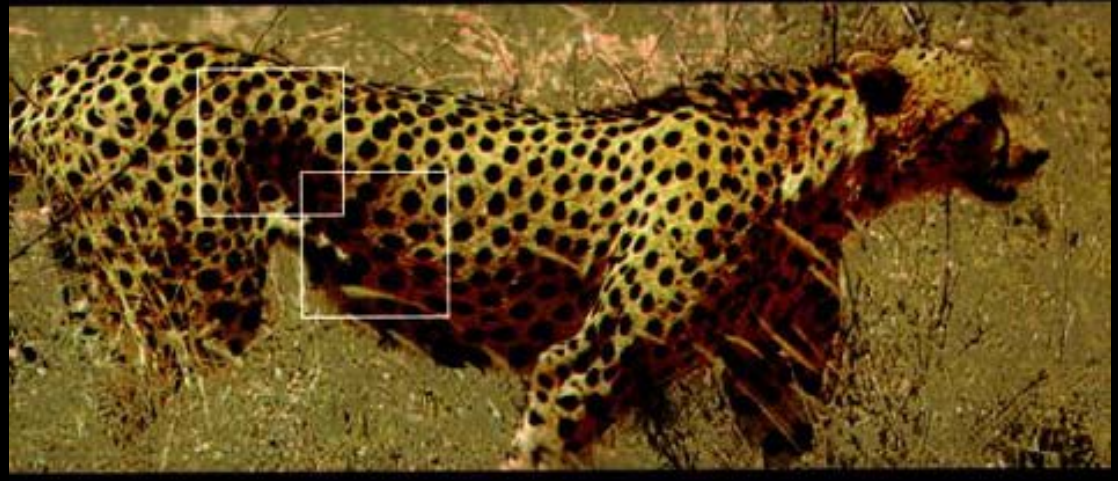


# Statistical Pattern Recognition

Second Edition

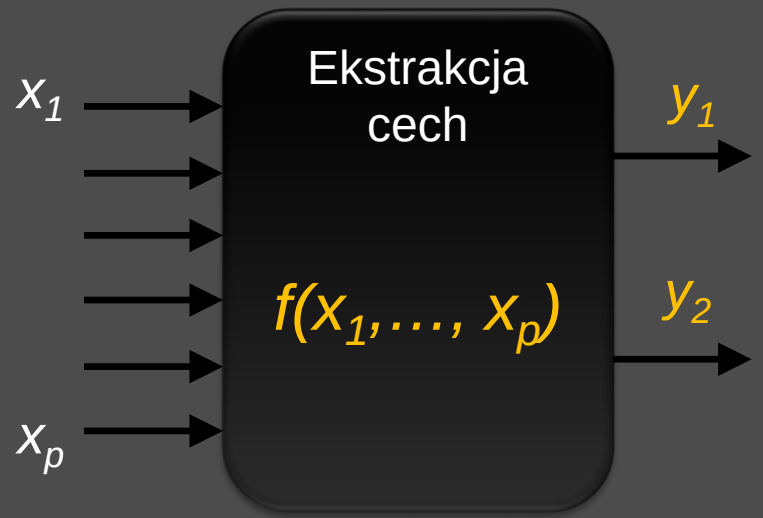
Andrew Webb

# Ekstrakcja cech

- Wprowadzenie
- Metody ekstrakcji cech
  - PCA (*Principal Component Analysis)* – Analiza składowych głównych
  - LDA (*Linear Discriminant Analysis)* – Liniowa analiza dyskryminacyjna
  - MDS (*Multidimensional Scaling*) – Skalowanie wielowymiarowe
- Przykład zastosowania

# Wprowadzenie

- Cel ekstrakcji: znaleźć transformację prowadzącą do takich cech, dla których zadanie selekcji cech jest łatwiejsze
  - określić transformację *p* pomiarów i dokonać selekcji cech w transformowanej przestrzeni

- Korzyści
  - uproszczenie klasyfikatora (klątwa wymiarowości)
  - zwiększenie jakości klasyfikacji i zdolności „uogólniania”
  - pozbycie się mało istotnej informacji
  - ułatwienie graficznej wizualizacji zbioru danych

- Określenia kryterium jakości ekstrakcji *J* prowadzi do zadań optymalizacji:
  - Ekstrakcja cech: dla ustalonej klasy transformacji szukamy takiej transformacji *A*, dla której $J\left(\tilde{A}\right)=\max_{A\in\mathcal{A}} J\left(A(\mathbf{x})\right)$, gdzie $\mathcal{A}$ jest zbiorem wszystkich możliwych transformacji dla przyjętej klasy.

    Wektorem cech jest wówczas wektor

    $$\mathbf{y}=\tilde{A}(\mathbf{x})$$

    Jeżeli transformacja jest liniowa, to $\mathbf{y}=\mathbf{A}^T\mathbf{x}$

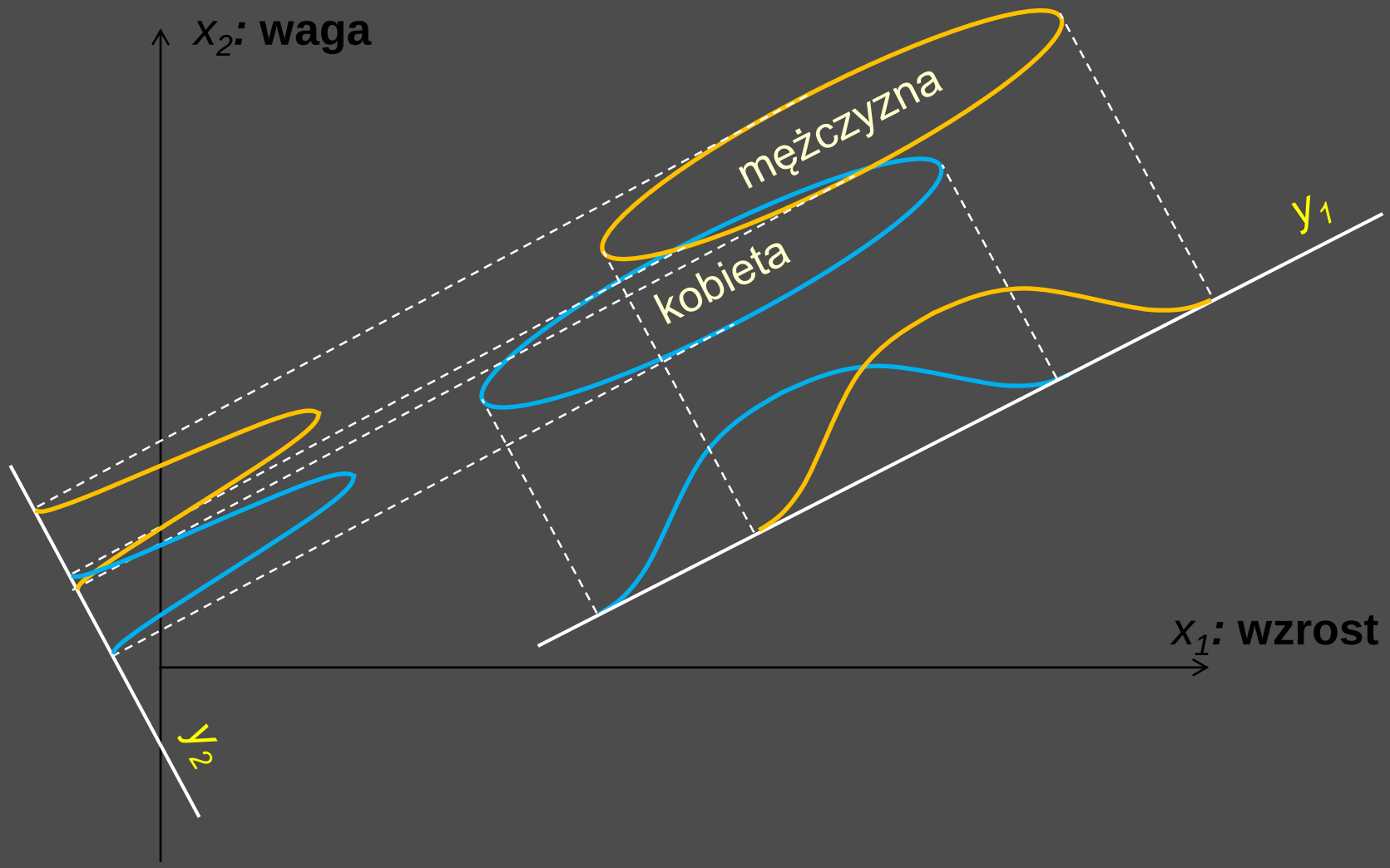
- Wyróżnić można dwie grupy kryteriów stosowanych przy opracowywaniu metod ekstrakcji cech:
  - Reprezentacja danych: celem ekstrakcji jest utworzenie dokładnej reprezentacji próbek w przestrzeni o zredukowanej liczbie wymiarów
  - Klasyfikacja: celem ekstrakcji jest podkreślenie informacji pozwalającej odseparować próbki pochodzące z różnych klas

# Metody ekstrakcji cech

Sformułowanie problemu

- Dane
  - zbiór *p*-wymiarowych wektorów cech
  - kryterium jakości ekstrakcji (ocena zdolności do reprezentacji danych w transformowanej przestrzeni lub ocena separowalności)
- Szukane: najlepsza (w sensie przyjętego kryterium) transformacja przestrzeni cech

- Przykład transformacji



# Ekstrakcja cech

PCA – *Principal Component Analysis*

- Liniowa transformacja cech
- Obrót układu współrzędnych
- Uporządkowanie nowych cech („składowych głównych”) ze względu na wariancję wzdłuż związanych z nimi osi
- Nowe cechy są kombinacją liniową oryginalnych cech i mogą nie mieć bezpośredniej interpretacji
- Ewentualna informacja o klasach / grupach nie jest brana pod uwagę

# PCA – *Principal Component Analysis*

- Dane są projektowane na przestrzeń rozpiętą na zmienionych osiach układu współrzędnych
- Osie nowego układu współrzędnych są ortogonalne
- Nowy układ współrzędnych ma początek w punkcie ciężkości zbioru danych pomiarowych
- Skala na poszczególnych osiach jest wyrażona w jednostkach będących odchyleniami standardowymi oryginalnych danych wzdłuż tych osi
- Kierunki kolejnych osi są tak dobierane, aby maksymalizować wariancję projektowanych na nie danych (wyjątkiem jest ostatnia oś, której kierunek jest zdeterminowany przez pozostałe osie)

# PCA – *Principal Component Analysis*

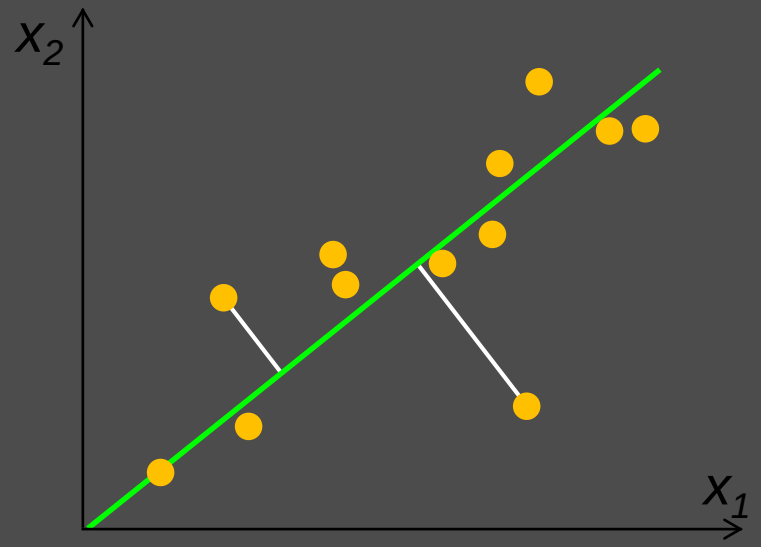
- Pierwsza „składowa główna” to kierunek zadany przez prostą, dla której suma kwadratów odległości od danych pomiarowych (czyli wariancja) jest minimalna

  Minimalizacja wariancji

  w kierunku prostopadłym

  do prostej = maksymalizacja

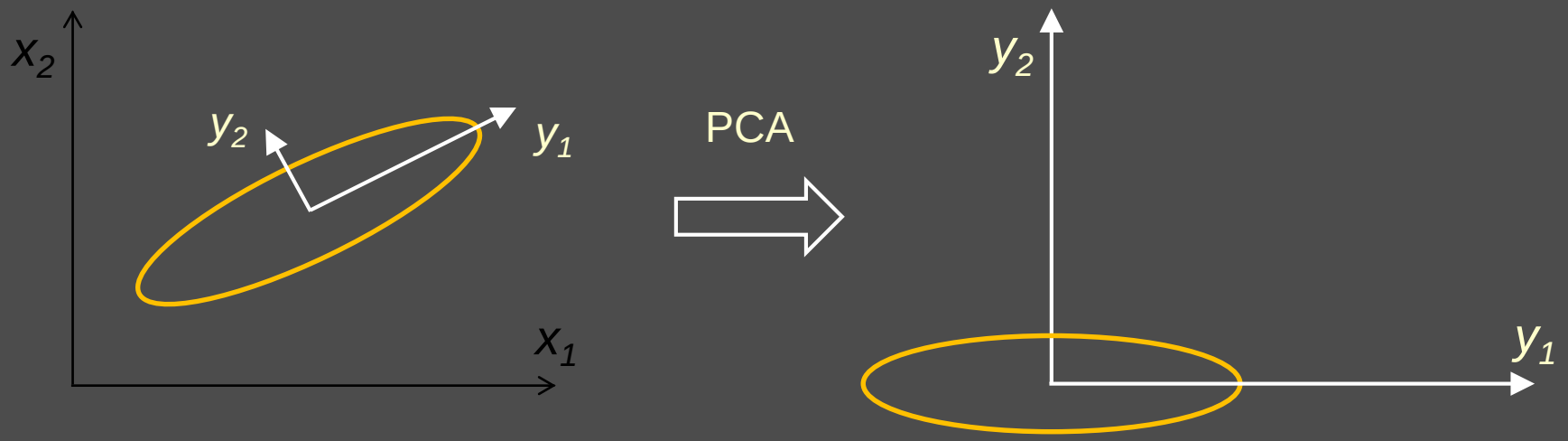
  wariancji wzdłuż tej prostej




- Druga składowa główna jest prostopadła do pierwszej składowej głównej i razem z nią tworzy płaszczyznę najlepiej aproksymującą (w sensie minimum wariancji) dane pomiarowe

# PCA – *Principal Component Analysis*

- Kolejne składowe główne są zdefiniowane w podobny sposób
- Ostatnia składowa główna jest określona jednoznacznie przez poprzednie składowe, gdyż musi być do nich ortogonalna.



# PCA – *Principal Component Analysis*

Zestaw oryginalnych cech: $x_1, x_2, \ldots, x_p$

Kombinacje liniowe cech: $y_1, y_2, \ldots, y_p$

$$y = \sum_{j=1}^{p} a_{ij} x_j$$

lub macierzowo

$$\mathbf{y} = \mathbf{A}^T \mathbf{x}$$

gdzie **x** i **y** to wektory zmiennych losowych, **A** jest macierzą współczynników

Szukamy takiej macierzy ortogonalnej transformacji **A**, które daje optymalne wartości wariancji kolejnych zmiennych $y_j$

# PCA – *Principal Component Analysis*

Rozpoczynamy analizę od pierwszej składowej głównej $y_1$:

$$y_1 = \sum_{j=1}^{p} a_{1j} x_j$$

Należy wyznaczyć wektor $\mathbf{a}_1 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \end{bmatrix}^T$ maksymalizujący wariancję zmiennej $y_1$, przy czym dla jednoznaczności rozwiązania przyjmujemy, że $\|\mathbf{a}_1\|^2 = \mathbf{a}_1^T \mathbf{a}_1 = 1$.

$$\begin{aligned} \operatorname{var}(y_1) &= \mathrm{E}\left[y_1^2\right] - \mathrm{E}\left[y_1\right]^2 = \mathrm{E}\left[\mathbf{a}_1^T \mathbf{x}\mathbf{x}^T \mathbf{a}_1\right] - \mathrm{E}\left[\mathbf{a}_1^T \mathbf{x}\right]\mathrm{E}\left[\mathbf{x}^T \mathbf{a}_1\right] = \\ &= \mathbf{a}_1^T \left(\mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right] - \mathrm{E}\left[\mathbf{x}\right]\mathrm{E}\left[\mathbf{x}^T\right]\right)\mathbf{a}_1 = \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1 = \operatorname{var}(y_1) \end{aligned}$$

# PCA – *Principal Component Analysis*

Maksymalizacja wariancji przy ograniczeniu na normę wektora **a** prowadzi do zadania optymalizacji z ograniczeniami. Zgodnie z metodą Lagrange'a:

$$L(\mathbf{a}_1) = \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1 - \lambda \mathbf{a}_1^T \mathbf{a}_1$$

Po różniczkowaniu ze względu na $\boldsymbol{a}_1$ i przyrównaniu do zera otrzymujemy:

$$\boldsymbol{\Sigma} \mathbf{a}_1 - \lambda \mathbf{a}_1 = 0$$

$$\boldsymbol{\Sigma} \mathbf{a}_1 = \lambda \mathbf{a}_1$$

Wynika z tego, że $\boldsymbol{a}_1$ musi być wektorem własnym macierzy $\boldsymbol{\Sigma}$ z wartością własną równą λ .

# PCA – *Principal Component Analysis*

Ponadto, maksymalizacja wariancji zmiennej $y_1$

$$\max\left\{\mathbf{a}_1^T \boldsymbol{\Sigma}\, \mathbf{a}_1\right\} = \max\left\{\lambda\, \mathbf{a}_1^T\, \mathbf{a}_1\right\} = \max \lambda$$

sprowadza się do wybrania wektora własnego $\boldsymbol{a}_1$, któremu odpowiada największa wartość własna macierzy $\boldsymbol{\Sigma}$. Wybór wektora własnego jest jednoznaczny, jeżeli wartość własna $\lambda$ nie jest wielokrotnym pierwiastkiem równania charakterystycznego:

$$\left|\boldsymbol{\Sigma} - \lambda \mathbf{I}\right| = 0$$

Oznaczmy przez $\lambda_1$ wartość własną związaną z pierwszą składową główną $\boldsymbol{y}_1$

# PCA – *Principal Component Analysis*

Drugą składową główną $y_2 = \mathbf{a}_2^T \mathbf{x}$ otrzymujemy maksymalizując – jak poprzednio – wariancję zmiennej $y_2$ przy ograniczeniu na normę składowej $\|\mathbf{a}_2\|^2 = \mathbf{a}_2^T \mathbf{a}_2 = 1$ oraz dodatkowo wymagając, aby składowa $y_2$ nie była skorelowana z $y_1$ :

$$\mathrm{E}[y_2 y_1] - \mathrm{E}[y_2]\mathrm{E}[y_1] = 0 \ ,$$

co jest równoważne z $\mathbf{a}_2^T \boldsymbol{\Sigma} \mathbf{a}_1 = 0$ .

Ze względu na fakt, że $\boldsymbol{a}_1$ jest wektorem własnym macierzy $\boldsymbol{\Sigma}$ , to również $\boldsymbol{a}_2$ jest ortogonalne do $\boldsymbol{a}_1$ :

$$\mathbf{a}_2^T \mathbf{a}_1 = 0$$

# PCA – *Principal Component Analysis*

Zgodnie z metodą Lagrange'a:

$$L(\mathbf{a}_2) = \mathbf{a}_2^T \boldsymbol{\Sigma} \mathbf{a}_2 - \lambda \mathbf{a}_2^T \mathbf{a}_2 - \eta \mathbf{a}_2^T \mathbf{a}_1$$

Po różniczkowaniu ze względu na $\boldsymbol{a}_2$ i przyrównaniu do zera otrzymujemy:

$$2\boldsymbol{\Sigma}\mathbf{a}_2 - 2\lambda \mathbf{a}_2 - \eta \mathbf{a}_1 = 0$$

Mnożenie obustronne przez $\boldsymbol{a}_1^T$ daje:

$$2\mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_2 - 2\lambda \mathbf{a}_1^T \mathbf{a}_2 - \eta \mathbf{a}_1^T \mathbf{a}_1 = 0$$

$$\eta \mathbf{a}_1^T \mathbf{a}_1 = 0$$

Z powyższego wynika, że $\eta = 0$ .

# PCA – *Principal Component Analysis*

Pochodna funkcji Lagrange'a przyjmuje postać:

$$2\boldsymbol{\Sigma}\mathbf{a}_2 - 2\lambda\,\mathbf{a}_2 = 0$$

$$\boldsymbol{\Sigma}\mathbf{a}_2 = \lambda\,\mathbf{a}_2$$

Wynika z tego, że $\boldsymbol{a}_2$ jest wektorem własnym macierzy $\boldsymbol{\Sigma}$ i – analogicznie jak poprzednio – odpowiadająca mu wartość własna $\lambda$ jest największą spośród pozostałych wartości własnych.

# PCA – *Principal Component Analysis*

- Podsumowując, procedura wyznaczania składowych głównych składa się z następujących kroków:

  1) Rozkład macierzy kowariancji **Σ** na wektory własne i obliczenie odpowiadających im wartości własnych

  2) Uporządkowanie wektorów własnych zgodnie z malejącymi wartościami własnymi: $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ ; $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

  3) Utworzenie macierzy transformacji

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_p \end{bmatrix}$$

# PCA – *Principal Component Analysis*

- Macierz transformacji jest wykorzystywana przy projekcji danych z przestrzeni oryginalnych cech na przestrzeń składowych głównych:

$$\mathbf{y} = \mathbf{A}^T \mathbf{x}$$

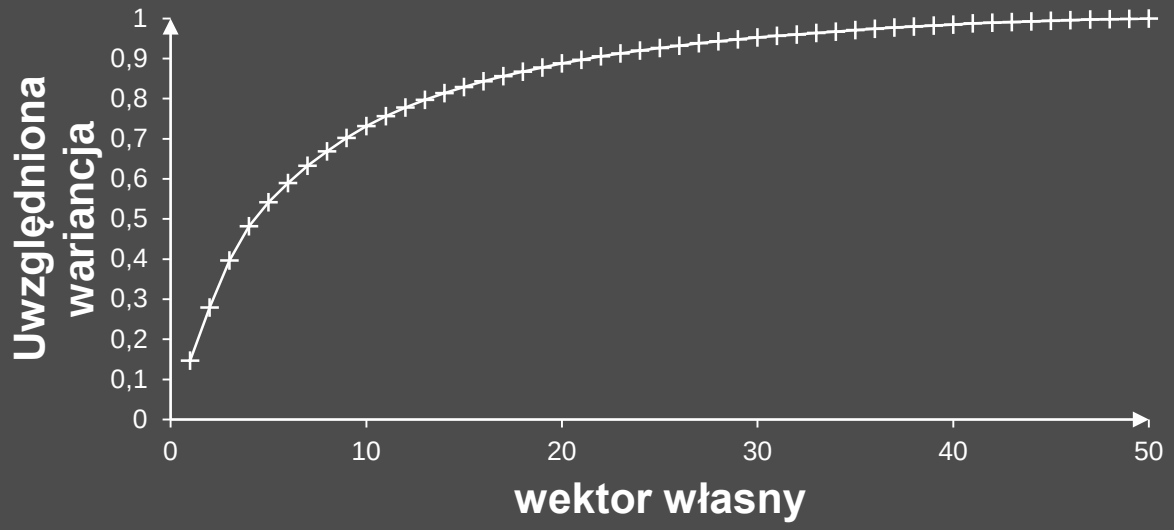
- Powyższe równanie przypisuje zaobserwowanemu losowemu wektorowi **x** składowe główne **y**. Na ogół, wartość oczekiwana **y** jest różna od zera. Aby była ona równa zeru, projekcja powinna zostać zdefiniowana jako: $\mathbf{y} = \mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu}),$

gdzie **μ** jest wartością oczekiwaną **x**. W praktyce, w miejsce **μ** stosuje się średnią **m** z próby.

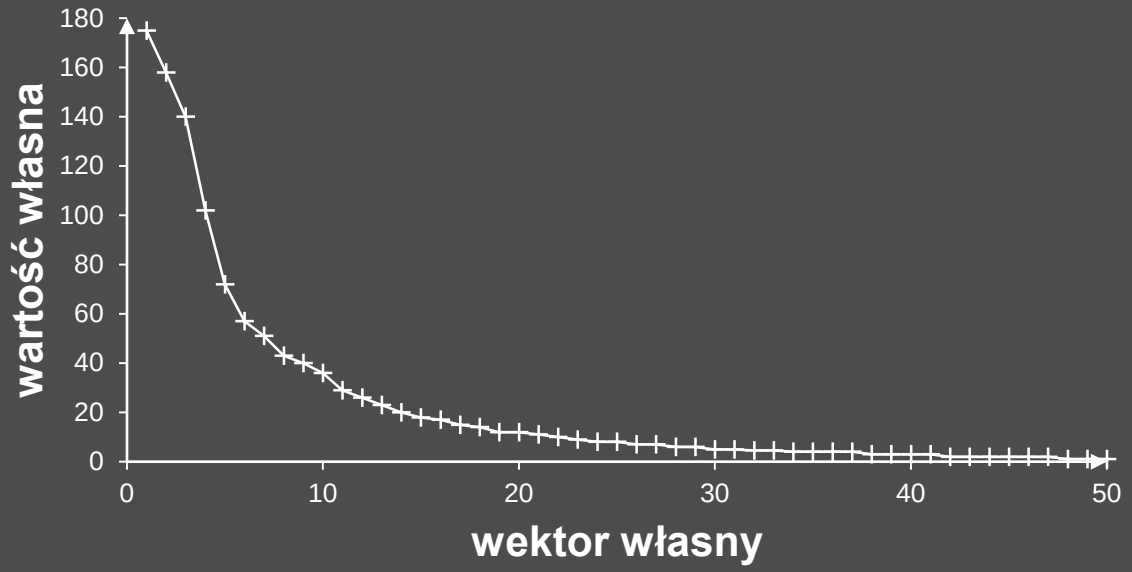
# PCA – *Principal Component Analysis*

- Ilość wariancji uwzględnionej przez *k* pierwszych składowych głównych jest równa:

$$\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$




Przykładowe dane



# PCA – *Principal Component Analysis*

- Selekcja cech w transformowanej przestrzeni: spośród *p* składowych głównych wybrać *d* takich, które uwzględniają najwięcej wariancji
- Macierz transformacji ograniczy się do *d* pierwszych wektorów własnych (wymiar macierzy transformacji to *p* x *d*):

$$\mathbf{A}_d = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_d \end{bmatrix}$$

- Projekcja do zredukowanej przestrzeni ma postać:

$$\mathbf{y}_d = \mathbf{A}_d^T \mathbf{x}$$

lub

$$\mathbf{y}_d = \mathbf{A}_d^T (\mathbf{x} - \boldsymbol{\mu})$$

# PCA – *Principal Component Analysis*

- Projekcja wstecz z transformowanej przestrzeni do oryginalnej ma postać: $\mathbf{x} = \left(\mathbf{A}^T\right)^{-1}\mathbf{y}$ lub $\mathbf{x} = \left(\mathbf{A}^T\right)^{-1}\mathbf{y} + \boldsymbol{\mu}$
- Jeżeli **A** jest macierzą symetryczną ($\mathbf{A}^T = \mathbf{A}^{-1}$), wówczas projekcja przyjmuje postać: $\mathbf{x} = \mathbf{A}\mathbf{y}$ lub $\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu}$
- Jeżeli dodatkowo nastąpiła redukcja wymiaru, to projekcję można zapisać jako: $\mathbf{x}_d = \mathbf{A}_d\mathbf{y}_d + \boldsymbol{\mu}$
- Jeżeli wstawimy do powyższego wzoru w miejsce $\boldsymbol{y}_d$ odpowiednie wyrażenie na projekcję, to otrzymamy:

$$\mathbf{x}_d = \mathbf{A}_d\mathbf{A}_d^T\left(\mathbf{x} - \boldsymbol{\mu}\right) + \boldsymbol{\mu}$$

# PCA – *Principal Component Analysis*

- Składowe główne zależą od skali (jednostki)!
- Przed przystąpieniem do analizy można ustandaryzować dane pomiarowe  taki sposób, aby wartości cech miały podobne zakresy
- Typową metodą standaryzacji jest przesunięcie wartości średniej do zera oraz unormowanie wariancji do jedności. Wówczas składowe główna można wyznaczyć na podstawie macierzy korelacji.

# PCA – *Principal Component Analysis*

- Zadanie: wyznaczyć składowe główne dla dwuwymiarowego zbioru:

  X={ (*1*,*2*), (*3*,*3*), (*3*,*5*), (*5*,*4*), (*5*,*6*), (*6*,*5*), (*8*,*7*), (*9*,*8*) }

- Rozwiązanie:
  - Macierz kowariancji: $\boldsymbol{\Sigma} = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$
  - Wartości własne są rozwiązaniami równania charakterystycznego:

$$\left|\boldsymbol{\Sigma} - \lambda \mathbf{I}\right| = 0 \Rightarrow \begin{vmatrix} 6.25-\lambda & 4.25 \\ 4.25 & 3.5-\lambda \end{vmatrix} = 0 \Rightarrow \begin{matrix} \lambda_1 = 9.34 \\ \lambda_2 = 0.41 \end{matrix}$$

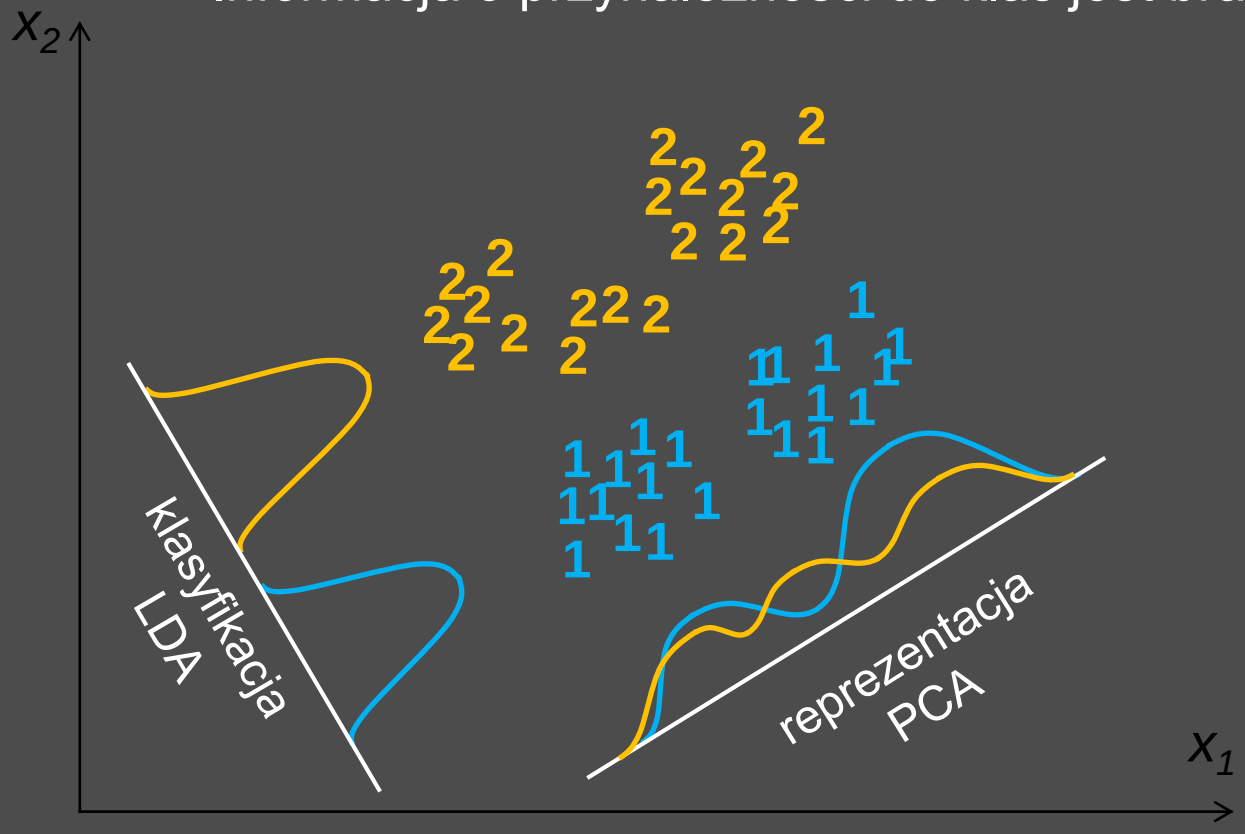
  - Wektory własne są rozwiązaniami układu:

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 a_{11} \\ \lambda_1 a_{12} \end{bmatrix} \Rightarrow \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$$

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 a_{21} \\ \lambda_2 a_{22} \end{bmatrix} \Rightarrow \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix}$$

# LDA – *Linear Discriminant Analysis*

- Celem transformacji jest podkreślenie informacji pozwalającej odseparować próbki pochodzące z różnych klas
- Informacja o przynależności do klas jest brana pod uwagę



# LDA – *Linear Discriminant Analysis*

- Przypadek z dwiema klasami
  - $n_1$ próbek należy do klasy $\omega_1$, $n_2$ próbek należy do klasy $\omega_2$ ( $n_1 + n_2 = n$ )

$$\left\{ \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T \right\}$$

  - Szukamy takiego kierunku $\boldsymbol{a}$, na który będziemy rzutować dane pomiarowe:

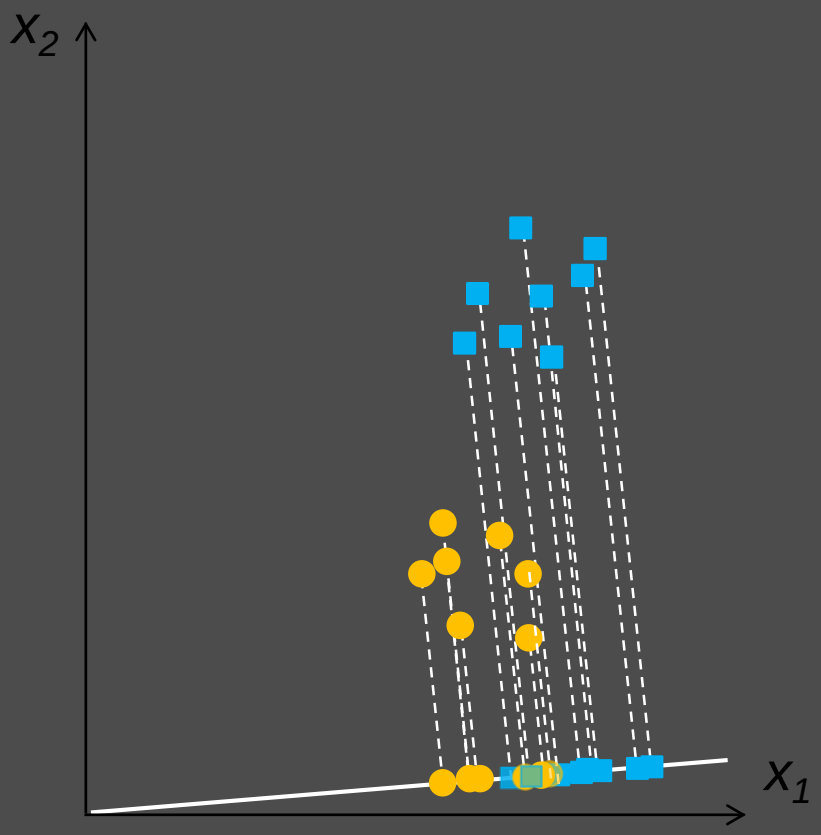
$$y = \mathbf{a}^T \mathbf{x}$$

$$\mathbf{x}^i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T$$

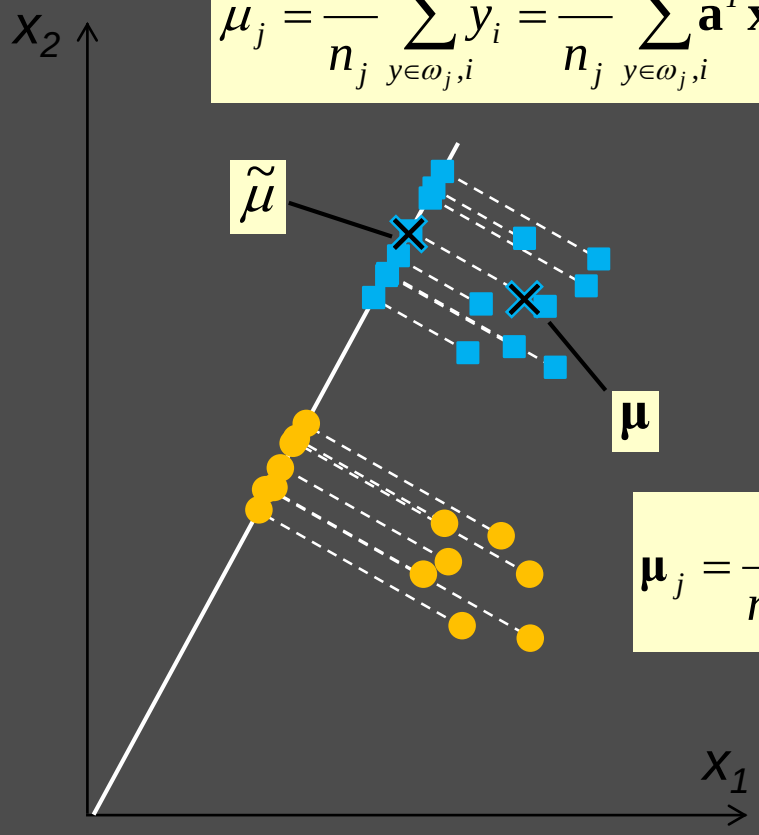
wymagając, aby po zrzutowaniu dane pochodzące z różnych klas były od siebie jak najbardziej odseparowane.

# LDA – *Linear Discriminant Analysis*

- Przykład w dwuwymiarowej przestrzeni cech




$$\tilde{\mu}_j = \frac{1}{n_j} \sum_{y \in \omega_j, i} y_i = \frac{1}{n_j} \sum_{y \in \omega_j, i} \mathbf{a}^T \mathbf{x}^i = \mathbf{a}^T \boldsymbol{\mu}_j$$

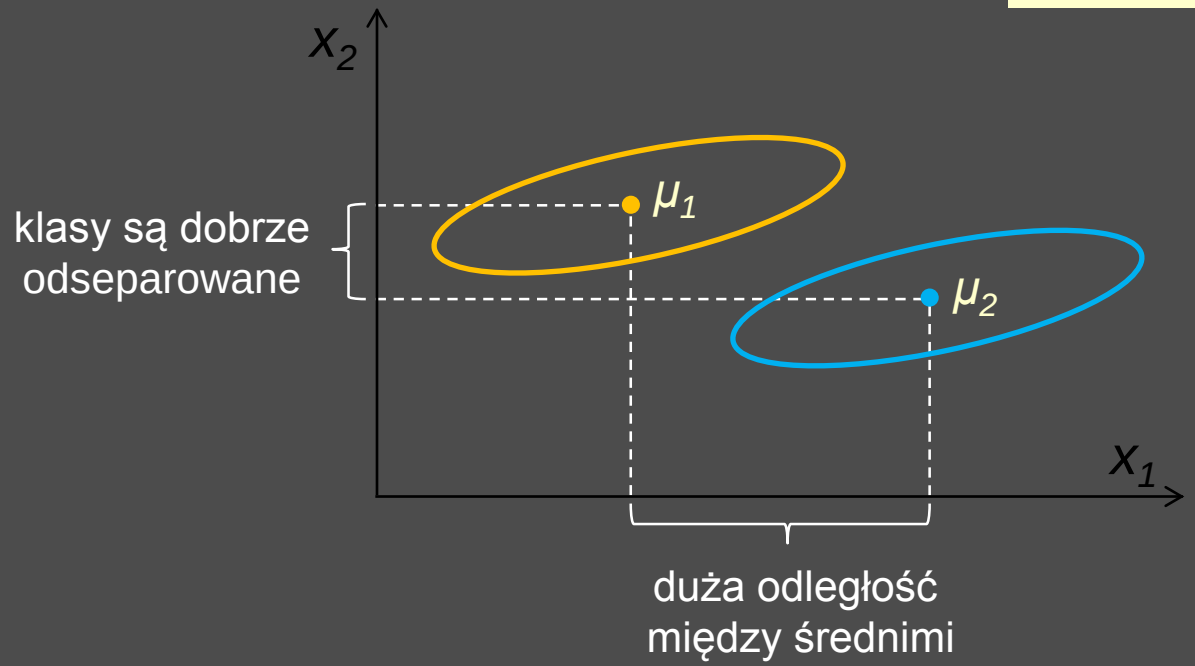



$$\boldsymbol{\mu}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in \omega_j, i} \mathbf{x}^i$$

# LDA – *Linear Discriminant Analysis*

- Jak mierzyć separowalność klas dla danych rzutowanych na kierunek ***a*** ?
- Propozycja: mierzyć odległość między średnimi po projekcji:

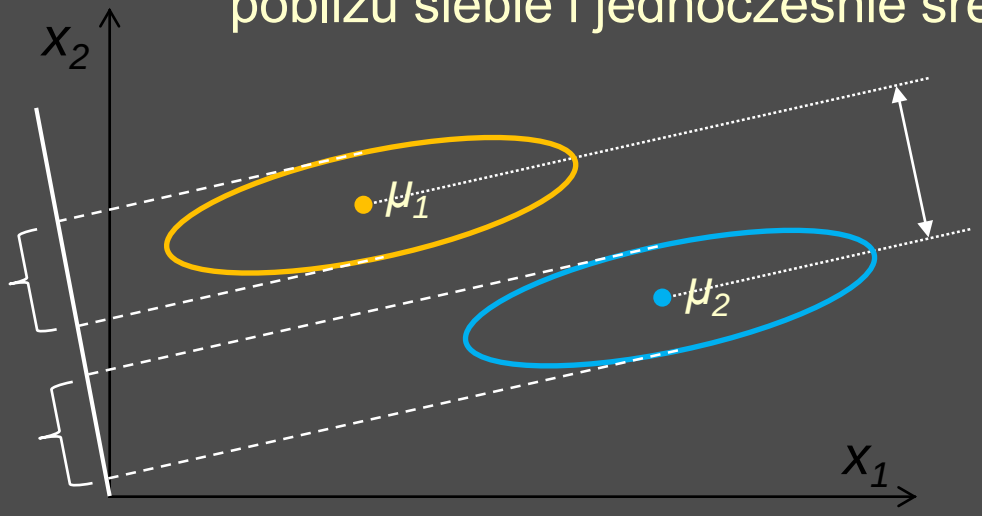
$$J(\mathbf{a}) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right|$$




Problem: nie bierzemy pod uwagę rozrzutu danych

# LDA – *Linear Discriminant Analysis*

- Propozycja Fishera: uwzględnić rozrzuty danych w klasach
- Rozrzut w klasie: $\tilde{s}_j^2 = \sum_{y \in \omega_j, i} \left( y_i - \tilde{\mu}_j \right)^2$
- Miara separowalności: $J(\mathbf{a}) = \frac{\left| \tilde{\mu}_1 - \tilde{\mu}_2 \right|}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{a}^T \mathbf{S}_B \mathbf{a}}{\mathbf{a}^T \mathbf{S}_W \mathbf{a}}$
- Szukamy transformacji, po której próbki z jednej klasy leżą w pobliżu siebie i jednocześnie średnie po transformacji leżą daleko



# LDA – *Linear Discriminant Analysis*

- Aby znaleźć maksimum funkcji $J(\mathbf{a})$, różniczkujemy względem $\mathbf{a}$ i przyrównujemy do zera:

$$\frac{dJ}{d\mathbf{a}} = \frac{d}{d\mathbf{a}}\left[\frac{\mathbf{a}^T\mathbf{S}_B\mathbf{a}}{\mathbf{a}^T\mathbf{S}_W\mathbf{a}}\right] = 0$$

$$\left[\mathbf{a}^T\mathbf{S}_W\mathbf{a}\right]\frac{d\left[\mathbf{a}^T\mathbf{S}_B\mathbf{a}\right]}{d\mathbf{a}} - \left[\mathbf{a}^T\mathbf{S}_B\mathbf{a}\right]\frac{d\left[\mathbf{a}^T\mathbf{S}_W\mathbf{a}\right]}{d\mathbf{a}} = 0$$

$$\left[\mathbf{a}^T\mathbf{S}_W\mathbf{a}\right]2\mathbf{S}_B\mathbf{a} - \left[\mathbf{a}^T\mathbf{S}_B\mathbf{a}\right]2\mathbf{S}_W\mathbf{a} = 0$$

# LDA – *Linear Discriminant Analysis*

- Następnie dzielimy obie strony przez $\boldsymbol{a}^T\boldsymbol{S}_W\boldsymbol{a}$:

$$\left[\frac{\mathbf{a}^T\mathbf{S}_W\mathbf{a}}{\mathbf{a}^T\mathbf{S}_W\mathbf{a}}\right]\mathbf{S}_B\mathbf{a}-\left[\frac{\mathbf{a}^T\mathbf{S}_B\mathbf{a}}{\mathbf{a}^T\mathbf{S}_W\mathbf{a}}\right]\mathbf{S}_W\mathbf{a}=0$$

$$\mathbf{S}_B\mathbf{a}-J\,\mathbf{S}_W\mathbf{a}=0$$

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{a}-J\,\mathbf{a}=0$$

co prowadzi w rezultacie do uogólnionego zagadnienia rozkładu na wektory i wartości własne: $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{a}=J\,\mathbf{a}$

- Rozwiązaniem jest:

$$\mathbf{a}^*=\arg\max_{\mathbf{a}}\left\{\frac{\mathbf{a}^T\mathbf{S}_B\mathbf{a}}{\mathbf{a}^T\mathbf{S}_W\mathbf{a}}\right\}=\mathbf{S}_W^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)$$

# LDA – *Linear Discriminant Analysis*

- Zadanie: wyznaczyć kierunek, który najlepiej separuje obserwacje:

  **X1**={ (*4,1*), (*2,4*), (*2,3*), (*3,6*), (*4,4*) }; **X2**={ (*9,10*), (*6,8*), (*9,5*), (*8,7*), (*10,8*) }

- Rozwiązanie:

$$\mathbf{S}_1 = \begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 2.6 \end{bmatrix}; \quad \mathbf{S}_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix}; \quad \mathbf{S}_B = \begin{bmatrix} 29.16 & 21.6 \\ 21.6 & 16 \end{bmatrix}$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 & 3.6 \end{bmatrix}; \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 8.4 & 7.6 \end{bmatrix}; \quad \mathbf{S}_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

  - Wektory własne są rozwiązaniami układu:

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{a} = \lambda\mathbf{a} \Rightarrow \left|\mathbf{S}_W^{-1}\mathbf{S}_B - \lambda\mathbf{I}\right| = 0 \Rightarrow \begin{vmatrix} 11.89-\lambda & 8.81 \\ 5.08 & 3.76-\lambda \end{vmatrix} = 0 \Rightarrow \lambda = 15.65$$

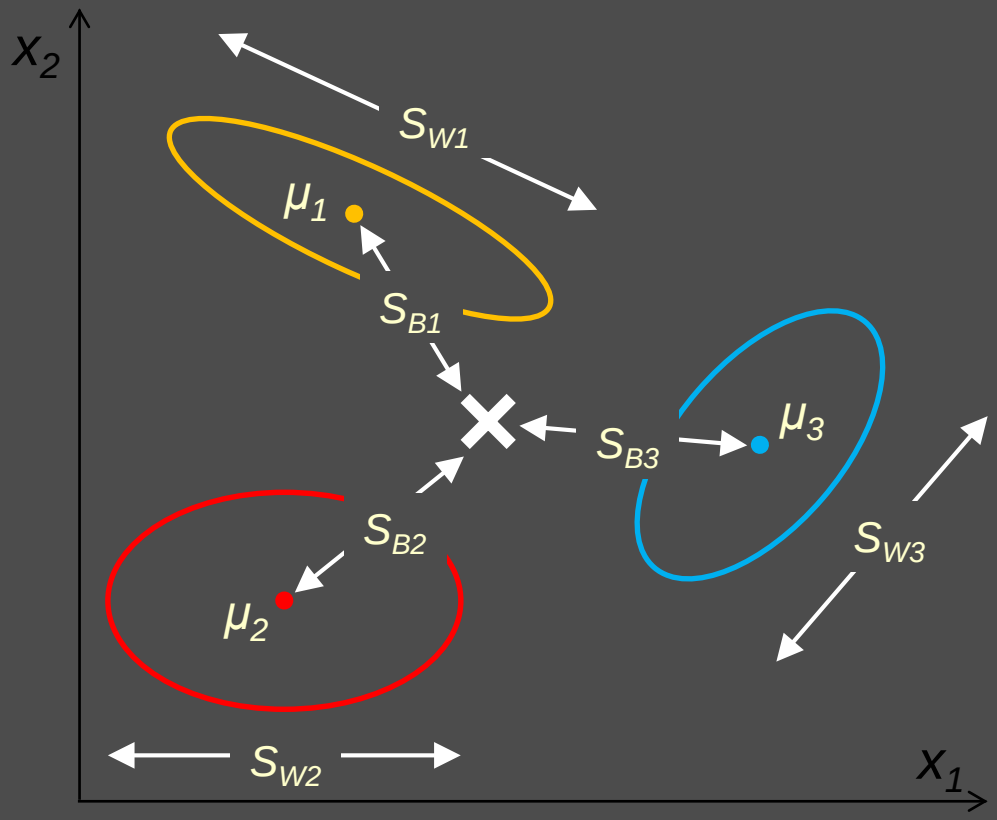
$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 15.65\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \Rightarrow \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$

  - lub bezpośrednio: $\mathbf{a}^* = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{bmatrix} -0.91 & -0.39 \end{bmatrix}^T$

# LDA – *Linear Discriminant Analysis*

- Uogólnienie na *C* klas: poszukujemy *C-1* wektorów $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{C-1}$ dających rzuty $y_1, y_2, \ldots, y_{C-1}$ :

$$\mathbf{y} = \mathbf{A}^T \mathbf{x}$$




Miara separowalności:

$$J(\mathbf{A}) = \frac{\mathbf{A}^T \mathbf{S}_B \mathbf{A}}{\mathbf{A}^T \mathbf{S}_W \mathbf{A}}$$

# LDA – *Linear Discriminant Analysis*

- Rozwiązaniem jest:

$$\mathbf{A}^* = \arg\max_{\mathbf{a}} \left\{ \frac{\mathbf{A}^T \mathbf{S}_B \mathbf{A}}{\mathbf{A}^T \mathbf{S}_W \mathbf{A}} \right\} = \left( \mathbf{S}_B - \lambda_j \mathbf{S}_W \right) \mathbf{a}_j^* = 0$$

- Kierunki związane z największą separowalnością to wektory własne odpowiadające największym wartościom własnym macierzy $\mathbf{S}_W^{-1}\mathbf{S}_B$ .

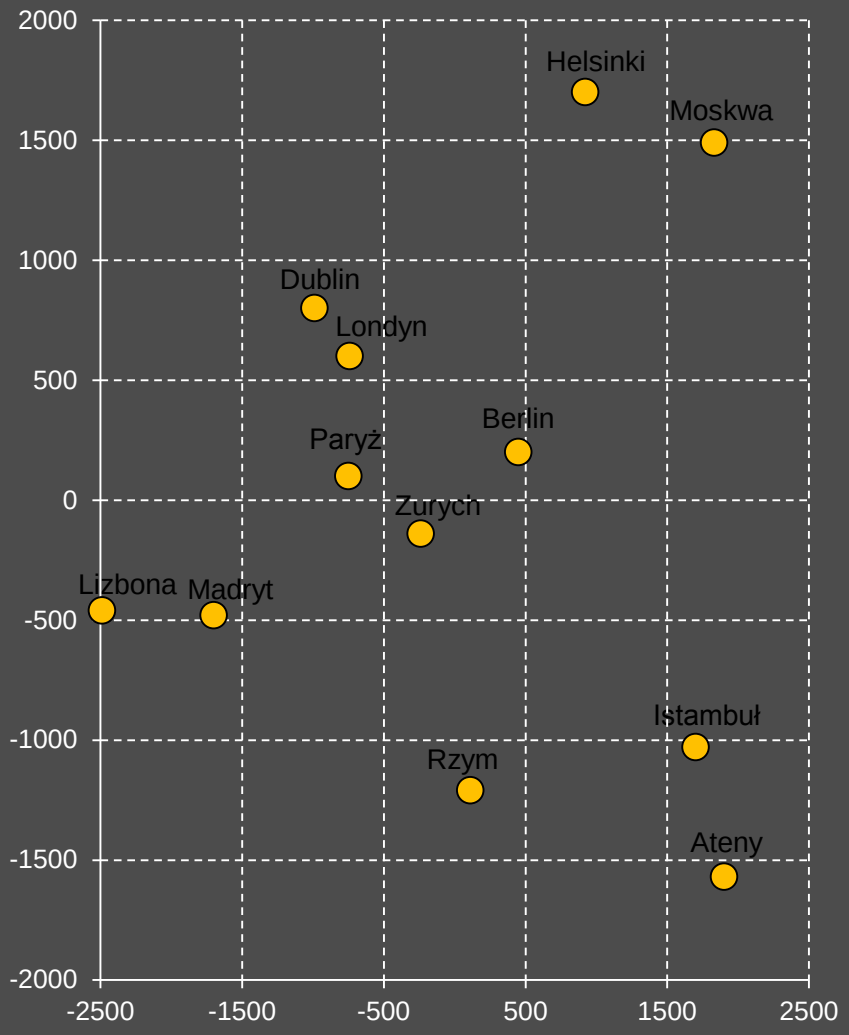
# MDS – *Multidimensional Scaling*

- Punktem wyjścia MDS jest macierz niepodobieństwa (*dissimilarity matrix*). Dla *p*-wymiarowego zbioru *N* próbek dane są wzajemne odległości $d_{ij}$ , $i,j=1,\ldots,N$. Macierz niepodobieństwa $\mathbf{D} = \left[ d_{ij} \right]$ ma wymiar *N* x *N*
- MDS polega na ulokowaniu próbek w przestrzeni o zadanej liczbie wymiarów w taki sposób, aby wzajemne odległości zostały w możliwie najlepszym stopniu zachowane
- MDS pozwala na zastosowanie nieliniowych transformacji cech
- Szukane: $\mathbf{X} = \left[ \mathbf{x}^1 \quad \mathbf{x}^2 \quad \cdots \quad \mathbf{x}^N \right]$

# MDS – *Multidimensional Scaling*

- Przykład

| | *Ateny* | *Berlin* | *Dublin* | *Helsinki* | ... |
|---|---|---|---|---|---|
| *Ateny* | 0 | 1803 | 2859 | 2469 | |
| *Berlin* | 1803 | 0 | 1322 | 1107 | |
| *Dublin* | 2859 | 1322 | 0 | 2031 | ... |
| *Helsinki* | 2469 | 1107 | 2031 | 0 | |
| ... | ... | | | | |



# MDS – *Multidimensional Scaling*

- Kwadrat odległości Euklidesowej między próbkami o numerach *s* i *r* jest równy:

$$d_{rs}^2 = \left\| \mathbf{x}^r - \mathbf{x}^s \right\|^2 = \sum_{j=1}^{p} \left( x_{rj} - x_{sj} \right)^2 = \sum_{j=1}^{p} x_{rj}^2 - 2\sum_{j=1}^{p} x_{rj} x_{sj} + \sum_{j=1}^{p} x_{sj}^2 = b_{rr} + b_{ss} - 2b_{rs}$$

gdzie $b_{rs} = \sum_{j=1}^{p} x_{rj} x_{sj} \Leftrightarrow \mathbf{B} = \mathbf{X}\mathbf{X}^T$ (*)

- W celu uzyskania jednoznacznego rozwiązania, umieszczamy średnią z danych pomiarowych w początku układu współrzędnych oraz przyjmujemy założenie:

$$\sum_{n=1}^{N} x_{nj}, \; j = 1,2,\ldots, p$$

# MDS – *Multidimensional Scaling*

- Sumując równanie (*) po *r*, *s*, oraz *r* i *s*, a także definiując

$$T = \sum_{n=1}^{N} b_{nn} = \sum_{n=1}^{N} \sum_{j=1}^{p} x_{nj}^2$$

otrzymujemy:

$$\sum_{r=1}^{N} d_{rs}^2 = T + N b_{ss}$$

$$\sum_{s=1}^{N} d_{rs}^2 = T + N b_{rr}$$

$$\sum_{r=1}^{N} \sum_{s=1}^{N} d_{rs}^2 = 2NT$$

- Definiujemy następnie:

$$d_{\bullet s}^2 = \frac{1}{N} \sum_{r=1}^{N} d_{rs}^2 \ , \ d_{r\bullet}^2 = \frac{1}{N} \sum_{s=1}^{N} d_{rs}^2 \ , \ d_{\bullet\bullet}^2 = \frac{1}{N^2} \sum_{r=1}^{N} \sum_{s=1}^{N} d_{rs}^2$$

# MDS – *Multidimensional Scaling*

- Z równania (*) mamy:

$$b_{rs} = \frac{1}{2}\left(d_{r\bullet}^2 + d_{\bullet s}^2 - d_{\bullet\bullet}^2 - d_{rs}^2\right)$$

- Wiadomo, że $\mathbf{B} = \mathbf{X}\mathbf{X}^T$ ,
- $\mathbf{X} = \mathbf{C}\mathbf{D}^{1/2}$ można użyć jako aproksymacji macierzy **X**, przy czym **C** jest macierzą, której kolumny są wektorami własnymi macierzy **B** a $\mathbf{D}^{1/2}$ jest macierzą diagonalną, której przekątna zawiera pierwiastki wartości własnych.
- Selekcji cech dokonujemy – podobnie jak w PCA – na podstawie wartości własnych

# MDS – *Multidimensional Scaling*

- Można pokazać, że wartości własne macierzy $\boldsymbol{X}\boldsymbol{X}^T$ ($N$ x $N$) są identyczne jak wartości własne macierzy $\boldsymbol{X}^T\boldsymbol{X}$ ($p$ x $p$), natomiast ich wektory własne powiązane są liniową transformacją.
- Powyższy fakt wskazuje na związek między MDS a PCA: PCA wykonane na macierzy korelacji (zamiast kowariancji) jest równoważne MDS przy ustandaryzowanej odległości Euklidesowej, gdzie każda cecha ma jednostkową wariancję.

# MDS – *Multidimensional Scaling*

- W ogólności, MDS polega na znalezieniu odwzorowania

$$\mathbf{y} = \mathbf{g}(\mathbf{x}; \boldsymbol{\theta})$$

z *p*-wymiarowej przestrzeni cech do *d*-wymiarowej, gdzie $\boldsymbol{\theta}$ jest wektorem parametrów.

- Rozważane odwzorowanie może być liniowe i wówczas:

$$\mathbf{y}_d = \mathbf{g}(\mathbf{x}; \mathbf{A}_d) = \mathbf{A}_d^T \mathbf{x}$$

- Zastosowanie nieliniowej transformacji $\mathbf{g}$ prowadzi do nieliniowych metod redukcji wymiaru.

# MDS – *Multidimensional Scaling*

- Dla typowych zbiorów danych macierz niepodobieństwa ***D*** może mieć zbyt duży wymiar z punktu widzenia przetwarzania
- MDS nie definiuje w sposób bezpośredni transformacji, która dla danego wektora **x** zwracałaby jego reprezentację **y** w transformowanej przestrzeni.

# Przykład zastosowania

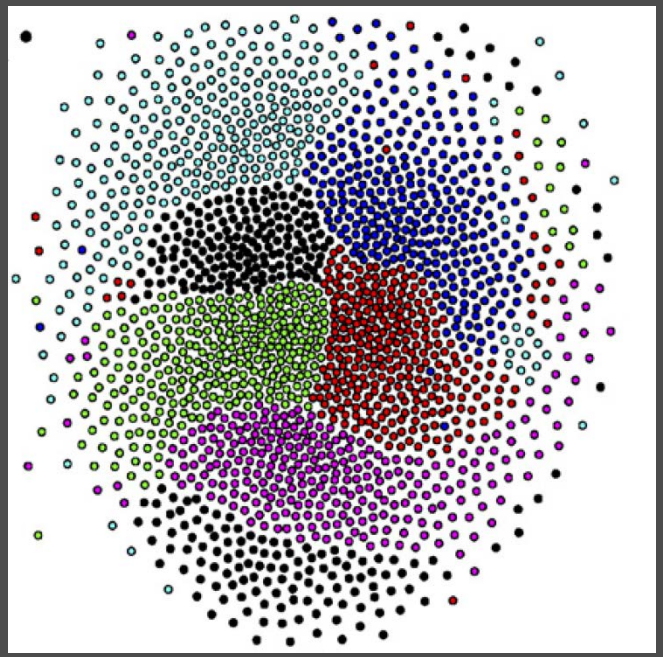
Ekstrakcja cech w systemie rozpoznawania mowy

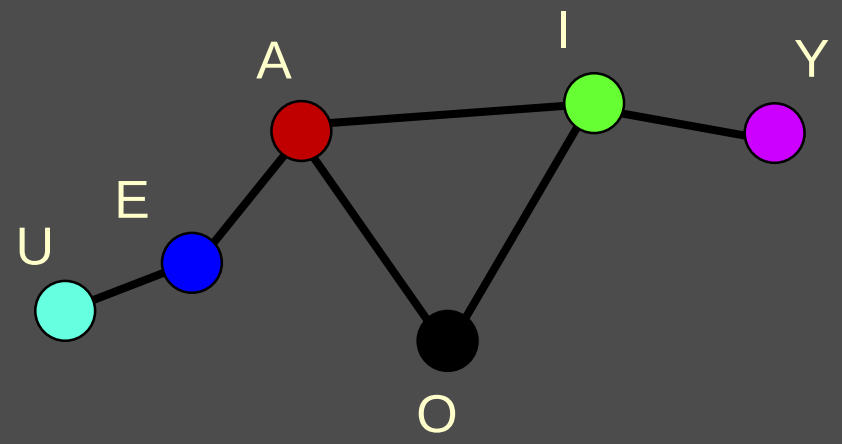
- Dane:
  - 150-ciu lektorów wymawiających dwukrotnie litery alfabetu (w sumie 7800 próbek)
  - Liczba cech: 617
  - Skupiono się na samogłoskach: `A`, `E`, `I`, `O`, `U`, `Y`
  - Redukcja wymiarów do dwóch

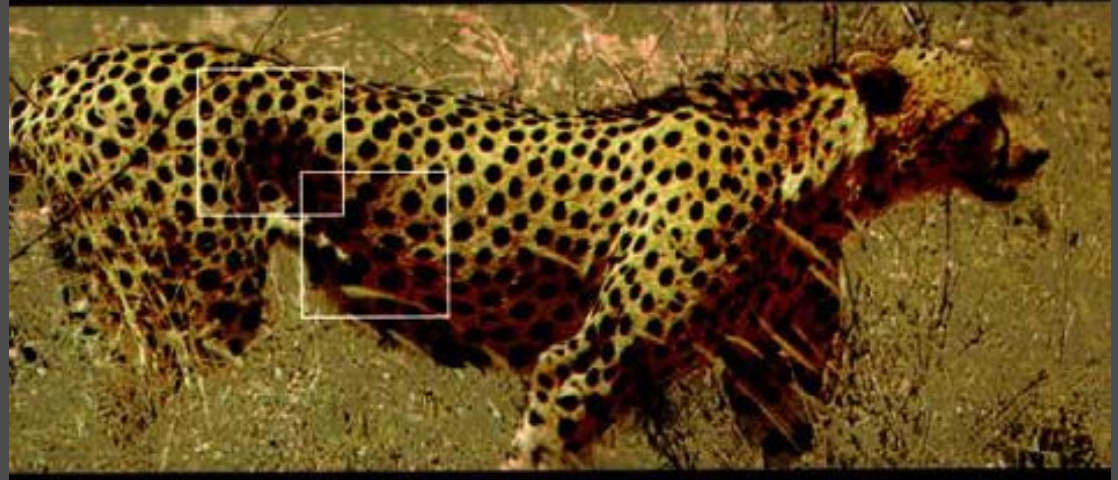
Michael Aupetit,
*Visualizing distortions and recovering topology in continuous projection techniques*,
Neurocomputing70 (2007) pp. 1304–1330

# Rozwiązanie problemu

- PCA
- LDA
- MDS – CCA (*Curvilinear Component Analysis*)







# Statistical Pattern Recognition

Second Edition

Andrew Webb