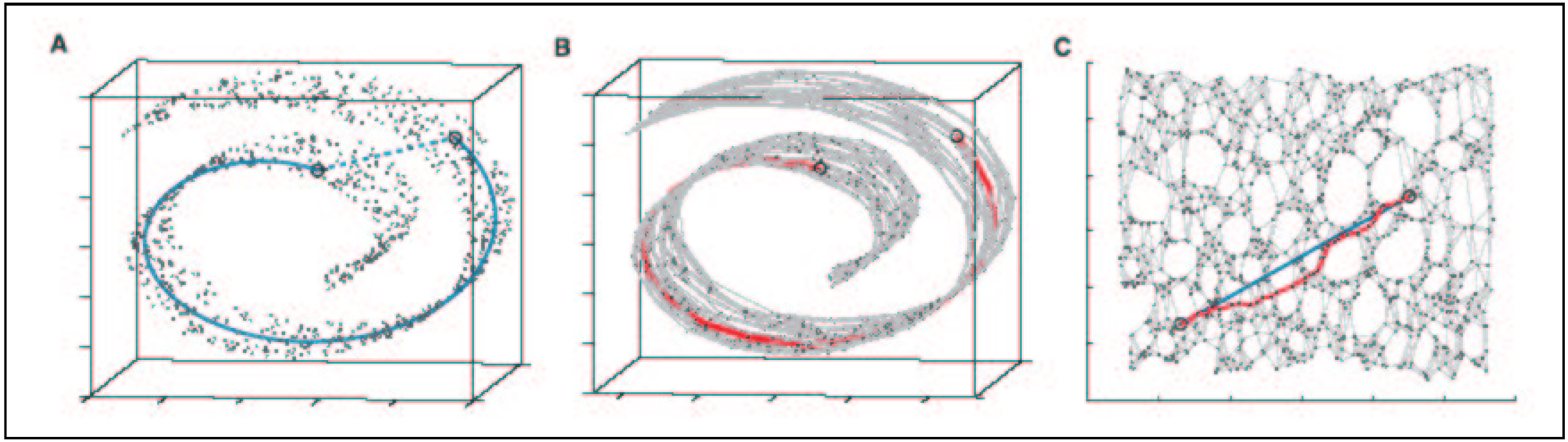



Omówione metody reprezentują podejście mające na celu **konstrukcję cech**, w którym $n$ cech wejściowych zostaje poddanych pewnej transformacji. W przypadku PCA transformacja ta jest przekształceniem liniowym, a nowe cechy odpowiadają zmienionej bazie, w której kierunki wektorów są tożsame z kierunkami największej wariancji w danych wejściowych. Przykładem metody wykorzystującej nieliniowe transformacje jest algorytm Isomap [5] opisany poniżej.

**Algorytm Isomap.** W 2000 roku Tenenbaum i de Silva zaproponowali algorytm Isomap do nieliniowej redukcji wymiaru (zobacz [5]). Jego celem jest znalezienie niskowymiarowej reprezentacji dla punktów leżących na nieliniowej powierzchni w przestrzeni o dużym wymiarze. Dla takich punktów liniowe metody redukcji wymiaru nie najlepiej zachowują relacje między punktami. Intuicyjny przykład daje rysunek 1. Głównym pomysłem, na którym opiera się algorytm Isomap jest wykorzystanie odległości ścieżkowych w grafie sąsiedztwa, do przybliżania geodezyjnych odległości miedzy punktami leżącymi na nieliniowej powierzchni. Algorytm składa się z następujących czterech kroków:

[5] J.B. Tenenbaum, V. de Silva, J.C. Langford, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, Science, 22 December, 2000.

1. Oblicz odległości euklidesowe między wszystkimi punktami.
2. Skonstruuj graf sąsiedztwa w następujący sposób:
   - Dowolne dwa punkty $a$ i $b$ są połączone krawędzią jeżeli punkt $a$ jest jednym z $k$ najbliższych sąsiadów punktu $b$ lub $b$ jest jednym z $k$ najbliższych sąsiadów punktu $a$ albo (w alternatywnej wersji algorytmu) odległość euklidesowa między punktami $a$ i $b$ jest nie większa niż ustalony parametr $\epsilon$.
   - Za odległość pomiędzy punktami połączonymi krawędzią przyjmuje się odległość euklidesową. Odległość pomiędzy nie połączonymi punktami incjalizowana jest na nieskończoność.
3. Oblicz najkrótsze ścieżki między wszystkimi punktami w grafie sąsiedztwa (np. przy użyciu algorytmu Dijkstry (por. [1])).
4. Wykonaj klasyczny algorytm skalowania wielowymiarowego (MDS) w celu znalezienia niskowymiarowego zanurzenia zachowującego odległości grafowe między punktami.



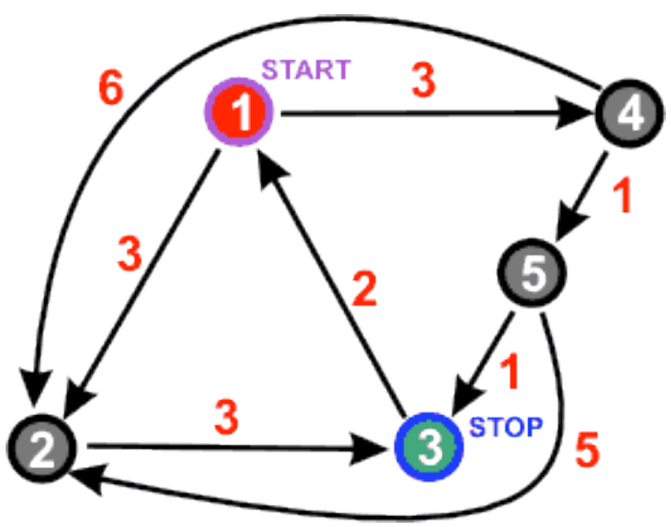

Rysunek 1: Dla wyróżnionych punktów leżących na nieliniowej powierzchni odległość euklidesowa nie opisuje poprawnie odległości geodezyjnej (rys. A). Rys. B przedstawia graf sąsiedztwa oraz ścieżkę między dwoma wyróżnionymi punktami. Na rys. C przedstawione jest zanurzenie w dwuwymiarową przestrzeń euklidesową najlepiej zachowujące odległości ściekowe między punktami. Ilustracja została zaczerpnięta z pracy [5].

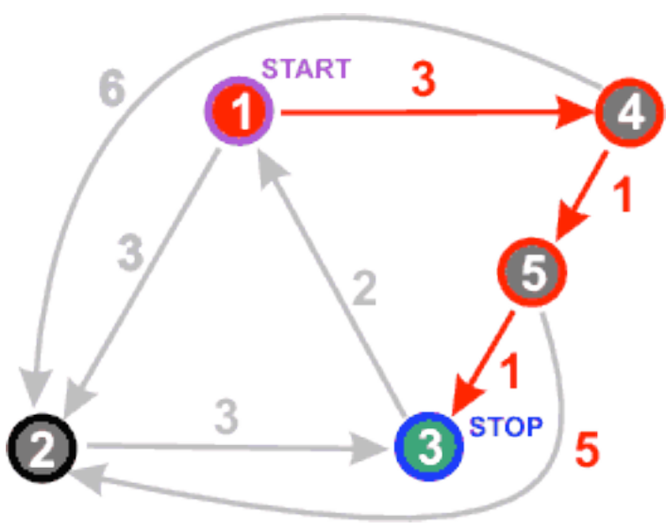
S (wierzchołki, dla których są policzone najkrótsze ścieżki)
Q (wierzchołki jeszcze nie przetworzone)

**Algorytm Dijkstry** znajduje w grafie wszystkie najkrótsze ścieżki pomiędzy wybranym wierzchołkiem a wszystkimi pozostałymi. Dodatkowo wylicza również koszt przejścia każdej z tych ścieżek. Bardzo ważnym założeniem w algorytmie Dijkstry są nieujemne wagi krawędzi. Jeśli krawędzie mogą przyjmować wagi ujemne, to **NIE WOLNO** stosować algorytmu Dijkstry.

prof. Edsger Dijkstra (1930-2002)

1972 - nagroda Turinga

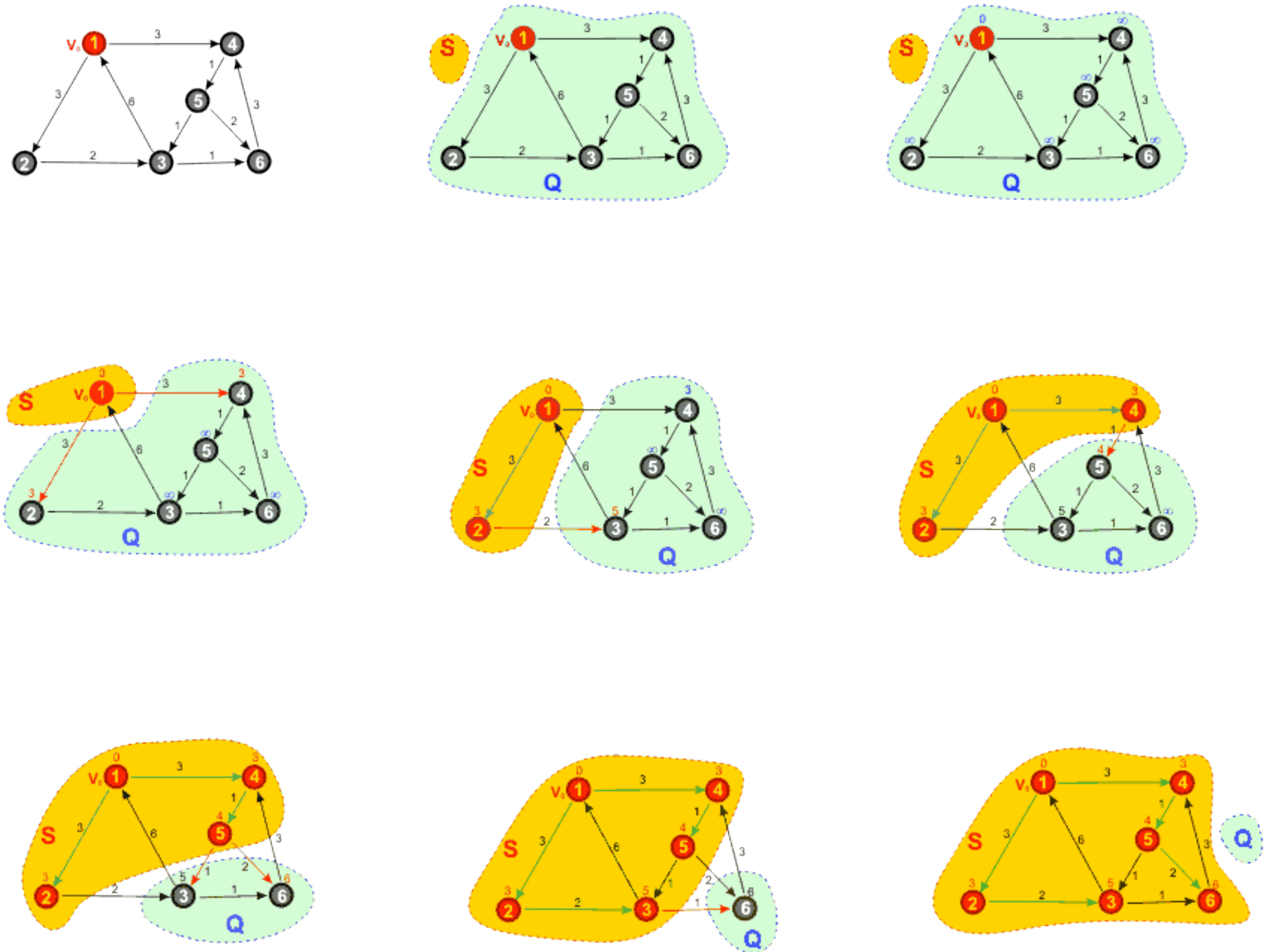


S (wierzchołki, dla których są policzone najkrótsze ścieżki)
Q (wierzchołki jeszcze nie przetworzone)

p(v) - poprzednik v
d(i) - koszt dojścia do i-tego

S ← Ø;  Q ← V
Dla i = 1,2,...,n :  p(i) ← 0;  d(i) ← ¥
$d(v_o)$ ← 0
Dopóki Q ≠ Ø wykonuj
u ← wierzchołek ze zbioru Q o najmniejszym d(u):
Q ← Q \ u;  S ← S U u:
Dla każdego v w Q, takiego że krawędź (u,v) jest w grafie, jeśli d(v) > d(u) + waga(u,v), to d(v) ← d(u) + waga(u,v);  p(v) ← u


S
Q

# Analiza składowych głównych -PCA

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{bmatrix}.$$

wartość oczekiwana i jej estymator:

$$E(\mathbf{x}) \equiv E\mathbf{x} = [E(x^{(1)}), \dots, E(x^{(p)})]^T$$

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i,$$

**Macierz kowariancji** opisuje rozproszenie rozkładu wielowymiarowego. Na diagonali macierzy kowariancji występują wariancje składowych wektora losowego, a na pozostałych pozycjach kowariancje między nimi:

$$\mathbf{\Sigma} \equiv Cov(\mathbf{x}) = E[(\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})^T] = [\sigma_{ij}]_{i,j=1}^{p},$$
$$\sigma_{ij} = E[(x^{(i)} - Ex^{(i)})(x^{(j)} - Ex^{(j)})].$$

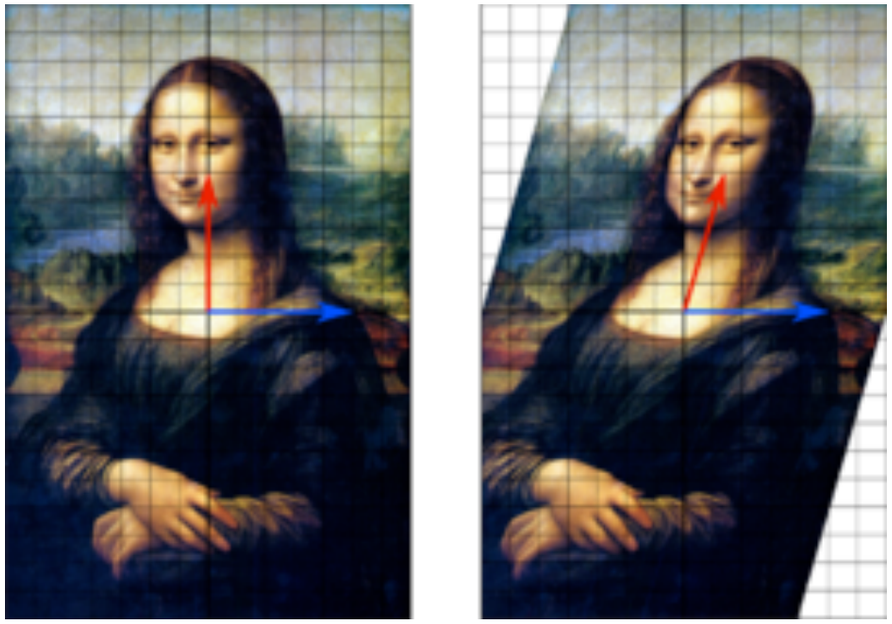
Macierz kowariancji estymujemy poprzez próbkową macierz kowariancji $\mathbf{\Sigma_X}$:

$$\mathbf{\Sigma_X} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

wektor własny

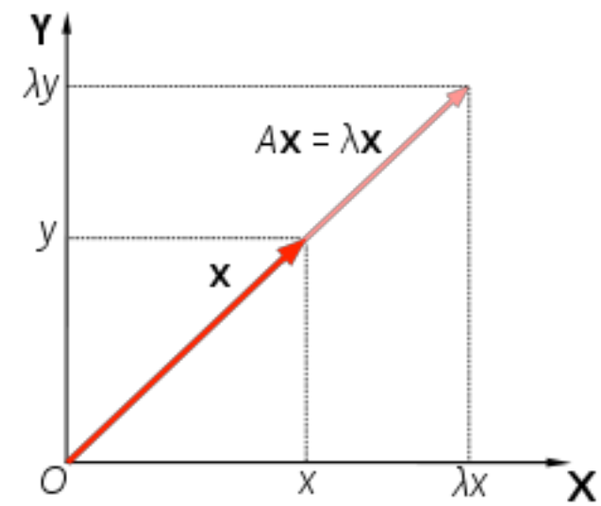
$$Av = \lambda v$$

wartość własna



który wektor jest wektorem własnym tego przekształcenia ?

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \qquad A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}.$$



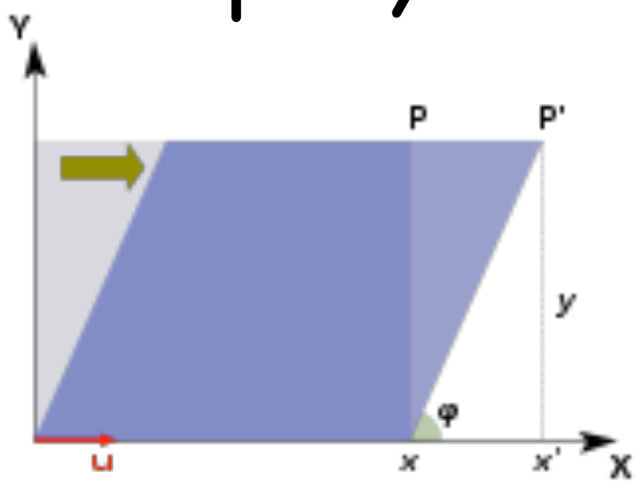

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \qquad \mathbf{x} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

$$A\mathbf{x} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 1 \cdot (-3) \\ 1 \cdot 3 + 2 \cdot (-3) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}.$$

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \cdot 0 + 1 \cdot 1 \\ 1 \cdot 0 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

# przykład: powinowactwo




przekształcenie

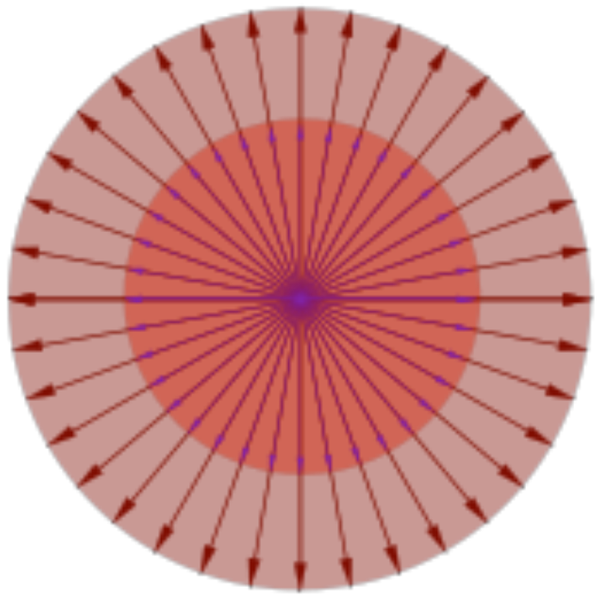
$$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$$

jego macierz

$$\mathbf{u}_1 = (1, 0)$$

jego wektor własny

przykład: podobieństwo



przekształcenie

$$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$$

jego macierz

jakie są jego wektory i wartości własne ???

# Analiza składowych głównych -PCA

Przy analizie składowych głównych będzie nas interesowała wariancja zmiennej losowej $\mathbf{a}^T\mathbf{x}$, czyli rzutu wektora losowego $\mathbf{x}$ na ustalowy wektor $\mathbf{a}$. Zobaczmy, że jest ona równa:

$$Var(\mathbf{a}^T\mathbf{x}) = E(\mathbf{a}^T\mathbf{x} - E(\mathbf{a}^T\mathbf{x}))^2 = E[\mathbf{a}^T(\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})^T\mathbf{a}] =$$
$$= \mathbf{a}^T Cov(\mathbf{x})\mathbf{a} = \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}.$$

# Analiza składowych głównych -PCA

Celem analizy jest wyznaczenie nowych nieskorelowanych zmiennych (składowych głównych wektora losowego), które będą miały największą możliwą wariancję. Nowych zmiennych będziemy szukać pośród kombinacji liniowych wektora losowego $\mathbf{x}$. Na początku znajdujemy wektor $\mathbf{v}_1$, taki że:

$$Var(\mathbf{v}_1^T\mathbf{x}) = \max_{\mathbf{a}\in\mathbb{R}^p, \mathbf{a}^T\mathbf{a}=1} \{Var(\mathbf{a}^T\mathbf{x})\} = \max_{\mathbf{a}\in\mathbb{R}^p, \mathbf{a}^T\mathbf{a}=1} \{\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}\}.$$

# Analiza składowych głównych -PCA

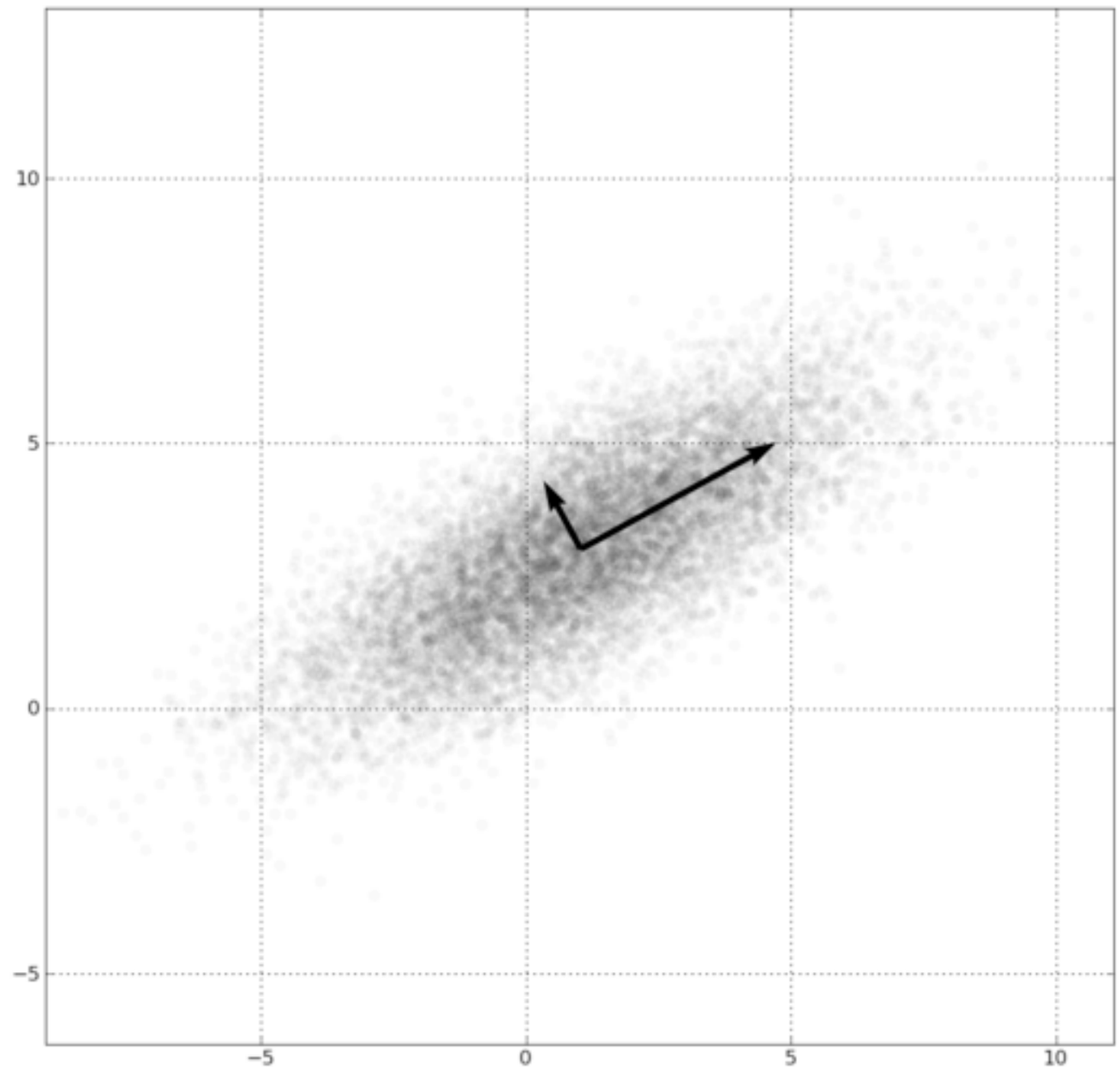
Poszukujemy więc takiego kierunku $\mathbf{v}_1$, by rzut ortogonalny wektora losowego $\mathbf{x}$ na ten kierunek dawał zmienną losową o maksymalnej wariancji. Pierwsza składowa główna dana jest przez:

$$\mathbf{v}_1^T(\mathbf{x}-\mathbf{m}),$$

gdzie $\mathbf{m}$ jest wartością oczekiwaną wektora losowego $\mathbf{x}$ (wektor losowy $\mathbf{x}$ centrujemy – nie zmienia to oczywiście wariancji). Kolejne główne składowe są zdefiniowane analogicznie, z tym że każda następna główna składowa ma być nieskorelowana ze wszystkimi wcześniejszymi. Wektor $\mathbf{v}_i$ $(i = 1, \ldots, p)$ nazywany wektorem współczynników $i$-tej składowej głównej jest wybrany tak by spełniał:

$$Var(\mathbf{v}_i^T\mathbf{x}) = \max_{\substack{\mathbf{a}\in\mathbb{R}^p,\ \mathbf{a}^T\mathbf{a}=1 \\ \forall_{1\leqslant h<i}\ E[(\mathbf{v}_h^T(\mathbf{x}-\mathbf{m})\mathbf{v}_i^T(\mathbf{x}-\mathbf{m}))]=0}} \{Var(\mathbf{a}^T\mathbf{x})\}.$$

Kolejne wektory $\mathbf{v}_i$ wyznaczają więc kolejne kierunki największej wariancji wektora losowego $\mathbf{x}$.




**Twierdzenie 1.** *Niech* $\mathbf{\Sigma}$ *będzie macierzą kowariancji wektora losowego* $\mathbf{x}$ *i niech* $\lambda_1 \geqslant \ldots \geqslant \lambda_p > 0$ *będą wartościami własnymi* $\mathbf{\Sigma}$. *Dla* $i = 1, \ldots, p$ *wektor* $\mathbf{v}_i$ *(i-ty wektor współczynników i-tej głównej składowej) dany jest poprzez i-ty (jednostkowy) wektor własny macierzy* $\mathbf{\Sigma}$, *odpowiadajacy wartości własnej* $\lambda_i$.

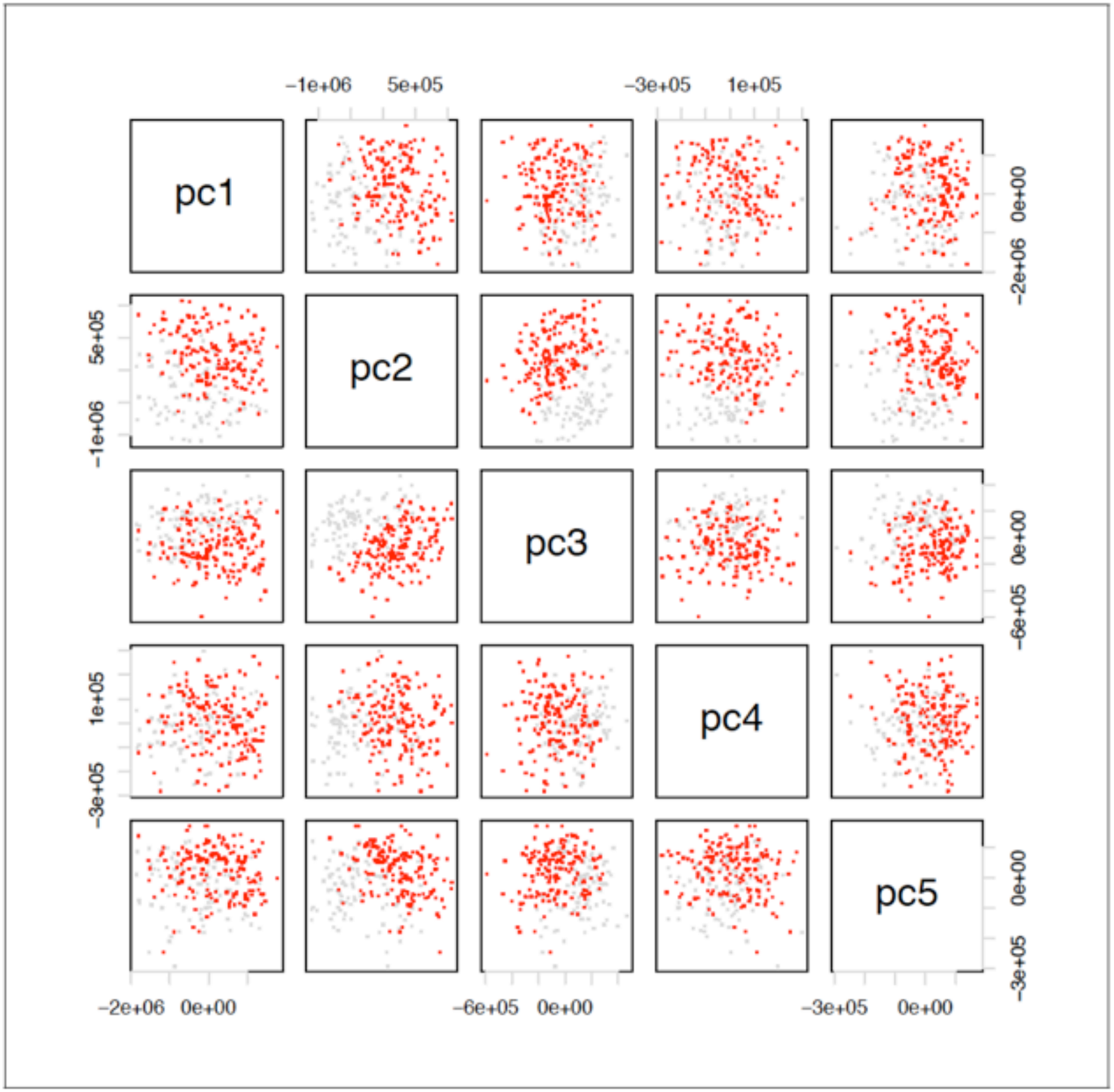
# Analiza składowych głównych
# -PCA

**Uwaga.** W założeniach Twierdzenia 1. przyjęliśmy, że wszystkie wartości własne macierzy kowariancji są dodatnie. Macierz kowariancji jest macierzą symetryczną i nieujemnie określoną więc jej wartości własne są rzeczywiste i nieujemne. Jeżeli niektóre z nich są równe zero, możemy ograniczyć się do właściwej podprzestrzeni $\mathbb{R}^p$, w której skupiony jest rozkład. Znajdziemy wtedy tyle głównych składowych ile jest dodatnich wartości własnych.

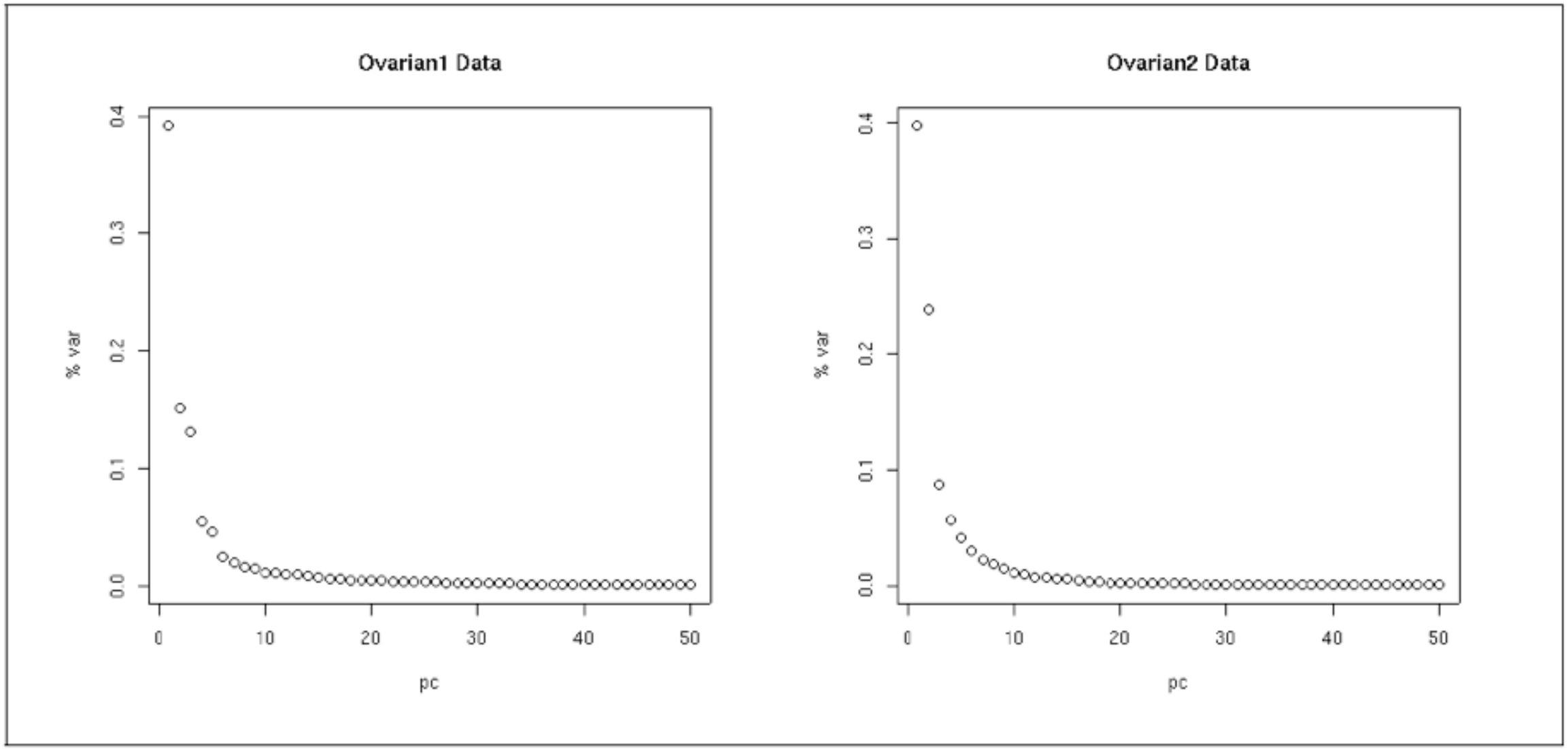
# Analiza składowych głównych -PCA

Najczęstszym zastosowaniem opisywanej metody jest redukcja wymiaru danych. Zadanie to polega na opisaniu danych o dużym wymiarze (dużej liczbie cech) przy pomocy mniejszej liczby cech, jednocześnie zachowując maksimum informacji . W przypadku PCA informacja ta jest mierzona wariancją. Analiza składowych głównych umożliwia opisanie wielowymiarowych danych przy pomocy małej liczby nieskorelowanych współrzędnych (wyznaczonych przez wektory własne macierzy kowariancji), zachowując rozrzut między danymi. Wymiar nowej przestrzeni będzie zależał od tego, jak dużą część wariancji będziemy chcieli zachować. Ponieważ macierz kowariancji jest symetryczna suma wyrazów na jej przekątnej (czyli suma poszczególnych wariancji) jest równa sumie wartości własnych tej macierzy, a każda składowa główna zachowuje wariancje równą odpowiadajacej jej wartości własnej, procent wariancji wektora wyjaśniony przez $k$ pierwszych składowych głównych można więc łatwo policzyć:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} 100\%.$$




Rysunek 1: Analiza składowych głównych dla zbioru danych z Clinical Proteomics Program Databank dotyczących raka jajnika. Rzutowanie obserwacji na układy współrzędnych wyznaczone przez pary pierwszych pięciu głównych składowych. Czerwone punkty odpowiadają chorym pacjentom, szare zdrowym. W tym przypadku rzutowanie na drugą i trzecią główną składową dobrze rozdziela obserwacje z różnych klas.




Rysunek 2: Procent wariancji danych wzdłuż kolejnych głównych składowych dla dwóch zbiorów danych z Clinical Proteomics Program Databank. Prawie cała wariancja jest zawarta w kilku pierwszych głównych składowych.

# Analiza składowych głównych -PCA

W praktyce zadanie PCA sprowadza się do znalezienia wektorów i wartości własnych próbkowej macierzy kowariancji $\mathbf{\Sigma_X}$. W przypadku gdy liczba cech przewyższa liczbę obserwacji ($p > n$), zamiast znajdywania wektorów własnych macierzy kowariancji $\mathbf{\Sigma_X}$ ($p \times p$) można rozwiązać mniejsze zadanie. Macierz kowariancji, można zapisać (pomijając czynnik stały) następująco:

$$\mathbf{\Sigma_X} = \mathbf{X'X'}^T,$$

gdzie $\mathbf{X'}$ jest macierzą powstałą poprzez odjęcie wartości średniej od kolumn macierzy $\mathbf{X}$:

$$\mathbf{X'} = [(\mathbf{x}_1 - \bar{\mathbf{x}}), \ldots, (\mathbf{x}_n - \bar{\mathbf{x}})].$$

Zadanie znajdywania wektorów własnych macierzy $\mathbf{X'X'}^T$ ($p \times p$) można zastąpić znajdywaniem wektorów własnych macierzy $\mathbf{X'}^T\mathbf{X'}$ ($n \times n$):

$$\mathbf{X'}^T\mathbf{X'}\mathbf{w_i} = \alpha_i \mathbf{w_i}.$$

Mnożąc obie strony równania przez macierz $\mathbf{X'}$ otrzymujemy wektory własne macierzy $\mathbf{X'X'}^T$, których szukamy:

$$(\mathbf{X'X'}^T)(\mathbf{X'w_i}) = \alpha_i(\mathbf{X'w_i}).$$

# selekcja cech

Innym podejściem do zadania redukcji wymiaru jest **selekcja cech**, polegająca na wyborze pewnego podzbioru spośród zbioru oryginalnych cech. W przypadku selekcji cech stosowanej w zadaniu klasyfikacji dokonujemy wyboru cech najlepiej dyskryminujących dwie populacje. Podsumujemy teraz korzyści, na jakie możemy liczyć stosując redukcję wymiaru danych.

- Redukcja wymiaru pozwala skupić się na cechach istotnych z punktu widzenia zadania klasyfikacji co pozwala z kolei zredukować efekt **przeuczenia**, polegający na zbytnim dopasowaniu klasyfikatora do danych treningowych.

- Poprzez eliminację szumu, czyli cech nieistotnych dla zadania klasyfikacji, zwiększamy skuteczność klasyfikatora.

- Dzięki redukcji wymiaru usuwamy problemy związane z wysokim wymiarem danych (problemy te były zasygnalizowane np. przy omawianiu metody najbliższych sąsiadów).

- Redukcja wymiaru danych pozwala w przypadku każdego klasyfikatora, na znaczną oszczędność czasową i pamięciową.

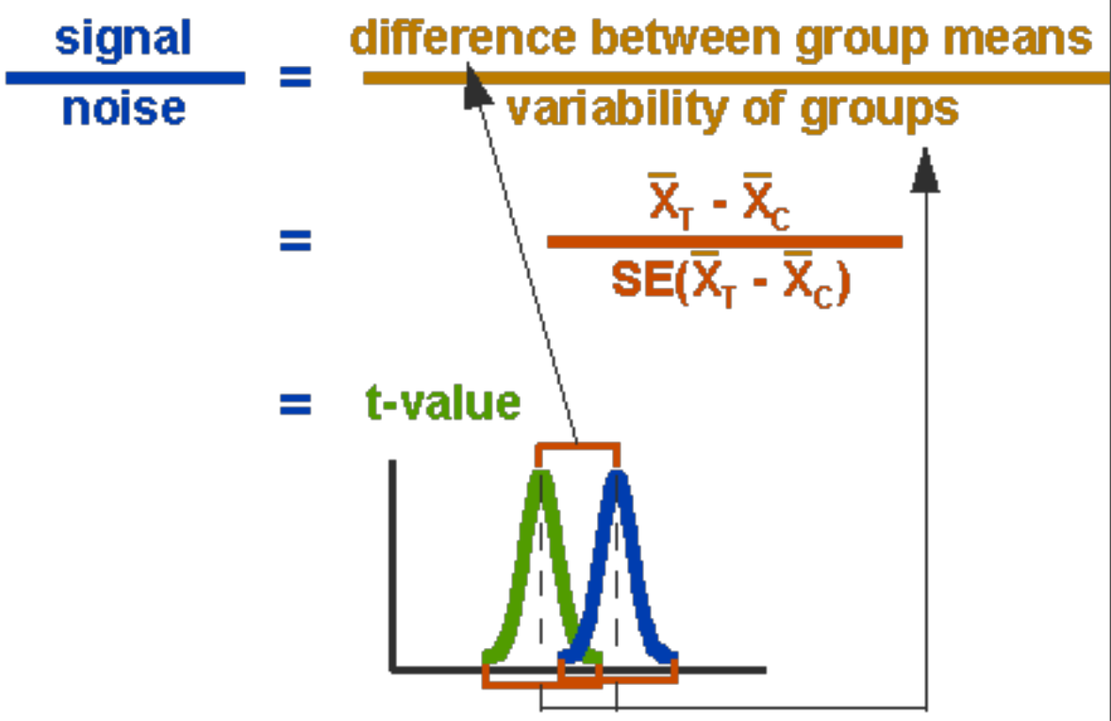
- Umożliwia lepszą interpretację sygnałów zidentyfikowanych w danych.

- W konkretnym przykładzie analizy danych mikromacierzowych, badanie wybranych 50 genów jest wykonalne (czasowo, finansowo, merytorycznie), natomiast analiza wszystkich (nawet 40000) próbek jest po prostu niemożliwa.

# 2 Metody wyboru cech dyskryminujących

Omówimy teraz popularne metody **selekcji cech**.

Zamiast szukać niskowymiarowego zanurzenia zachowujące odległości uwzględniające wszystkie zmienne, będziemy wybierać tylko te spośród oryginalnych zmiennych, które najlepiej (według pewnego testu) różnicują obserwacje z różnych klas. Umożliwia to jednocześnie bezpośrednią identyfikację potencjalnych biomarkerów.




## 2.1 Test t

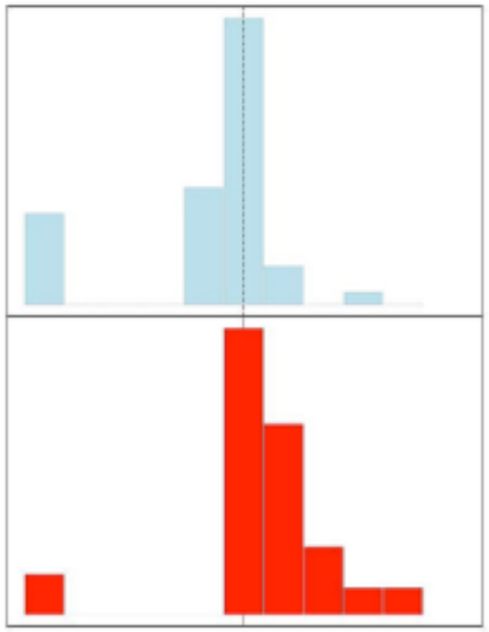
Jest to test na zgodność średnich dwóch prób losowych. Statystyka testowa dana jest wzorem:

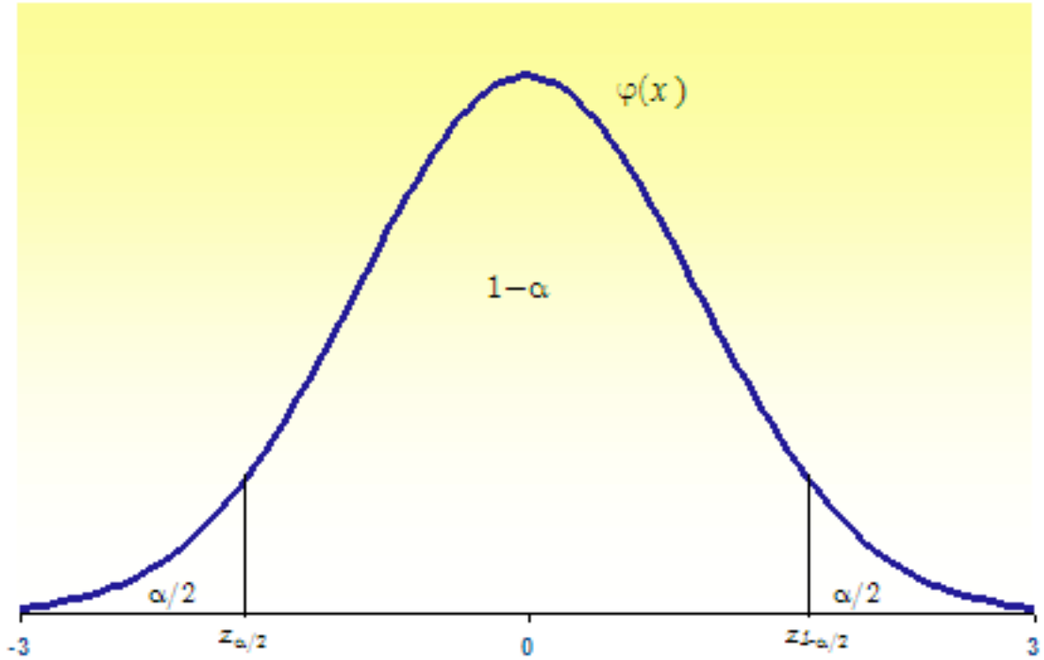
$$t(X_a, X_b) = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{\frac{\sigma_a}{n_a} + \frac{\sigma_b}{n_b}}},$$

gdzie $\overline{X}_a$ i $\overline{X}_b$ to wartości średnie, a $\sigma_a$ oraz $\sigma_b$ to wariancje prób $X_a$ i $X_b$ o licznościach $n_a$ i $n_b$. Wybieramy cechy o największej wartości statystyki.

## 2.2 Algorytm PPC

Algorytm PPC (*ang. Peak Probability Contrasts* (por. [6]) stosuje inne podejście do wyboru dyskryminujących cech. Omówimy je na przykładzie analizowanych w naszej grupie danych spektrometrycznych (każdy pacjent opisany jest przez *spektrum*, czyli wektor rzeczywisty reprezentujący ilości poszczególnych peptydów w surowicy krwi.) Peptydy odpowiadają wierzchołkom zidentyfikowanym w spektrum, a ich ilość jest proporcjonalna do wysokości wuerzchołka.



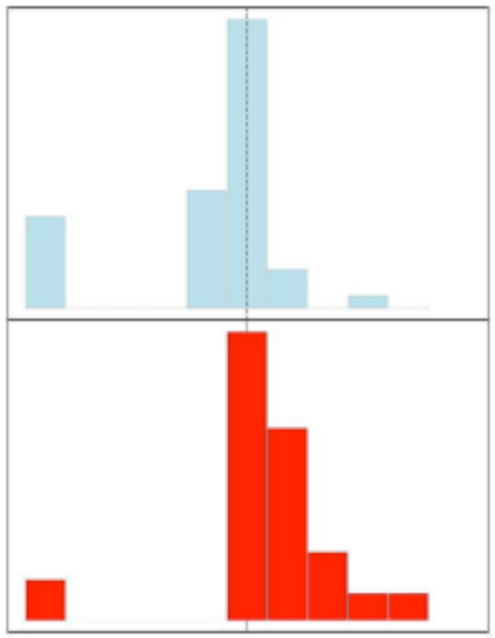



Zastosowane w algorytmi PPC podejście polega na wybraniu osobno dla każdej cechy (czyli peptydu) punktu przecięcia (ustalonej ilości peptydu) najlepiej różnicującego dwie grupy (np. chorych i zdrowych). Przecięć szuka

się wśród ustalonej liczby kwantyli. Przypomnijmy, że kwantylem rzędu $p$ ciągłej zmiennej losowej $X$ nazwiemy liczbę $q_p$ spełniającą warunek:

$$P(X < q_p) = p$$

Wybór przecięcia wśród kwantyli jest uzasadniony przez fakt, że wysokości wierzchołków (odpowiadające ilości danego peptydu) rozkładają się bardzo nierównomiernie. Kwantyl najlepiej rozdzielający wysokości wierzchołków z różnych klas jest wybrany następująco:



- Niech $q_{\alpha i}$ będzie kwantylem rzędu $\alpha$ zmiennej oznaczającej wysokość wierzchołka na $i$-tej pozycji w spektrum.
- Dla danych dwóch klas $G_1$, $G_2$, o licznościami $n_1$,$n_2$, niech $p_{il}(\alpha)$ będzie proporcją obserwacji w klasie $l$, w których wysokość na pozycji $i$-tej jest większa niż $q_{\alpha i}$:

$$p_{il}(\alpha) = \sum_{j \in G_l} I[x_{ij} > q_{\alpha i}]/n_l \quad l = 1, 2,$$

  gdzie $x_{ij}$ oznacza wysokość wierzchołka na $i$-tej pozycji widma u $j$-tego pacjenta, a $I[\cdot]$ jest indykatorem, równym 1 jeżeli wyrażenie jest prawdziwe i 0 w.p.p..
- Dla każdego $i$ spośród rozpatrywanych rzędów kwantyli $\alpha_1 \ldots \alpha_h$ wybierz rząd $\hat{\alpha}_i$ maksymalizujący po $\alpha$ wyrażenie $|p_{i2}(\alpha) - p_{i1}(\alpha)|$. Optymalny kwantyl (punkt przecięcia) dla $i$-tej cechy oznaczymy przez $\hat{q}_i = q_{\hat{\alpha}_i i}$. Częstości występowania wierzchołków o wysokości większej niż $\hat{q}_i$ w poszczególnych klasach oznaczymy przez $\hat{p}_{i1} = p_{i1}(\hat{\alpha}_i)$ oraz $\hat{p}_{i2} = p_{i2}(\hat{\alpha}_i)$.

W ten sposób możemy ustalić ranking cech (peptydów) według malejących wartości $|\hat{p}_{i2} - \hat{p}_{i1}|$.

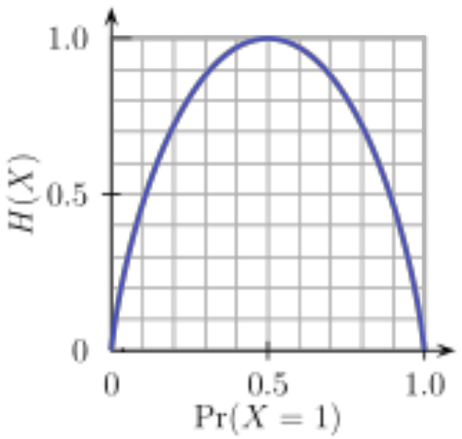
# ENTROPIA

$X$ zmienna losowa

przyjmujaca wartości $\{x_1, \ldots x_n\}$

$$H(X) = E(I(x)) = \sum_{i=1}^{n} p(x_i) I(x_i)$$

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$



## 2.3 Miary zróżnicowania między rozkładami prawdopodobieństwa

Dr. Solomon Kullback

Algorytm PPC nie bierze pod uwagę różnicy w wysokościach wierzchołków, które znajdują się powyżej lub poniżej ustalonego przecięcia. Zamiast wyznaczać pojedynczy punkt przecięcia, możemy porównywać rozkłady wysokości wierzchołków w zależności od klasy. Służą do tego miary zróżnicowania rozkładów prawdopodobieństwa. Jedną z najbardziej znanych miar jest odległość Kullbacka-Leiblera[1] zdefiniowana (w przypadku dyskretnym) następująco:

Dr. Richard A. Leibler

$$D_{KL}(p_1||p_2) = \sum_x p_1(x) log_2 \frac{p_1(x)}{p_2(x)},$$

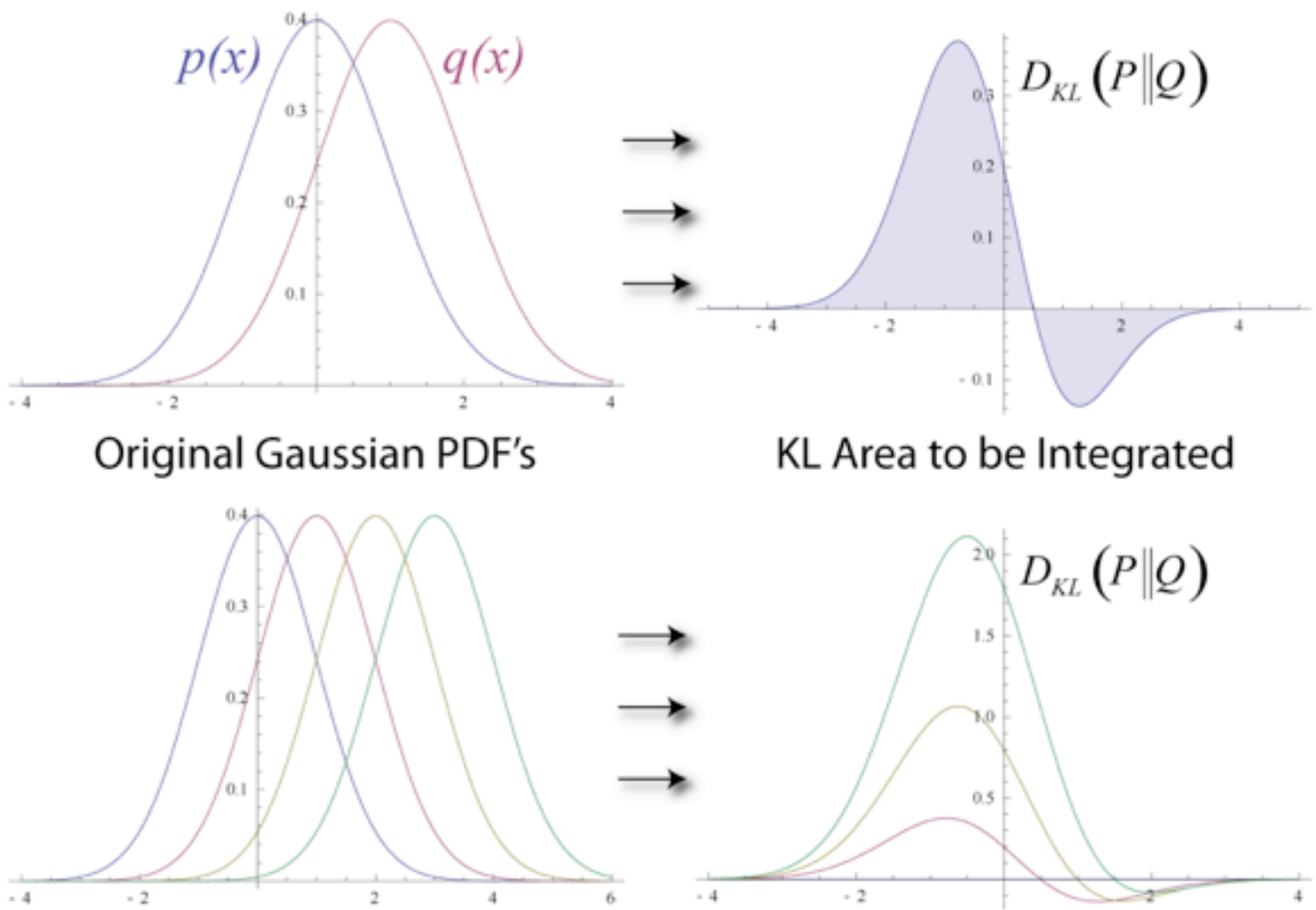
gdzie $p_1$ i $p_2$ są rozkładami zmiennej losowej $X$, a $x$ przebiega zbiór wartości tej zmiennej. Odległość Kullbacka-Leiblera jest nieujemna i równa 0 wtedy i tylko wtedy, gdy $p_1 \equiv p_2$. Miara ta jest jednak niesymetryczna. Chcąc

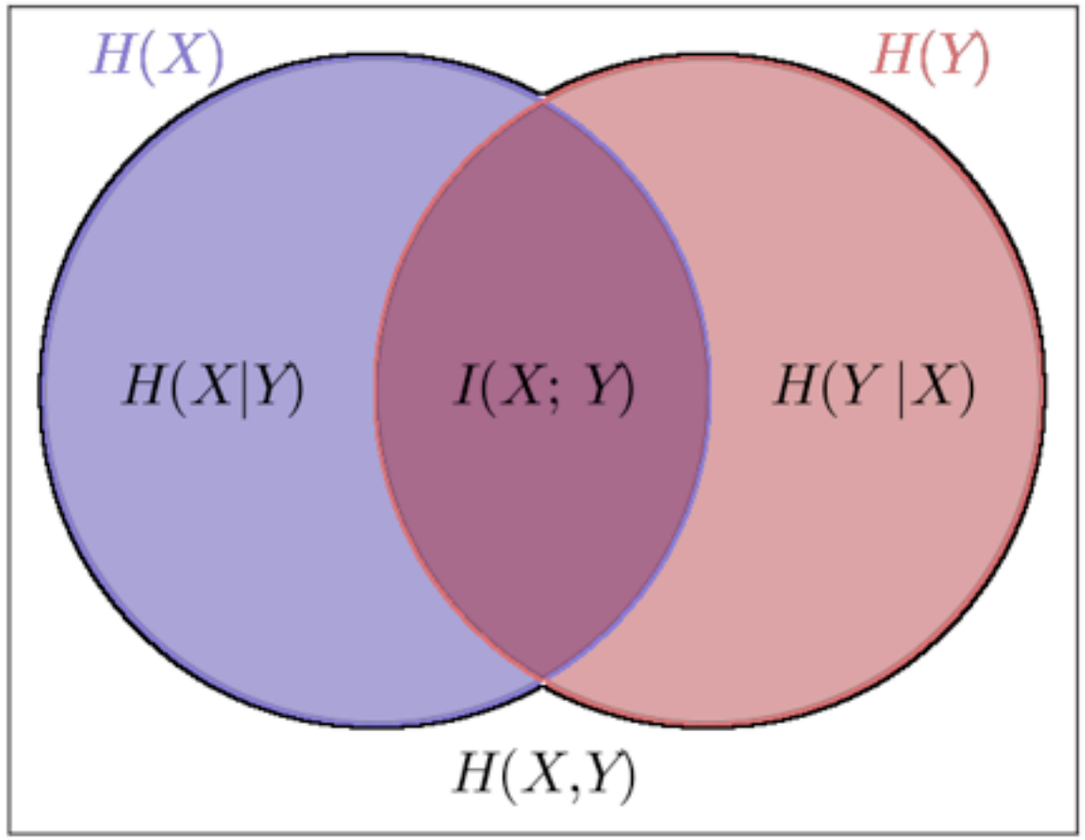
[1] Odległość Kullbacka-Leiblera mimo przyjętej nazwy nie jest oczywiście metryką.

jednakowo traktować obydwa rozkłady, możemy zastosować miarę *resistor-average* zaproponowaną przez D.H. Johnsona i S. Sinanovića. Jest to średnia harmoniczna odległości Kullbacka-Leiblera:

$$\frac{1}{D_{RA}(p_1, p_2)} = \frac{1}{D_{KL}(p1||p2)} + \frac{1}{D_{KL}(p2||p1)}.$$

Jeżeli $p_1 \equiv p_2$, to przyjmujemy $D_{RA}(p_1, p_2) = 0$.






## 2.4 Wzajemna informacja dwóch źródeł

Inną metodą wyboru istotnych cech jest zbadanie, jak dobrze każda z nich objaśnia zmienną decyzyjną. Można w tym celu zastosować współczynnik wzajemnej informacji wywodzący się z teorii informacji Shanona.

Niech źródło $A$ nadaje komunikaty ze zbioru $\Omega_1 = \{x_1, \ldots, x_m\}$, a źródło $B$ ze zbioru $\Omega_2 = \{y_1, \ldots, y_n\}$. Niech $p_1(x)$ oznacza prawdopodobieństwo otrzymania komunikatu $x$ ze źródła $A$, a $p_2(y)$ prawdopodobieństwo otrzymania komunikatu $y$ ze źródła $B$. Dalej, niech $p(x, y)$ oznacza prawdopodobieństwo otrzymania pary komunikatów $(x, y)$ odpowiednio ze źródeł $A$ i $B$. Wtedy współczynnik wzajemnej informacji źródeł $A$ i $B$ zdefiniowany jest następująco:

$$I(A, B) = \sum_x \sum_y p(x, y) \log_2 \frac{p(x, y)}{p_1(x) p_2(y)}$$

Współczynnik wzajemnej informacji jest nieujemny i równy 0 wtedy i tylko wtedy, gdy źródła nadają komunikaty niezależnie od siebie tzn. $p(x, y) = p_1(x)p_2(y)$. Wzajemna informacja jest także nie większa niż entropia każdego z dwóch źródeł.

W naszym przypadku będziemy badać wzajemną informację wektora wartości danej cechy w różnych próbkach oraz wektora wyznaczającego klasyfikację tych próbek.

z miary zmienności $V$, średnią zmienność warunkową zmiennej losowej $Y$ pod warunkiem $X$ opiszemy przez:

$$E[V(Y|X)] = \sum_x p(x)V(Y|x),$$

gdzie $V(Y|x)$ wyraża zmienność $Y$ pod warunkiem, że $X = x$:

$$V(Y|x) = 1 - \sum_y p(y|x)^2.$$

W związku z powyższym stopień zmniejszenia zmienności zmiennej $Y$ przy znajomości zmiennej $X$ nazywany współczynnikiem $\tau$ Goodmana-Kruskala zdefiniujemy następująco:

$$\tau_{GK} = \frac{V(Y) - E[V(Y|X)]}{V(Y)}$$

Współczynnik $\tau_{GK}$ osiąga wartość najmniejszą (równą 0) wtedy i tylko wtedy, gdy zmienne $X$ i $Y$ są niezależne, natomiast wartość największą (równą 1), gdy dla każdego $x$ przebiegającego zbiór wartości zmiennej $X$ istnieje $y$ (wartość zmiennej Y) taki, że $p(y|x) = 1$. W praktyce będziemy za zmienną $Y$ przyjmować zmienną decyzyjną, czyli zmienną oznaczającą klasę obserwacji, natomiast za zmienną $X$ badaną cechę.

## 2.6 Korelacja

Ostatnią opisaną przez nas metodą jest ustalenie rankingu cech według wartości bezwzględnej korelacji między badanymi cechami, a zmienną decyzyjną. Współczynnik korelacji próbkowej $\tau_P$ ma postać:

$$\tau_P = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}},$$

gdzie $\overline{x}$ i $\overline{y}$ oznaczają średnie w próbach $x_1, \ldots, x_n$ oraz $y_1, \ldots, y_n$.

# 3 Testowanie hipotez statystycznych

Przypomnijmy, że filtrujemy cechy sortując je względem wartości pewnej statystyki (jak w przypadku testu t). Chcielibyśmy znać próg dla tej wielkości pozwalający określić cechy o wartościach statystyki powyżej tego progu jako **statystycznie istotne**. W tym celu musimy rozwiązać podstawowe w statystyce zadanie **testowania hipotez**. Załóżmy na początku, że interesuje nas tylko jedna cecha. Klasyczne (nie bayesowskie) podejście do testowania hipotez sprowadza się do następujących kroków:

1. Sformułuj **hipotezę zerową** $H_0$ oraz **hipotezę alternatywną** $H_1$. Najczęściej hipoteza zerowa odpowiada sytuacji nieciekawej, średniej, nie wyróżniającej się cechy. Natomiast odrzucenie hipotezy zerowej, równoznaczne przyjęciu hipotezy alternatywnej sugeruje, że rozważana cecha w istotny sposób dyskryminuje dwie populacje. Ponieważ decyzję o przyjęciu lub odrzucecniu hipotezy podejmujemy na podstawie danych, które traktujemy jako próbę losową, czyli realizację pewnego procesu losowego mamy niezerową szansę pomyłki. Odrzucenie poprawnej hipotezy zerowej określamy jako **błąd typu I**, natomiast przyjęcie fałszywej hipotezy zerowej nazywamy **błędem typu II**. Łatwo dostrzec, że obydwie te wielkości są ze sobą powiązane i dlatego musimy wybrać, którą z nich chcemy kontrolować. W zastosowaniach najczęściej konsekwencję różnych typów błędów są niesymetryczne i zazwyczaj przyjmuje się, że interesuje nas utrzymanie błędów typu I na odpowiednio niskim poziomie $\alpha$ (np. $\alpha = 1\%$ lub $\alpha = 5\%$).

2. Ustal poziom $\alpha$ dla błędów typu I.

3. Sformułuj odpowienią statystykę testową, czyli obliczaną na podstawie danych wielkość, której wartość będzie odpowiadała za przyjęcie bądź odrzucenie hipotezy zerowej. Jest to bardzo ważny krok i łatwo się zgodzić, że wybór kiepskiej statystyki zaważy na jakości testu.

4. Określ, które wartości statystyki testowej prowadzą do odrzucenia hipotezy zerowej. Wybór tych wartości jest taki, aby kontrolować poziom błędów $\alpha$ założony w kroku 2. W tym celu przydatne okazuje się pojęcie **p-wartości** (*ang. p-value*). Dla danej wartości statystyki testowej p-wartość jest zdefiniowana, jako prawdopodobieństwo uzyskania tej lub bardziej ekstremalnej wartości przy założeniu hipotezy zerowej. Jeśli tak policzona p-wartość jest mniejsza niż zakładany poziom błędów $\alpha$, hipoteza zerowa zostaje odrzucona.

5. W ostatnim kroku analizujemy dostępne dane i sprawdzamy, czy wartość statystyki testowej odpowiada p-wartości pozwalającej nam odrzucić hipotezę zerową.

## 3.1 Przykład: hipoteza o zróżnicowanej ekspresji

Załóżmy, że porównujemy poziom ekspresji pewnego genu w dwóch populacjach komórek (np. pochodzących od $m$ zdrowych i $n$ chorych dawców). Niech zmienne $X_{11}, X_{12}, \ldots, X_{1m}$ oznaczają poziom ekspresji w zdrowych komórkach natomiast $X_{21}, X_{22}, \ldots, X_{2n}$ opisują populację chorych komórek. Zakładamy, że pomiary dotyczące poziomów eskpresji są niezależne i pochodzą z rozkładu normalnego o nieznanej wariancji $\sigma^2$ identycznej w obydwu grupach oraz nieznanych wartościach oczekiwanych $\mu_1$ oraz $\mu_2$. Naturalna hipoteza zerowa mówi, że obydwie rozważane wartości oczekiwane są idenyczne, czyli $\mu_1 = \mu_2 = \mu$. Natomiast hipoteza alternatywna mówi, że są różne $\mu_1 \neq \mu_2$, czyli badany gen w istotny sposób różnicuje dwie populacje. Okazuje, się że adekwatną statystyką w tym zadaniu jest statystyka t (używana w teście t) omówiona przy okazji selekcji cech.

# 4 Jednoczesne testowanie wielu hipotez

Wróćmy teraz do naszych wielowymiarowych danych. Filtrując cechy np. przy pomocy testu t chcielibyśmy zredukować wymiar danych poprzez wybór cech które w statystycznie istotny sposób różnicują dwie populacje. Zauważmy, że nasze zadanie jest równoważne problemowi jednoczesnego testowania wielu tysięcy hipotez zerowych: $H_1, H_2, \ldots, H_m$. Oznaczmy przez $R$ liczbę odrzuconych hipotez (czyli np. w przypadku mikromacierzy liczbę genów o różnicującej ekspresji). Mamy następującą sytuację:

| | # przyjętych $H_0$ | # odrzuconych $H_0$ | $\sum$ |
|---|---|---|---|
| # prawdziwych $H_0$ | $U$ | $V$ | $m_0$ |
| # fałszywych $H_0$ | $T$ | $S$ | $m_1$ |
| $\sum$ | $m - R$ | $R$ | $m$ |

gdzie $R$ jest obserwowaną zmienną losową, $m_0$ oraz $m_1$ nieznanymi parametrami, podobnie jak $U$, $V$, $T$ oraz $S$ są nieobserwowanymi zmiennymi losowymi.

Opiszemy teraz jak uogólnia się zadanie kontroli błędów typu I w problemie testowania wielu hipotez. W przypadku pojedynczego testu (hipotezę zerową oznaczamy tutaj dość myląco przez $H_1$) potrafiliśmy policzyć wielkość $c_\alpha$, taką że:

$$\Pr(|T_1| \geq c_\alpha | H_1) \leq \alpha$$

gdzie $T_1$ jest wartością statystyki testowej. Odrzucaliśmy hipotezę $H_1$ jesli $|T_1| \geq c_\alpha$. Najczęsciej stosowane uogólnienia tego podejścia są następujące (por. [2]):

[2] Dudoit,S., Shaffer,J.P. and Boldrick,J.C. *Multiple Hypothesis Testing in Microarray Experiments*, Statistical Science, Vol. 18, No. 1, p. 71-103.

| | # przyjętych $H_0$ | # odrzuconych $H_0$ | $\sum$ |
|---|---|---|---|
| # prawdziwych $H_0$ | $U$ | $V$ | $m_0$ |
| # fałszywych $H_0$ | $T$ | $S$ | $m_1$ |
| $\sum$ | $m - R$ | $R$ | $m$ |

- **PCER** (*ang. Per-comparison error rate*), miara jest zdefiniowana jako średnia z wartości oczekiwanej błędów typu I, czyli:

$$\mathrm{PCER} = \frac{E(V)}{m}$$

| | # przyjętych $H_0$ | # odrzuconych $H_0$ | $\sum$ |
|---|---|---|---|
| # prawdziwych $H_0$ | $U$ | $V$ | $m_0$ |
| # fałszywych $H_0$ | $T$ | $S$ | $m_1$ |
| $\sum$ | $m - R$ | $R$ | $m$ |

- **PFER** (*ang. Per-family error rate*), odpowiada oczekiwanej liczbie błędów typu I:

$$\mathrm{PFER} = E(V)$$

| | # przyjętych $H_0$ | # odrzuconych $H_0$ | $\sum$ |
|---|---|---|---|
| # prawdziwych $H_0$ | $U$ | $V$ | $m_0$ |
| # fałszywych $H_0$ | $T$ | $S$ | $m_1$ |
| $\sum$ | $m - R$ | $R$ | $m$ |

- **FWER** (*ang. Family-wise error rate*), jest zdefiniowana jako prawdopodobieństwo co najmniej jednego błędu typu I:

$$\mathrm{FWER} = \Pr(V \geq 1)$$

| | # przyjętych $H_0$ | # odrzuconych $H_0$ | $\sum$ |
|---|---|---|---|
| # prawdziwych $H_0$ | $U$ | $V$ | $m_0$ |
| # fałszywych $H_0$ | $T$ | $S$ | $m_1$ |
| $\sum$ | $m - R$ | $R$ | $m$ |

- **FDR** (*ang. False discovery rate*), definiujemy jako oczekiwaną proporcję błędów typu I pomiędzy odrzuconymi hipotezami zerowymi (jest to procent fałszywych pozytywów, czyli cech uznanych niesłusznie za istotne):

$$\mathrm{FDR} = E(Q)$$

$$Q = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases}$$

Omówmy bardziej szczegółowo tylko dwie ostatnie miary jako najczęściej stosowane w bioinformatyce. W przypadku FWER stosujemy tzw **poprawkę Bonferroniego** i odrzucamy hipotezę zerową $H_j$ $(j = 1, 2, \ldots, m)$ jeśli odpowiednia p-wartość jest mniejsza bądź równa $\frac{\alpha}{m}$ (gdzie $\alpha$ jest dopuszczalnym procentem błędów I typu w pojedynczym teście).

Sensowność poprawki uzasadnia następujące rozumowanie: załóżmy, że poprawne hipotezy zerowe to $H_1, \ldots H_{m_0}$. Niech $P_j$ będzie zmienną losową opisującą p-wartość dla hipotezy $H_j$, natomiast $\tilde{P}_j = \min(mP_j, 1)$ jej „poprawionym" odpowiednikiem. Mamy:

$$\mathrm{FWER} = \Pr(V \geq 1) = \Pr(\cup_{j=1}^{m_0}\{\tilde{P}_j \leq \alpha\}) \leq \sum_{j=1}^{m_0} \Pr(\tilde{P}_j \leq \alpha) \leq$$

$$\leq \sum_{j=1}^{m_0} \Pr(P_j \leq \frac{\alpha}{m}) = \frac{m_0 \alpha}{m}$$

Ostatnia równość wynika z obserwacji, że $P_j$ mają rozkład jednostajny na odcinku $(0, 1)$ przy założeniu hipotezy zerowej.

Problem związany z poprawką Bonferroniego jest taki, że ograniczając liczbę fałszywych pozytywów przez obniżenie progu p-wartości do $\frac{\alpha}{m}$ drastycznie zwiększamy próg statystyki testowej, co z kolei prowadzi do wielu fałszywych negatywów, czyli przyjęcia niepoprawnej hipotezy zerowej, a więc nie wykrycia istotnych cech.

## 4.1 Ocena istotności cech przy użyciu testu FDR

Opiszemy teraz praktyczną implementację metody FDR (False Discovery Rates). Dla danego poziomu $t$ oznaczymy przez $T$ liczbę cech, dla których wartość badanego testu przekracza $t$. Następnie wielokrotnie ($R$ razy) permutujemy etykiety klas obserwacji, za każdym razem obliczając wartość testu dla każdej cechy oraz oznaczając liczbę cech, dla których wartość testu przekracza $t$ przez $T_i^*$ $(i = 1, \ldots, R)$. Wartość FDR dla poziomu $t$ będziemy estymować wzorem:

$$\widehat{FDR}(t) = \frac{\sum_{i=1}^{R} T_i^* / R}{T}.$$

Wyznaczamy taki poziom $\hat{t}$, dla którego FDR jest niskie (np. 0.05). Cechy, dla których wartość testu przewyższa $\hat{t}$ będziemy traktować jako istotne.