ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

MU

IBA

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

VIII.
ANALÝZA HLAVNÍCH KOMPONENT

ZAČÍNÁME

ANALÝZA HLAVNÍCH KOMPONENT

PRINCIPAL COMPONENT ANALYSIS (PCA)

ROZKLAD PODLE VLASTNÍCH ČÍSEL

SINGULAR VALUE DECOMPOSITION (SVD)

Karhunenova-Loevova transformace

ZAČÍNÁME

- **extrakce příznaků** - hledání zobrazení (optimálního) Z, které transformuje původní m rozměrný prostor (obraz) na prostor (obraz) n rozměrný ($m \geq n$);
- nalezení vhodné transformace – potřeba optimalizačního kritéria:
  - obrazy v novém prostoru budou aproximovat původní obrazy ve smyslu minimální střední kvadratické odchylky;
  - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby

ZAČÍNÁME

- ☑ aby byla úloha řešitelná, hledáme zobrazení v oboru lineárních zobrazení

ZAČÍNÁME

- ☑ aby byla úloha řešitelná, hledáme zobrazení v oboru lineárních zobrazení

Jak poznáme lineární zobrazení?

ZAČÍNÁME

- ☑ aby byla úloha řešitelná, hledáme zobrazení v oboru lineárních zobrazení

Jak poznáme lineární zobrazení?

$$\mathbf{x} = \mathbf{Z}.\mathbf{y}$$

TEORIE

- předpokládejme, že je dáno K obrazů a nechť existuje m příznakových veličin, které tyto obrazy charakterizují. Tedy k-tý obraz je vyjádřen m rozměrným sloupcovým vektorem $\mathbf{y}_k \in \Upsilon^m$, k=1,…,K.
- aproximujme nyní kterýkoliv obraz $\mathbf{y}_k$ lineární kombinací n ortonormálních vektorů $\mathbf{e}_i$ $(m \geq n)$

$$x_k = \sum_{i=1}^{n} c_{ki} e_i . \qquad (☺)$$

TEORIE

- 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, tj. platí

$$c_{ki} = {}^T\mathbf{y}_k . \mathbf{e}_i .$$

- použijeme-li jako kritérium minimální střední kvadratické odchylky, pak je

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

TEORIE

- pak pomocí dříve uvedených vztahů pro $\mathbf{x}_k$ a c_{ki} dostaneme

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

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

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

(je tedy závislá na volbě bázového systému $\mathbf{e}_i$)

TEORIE

- ☑ diskrétní konečný rozvoj podle vztahu (☺) s bázovým systémem $\mathbf{e}_i$, optimálním podle kritéria minimální střední kvadratické chyby nazýváme diskrétní Karhunenův – Loevův rozvoj;
- ☑ aby střední kvadratická odchylka podle výše uvedeného vztahu byla minimální, musí být odečítaná hodnota na pravé straně rovnice maximální.

TEORIE

- musíme tedy maximalizovat výraz

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

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).

TEORIE

- ☑ 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 charakteristické vektory, lze dokázat, výe uvedený výraz dosahuje maxima, jestliže platí

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

a pro velikost maxima je

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

TEORIE

- pro minimální střední kvadratickou odchylku tedy platí

$$\varepsilon_{min}^2 = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{y}_k\right\|^2 - \sum_{i=1}^{n}\lambda_i =$$

$$= tr(\kappa(\mathbf{y})) - \sum_{i=1}^{n}\lambda_i = \sum_{i=n+1}^{m}\lambda_i$$

TEORIE

- v některých případech je vhodnější vektory $\mathbf{y}_k$ před aproximací centrovat se střední hodnotou

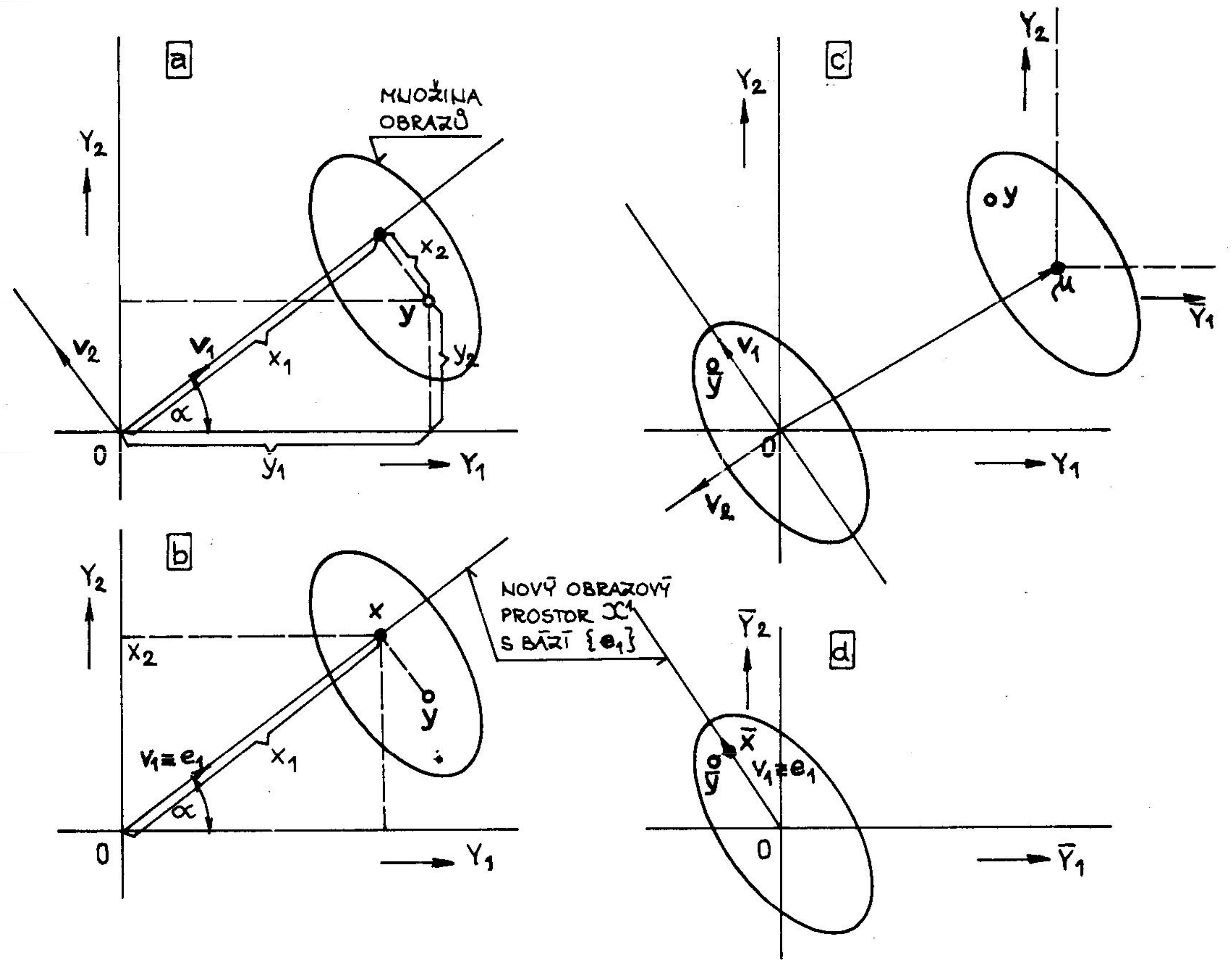
$$\boldsymbol{\mu} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_k$$

a místo s obrazem $\mathbf{y}_k$ počítáme s jeho centrovanou verzí $\overline{\mathbf{y}}_k = \mathbf{y}_k - \boldsymbol{\mu}$.

Postup výpočtu se nemění, ale místo autokorelační matice používáme disperzní matici ve tvaru

$$D(\mathbf{y}) = \frac{1}{K}\sum_{k=1}^{K}\overline{\mathbf{y}}_k.^T\,\overline{\mathbf{y}}_k. \quad \text{Platí} \quad \kappa(\mathbf{y}) = D(\mathbf{y}) + \boldsymbol{\mu}.^T\,\boldsymbol{\mu}.$$

GEOMETRICKÁ INTERPRETACE



VLASTNOSTI

- 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.

ROZDĚLENÍ DO TŘÍD

Jak se změní podmínky, když obrazy **y** budou platit, které budou vymezeny jako části spojitého obrazového prostoru $\mathcal{Y}^m$?

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

V tom případě autokorelační matice bude

$$\kappa(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\mathcal{Y}^m} y.^T y.p(\mathbf{y} \mid \omega_r).dy = \int_{\mathcal{Y}^m} y.^T y.p(\mathbf{y}).dy$$

ROZDĚLENÍ DO TŘÍD

- disperzní matice

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

kde

$$\boldsymbol{\mu}_r = \int_{\Upsilon^m} \mathbf{y}.p(\mathbf{y}\,|\,\omega_r).dy$$

nebo vztahem

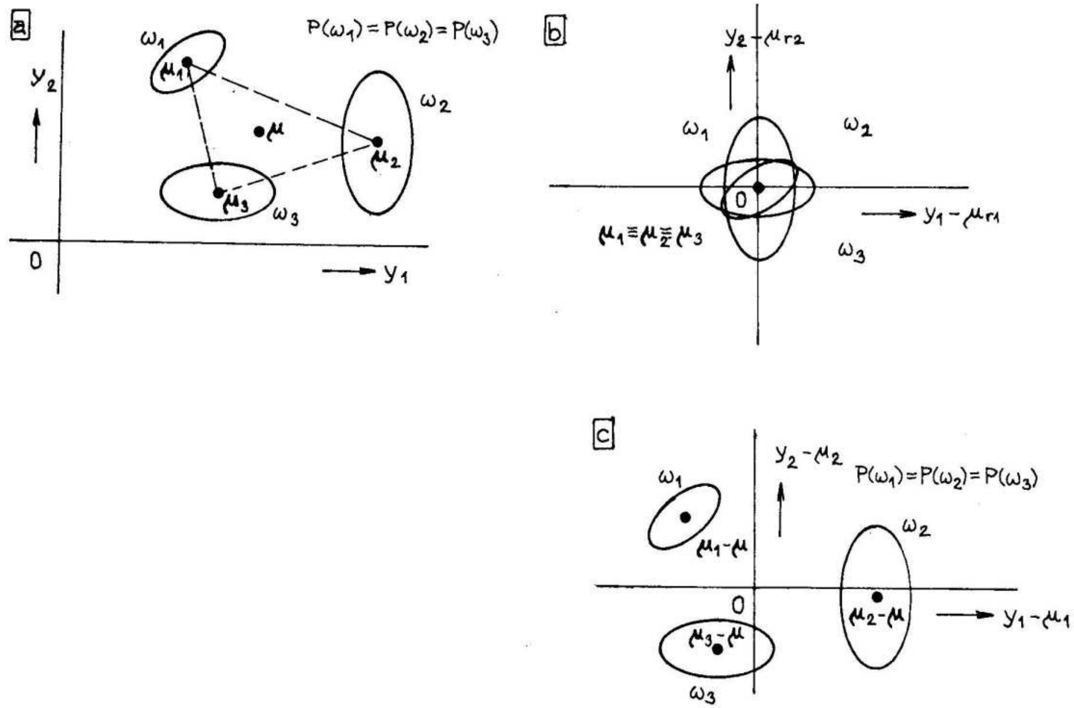
$$D^0(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} (\mathbf{y}-\boldsymbol{\mu}).^T(\mathbf{y}-\boldsymbol{\mu}).p(\mathbf{y}\,|\,\omega_r).dy =$$

$$= \int_{\Upsilon^m} (\mathbf{y}-\boldsymbol{\mu}).^T(\mathbf{y}-\boldsymbol{\mu}).p(\mathbf{y}).dy$$

ROZDĚLENÍ DO TŘÍD

kde střední hodnota **μ** je vážený průměr středních hodnot všech tříd, tj.

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



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“

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ