

# Bi8600: Vícerozměrné metody

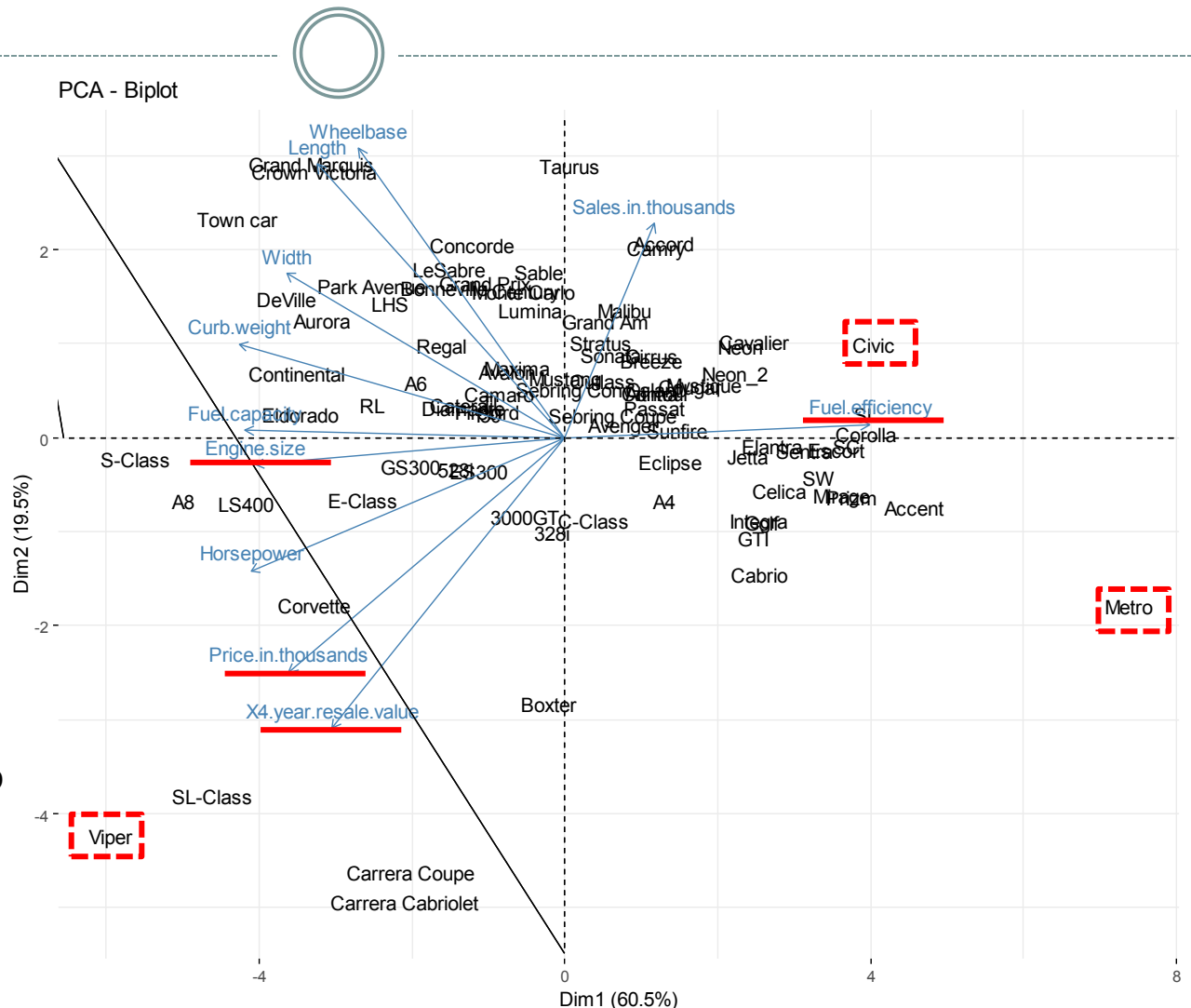
## 4. cvičení



**Opakování**  
**Ordinační metody**  
**Korespondenční analýza**  
**Nemetrické škálování**  
**Diskriminační analýza**

# Opakování

- Jaká je efektivita paliva a velikost motoru modelu Viper?
- Jaká je efektivita paliva a velikost motoru modelu Metro?
- Jaká je korelace efektivity paliva a velikosti motoru?
- Víme-li, že prodejní cena modelu Civic je po 4 letech nízká, jaká bude cena (Price.in.thousands) tohoto modelu?



# Korespondenční analýza - otázky



- Korespondenční analýza je nástroj pro hodnocení vztahů mezi ... a ... datové matice.
- Co popisuje vlastní číslo v korespondenční analýze?
- Co značí vysoká hodnota inercie? V jaké situaci bude hodnota inercie nízká?
- Vyberte, co lze interpretovat z biplotu korespondenční analýzy:
  - 1) vztah objektů
  - 2) vztah proměnných
  - 3) vztah objektů a proměnných
- Jaký maximální počet nových os může vzniknout?

# Korespondenční analýza



- Analogie k PCA
- Vstupní data = agregované údaje objektů/vzorků (průměry, počty)
- Výpočet = analýza vlastních čísel na matici chi-kvadrát hodnot.

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^n [X_{ij} - E_{ij}]^2 / E_{ij}$$

- CA přerozděluje inercií, vysoká inercie – silná vazba mezi řádky a sloupci
- Využití: nejčastěji data abundancí (ekologii), dotazníkové studie
- Nevýhoda: upřednostňuje unikátní málo četné kombinace

## R packages

- `CA()` [*FactoMineR* package],
- `ca()` [*ca* package],
- `dudi.coa()` [*ade4* package],
- `corresp()` [*MASS* package],
- and `epCA()` [*ExPosition* package]

# Korespondenční analýza – interpretace biplotu I.



- Pozice objektů (vzorky, v obrázku plná kolečka) a proměnných (druhy, prázdné čtverečky) v biplotu korespondenční analýzy interpretujeme následujícím způsobem:

2. Druhy, které se vyskytovaly spolu ve vzorcích, budou v ordinačním diagramu umístěny poblíž sebe (C, D).

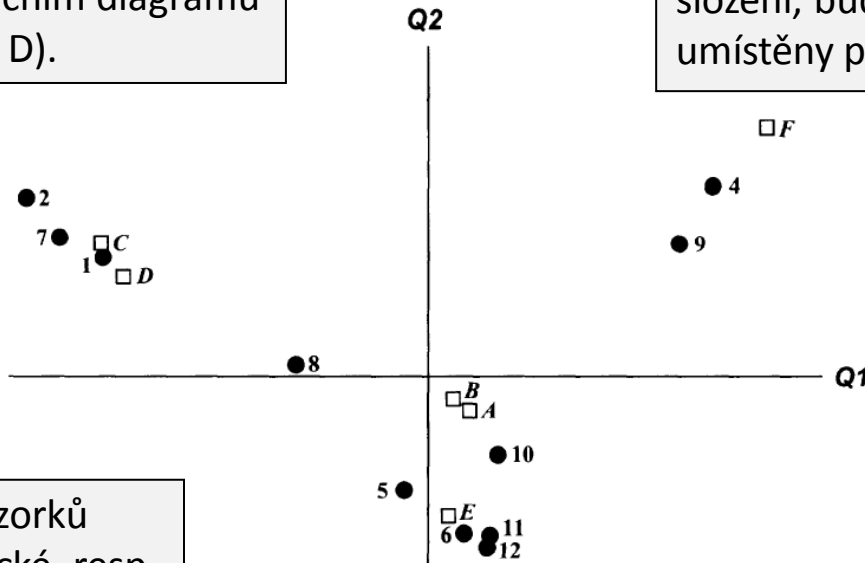
Druhy, které se vyskytovaly v jiných vzorcích, budou v diagramu umístěny dále od sebe (E, F).

3. Druhy umístěny poblíž vzorků byly pro tyto vzorky typické, resp. se vyskytovaly pouze v nich (1-C).

Když se druh v daném vzorku nevyskytoval, budou od sebe druh a vzorek v ordinačním diagramu vzdáleny (1-F).

1. Vzorky, které mají podobné druhové složení, budou v ordinačním diagramu umístěny poblíž sebe (4, 9).

Vzorky, které nemají společné druhy, budou v ordinačním diagramu umístěny dále od sebe (1, 9).



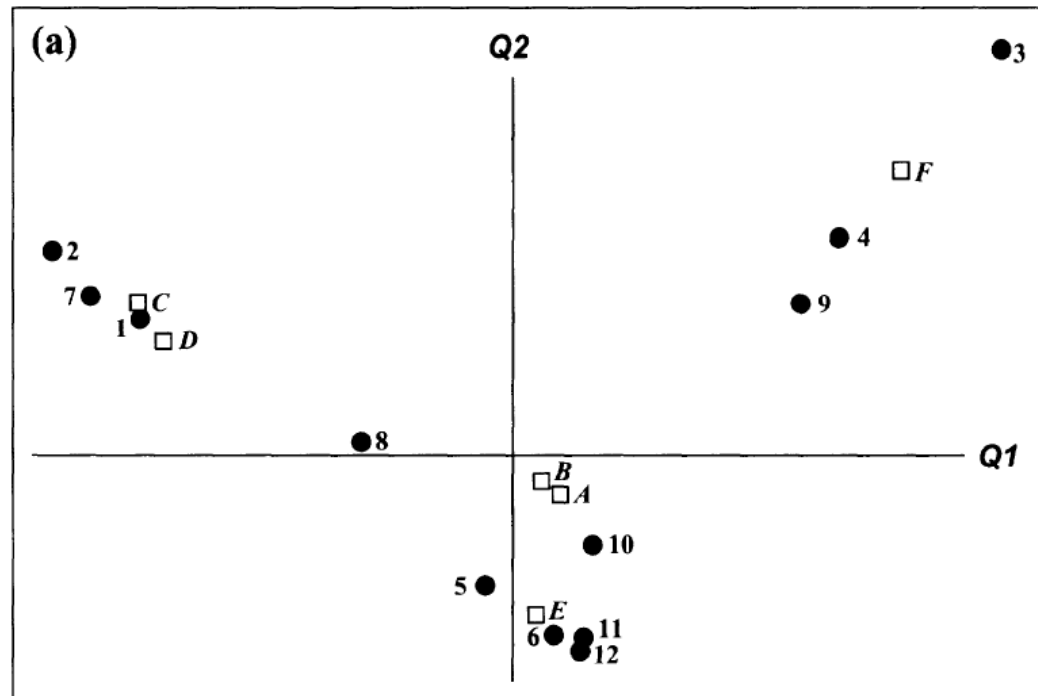
Kenkel et al. (2002)

Body poblíž středu ordinačního diagramu nemají výrazný profil (B, A).

# Korespondenční analýza – interpretace biplotu II.



- Interpretujte biplot z korespondenční analýzy:
  - a) Vztah vzorku 2 vs. 7 a 2 vs. 9.
  - b) Které druhy se vyskytovaly ve stejných a které v odlišných vzorcích?
  - c) Ve kterém vzorku je nejvíce přítomný druh E a C?



Kenkel et al. (2002)

# Nemetrické škálování (NMDS)



- Jaký je princip a základní výstup ne/metrického škálování?
- Jaký je rozdíl mezi metrickým a nemetrickým škálováním?
- Jaké jsou předpoklady NMDS?

# Diskriminační analýza – proč?



- Jak se liší diskriminační analýza od shlukové analýzy? („unsupervised“ vs. „supervised“)
- Doplňte: „Nové osy diskriminační analýzy jsou tvořeny tak, aby ... “
- Co vyjadřuje vlastní číslo osy diskriminační analýzy?
- Jaké jsou předpoklady diskriminační analýzy?



# Diskriminační analýza – OMEZENÍ?



- Předpoklad vícerozměrného normálního rozdělení prediktorů v každé ze skupin.
- Citlivá na přítomnost odlehlých hodnot.
- Citlivá na redundantní proměnné v modelu.
- Rovnice modelu je v základní verzi lineární a tedy i hodnocený problém musí mít lineární řešení.

# Diskriminační analýza - cíle



1. **Vytvoření zástupných proměnných**, které nejlépe odliší skupiny objektů.
2. **Vytvoření pravidla pro klasifikaci** objektů do skupin.
  - a) Identifikace proměnných diskriminujících mezi předem danými skupinami objektů.
  - b) Vyhodnocení klasifikace pro objekty, u kterých známe zařazení do skupin.
3. **Klasifikace** nových objektů do skupin.

## Využití:

- v antropologii pro klasifikaci koster,
- v medicíně k určení rizikovosti pacientů,
- ve finančnictví k předvídání krachů firem,
- v biologii ke klasifikaci rostlin,
- v sociologii u psychologických testů.



Iris Versicolor

Iris Setosa

Iris Virginica

# Výběr proměnných do modelu



- Výběr provádíme na základě:
  1. Expertní znalosti proměnných (zohledňujeme např. finanční zátěž, chybovost měření, vyplněnost).
  2. Pozorovaných dat (hodnotíme korelace proměnných, přínos unikátní informace - % rozptylu, které popisuje, příspěvek k diskriminaci, atd. ).
  3. Dopředné/zpětné eliminace (proměnné jsou postupně přidávány/odebírány tak, aby došlo k významnému „zlepšení“ modelu).

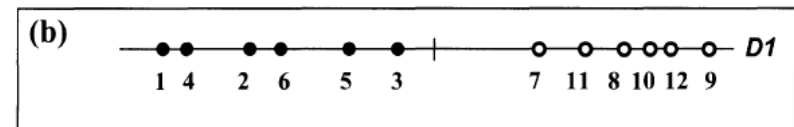
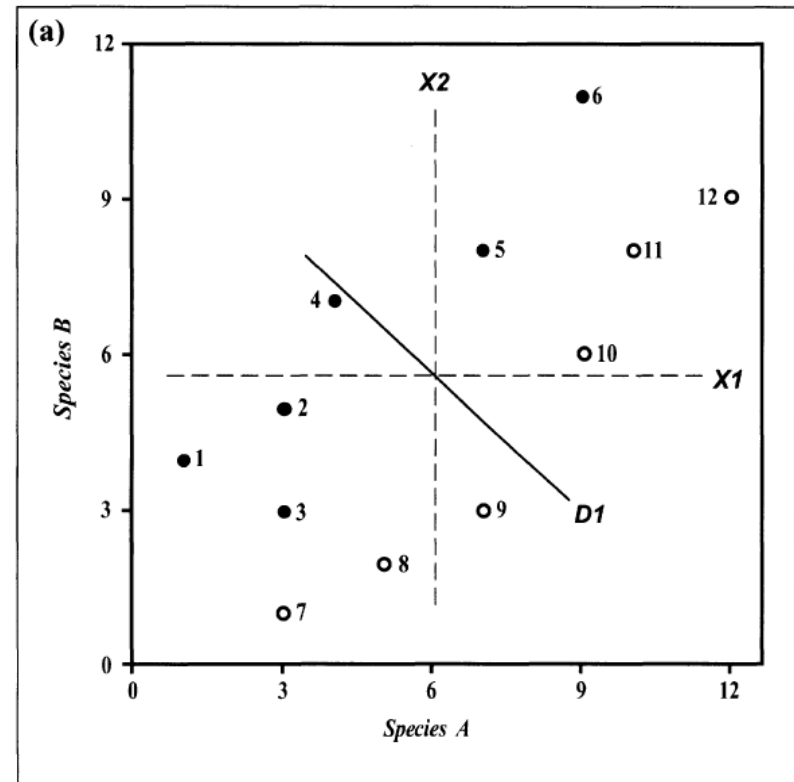
# Diskriminační analýza – algoritmus I



## 2 fáze výpočtu:

### 1. Vytvoření kanonických os

- Z původně vysokého počtu parametrů vytvoříme nové osy, které odliší shluky v datech.
- Pomocí vlastních čísel opět vybíráme počet os, které nejlépe popisují rozdíl mezi skupinami.
- Osy nejsou v prostoru původních proměnných ortogonální (jako tomu bylo u PCA).
- Maximální počet os je roven počtu skupin mínus jedna.



Kenkel et al. (2002)

# Diskriminační analýza – algoritmus II



## 2 fáze výpočtu:

### 2. Klasifikace objektů do skupin.

- Na vstupu definujeme apriorní pravděpodobnosti zařazení objektů do skupin.
- Pro každý objekt je spočítána vzdálenost od centroidu dané skupiny.
- Kombinací apriorní pravděpodobnosti a Mahalanobisovy vzdálenosti jsou spočítány posteriorní pravděpodobnosti zařazení objektu do dané skupiny.
- Pro každou ze skupin je definována diskriminační funkce. Při klasifikaci nových objektů zařadíme objekt do té skupiny, kde diskriminační funkce nabývá maxima.

# Výstup diskriminační analýzy



➤ **Popis významu proměnných v modelu:**

- a) Wilksovo lambda modelu,
- b) Wilksovo lambda proměnných,
- c) Parciální lambda,
- d) Tolerance.

➤ **Kanonická analýza:**

- a) Vlastní vektory,
- b) Vlastní čísla.

➤ **Klasifikace objektů:**

- a) Apriorní pravděpodobnost,
- b) Mahalanobisova vzdálenost,
- c) Diskriminační funkce,
- d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



- **Popis významu proměnných v modelu:**
  - a) Wilksovo lambda modelu - analogické s ANOVA – hodnotí podíl vnitroskupinového a celkového rozptylu (rozsah: 0–1; hodnoty blízké nule značí dobrou diskriminaci skupin),
  - b) Wilksovo lambda proměnných,
  - c) Parciální lambda,
  - d) Tolerance.
  
- **Kanonická analýza:**
  - a) Vlastní vektory,
  - b) Vlastní čísla.
  
- **Klasifikace objektů:**
  - a) Apriorní pravděpodobnost,
  - b) Mahalanobisova vzdálenost,
  - c) Diskriminační funkce,
  - d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



## ➤ Popis významu proměnných v modelu

- a) Wilksovo lambda modelu,
- b) Wilksovo lambda proměnných - wilksovo lambda celého modelu při vyřazení dané proměnné (naopak: čím větší, tím je proměnná důležitější pro diskriminaci),
- c) Parciální lambda,
- d) Tolerance.

## ➤ **Kanonická analýza:**

- a) Vlastní vektory,
- b) Vlastní čísla.

## ➤ **Klasifikace objektů:**

- a) Apriorní pravděpodobnost,
- b) Mahalanobisova vzdálenost,
- c) Diskriminační funkce,
- d) Posteriorní pravděpodobnost.



# Výstup diskriminační analýzy



- **Popis významu proměnných v modelu:**
  - a) Wilksovo lambda modelu,
  - b) Wilksovo lambda proměnných,
  - c) **Parciální lambda**: unikátní příspěvek dané proměnné k diskriminaci (čím nižší je hodnota, tím větší unikátní diskriminační sílu prediktor nese),
  - d) Tolerance.
  
- **Kanonická analýza:**
  - a) Vlastní vektory,
  - b) Vlastní čísla.
  
- **Klasifikace objektů:**
  - a) Apriorní pravděpodobnost,
  - b) Mahalanobisova vzdálenost,
  - c) Diskriminační funkce,
  - d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



- **Popis významu proměnných v modelu:**
  - a) Wilksovo lambda modelu,
  - b) Wilksovo lambda proměnných,
  - c) Parciální lambda,
  - d) **Tolerance:** unikátní variabilita proměnné nevysvětlená ostatními proměnnými v modelu ( $1 - \text{tolerance} = R^2$  variabilita proměnné, kterou lze vysvětlit kombinací ostatních proměnných).
  
- **Kanonická analýza:**
  - a) Vlastní vektory,
  - b) Vlastní čísla.
  
- **Klasifikace objektů:**
  - a) Apriorní pravděpodobnost,
  - b) Mahalanobisova vzdálenost,
  - c) Diskriminační funkce,
  - d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



- **Popis významu proměnných v modelu:**
  - a) Wilksovo lambda modelu,
  - b) Wilksovo lambda proměnných,
  - c) Parciální lambda,
  - d) Tolerance.
  
- **Kanonická analýza:** vytváří nové osy tak, aby jejich diskriminační funkce byla co největší (počet nových os =  $\min(\text{počet skupin, počet proměnných}) - 1$ )
  - a) **Vlastní vektory**: určují směr nových os (definovány jako lineární kombinace proměnných v modelu).
  - b) **Vlastní čísla**: popisují podíl variability mezi a v rámci skupin objektů na nových osách. Osy s nízkou hodnotou vlastního čísla nepřispívají k popisu rozdílu mezi skupinami.
  
- **Klasifikace objektů:**
  - a) Apriorní pravděpodobnost,
  - b) Mahalanobisova vzdálenost,
  - c) Diskriminační funkce,
  - d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



➤ **Popis významu proměnných v modelu:**

- a) Wilksovo lambda modelu,
- b) Wilksovo lambda proměnných,
- c) Parciální lambda,
- d) Tolerance.

➤ **Kanonická analýza:**

- a) Vlastní vektory,
- b) Vlastní čísla.

➤ **Klasifikace objektů:**

- a) Apriorní pravděpodobnost: pravděpodobnost výskytu objektu ve shluku (rovnoměrná/proporcionální/nastavená uživatelem na základě znalostí dané problematiky)
- b) Mahalanobisova vzdálenost,
- c) Diskriminační funkce,
- d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



➤ **Popis významu proměnných v modelu:**

- a) Wilksovo lambda modelu,
- b) Wilksovo lambda proměnných,
- c) Parciální lambda,
- d) Tolerance.

➤ **Kanonická analýza:**

- a) Vlastní vektory,
- b) Vlastní čísla.

➤ **Klasifikace objektů:**

- a) Apriorní pravděpodobnost,
- b) **Mahalanobisova vzdálenost:** Používána pro popis vzdáleností objektů od centroidů skupin a následně pro výpočet posteriorních pravděpodobností,
- c) Diskriminační funkce,
- d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



➤ **Popis významu proměnných v modelu:**

- a) Wilksovo lambda modelu,
- b) Wilksovo lambda proměnných,
- c) Parciální lambda,
- d) Tolerance.

➤ **Kanonická analýza:**

- a) Vlastní vektory,
- b) Vlastní čísla.

➤ **Klasifikace objektů:**

- a) Apriorní pravděpodobnost,
- b) Mahalanobisova vzdálenost,
- c) **Diskriminační funkce:** pro každou skupinu jedna rovnice, objekt je zařazen do skupiny s maximální hodnotou klasifikační funkce.
- d) Posteriorní pravděpodobnost.

# Výstup diskriminační analýzy



➤ **Popis významu proměnných v modelu:**

- a) Wilksovo lambda modelu,
- b) Wilksovo lambda proměnných,
- c) Parciální lambda,
- d) Tolerance.

➤ **Kanonická analýza:**

- a) Vlastní vektory,
- b) Vlastní čísla.

➤ **Klasifikace objektů:**

- a) Apriorní pravděpodobnost,
- b) Mahalanobisova vzdálenost,
- c) Diskriminační funkce,
- d) **Posterioerní pravděpodobnost:** pravděpodobnost klasifikace objektu do dané skupiny (kombinace Mahalanobisových vzdáleností objektů od centroidů shluků s apriorní pravděpodobností).