ANALÝZA A KLASIFIKACE DAT

RNDr. Eva Janoušová

VIII.
ANALÝZA HLAVNÍCH KOMPONENT

ÚVOD – EXTRAKCE PŘÍZNAKŮ

- jedním z principů výběru příznaků
- transformace původních příznakových proměnných na menší počet jiných příznakových proměnných ⇒ tzn. hledání (optimálního) zobrazení Z, které transformuje původní m-rozměrný prostor (obraz) na prostor (obraz) n-rozměrný ($m \geq n$)
- pro snadnější řešitelnost hledáme zobrazení Z v oboru lineárních zobrazení

ÚVOD – EXTRAKCE PŘÍZNAKŮ

- 3 kritéria pro nalezení optimálního zobrazení Z:
  - obrazy v novém prostoru budou aproximovat původní obrazy ve smyslu minimální střední kvadratické odchylky
  - rozložení pravděpodobnosti veličin v novém prostoru budou splňovat podmínky kladené na jejich pravděpodobnostní charakteristiky
  - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby

ÚVOD – EXTRAKCE PŘÍZNAKŮ

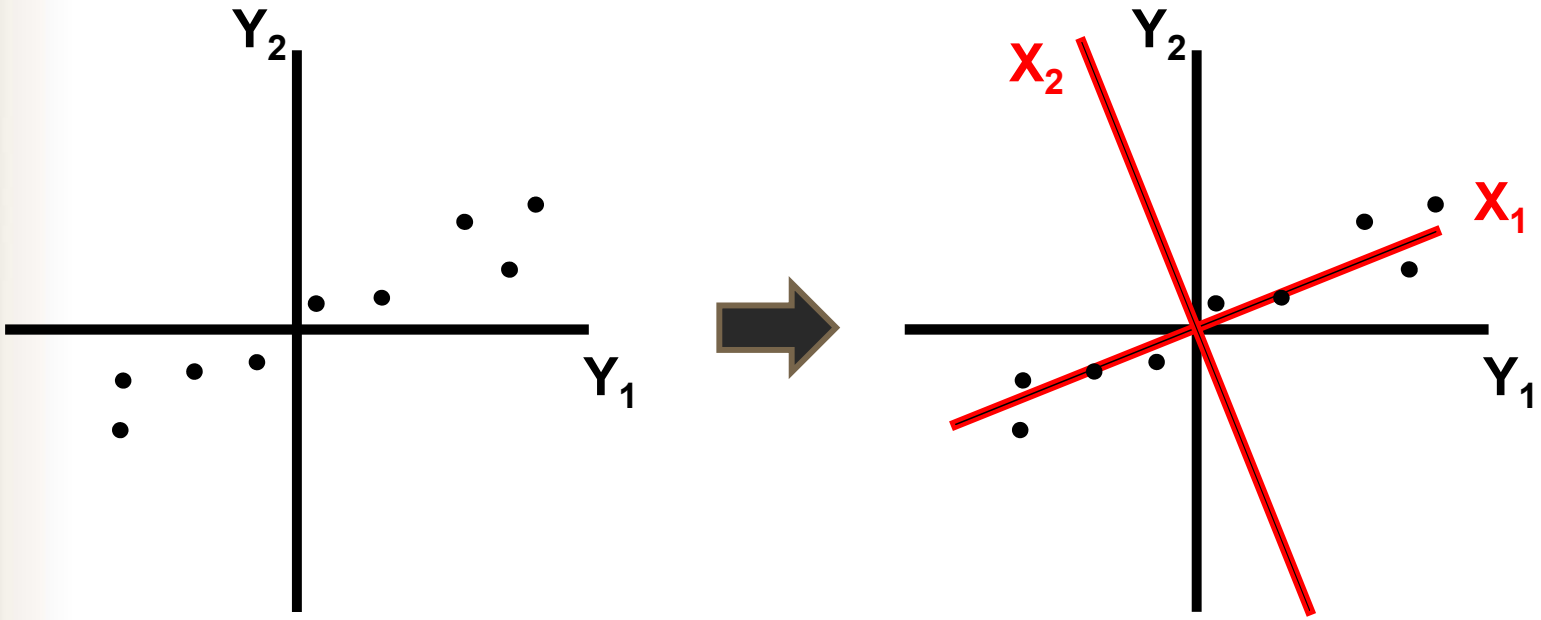
- 3 kritéria pro nalezení optimálního zobrazení Z:
  - **obrazy v novém prostoru budou aproximovat původní obrazy ve smyslu minimální střední kvadratické odchylky**
  - rozložení pravděpodobnosti veličin v novém prostoru budou splňovat podmínky kladené na jejich pravděpodobnostní charakteristiky
  - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby

ANALÝZA HLAVNÍCH KOMPONENT

- PCA – Principal Component Analysis
- osnova:
  - opakování učiva z Vícerozměrných statistických metod
  - jiný (obecnější) pohled na PCA
  - příklad – výpočet PCA krok po kroku
  - PCA při rozdělení obrazů do klasifikačních tříd
  - rozšiřující poznatky o PCA

PCA - OPAKOVÁNÍ

- snaha redukovat počet proměnných nalezením nových latentních proměnných (hlavních komponent) vysvětlujících co nejvíce variability původních proměnných
- nové proměnné (X_1, X_2) lineární kombinací původních proměnných (Y_1, Y_2)



PCA - OPAKOVÁNÍ

- vstup do PCA:
  - kovarianční matice
  - matice korelačních koeficientů
- hlavní komponenty odpovídají vlastním vektorům kovarianční matice (či matice korelačních koef.)
- variabilita vysvětlená příslušnou komponentou odpovídá vlastním číslům
- vlastní vektory seřazeny podle vlastních hodnot (sestupně) $\Rightarrow$ vybráno prvních *n* komponent vyčerpávajících nejvíce variability původních dat
- předpoklady: kvantitativní proměnné s normálním rozdělením

PCA – JINÝ (OBECNĚJŠÍ) POHLED

- dáno *K* obrazů charakterizovaných *m* příznakovými proměnnými (nerozdělenými do klasifikačních tříd)

příznaky

obrazy	$\mathbf{p}_1$ $\mathbf{p}_2$ … $\mathbf{p}_m$
$\mathbf{y}_1$	
$\mathbf{y}_2$	
…	
$\mathbf{y}_K$	

- aproximujme nyní kterýkoliv obraz $\mathbf{y}_k$ lineární kombinací *n* ortonormálních vektorů $\mathbf{e}_i$ ($n \leq m$)

$$\mathbf{x}_k = \sum_{i=1}^{n} c_{ki} e_i$$

- koeficienty c_{ki} lze považovat za velikost i-té souřadnice vektoru $\mathbf{y}_k$ vyjádřeného v novém systému souřadnic s bází $\mathbf{e}_i$, i=1,2,…,n

$$c_{ki} = \mathbf{y}_k^T e_i$$

PCA – KRITÉRIUM MINIMÁLNÍ STŘEDNÍ KVADRATICKÉ ODCHYLKY

- nalezení optimálního zobrazení pomocí **kritéria minimální střední kvadratické odchylky**:

$$\varepsilon_k^2 = \left\| \mathbf{y}_k - \mathbf{x}_k \right\|^2$$

- vztah lze pomocí dříve uvedených vztahů upravit na:

$$\varepsilon_k^2 = \left\| \mathbf{y}_k \right\|^2 - \sum_{i=1}^{n} c_{ki}^2$$

- střední kvadratická odchylka pro všechny obrazy $\mathbf{y}_k$, k=1,…,K je

$$\varepsilon^2 = \frac{1}{K} \sum_{k=1}^{K} \varepsilon_k^2 = \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{y}_k \right\|^2 - \sum_{i=1}^{n} {}^T\mathbf{e}_i \left[\frac{1}{K} \sum_{k=1}^{K} \mathbf{y}_k .^T \mathbf{y}_k \right] . \mathbf{e}_i$$

PCA – KRITÉRIUM MINIMÁLNÍ STŘEDNÍ KVADRATICKÉ ODCHYLKY

- musíme zvolit bázový systém $\mathbf{e}_i$ tak, aby střední kvadratická odchylka ε^2 byla minimální

- diskrétní konečný rozvoj podle vztahu $\mathbf{x}_k = \sum_{i=1}^{n} c_{ki} e_i$ s bázovým systémem $\mathbf{e}_i$, optimálním podle kritéria minimální střední kvadratické chyby, nazýváme diskrétní Karhunenův – Loevův rozvoj

PCA – KRITÉRIUM MINIMÁLNÍ STŘEDNÍ KVADRATICKÉ ODCHYLKY

- střední kvadratická odchylka

$$\varepsilon^2 = \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k^2 = \frac{1}{K}\sum_{k=1}^{K}\|\mathbf{y}_k\|^2 - \sum_{i=1}^{n}{}^T\mathbf{e}_i\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_k.{}^T\mathbf{y}_k\right].\mathbf{e}_i$$

je minimální, když je maximální výraz

$$\sum_{i=1}^{n}{}^T\mathbf{e}_i.\kappa(\mathbf{y}).\mathbf{e}_i\,, \quad \text{kde} \quad \kappa(\mathbf{y}) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_k.\mathbf{y}_k^T$$

je autokorelační matice řádu *m*. Protože je symetrická a semidefinitní, jsou její vlastní čísla λ_i, i=1,…,m, reálná a nezáporná a vlastní vektory $\mathbf{v}_i$, jsou buď ortonormální, nebo je můžeme ortonormalizovat (v případě násobných vlastních čísel).

PCA – KRITÉRIUM MINIMÁLNÍ STŘEDNÍ KVADRATICKÉ ODCHYLKY

- uspořádáme-li vlastní čísla sestupně podle velikosti, tj. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$

  a podle toho očíslujeme i odpovídající vlastní vektory, lze dokázat, výše uvedený výraz dosahuje maxima, jestliže platí

$$\mathbf{e}_i = \mathbf{v}_i,\ i=1,\ldots,n$$

  a pro velikost maxima je

$$\max \sum_{i=1}^{n} \mathbf{e}_i^T . \kappa(\mathbf{y}) . \mathbf{e}_i = \sum_{i=1}^{n} \lambda_i$$

- pak pro minimální střední kvadratickou platí

$$\varepsilon_{\min}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{y}_k\|^2 - \sum_{i=1}^{n} \lambda_i = tr(\kappa(\mathbf{y})) - \sum_{i=1}^{n} \lambda_i = \sum_{i=n+1}^{m} \lambda_i$$

PCA – VSTUPNÍ MATICE

- **autokorelační matice** – data nejsou nijak upravena (zohledňována průměrná hodnota i rozptyl původních dat)
- **kovarianční (disperzní) matice** – data centrována (od každé příznakové proměnné odečtena její střední hodnota) – zohledňován rozptyl původních dat
- **matice korelačních koeficientů** – data standardizována (odečtení středních hodnot a podělení směrodatnými odchylkami) – použití pokud mají proměnné různá měřítka

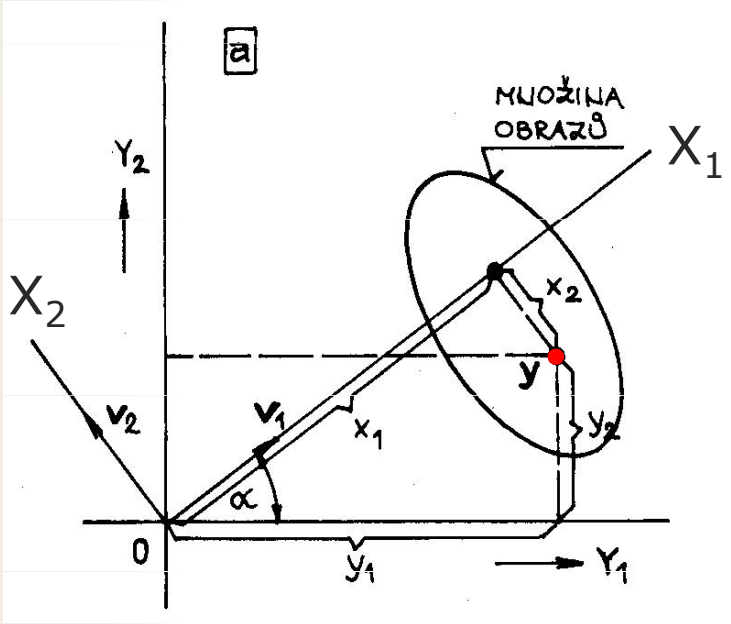
PCA – VSTUPNÍ MATICE

- **autokorelační matice** – data nejsou nijak upravena (zohledňována průměrná hodnota i rozptyl původních dat)
- **kovarianční (disperzní) matice** – data centrována (od každé příznakové proměnné odečtena její střední hodnota) – zohledňován rozptyl původních dat
- **matice korelačních koeficientů** – data standardizována (odečtení středních hodnot a podělení směrodatnými odchylkami) – použití pokud mají proměnné různá měřítka
- každou úpravou původních dat ale přicházíme o určitou informaci!

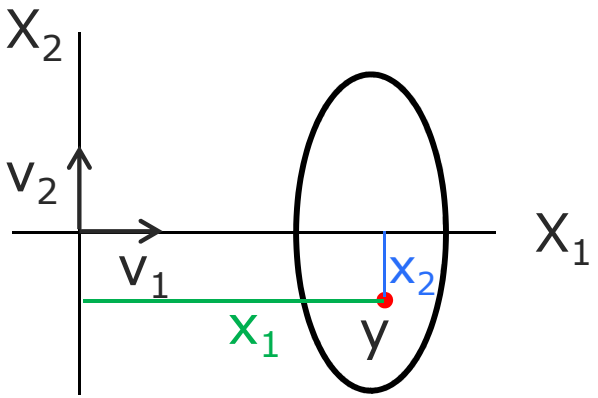
PCA – VLASTNOSTI KARHUNENOVA-LOEVOVA ROZVOJE

- při daném počtu *n* členů rozvoje poskytuje ze všech možných aproximací nejmenší střední kvadratickou odchylku;
- při použití disperzní matice jsou transformované souřadnice nekorelované; pokud se výskyt obrazů řídí normálním rozložením zajišťuje nekorelovanost i jejich nezávislost;
- vliv každého členu uspořádaného rozvoje se zmenšuje s jeho pořadím;
- změna požadavků na velikost střední kvadratické odchylky nevyžaduje přepočítávat celý rozvoj, nýbrž jen změnit počet jeho členů.

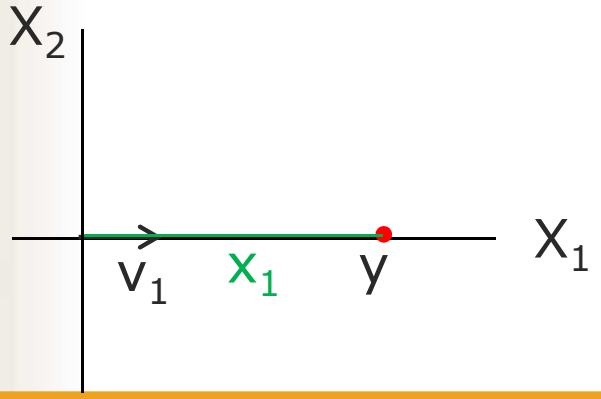
PCA – GEOMETRICKÁ INTERPRETACE



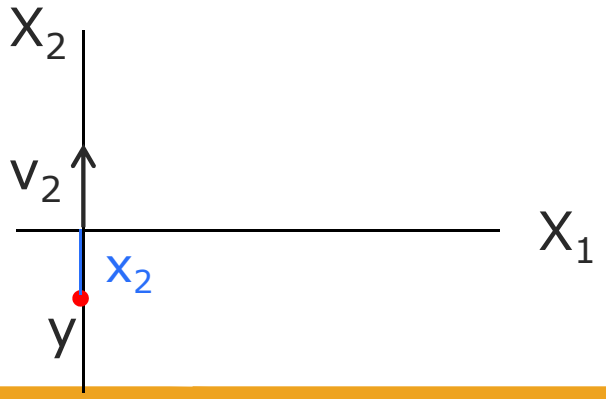
použití obou hlavních komponent



použití 1. hlavní komponenty



použití 2. hlavní komponenty



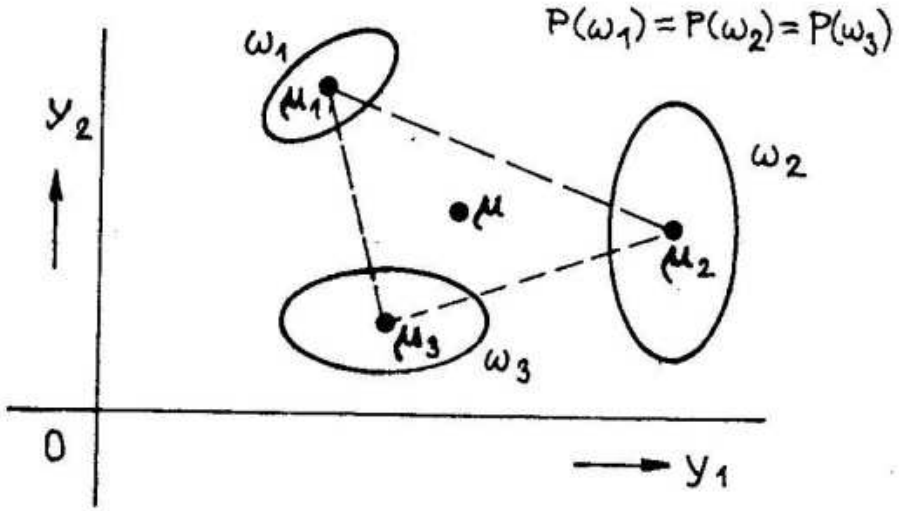
PCA - PŘÍKLAD

- data:

A	101	16
B	105	18
C	103	42
D	98	23
E	93	6

PCA – ROZDĚLENÍ DO TŘÍD

- Výskyt obrazů v jednotlivých klasifikačních třídách bude popsán podmíněnými hustotami pravděpodobnosti $p(\mathbf{y}|\omega_r)$, r=1,2,…,R a apriorní pravděpodobnost klasifikačních tříd bude $P(\omega_r)$.



- V tom případě autokorelační matice bude

$$\kappa(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} y.y^T.p(\mathbf{y}\,|\,\omega_r).d\mathbf{y} = \int_{\Upsilon^m} y.^T y.p(\mathbf{y}).d\mathbf{y}$$

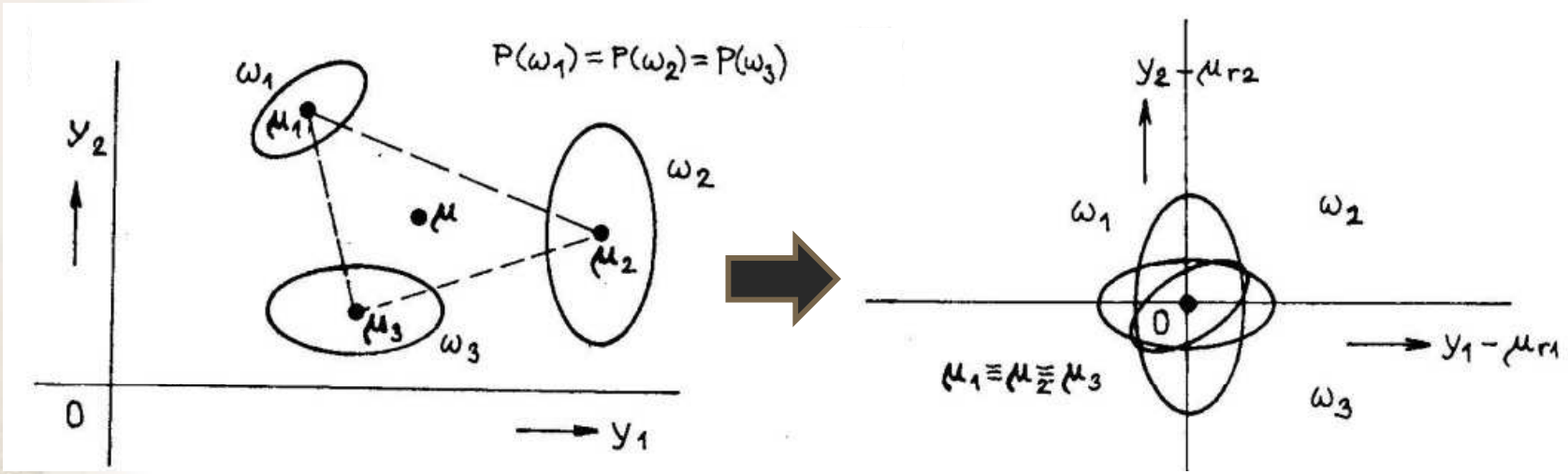
PCA – ROZDĚLENÍ DO TŘÍD

- disperzní matice – vztah 1:

$$D^1(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\mathcal{Y}^m} (\mathbf{y}-\boldsymbol{\mu}_r).(\mathbf{y}-\boldsymbol{\mu}_r)^T.p(\mathbf{y}\,|\,\omega_r)d\mathbf{y}$$

kde $\boldsymbol{\mu}_r = \int_{\mathcal{Y}^m} \mathbf{y}.p(\mathbf{y}\,|\,\omega_r)d\mathbf{y}$

- rozlišení klasifikačních tříd jen podle disperze
- transformované příznak. proměnné nekorelované



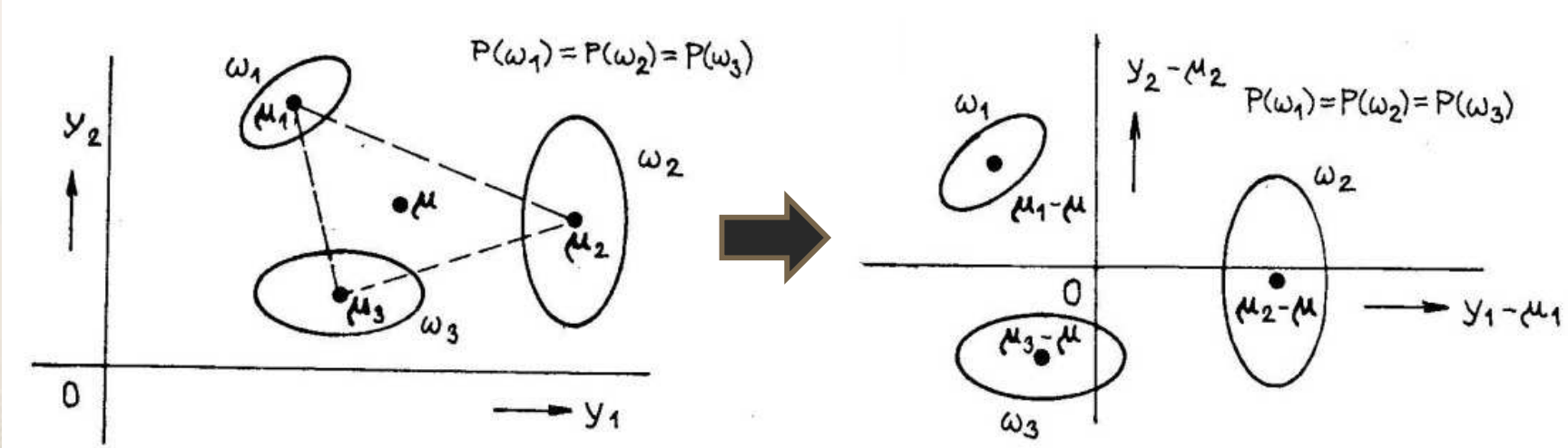
PCA – ROZDĚLENÍ DO TŘÍD

- disperzní matice – vztah 2:

$$D^0(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} (\mathbf{y}-\boldsymbol{\mu}).(\mathbf{y}-\boldsymbol{\mu})^T.p(\mathbf{y}\,|\,\omega_r).d\mathbf{y} = \int_{\Upsilon^m} (\mathbf{y}-\boldsymbol{\mu}).(\mathbf{y}-\boldsymbol{\mu})^T.p(\mathbf{y}).d\mathbf{y}$$

kde $\boldsymbol{\mu} = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} \mathbf{y}.p(\mathbf{y}\,|\,\omega_r)d\mathbf{y} = \int_{\Upsilon^m} \mathbf{y}.p(\mathbf{y})d\mathbf{y}$

- neodstraňuje vliv středních hodnot obrazů v jednotlivých třídách – použití pokud jsou stř. h. výrazně odlišné a nesou velké množství informace



PCA – ROZŠIŘUJÍCÍ POZNATKY

- ☑ výpočet PCA, když je $m >> K$
- ☑ souvislost se singulárním rozkladem (SVD – Singular Value Decomposition)

Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE"

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ