



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

RNDr. Danka Haruštiaková, PhD., RNDr. Jiří Jarkovský, PhD.,  
Mgr. Simona Littnerová

# Vícerozměrné statistické metody v biologii

Leden 2010



Příprava a vydání těchto učebních textů byly podporovány projektem ESF č.  
CZ.1.07/2.2.00/07.0318

„Víceborová inovace studia Matematické biologie“ a státním rozpočtem České republiky.

Vícerozměrné statistické metody představují velice užitečný nástroj pro uchopení, zjednodušení a vizualizaci velmi složitých dat. Použitelnost těchto metod v přírodních vědách je velmi široká, často se s nimi setkáváme nejenom v ekologii, experimentální biologii, antropologii, environmentální chemii, ale i v geografii a geologii. Zpracování rozsáhlých biologických a hlavně ekologických dat se bez znalosti vícerozměrných statistických metod již neobejde. Na druhou stranu mohou v případě nesprávného užití vést k zavádějícím výsledkům, jejichž chybnost nemusí být ovšem na první pohled zřejmá, protože je skryta za složitou strukturou dat a komplikovaností výpočtu. Znalost vícerozměrných statistických metod se tak stala nutnou součástí biologického vzdělání.

Cílem tohoto učebního textu není podrobný teoretický výklad jednotlivých typů vícerozměrných analýz, ale ve stručné a přehledné formě představit postupy analýz, objasnit základy jejich využití včetně potenciálně slabých míst a poskytnout návody ke správné interpretaci výsledků.

Dostupnost nových studijních materiálů, kterých je v současné době stále nedostatek, by měla přispět k zvýšení odbornosti studentů biologie i dalších přírodovědných oborů.

Jednotlivé představené metody vícerozměrné analýzy dat je možné použít při studiu různých biologických objektů. Vzhledem k zaměření autorů pocházejí ovšem uváděné příklady zejména z oblasti ekologie živých organismů.

Česká a ani anglická terminologie používaná v dostupné literatuře není zcela stabilizovaná a často se stává, že tytéž metody jsou v různých učebnicích a statistických programech uváděny různými názvy. Z tohoto důvodu uvádíme anglické názvy metod kurzivou a další ekvivalenty názvů v závorkách.

Na tomto místě bychom rádi poděkovali za připomínky recenzentům, jejichž poznámky výrazně zlepšily kvalitu těchto učebních textů.

Příprava a vydání těchto učebních textů byly podporovány projektem ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“ a státním rozpočtem České republiky.

V Brně, 2011,

autoři

# 1 Úvod

## 1.1 Smysl a cíle vícerozměrné analýzy dat

Veškerý svět kolem nás je vícerozměrný. Kromě vnímání třírozměrného tvaru můžeme každý objekt popsat celou řadou dalších charakteristik, jako je třeba barva, hmotnost, chuť atd. Přes tuto skutečnost, kterou vnímáme každý den, je pro nás ovšem problémem představit si tento stav popsáný ve formě datové tabulky a nebo jej dokonce nějakým způsobem popsat jinému člověku – nastává zde tedy místo pro speciální typ analýzy, tedy vícerozměrnou analýzu.

Metody vícerozměrné analýzy jsou velmi užitečným prostředkem pro explorativní analýzu složitých dat.

Ačkoliv klasická statistika zná řadu způsobů popisu jednotlivých měřených nebo pozorovaných proměnných, je pro nás v případě hodnocení velkého množství proměnných velmi obtížné si tyto výstupy poskládat do jednoduchého obrazu vedoucího k pochopení podstaty. Právě vícerozměrná analýza dat je nástrojem sloužícím k usnadnění tohoto procesu a její přínos lze shrnout následovně:

- nalezení smysluplných pohledů na data popsaná velkým množstvím proměnných,
- nalezení a popsání skrytých vazeb mezi proměnnými a tak zjednodušení jejich struktury,
- jednoduchá vizualizace dat, kdy v jediném grafu se skrývá informace např. z 20 proměnných,
- umožnění a/nebo zjednodušení a interpretace dat na základě jejich zjednodušení a vizualizace vícerozměrnou analýzou.

Ačkoliv je v případě vícerozměrných analýz používána celá řada matematických postupů, jedno mají všechny tyto analýzy společné – hledají, které naměřené proměnné nebo objekty spolu nějakým způsobem souvisí a které je tedy možné jako podobné sloučit a tak snížit složitost naměřených dat.

Na tomto místě musíme uvést i nevýhody vícerozměrné analýzy dat.

Zjednodušení vícerozměrného problému je možné pouze tehdy, kdy existuje vazba mezi naměřenými proměnnými. Pokud by mezi nimi žádná vazba neexistovala, nebo byla velmi slabá, nemá smysl vícerozměrné metody používat.

Dalším kamenem úrazu může být nesprávné použití metody, které může vést k zavádějícím výsledkům. Při zpracovávání vícerozměrných dat ovšem nemusí být tato chyba patrná, protože je zakryta složitou strukturou dat a náročností výpočtu.

Příklady užití vícerozměrných metod můžeme najít v různých oblastech, nejen v přírodovědných oborech, ale také v sociologii, ekonomii i marketingu. Z oblasti biologických věd můžeme zmínit aplikace v ekologii, ekotoxikologii, taxonomii, etologii, antropologii atd.

Konkrétně z ekologie můžeme uvést využití mnohorozměrných metod např. při hodnocení vlivu environmentálních změn na biologická společenstva, klasifikaci vegetačních i půdních společenstev, atd.

## 1.2 Statistické software pro vícerozměrnou analýzu dat

V současnosti je k dispozici mnoho nástrojů ke zpracování a analýze mnohorozměrných dat. Nejrozšířenější a nejpoužívanější software pro vícerozměrnou analýzu uvádíme níže.

**Systém R** je volně dostupný software (<http://www.R-project.org>) pro zpracování dat a jejich analýzu s grafickými výstupy. Výhodou tohoto systému jsou algoritmy, které zatím v komerčních softwarových nástrojích nejsou tolik rozšířené. Systém R na rozdíl od jiných softwarů nabízí např. hodnocení výsledků shlukování ve formě tzv. Silhouette plot.

**SPSS** je běžný komerční software s rozšířenými možnostmi zpracování dat a jejich analýzy. Vícerozměrné metody jsou součástí tohoto softwaru, pro specifické potřeby biologa ovšem nemusí vždy postačovat.

**Statistica for Windows** je běžný komerční software na analýzu a zpracování dat s hezkými grafickými výstupy. Metody vícerozměrné analýzy jsou součástí tohoto softwaru, ovšem na rozdíl od specializovaných nástrojů je v něm omezené množství možných nastavení vícerozměrných analýz.

**Syntax 2000** je software zaměřený na analýzu ekologických a taxonomických dat. Poskytuje metody hierarchického shlukování, nehierarchického shlukování a ordinace. Výhodou tohoto softwarového nástroje je množství možných nastavení analýz, které v běžných komerčních softwarech nejsou k dispozici.

**Canoco for Windows 4.5** s dalšími aplikacemi je soubor nástrojů specializovaný na analýzu ekologických dat se zvláštním zaměřením na ordinační metody. K dispozici jsou všechny běžné ordinační metody, jejich kanonické formy, i hybridní formy. U kanonických ordinačních metod poskytuje možnost statisticky testovat významnost všech nezávislých proměnných a také kanonických os. V aplikaci **Canoco console 4.5** má uživatel další možnosti nastavení. Aplikace **CanoDraw for Windows** poskytuje hezké grafické výstupy analýz, které lze snadno upravovat.

**PAST** je volně dostupný software (<http://folk.uio.no/ohammer/past/>) vyvinutý původně pro analýzu paleontologických dat s rozsáhlou nabídkou méně obvyklých vícerozměrných analýz, včetně analýzy tvarů. Další výhodou je i nabídka metod pro analýzu biodiverzity, která z software PAST činí univerzální nástroj analýzy ekologických dat.

### 1.3 Parametrická a neparametrická vícerozměrná statistika

Vícerozměrná statistická analýza se řídí stejnými zákonitostmi jako klasická jednorozměrná analýza a řada jejích metod je citlivá na předpoklady o rozložení, přítomnost odlehklých hodnot apod.

Klasickým příkladem je provázanost analýzy hlavních komponent s parametrickou kovariancí nebo korelací, kdy přítomnost odlehklé hodnoty vede k vysoké hodnotě korelace a její významnosti i když zbývající data nevykazují žádný vztah. V případě analýzy hlavních komponent tato situace vede k tomu, že první, nejdůležitější, faktorová osa ukazuje pouze informaci o přítomnosti odlehklé hodnoty v datech a nijak nepřispívá k pochopení zdrojů variability dat. Naproti tomu některé vícerozměrné metody lze považovat za velmi robustní a analogické neparametrickým přístupům klasické statistiky (např. některé shlukovací algoritmy).

Z těchto důvodů je při výpočtu vícerozměrných analýz třeba věnovat odpovídající pozornost ověření předpokladů, které jsou v rámci učebního textu také u jednotlivých metod uvedeny.

## 2 Datové podklady

Podkladem každé vícerozměrné analýzy je vždy tabulka (Tabulka 1.1) obsahující v řádcích jednotlivé měřené objekty (např. lokality, vzorky, respondenty) a ve sloupcích proměnné měřené na těchto objektech. Každá proměnná představuje jeden rozměr objektu.

**Tabulka 1.1** Ukázka datové tabulky.

Vzorek	Půdní typ	<i>Quercus</i> (B-B stupnice)*	Teplota vzduchu (°C)	Srážky (měsíční úhrn mm)
1	jíl	2	21	25
2	jíl	1	18	10
3	jíl	2	19	30
4	rašelina	1	20	62
5	písek	4	17	8
6	písek	3	21	4
...	...	...	...	...

\* Braun-Blanquetova stupnice

### 2.1 Typy dat

Data je možné měřit v následujících stupnicích (škálách):

- **Nominální stupnice** (*nominal scale*)

Tato stupnice je kvalitativní. Hodnoty nemají mezi sebou žádný vztah, platí zde pouze rovnost a nerovnost. Jako příklad lze uvést proměnnou půdní typy, která nabývá hodnot „jíl“, „rašelina“, „písek“. Kódy přiřazené k těmto hodnotám (např. „1“, „2“, „3“) pouze označují dané hodnoty a neplatí mezi nimi vztah „větší“ a „menší“. Specifické postavení mezi znaky zaznamenávanými na nominální stupnici mají znaky binární – tyto nabývají pouze dvou hodnot (např. proměnná pohlaví: muž, žena).

- **Pořadová stupnice** (*ordinal scale*)

Pro hodnoty na pořadové stupnici kromě rovnosti a nerovnosti lze určit také vztah menší a větší. Příkladem proměnné měřené na této škále je abundance rostlin měřená na Braun-Blanquetova stupnici, která pokryvnost rostlinných taxonů hodnotí na 7-stupňové škále. Možné hodnoty nebo kódy této stupnice lze seřadit od nejnižší abundance po nejvyšší. Ovšem nelze určit, zda rozdíl mezi hodnotami „1“ a „2“ je větší nebo menší než rozdíl mezi hodnotami „4“ a „5“.

- **Intervalová stupnice** (*interval scale*)

Kromě vlastností předchozích dvou stupnic je zde možné také sčítání a odečítání. Na rozdíl od pořadové stupnice lze zde vyjádřit míru rozdílu mezi objekty. Intervalová stupnice ovšem nemá přirozený nulový bod. Příkladem je teplota měřena v stupních Celsia. Rozdíl 5 stupňů znamená to stejné přes celou stupnici. Hodnota 0 je reálná teplota; lze určit rozdíl mezi hodnotou 0 a 5 stupňů, nelze ovšem určit, kolikrát je hodnota 5 vyšší než hodnota 0.

- **Poměrová stupnice** (*ratio scale*)

Poměrová stupnice dovoluje vyjádřit poměr mezi hodnotami. Tato stupnice má přirozený nulový bod, lze proto určit poměr (např. hodnoty délky, plochy nebo objemu).

Z hlediska statistického zpracování dat můžeme proměnné rozdělit na:

1. **kvalitativní** (*qualitative*)
  - a. **binární** (*binary*, dvoustavové, alternativní) – nabývají pouze dvou hodnot, většinou je kódujeme 0 a 1 (např. přítomnost nebo nepřítomnost určitého živočišného druhu)
  - b. **vícestavové** (*multistate*) – nabývají vícero hodnot, např. výše uvedené typy půd
2. **semikvantitativní** (*semiquantitative*) – do této skupiny patří proměnné, jejichž hodnoty jsou vyjádřeny pomocí odhadové stupnice, která nemá konstantní rozdíly mezi sousedícími hodnotami (např. Braun-Blanquetova stupnice pokryvnosti)
3. **kvantitativní** (*quantitative*) – proměnné lze vyjádřit měřitelnou stupnicí, na níž jsou konstantní rozdíly mezi jednotkami
  - a. **nespojitý, diskrétní** (*discontinuous, discrete*) – proměnné, které nabývají pouze určité numerické hodnoty (např. počet květů)
  - b. **spojitý, kontinuální** (*continuous*) – proměnné, které mohou nabývat nekonečného počtu hodnot mezi dvěma pevnými body dané stupnice (např. výška stromů, koncentrace rtuti v půdě, apod.).

V analýzách je problematické použití vícestavových kvalitativních proměnných. Alternativní možností jak pracovat s takovými daty je jejich převedení do umělých binárních proměnných, tzv. dummy variables, kde každý stav převedeme na novou binární proměnnou kódovanou 0 a 1, kde 1 znamená přítomnost daného stavu.

## 2.2 Možné problémy dat a jejich řešení

Různé metody vícerozměrné analýzy kladou několik požadavků na vstupní data. V prvním řadě všechny metody vyžadují úplné datové matice bez chybějících dat. Některé metody jsou dostatečně robustní ve vztahu k odchylkám od normálního rozložení dat, některé metody vyžadují mnohorozměrné normální rozložení dat. Tento problém lze vyřešit vhodnou transformací dat. V některých případech mají měřené proměnné různé jednotky, často se řádově liší a tak je vhodné převést proměnné na stejné měřítko. K tomu slouží standardizace dat.

### 2.2.1 Chybějící data

V případě, že některé hodnoty není možné určit nebo naměřit, je nutné tyto situace ošetřit. K tomu je několik možností.

1. Objekty, ve kterých hodnoty chybí, můžeme vypustit. Toto řešení je vhodné tehdy, když jsou chybějící data pouze v několika málo objektech.
2. Proměnné, u kterých hodnoty chybí, můžeme vypustit, pokud jich není mnoho a nejde o klíčové proměnné.
3. Chybějící hodnoty můžeme doplnit a to různými metodami:
  - a. doplnění průměru z hodnot, které jsou k dispozici,
  - b. dopočítání chybějících hodnot pomocí mnohonásobného regresního modelu za použití objektů bez chybějících hodnot.

Tyto metody ovšem způsobí duplikaci informace, kterou již známe a dochází tím ke snížení počtu nezávislých pozorování v datech, čili stupňů volnosti. Takto upraveným objektům je pak možné přiřadit menší statistickou váhu.

### 2.2.2 Transformace dat

Transformace je možná několika způsoby. K transformaci se používají konstanty a funkce nezávislé na analyzovaných datech.

**Lineární transformace** (např. násobení hodnot proměnné konstantou) nemění výsledky analýzy v případech, že jde o analýzu kvalitativního vztahu proměnných (např. korelace); v případě, že je důležitá absolutní hodnota proměnné dochází k vážení jejího významu v analýze.

Dalším příkladem je **adjustace proměnné** na vliv jiných proměnných pomocí jejich lineární kombinace (např. adjustace hladiny hemoglobinu na věk pacientů). Tato úprava mění i interpretaci výsledné proměnné.

Většina transformací, které se používají v biologii, jsou **nelineární transformace**. Tyto transformace mění rozdělení dat.

#### Logaritmická transformace

$$y_{ij} = \log_c x_{ij} \quad \text{nebo (když jsou přítomny nuly)} \quad y_{ij} = \log_c (x_{ij} + 1) \quad (2.1)$$

Tato transformace se často používá ze čtyř různých důvodů:

- k získání statisticky vhodných vlastností normálního rozložení u proměnných s log-normálním rozložením,
- k dosažení homogenity rozptylu
- k linearizaci vztahu proměnných
- k přiřazení menší váhy dominantním proměnným, čili ke zvýraznění kvalitativní stránky dat.

#### Odmocninová transformace

$$y_{ij} = \sqrt{x_{ij}} \quad (2.2)$$

Proměnné nesmí dosahovat nulových hodnot, a proto se někdy používá ve tvaru:

$$y_{ij} = \sqrt{x_{ij} + 0.5} \quad (2.3)$$

Tato transformace se používá:

- před analýzou proměnných s Poissonovým rozdělením (např. počet jedinců určitého druhu získaných z jedné pasti za určitou časovou jednotku),
- k přiřazení nižší váhy dominantním proměnným.

#### Arkussinová transformace

Používá se v kombinaci s odmocninovou transformací.



$$y_{ij} = \arcsin \sqrt{x_{ij}} \quad (2.4)$$

Předpokládá, že data jsou měřena v intervalu  $<0,1>$ .

Používá se na úpravu procentuálních hodnot vyjadřených v intervalu  $<0,1>$  (např. vegetační pokryvnosti druhů).

### Exponenciální transformace

$$y_{ij} = a^{x_{ij}} \quad (2.5)$$

Když  $a$  je reálné číslo větší než 1, jsou zvýrazněny dominantní proměnné, pro hodnoty  $a < 1$  se běžně nepoužívá.

### Transformace na ordinální škálu

Hodnoty proměnných jsou převedeny do tříd. Čím vyšší je číslo třídy, tím vyšší byla původní hodnota. Ovšem stejné číslo třídy nemusí vždy znamenat stejnou hodnotu původní proměnné. Intervaly tříd nemusí být stejné.

Typickou transformací na ordinální škálu je použití Braun-Blanquetovy stupnice při kvantifikaci pokryvnosti vegetace (Tabulka 2.1).

**Tabulka 2.1** Braun-Blanquetova stupnice pokryvnosti vegetačních druhů.

stupeň	popis	kód
r	druh velmi vzácný, jen 1-3 drobné exempláře	1
+	pokryvnost nižší než 1 %	2
1	pokryvnost 1 - 5 %	3
2	pokryvnost 5 - 25 %	4
3	pokryvnost 25 - 50 %	5
4	pokryvnost 50 - 75 %	6
5	pokryvnost 75 - 100 %	7

Extrémem je binarizace – transformace na prezenci a absenci.

$$y_{ij} = 0 \quad \text{když} \quad x_{ij} = 0 \quad y_{ij} = 1 \quad \text{když} \quad x_{ij} > 0 \quad (2.6)$$

Transformací na ordinální škálu se vždy strácí část informace. V některých případech je ovšem tato transformace jediná možnost jak dosáhnout srovnatelnosti dat (např. třídy ekologického stavu).

Je ovšem velmi výhodné sbírat data v terénu na ordinální škále, tak jak je to např. běžné v botanickém monitoringu.

### 2.2.3 Standardizace dat

Ke standardizaci se používají statistiky odvozené z analyzovaného souboru dat (rozpětí, směrodatná odchylka, průměr, maximum atd.). Proměnné se tímto postupem provádějí na stejné měřítko, čili přestává záležet na skutečném rozměru příslušné proměnné. K nejčastějším úpravám patří centrování a standardizace směrodatnou odchylkou.

### Standardizace rozpětím

$$y_{ij} = \frac{x_{ij} - \min_j \{x_{ij}\}}{\max_j \{x_{ij}\} - \min_j \{x_{ij}\}} \quad (2.7)$$

Doporučuje se použít v případech, kdy jsou sice proměnné měřeny ve stejném měřítku, ovšem mezi jejich hodnotami jsou velmi velké rozdíly.

### Centrování

Při centrování je od původní hodnoty pouze odečítán průměr proměnné, tj. od prvků sloupce se odečte jejich sloupcový aritmetický průměr.

$$y_{ij} = x_{ij} - \bar{x}_j \quad (2.8)$$

### Standardizace směrodatnou odchylkou

Pod pojmem standardizace většinou rozumíme úpravu hodnot proměnné tak, aby standardizovaná proměnná měla nulový průměr a rozptyl roven jedné. Nová hodnota se získá odečtením sloupcového průměru od původní hodnoty a podělením sloupcovou střední hodnotou. Výpočtem dostáváme tzv. z-skóre.

$$y_{ij} = z = \frac{x_{ij} - \bar{x}_j}{s_{x_j}} \quad (2.9)$$

V další části jsou představeny metody standardizace ekologických dat, které se používají zejména ve shlukové analýze. Standardizace je definována jako použití určitého standardu pro všechny proměnné (v ekologických studiích jde např. o druhy) nebo objekty (vzorky, lokality) před spočítáním (ne)podobností nebo před aplikací analýzy.

### Standardizace na celkový součet řádku

Hodnoty proměnných v objektu se sečtou a každá hodnota je vydělena tímto součtem.

V ekologických studiích se takto určí relativní abundance (dominance) druhů. Je potřebné používat tuto standardizaci opatrně v případě, že jsou součty řádků velmi rozdílné, protože vzácné druhy se objevují až ve vzorcích s vysokým počtem jedinců.

$$y_{ij} = \frac{x_{ij}}{\sum_i x_{ij}} \quad (2.9)$$

### Standardizace na celkový součet sloupce

Pro každý sloupec (proměnná) je určen součet přes všechny objekty. Původní hodnoty jsou pak poděleny sloupcovým součtem.

V ekologických studiích, kde proměnné představují jednotlivé druhy, tímto způsobem získáme frekvence druhů v objektech.

Tato standardizace silně nadváží vzácné druhy a podváží běžné druhy. Proto se tato standardizace doporučuje pouze tehdy, když se frekvence druhů v tabulce velmi neliší. Tato standardizace bývá používána v případech, když se v seznamu druhů vyskytují různé trofické úrovně, protože vyšší trofické úrovně jsou méně zastoupeny (a proto může vyhovovat jejich nadvážení).

$$y_{ij} = \frac{x_{ij}}{\sum_j x_{ij}} \quad (2.10)$$

### Standardizace na maximum řádku

Všechny hodnoty v řádku jsou poděleny maximální hodnotou dosaženou u některé proměnné v řádku.

Tato standardizace je aplikovaná ze stejného důvodu jako standardizace na celkový součet řádku. Je méně citlivá na počet proměnných, je ovšem potřeba užívat ji opatrně v případech, když jsou veliké rozdíly ve vyrovnanosti vzorků.

$$y_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}} \quad (2.11)$$

### Standardizace na maximum sloupce

Všechny hodnoty v sloupci jsou poděleny maximální hodnotou sloupce.

Tato standardizace je v ekologických studiích doporučovaná podobně jako standardizace na celkový součet sloupce, když jsou přítomny různé trofické úrovně.

$$y_{ij} = \frac{x_{ij}}{\max_j \{x_{ij}\}} \quad (2.12)$$

### Standardizace na jednotkovou délku vektoru řádku

Podělením hodnot proměnných u objektu odmocninou sumy čtverců hodnot se všechny vektory objektů zobrazí na jednotkové kružnici prostoru tvořeného proměnnými (v ekologických studiích jde o druhy). Euklidovské vzdálenosti se touto standardizací redukuje na tětivové vzdálenosti (*chord distance*).

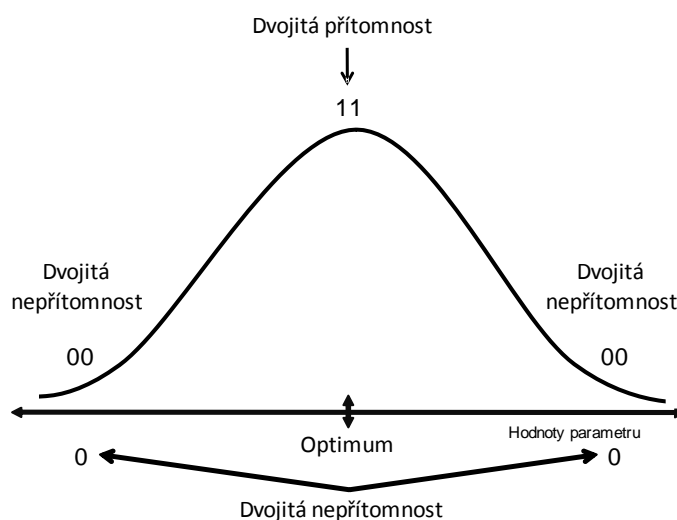
$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2}} \quad (2.13)$$

#### 2.2.4 Problém dvou nul (*double zero problem*)

Problém dvou nul je častým problémem v ekologických studiích. Vyskytuje se u proměnných, kde nula znamená nepřítomnost a ne normální hodnotu stupnice. Typickým příkladem jsou početnosti (abundance) druhů.. Druhy jsou známy unimodální distribucí niky podél enviromentálního gradientu. Jestliže se druh na porovnávaných objektech (např. lokalitách) vyskytuje, indikuje to jejich podobnost, není-li však zastoupen na žádné může to

být např. způsobeno tím, že environmentální podmínky nik obou lokalit jsou buď „vyšší“ než optimální nika anebo má jedna z nich „vyšší“ a druhá „nižší“ než vlastnosti optimální niky. Proto je lépe nedělat ekologické závěry ze společné absence druhu na porovnávaných objektech (Obrázek 2.1). Tento problém se samozřejmě netýká pouze binárních dat prezence/absence, ale i kvantitativní analýzy absence/početnost. Problém dvou nul je častým problémem vícerozměrné analýzy v ekologii. Z tohoto důvodu není také vhodné analyzovat složení společenstev pomocí analýzy hlavních komponent PCA, která je na tento problém citlivá.

V praxi to znamená vybrat pro analýzu takovýchto dat pouze vhodné metody (např. asymetrické koeficienty podobnosti, korespondenční analýza) neovlivněné tímto problémem.



**Obrázek 2.1** Problém dvou nul (double-zero problem). Dvojitá nepřítomnost není stejná jako dvojitá přítomnost.

### 3 Vícerozměrné normální rozdělení

Použitelnost mnohých klasických statistických metod a postupů vyžaduje předpoklad o normálním rozdělení sledovaných proměnných. Podmínka normality vyplývá z toho, že metody založené na tomto předpokladu mohou využít kompletní matematický aparát schovaný za danou statistickou metodou. Tyto metody jsou také relativně snadno pochopitelné a se získanými řešeními se dobře pracuje. Ovšem v reálném světě bývá obtížné předpoklad o normálním rozložení dodržet, v mnohých oblastech přírodních a mnohdy i technických oborů není tento předpoklad samozřejmostí.

Předpokládejme však normalitu a předpoklad o jedné normálně rozložené náhodné proměnné můžeme rozšířit na předpoklad simultánního normálního rozložení dvou a více náhodných proměnných. Některé vícerozměrné postupy a metody vycházejí z předpokladu vícerozměrného normálního rozdělení. Vícerozměrné normální rozdělení může být také velmi užitečnou aproximací různých jiných simultánních rozdělení.

Vícerozměrné normální rozdělení je rozšířením jednorozměrného normálního rozložení pro více jak jednu náhodnou proměnnou ( $p \geq 2$ ). Náhodný vektor  $\mathbf{x}$  má vícerozměrné normální rozložení má-li jeho hustota pravděpodobnosti tvar

$$f(\mathbf{x}) = 2\pi^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right) \quad (3.1)$$

kde  $\boldsymbol{\mu}$  je vektor  $p$  středních hodnot (vektor průměrů) proměnných  $X_1, X_2, \dots, X_p$ ,  $\Sigma$  je kovariační matice (matice složená ze směrodatných odchylek).

Vícerozměrné normální rozložení má tyto vlastnosti:

1. lineární kombinace prvků  $\mathbf{x}$  mají normální rozložení
2. všechny podmnožiny  $\mathbf{x}$  mají normální rozložení
3. nekorelovanost náhodných proměnných z  $\mathbf{x}$  znamená jejich nezávislost
4. všechna podmíněná rozdělení jsou normální

Pro jednorozměrné normální rozložení má předešlý vzorec tvar

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (3.2)$$

V exponentu je čtverec vzdálenosti  $u^2 = \left(\frac{x-\mu}{\sigma}\right)^2$ , tedy vzdálenosti  $x$  od střední hodnoty  $\mu$  kde jednotkou vzdálenosti je  $\sigma$ .

Pro vícerozměrné normální rozložení můžeme chápat kvadratickou formu v exponentu jako čtverec vzdálenosti vektoru  $\mathbf{x}$  od vektoru  $\boldsymbol{\mu}$ , ve kterém je obsažena informace z kovarianční matice

$$C^2 = (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$$

$C$  je Mahalanobisova vzdálenost, pro zvolenou hodnotu  $f(\mathbf{x})$  její čtverec je geometricky plocha elipsoidu se středem  $\boldsymbol{\mu}$  a osami  $c\sqrt{\lambda_j} \mathbf{v}_j$  pro  $j = 1, 2, \dots, p$ , kde  $\lambda_j$  jsou vlastní čísla matice  $\Sigma$  a  $\mathbf{v}_j$  jsou vlastní vektory této matice.

$$C^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(p)$$

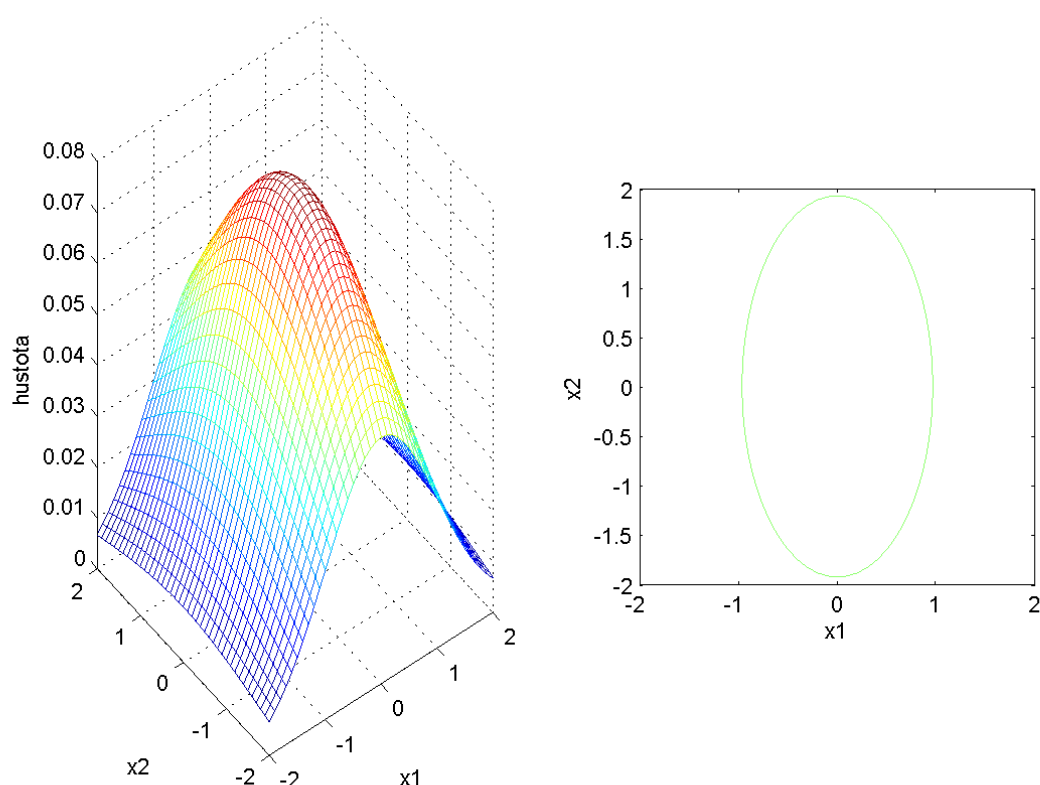
Dvourozměrné normální rozložení je speciální případ  $p$ -rozměrného normálního rozdělení pro  $p = 2$ . Jedná se o vhodné ilustrační schéma obecného případu. Máme dvě náhodné veličiny  $X_1$  a  $X_2$  se středními hodnotami  $\mu_1$  a  $\mu_2$ , s rozptyly  $\sigma_1^2$ ,  $\sigma_2^2$  a s kovariancí  $\sigma_{12}$ , pak je možné determinant kovarianční matice  $\boldsymbol{\Sigma}$  možné vyjádřit jako  $\sigma_1^2 \sigma_2^2 (1 - \rho^2)$  kde  $\rho$  je korelační koeficient definovaný jako  $\frac{\sigma_{12}}{\sigma_1 \sigma_2}$ . Tento determinant je roven nule když  $\rho = 1$ .

Podmíněné rozdělení  $X_1|x_2$  je normální se střední hodnotou  $\beta_0 + \beta_1 x_2$  a rozptylem  $\sigma_1^2(1 - \rho^2)$

$$\beta_1 = \frac{\sigma_{12}}{\sigma_2^2} \quad \beta_0 = \mu_1 - \beta_1 \mu_2$$

Podmíněné rozdělení  $X_1|x_2$  závisí lineárně na  $x_2$ . Rozptyl  $X_1$  nezávisí na  $x_2$ . Pro dvourozměrné normální rozdělení můžeme elipsy konstantní hustoty znázornit graficky (Obrázek 3.1).

$$f(x_1, x_2) = \text{konst.}$$



**Obrázek 3.1** Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty,  $\mu_1 = \mu_2 = 0$ ,  $\sigma_1 = 1$ ,  $\sigma_2 = 2$ ,  $\rho = 0$  (podle Tvrđík 2003).

### 3.1 Vícerozměrné charakteristiky rozdělení

Základní charakteristikou vícerozměrného rozdělení je vektor středních hodnot (vektor průměrů)

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix}$$

a kovariační matice

$$\Sigma = \text{var}(\mathbf{X}) = \text{cov}(\mathbf{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_1 \sigma_2 & \cdots & \sigma_1 \sigma_p \\ \sigma_2 \sigma_1 & \sigma_2^2 & \cdots & \sigma_2 \sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p \sigma_1 & \sigma_p \sigma_2 & \cdots & \sigma_p^2 \end{pmatrix}$$

kde  $\sigma_{ij}$  je kovariance dvou náhodných veličin, tj.

$$\sigma_{ij} = \text{cov}(X_i, X_j) = E(X_i - E(X_i))(X_j - E(X_j)) \quad (3.3)$$

a  $\sigma_{ii} = \sigma_i^2$  je rozptyl  $\text{var}(X_i)$ . Kovarianční matice je symetrická, neboť  $\sigma_{ij} = \sigma_{ji}$ .

#### 3.1.1 Medoid

Medoid je reprezentativní objekt datového souboru nebo shluku v datech, jehož průměr podobnosti od všech ostatních objektů v datech nebo ve shluku je minimální. Medoid má podobný význam jako průměr nebo centroid, jen je vždy reprezentován reálným objektem z datového souboru. Medoid bývá nejčastěji používán tam, kde není definován průměr nebo centroid (např. tři a vícerozměrný prostor). Tento termín se používá při shlukové analýze.

### 3.2 Wishartovo rozdělení

Uvažujeme v nezávislých náhodných vektorů  $\mathbf{u}_i$ ,  $i = 1, 2, \dots, v$ , vesměs s rozdělením  $N_p(\mathbf{o}_p, \Sigma)$ . Potom náhodná matice  $\mathbf{A} = \sum_{i=1}^v \mathbf{u}_i \mathbf{u}_i^T$  má  $p$ -rozměrné Wishartovo rozdělení se v stupni volnosti, tedy  $\mathbf{A} \sim W_p(\nu, \Sigma)$ .

Při odvození některých důležitých algoritmů ve vícerozměrné statistické analýze se uplatňuje dále uvedená vlastnost Wishartova rozdělení.

Součet nezávislých náhodných matic s Wishartovým rozdělením se shodnou střední hodnotou je rovněž Wishartovo rozdělení se stejnou střední hodnotou, přičemž stupně volnosti se sčítají.

$$\left. \begin{aligned} \mathbf{A} &= \mathbf{A}_1 + \mathbf{A}_2 + \dots + \mathbf{A}_H \\ \mathbf{A}_h &\sim W_p(\nu_h, \Sigma), h = 1, 2, \dots, H \end{aligned} \right\} \longrightarrow \mathbf{A}_h \sim W_p\left(\sum_{h=1}^H \nu_h, \Sigma\right) \quad (3.4)$$

Součtová věta pro Wishartovo rozdělení připomíná součtovou větu pro chí-kvadrát, jehož je Wishartovo rozdělení vícerozměrným zobecněním.

### 3.3 Hotellingovo rozdělení

Uvažujme regulární čtvercovou matici  $A$   $p$ -tého řádu a rozdělením  $W_p(\nu, \Sigma)$  a na  $A$  nezávislý  $p$ -položkový vektor  $a$  s rozdělením  $N_p(\mathbf{o}_p, \Sigma/c)$ . Potom kvadratická forma

$$Q_1 = c \mathbf{a}^T A^{-1} \mathbf{a}$$

má Hotellingovo rozdělení  $T^2(p, \nu - p + 1)$ .

V jednorozměrném normálním rozdělení se při testování hypotéz o střední hodnotě používá statistika (jednovýběrový t-test)

$$X \sim N(\mu, \sigma^2) \longrightarrow \frac{\bar{x} - \mu}{\sqrt{\frac{s^2(x)}{n}}} \sim t(n-1) \quad (3.5)$$

Druhou mocninu této statistiky můžeme upravit a zapsat ve tvaru  $t^2 = n(\bar{x} - \mu) [s^2(x)]^{-1} (\bar{x} - \mu)$ . Tento výraz odpovídá  $p$ -rozměrné statistice, vhodné k úsudku o  $\mu$ , která má Hotellingovo rozdělení  $T^2$  s  $p$  a  $n-p$  stupni volnosti, jedná se tedy o zobecnění  $t$ -rozdělení pro  $p$ -rozměrný prostor. Můžeme tedy psát

$$\mathbf{x} \sim N_p(\mu, \Sigma) \longrightarrow n(\mathbf{x} - \mu)^T S^{-1} (\mathbf{x} - \mu) \sim T^2(p, n-p). \quad (3.6)$$

Obdobným způsobem lze také získat zobecněný dvou výběrový t-test pro  $p$ -rozměrný prostor (Hotellingův test). Pak daná testová statistika má tvar

$$T^2 = \frac{n_1 n_2}{n} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta)^T S^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta), \quad (3.7)$$

kde  $\delta = \mu_1 - \mu_2$  (nejčastěji  $\delta = 0$ ), má opět Hotellingovo rozdělení s parametry  $p, n - p - 1$ .



## 4 Základy maticové algebry

Teoretickým základem libovolných vícerozměrných analýz je práce s maticemi.

Mnohorozměrná data jsou sbírána jako pozorování objektů popsaných několika proměnnými. Data mohou být zaznamenána v tabulce, ve které je každý objekt  $i$  (např. vzorek, lokalita, pozorování, pacient) reprezentován řádkem a ve které každý sloupec  $j$  představuje proměnnou  $y_j$  (např. druh přítomný ve vzorku, fyzikální nebo chemická proměnná, diagnóza, atd.). V každé buňce tabulky se nachází stav  $ij$  proměnné  $j$ , která se týká objektu  $i$ . Tuto tabulku nazýváme matice. Když označíme počet řádků matice (objekty)  $n$  a počet sloupců (proměnné)  $p$ , její rozměr je  $n \times p$ . (Obrázek 4.1).

	proměnná 1	proměnná 2	...	proměnná $j$	...	proměnná $p$
objekt 1	$y_{11}$	$y_{12}$	...	$y_{1j}$	...	$y_{1p}$
objekt 2	$y_{21}$	$y_{22}$	...	$y_{2j}$	...	$y_{2p}$
...	...	...		...		...
objekt $i$	$y_{i1}$	$y_{i2}$	...	$y_{ij}$	...	$y_{ip}$
...	...	...		...		...
objekt $n$	$y_{n1}$	$y_{n2}$	...	$y_{nj}$	...	$y_{np}$

Obrázek 4.1 Ukázka matice rozměru  $n \times p$ .

Tuto matici lze otočit tak, aby proměnné byly v řádcích a objekty ve sloupcích. Jde o transponování matice.

Ne vždy je jednoznačné, co jsou objekty a co proměnné. Například v ekologii mohou být různé lokality (objekty) sledovány s ohledem na druhy (proměnné), které se na nich vyskytují. Ovšem v behaviorálních studiích nebo v taxonomii hmyzu jistého rodu můžou být objekty dané druhy hmyzu a proměnnými různé lokality, které představují ekologické niky.

Objekty a proměnné lze jednoznačně rozlišit, když si uvědomíme, že počet objektů (na rozdíl od proměnných) lze teoreticky zvyšovat do nekonečna.

Mnohorozměrnými postupy lze analyzovat:

- vztahy mezi proměnnými pro soubor objektů (R mode analýza),
- vztahy mezi objekty pro soubor proměnných (Q mode analýza).

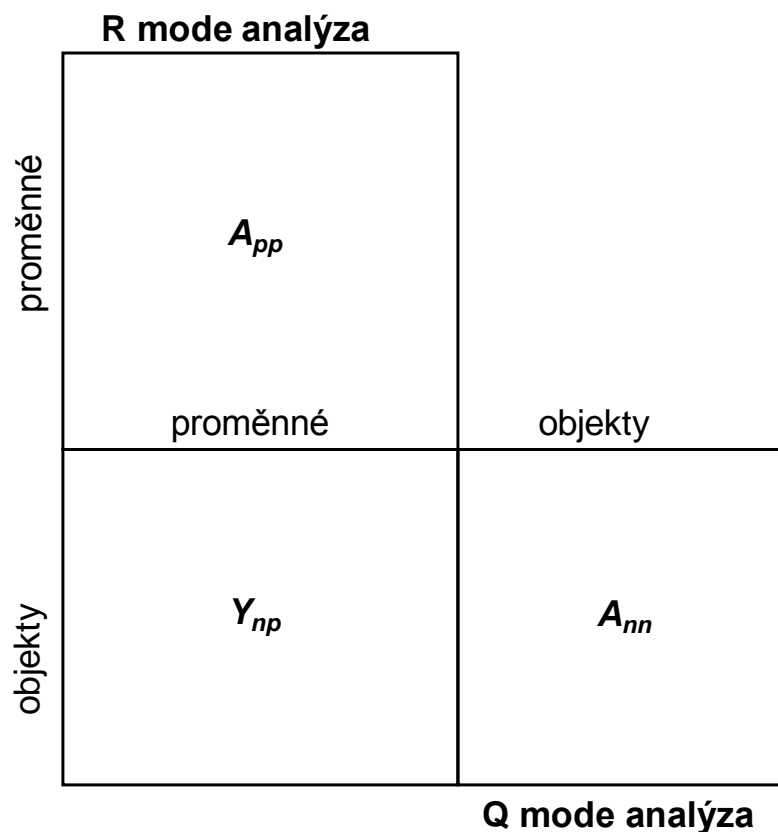
Matematické postupy aplikované při Q mode analýze jsou jiné než při R mode analýze. Např. korelační koeficient můžeme použít při sledování vztahů mezi proměnnými, nelze je ovšem použít pro vztah dvou objektů. Tady se používají jiné míry asociace, např. míry podobnosti.

Výše uvedenou matici rozměru  $n \times p$  můžeme zapsat ve tvaru

$$Y = [y_{ij}] = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2p} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{np} \end{bmatrix}.$$

## 4.1 Asociační matice

Asociační matice je v typickém případě čtvercová symetrická matice, kde sloupce a řádky odpovídají proměnným/objektům původní  $n \times p$  matice, průsečík řádků a sloupců obsahuje měřítko (metriku) vztahu mezi příslušnými proměnnými/objekty. Typ použité metriky se řídí typem dat (spojitá a nespojitá kvantitativní data, kategoriální data, binární data) a typem analýzy. Q mode analýza se snaží popsat vzájemnou pozici objektů v  $n$ -rozměrném prostoru. Typické je tedy použití metrik vzdálenosti a podobnosti. R mode analýza se snaží popsat vztahy mezi proměnnými, a tak je typické použití korelace a kovariance a dalších metrik závislostí (Obrázek 4.2). Některá data je samozřejmě možné sledovat jak z pozice objektů, tak proměnných (např. druhy použité jako proměnné odběrů a odběry použité jako proměnné v analýze taxonů).



**Obrázek 4.2** Původní data tvořila matice  $Y_{np}$  rozměru  $n$  (objekty)  $\times$   $p$  (proměnné). Z této matice lze vytvořit dvě asociační matice  $A_{pp}$  (proměnné  $\times$  proměnné) a  $A_{nn}$  (objekty  $\times$  objekty) (podle Legendre, Legendre 1998).

Asociační matici mezi proměnnými označíme  $A_{pp} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & \dots & \dots & \dots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix}$ ,

asociační matici mezi objekty  $A_{nn} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$ .

Asociační matice jsou nejčastěji symetrické, tj.  $a_{ij} = a_{ji}$ .

U asociační matice mezi objekty  $A_{nn}$  jsou hodnoty na diagonále  $a_{ii}$  rovny nule (když je mírou asociace vzdálenost) nebo jedné (když je mírou asociace podobnost).

U asociační matice mezi proměnnými  $A_{pp}$ , kde je mírou asociace korelace, jsou hodnoty na diagonále  $a_{ii}$  rovny jedné.

## 4.2 Speciální matice

Matice se stejným počtem řádků a sloupců je **čtvercová**. Jak uvidíme dále, pouze pro takovou matici můžeme vypočítat determinant, inverzní matice, vlastní hodnoty (*eigenvalues*) a vlastní vektory (*eigenvectors*). Tyto operace můžou být provedeny na asociační matici, která je vždy čtvercová.

$B_{nn} = [b_{ij}] = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix}$  je čtvercová matice řádu  $n$ .

**Diagonální matice** je čtvercová matice, která má všechny prvky neležící na diagonále nulové.

Např. matice  $\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$  je diagonální.

Diagonální matice, ve které jsou diagonální prvky rovny jedné, se nazývá **jednotková matice**.

$I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$

Jednotková matice má v maticové algebře stejnou roli jako jednotka v běžné algebře, tj. představuje neutrální prvek při násobení ( $\mathbf{I} * \mathbf{B} = \mathbf{B} * \mathbf{I} = \mathbf{B}$ ).

Podobně **skalární matice** je diagonální matice formy

$$\begin{bmatrix} k & 0 & \dots & 0 \\ 0 & k & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & k \end{bmatrix} = kI, \text{ kde jsou diagonální prvky identické. Tato matice představuje}$$

jednotkovou matici vynásobenou skalárem (konstantou).

Matice, jejíž všechny prvky jsou nulové, se nazývá **nulová matice**  $0 = [0]$  a je neutrálním prvkem při sčítání.

Čtvercová matice, jejíž prvky pod nebo nad diagonálou jsou nulové, se nazývá **triangulární (trojúhelníková) matice**. Např.  $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$  je triangulární matice. Diagonální

matice jsou také triangulární.

### Transponovaná matice

Transponovanou matici původní matice  $\mathbf{B}$  rozměru  $n \times p$  označíme  $\mathbf{B}'$ . Její formát bude  $p \times n$  a platí, že  $b'_{ij} = b_{ji}$ . Jednoduše řečeno, řádky jedné matice jsou sloupce druhé matice. Např. transponovaná matice k matici

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} \text{ je } B' = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}.$$

Čtvercová matice, u které platí, že je rovna své transponované matici ( $\mathbf{B} = \mathbf{B}'$ ), se nazývá **symetrická**. Platí, že  $b_{ij} = b_{ji}$ .

$$\text{Např. matice } \begin{bmatrix} 1 & 4 & 5 \\ 4 & 2 & 6 \\ 5 & 6 & 3 \end{bmatrix} \text{ je symetrická.}$$

## 4.3 Vektory a normalizace

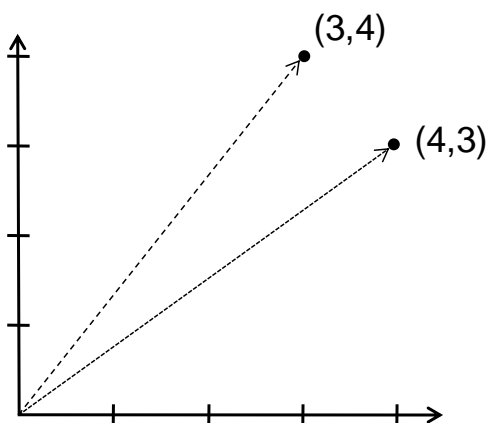
Sloupcová matice rozměru  $n \times 1$  se nazývá **vektor**.

Vektor zapíšeme následujícím způsobem:

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}$$

Vektor je definován jako uspořádaná  $n$ -tice reálných čísel, kde těchto  $n$  hodnot představuje souřadnice bodu v  $n$ -rozměrném Euklidovském prostoru.

Například, vektor  $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$  je uspořádaná dvojice reálných čísel (4, 3), kterou můžeme zakreslit do Euklidovského prostoru (obrázek 4.3).



**Obrázek 4.3** Zobrazení dvou vektorů v dvourozměrném prostoru. Obrázek dobře ilustruje rozdíl mezi vektory  $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$  a  $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$  (podle Legendre, Legendre 1998).

Délku každého vektoru je možné spočítat pomocí Pythagorovy věty. Například, délka vektoru  $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$  je  $\sqrt{4^2 + 3^2} = 5$ . Je to také délka vektoru  $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ .

K porovnání různých vektorů a také jejich směru slouží **normalizace**, tj. vydělení každého prvku vektoru jeho délkou. Normalizace vektoru  $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$  je  $\begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix}$ .

Délka normalizovaného vektoru je rovna jedné.

Normalizovaný vektor původního vektoru  $b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}$  můžeme zapsat jako

$$\begin{bmatrix} b_1 / \sqrt{b_1^2 + b_2^2 + \dots + b_n^2} \\ b_2 / \sqrt{b_1^2 + b_2^2 + \dots + b_n^2} \\ \dots \\ b_n / \sqrt{b_1^2 + b_2^2 + \dots + b_n^2} \end{bmatrix} = \frac{1}{\sqrt{b_1^2 + b_2^2 + \dots + b_n^2}} \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}$$

#### 4.4 Sčítání a násobení matic

Sčítat lze pouze matice stejného rozměru. Sčítání dvou matic pak spočívá ve sčítání příslušných prvků.

$$\mathbf{A} + \mathbf{B} = \mathbf{C}, \text{ kde } c_{ij} = a_{ij} + b_{ij} \quad (4.1)$$

Např.:

$$\begin{bmatrix} 1 & 5 \\ 14 & 2 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 15 & 20 \\ 10 & 8 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} 16 & 25 \\ 24 & 10 \\ 3 & 5 \end{bmatrix}$$

**Sčítání matic** má tyto vlastnosti:

- Kumulativnost:  $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- Asociativnost:  $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
- Distributivnost:  $(c + d)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$ ;  $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$
- Neutrálnost nuly – součet matice  $\mathbf{A}$  a nulové matice (obě stejného rozměru) se rovná matici  $\mathbf{A}$ :  $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$
- Opačná matice k matici  $\mathbf{A}$  se značí  $-\mathbf{A}$  a platí  $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$ . Existence opačné matice umožňuje odčítání dvou matic:  $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$ .

Odčítání matic vyjadřujeme operací sčítání:

$$\begin{bmatrix} 2 & 5 & 1 \\ 2 & -5 & 0 \end{bmatrix} - \begin{bmatrix} 2 & 8 & 1 \\ -5 & 6 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 5 & 1 \\ 2 & -5 & 0 \end{bmatrix} + \begin{bmatrix} -2 & -8 & -1 \\ 5 & -6 & -2 \end{bmatrix} = \begin{bmatrix} 0 & -3 & 0 \\ 7 & -11 & -2 \end{bmatrix}$$

**Násobení matice číslem** je velmi jednoduchá operace: každý prvek matice se násobí daným číslem (skalárem). Např.:

$$3 \cdot \begin{bmatrix} 1 & 4 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 12 \\ 9 & 15 \end{bmatrix}$$

**Násobení matice číslem** má tyto vlastnosti:

- $1\mathbf{A} = \mathbf{A}$
- Když  $c, d$  jsou reálná čísla, tak  $c(d\mathbf{A}) = (c \cdot d)\mathbf{A}$

**Násobení matic** je možné pouze mezi maticemi, pro které platí, že počet sloupců první matice je stejný jak počet řádků druhé matice. Výsledná matice má pak stejný počet řádků jako první matice a stejný počet sloupců jako druhá matice.

$$\text{Např. } A = \begin{bmatrix} 1 & 0 & 2 \\ 3 & 1 & 1 \\ 1 & 2 & 1 \\ -1 & 3 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & -1 \end{bmatrix}$$

$$C = A \cdot B = \begin{bmatrix} 1+0+6 & 2+0-2 \\ 3+2+3 & 6+1-1 \\ 1+4+3 & 2+2-1 \\ -1+6+6 & -2+3-2 \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 8 & 6 \\ 8 & 3 \\ 11 & -1 \end{bmatrix}$$

Prvek  $c_{ij}$  výsledné matice je skalár řádku  $i$  z matice **A** a sloupce  $j$  z matice **B**:

$$c_{ij} = a_i \cdot b_j = \begin{bmatrix} a_{i1} & a_{i2} & \dots & a_{ip} \end{bmatrix} \cdot \begin{bmatrix} b_{1j} \\ b_{2j} \\ \dots \\ b_{pj} \end{bmatrix} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{ip}b_{pj} \quad (4.2)$$

Pro násobení matic platí následující:

- Dvě matice je možné spolu násobit pouze když první matice má tolik sloupců, kolik má druhá matice řádků.
- O součinu **AB** hovoříme, že matici **A** násobíme maticí **B** zprava, matici **B** násobíme maticí **A** zleva.
- Dvě čtvercové matice stejného rozměru můžeme násobit mezi sebou v libovolném pořadí.
- Součin matice a její příslušné transponované matice je vždy možný. **B** · **B'** a také **B'** · **B** vždy existují.
- **B** · **B** (tedy druhá mocnina matice **B**) existuje, pouze když je matice **B** čtvercová.
- Násobení matic není kumulativní. **AB** ≠ **BA**. Když existuje součin matic **A** a **B**, neznamená to, že existuje součin matic **B** a **A**.
- Asociativnost: **A(BC)** = **(AB)C**.
- Distributivnost: **A(B + C)** = **AB + AC**, **(A + B)C** = **AC + BC**.
- **[AB]'** = **B' · A'** a **[ABCD...]'** = **...D' · C' · B' · A'**.

## 4.5 Determinant matice

**Determinant matice** je číslo definované pouze pro čtvercové matice. Determinant matice **A** označíme **|A|**. Pro toto číslo platí:

$$|A| = \sum (-1)^I a_{1j_1} \cdot a_{2j_2} \cdot \dots \cdot a_{nj_n}, \quad (4.3)$$

kde počet sčítanců je  $n!$  a  $I$  je počet inverzí v permutaci  $(j_1, j_2, \dots, j_n)$  prvků  $1, 2, \dots, n$ .

Determinant matice druhého řádu se vypočítá jednoduše:

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \quad (4.4)$$

Např.  $\begin{vmatrix} 2 & 5 \\ 1 & 3 \end{vmatrix} = 2 \cdot 3 - 5 \cdot 1 = 6 - 5 = 1$ .

Získané číslo je složeno z  $2! = 2$  součinů, každý z nich obsahuje pouze jeden a jeden prvek z každého řádku a sloupce matice.

Determinant matice třetího řádu můžeme vypočítat podle Sarrusova pravidla (platí pouze pro  $n = 3$ ):

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} =$$

$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{23}a_{32}a_{11} \quad (4.5)$$

Např.  $\begin{vmatrix} 1 & 3 & -2 \\ -3 & 0 & 1 \\ 2 & 5 & 6 \end{vmatrix} = 1 \cdot 0 \cdot 6 + 3 \cdot 1 \cdot 2 + (-3) \cdot 5 \cdot (-2) - (-2) \cdot 0 \cdot 2 - 3 \cdot (-3) \cdot 6 - 1 \cdot 5 \cdot 1 = 85$

Determinant  $n$ -tého stupně vypočítáme pomocí rozvoje determinantu  $n$ -tého stupně, tj. postupným snižováním stupně determinantu vynecháním  $i$ -tého řádku a  $j$ -tého sloupce. Takto determinant např. pátého řádu snížíme na čtvrtý stupeň a dále na třetí stupeň, který vypočítáme podle Sarrusova pravidla.

$A$  je matice čtvrtého stupně.  $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$

Determinant této matice je pak:

$$|A| = a_{11}|A_{11}| - a_{21}|A_{21}| + a_{31}|A_{31}| - a_{41}|A_{41}|, \quad (4.6)$$

kde determinant  $|A_{11}|$  je determinantom submatice  $A_{11}$ , kterou získáme z matice  $A$  vynecháním prvního řádku a prvního sloupce:



$$|A_{11}| = \begin{vmatrix} a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{vmatrix},$$

podobně vynecháním druhého řádku a prvního sloupce dostaneme  $A_{21}$ , atd.

$$\text{Např. } A = \begin{bmatrix} 1 & 4 & 0 & 3 \\ 2 & -1 & 1 & 5 \\ 0 & 4 & 1 & 4 \\ 3 & 5 & 9 & 2 \end{bmatrix}$$

$$|A| = 1 \cdot \begin{vmatrix} -1 & 1 & 5 \\ 4 & 1 & 4 \\ 5 & 9 & 2 \end{vmatrix} - 2 \cdot \begin{vmatrix} 4 & 0 & 3 \\ 4 & 1 & 4 \\ 5 & 9 & 2 \end{vmatrix} + 0 \cdot \begin{vmatrix} 4 & 0 & 3 \\ -1 & 1 & 5 \\ 5 & 9 & 2 \end{vmatrix} - 3 \cdot \begin{vmatrix} 4 & 0 & 3 \\ -1 & 1 & 5 \\ 4 & 1 & 4 \end{vmatrix}$$

$$|A| = 1 \cdot 201 - 2 \cdot (-43) + 0 \cdot (-214) - 3 \cdot (-19) = 201 + 86 + 0 + 57 = 344$$

Vlastnosti determinantu čtvercové matice pro  $n \geq 2$ :

- Hodnota determinantu se nezmění když zaměníme jeho řádky za sloupce a naopak, tj. determinant matice a její transpozice je stejný:  $|\mathbf{A}| = |\mathbf{A}'|$ .
- Hodnota determinantu se nezmění, když připočítáme k libovolnému řádku libovolnou lineární kombinaci jiných řádků.
- Když zaměníme mezi sebou dva řádky (sloupce), determinant změní znaménko.
- Když jsou dva řádky (sloupce) matice stejné, determinant je nula.
- Když jsou dva řádky (sloupce) matice lineárně závislé, determinant je nula.
- Když se všechny prvky některého řádku (sloupce) rovnají nule, determinant je nula.
- Determinant trojúhelníkové matice (a také diagonální matice) je součinem prvků na diagonále.

## 4.6 Hodnost matice

Čtvercová matice je tvořena  $n$  vektory (řádky nebo sloupce), které můžou být, nebo nemusí být lineárně nezávislé. Dva vektory jsou lineárně závislé, když prvky jednoho jsou násobkem prvků druhého vektoru.

$$\text{Např. vektory } \begin{bmatrix} -4 \\ -6 \\ -8 \end{bmatrix} \text{ a } \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \text{ jsou lineárně závislé, protože } \begin{bmatrix} -4 \\ -6 \\ -8 \end{bmatrix} = -2 \cdot \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}.$$

Podobně, vektor je lineárně závislý na dvou dalších (vzájemně nezávislých) vektorech, když jsou jeho prvky lineární kombinací prvků těchto dvou vektorů.

**Hodnost matice** (označíme  $h$ ) je definována jako počet lineárně nezávislých řádků (nebo sloupců) matice.

Maticí, jejíž hodnost je menší, než její stupeň ( $h < n$ ), nazýváme **singulární**. Její determinant je roven nule  $|\mathbf{A}| = 0$ .

Matice, které hodnost je rovna jejímu stupni ( $h = n$ ), je **regulární** a její determinant je různý od nuly  $|\mathbf{A}| \neq 0$ .

Hodnost matice se nezmění, když

- Vyměníme pořadí řádků nebo řádky za sloupce.
- Vynásobíme některé řádky nenulovým číslem.
- K libovolnému řádku připočítáme lineární kombinaci jiných řádků matice.
- V matici vynecháme řádek, který je lineární kombinací těch, které zůstaly v matici.
- Přidáme k matici řádek, který je lineární kombinací řádků matice.

Hodnost matice můžeme vypočítat pomocí elementárních úprav, a to tak, abychom pod diagonálou matice dostali nuly.

Elementárními úpravami matic rozumíme:

- Výměnu dvou řádků.
- Připočítání  $k$ -násobku jednoho řádku k jinému řádku matice ( $k \neq 0$ ).
- Násobení některého řádku nenulovým číslem.

Např. v matici  $\begin{bmatrix} 1 & 4 & 2 \\ 0 & 1 & 4 \\ 2 & 9 & 3 \end{bmatrix}$  vynásobíme první řádek číslem  $(-2)$  a připočítáme jej k třetímu řádku. Dostaneme  $\begin{bmatrix} 1 & 4 & 2 \\ 0 & 1 & 4 \\ 0 & 1 & -1 \end{bmatrix}$ . Pak násobíme druhý řádek číslem  $(-1)$  a připočítáme k třetímu řádku. Dostaneme  $\begin{bmatrix} 1 & 4 & 2 \\ 0 & 1 & 4 \\ 0 & 0 & -5 \end{bmatrix}$ . Výsledkem jsou tři lineárně nezávislé řádky.

Hodnost matice  $h = 3$ .

## 4.7 Inverzní matice

V maticové algebře neexistuje dělení matic. Lze jej ovšem nahradit násobením matice tzv. inverzní maticí. Inverzní matici matice  $\mathbf{A}$  značíme  $\mathbf{A}^{-1}$ .

Když inverzní matice existuje, je jedinečná a pro čtvercové matice platí, že  $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ .

Inverzní matice existuje pouze pro regulární matici, tj. když její determinant je různý od nuly. Když má čtvercová matice nulový determinant, jedná se o singulární matici a nedá se pro ni sestavit inverzní matice.

Pro obdélníkovou matici lze sestavit tzv. pseudoinverzní matici.

Inverzní matice má tyto vlastnosti:

- $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$
- $[\mathbf{A}^{-1}]^{-1} = \mathbf{A}$
- $[\mathbf{A}^T]^{-1} = [\mathbf{A}^{-1}]^T$
- $[\mathbf{AB}]^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- pro symetrickou matici (kde  $\mathbf{A}^T = \mathbf{A}$ ) platí:  $[\mathbf{A}^{-1}]^T = \mathbf{A}^{-1}$

- když  $\mathbf{A}^{-1} = \mathbf{A}'$ ,  $\mathbf{A}$  je ortogonální matice (matice, jejíž normalizované vektory jsou ortogonální, tj. vzájemně kolmé) a  $\mathbf{A}\mathbf{A}' = \mathbf{I}$

Inverzní matici  $\mathbf{A}^{-1}$  k dané čtvercové matici  $\mathbf{A}$  lze vypočítat pomocí Gauss – Jordanovy eliminační metody.

Postup je následující:

1. Sestavíme matici  $\mathbf{B}$  složenou z původní matice  $\mathbf{A}$  a jednotkové matice  $\mathbf{I}$ .

$$\mathbf{B} = [\mathbf{A}|\mathbf{I}] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 1 \end{bmatrix}$$

2. Elementárními úpravami matic (záměna řádků, připočítání  $k$ -násobku jednoho řádku k jinému řádku, násobení některého řádku nenulovým číslem) převedeme matici  $\mathbf{B}$  do tvaru, kdy jednotková matice  $\mathbf{I}$  bude vlevo. Tak získáme inverzní matici  $\mathbf{A}^{-1}$  v pravé polovině upravené matice.

$$\text{Např. } \mathbf{A} = \begin{bmatrix} 1 & 3 & 0 \\ -1 & -4 & 1 \\ 0 & 3 & 3 \end{bmatrix}$$

$$[\mathbf{A}|\mathbf{I}] = \begin{bmatrix} 1 & 3 & 0 & 1 & 0 & 0 \\ -1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 3 & 3 & 0 & 0 & 1 \end{bmatrix}$$

Matici jsme upravili těmito operacemi: k druhému řádku jsme připočítali první řádek; druhý řádek jsme vynásobili číslem -1; od třetího řádku jsme odpočítali trojnásobek druhého řádku; třetí řádek jsme vynásobili číslem 1/6; k druhému řádku jsme připočítali třetí řádek; od prvního řádku jsme odpočítali trojnásobek druhého řádku.

Výsledkem je upravená matice s jednotkovou maticí vlevo a inverzní maticí vpravo.

$$[\mathbf{I}|\mathbf{A}^{-1}] = \begin{bmatrix} 1 & 0 & 0 & \frac{5}{2} & \frac{3}{2} & -\frac{1}{2} \\ 0 & 1 & 0 & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{6} \\ 0 & 0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} \frac{5}{2} & \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}$$

Inverze je užitečná v mnoha aplikacích; typickým příkladem využití inverzní matice je řešení systémů rovnic nebo výpočet regresních modelů.

## 4.8 Vlastní hodnoty a vlastní vektory matice

Determinant a inverzní matice jsou užitečné při hledání ortogonální formy pro ne-ortogonální symetrickou matici. Zopakujme si, že ortogonální matice je matice, jejíž normalizované vektory jsou vzájemně kolmé a platí pro ni  $\mathbf{A}^{-1} = \mathbf{A}'$ .

Řešení tohoto problému je podstatou faktorové analýzy, které se budeme věnovat později. Tato metoda umožňuje redukovat velké množství proměnných vzájemně svázaných na menší počet nezávislých proměnných vysvětlujících lépe rozptýl dat než původní proměnné.

Matematický princip této metody spočívá ve výpočtu **vlastních čísel** (*eigenvalues*) a **vlastních vektorů** (*eigenvectors*) matice.

Ke čtvercové matici  $\mathbf{A}$  (ve většině případů jde již o symetrickou asociační matici) hledáme jinou matici  $\mathbf{\Lambda}$ , ekvivalentní k  $\mathbf{A}$ , která má nenulové prvky pouze na diagonále. Matici  $\mathbf{\Lambda}$  nazýváme maticí vlastních hodnot. Tyto jsou na sobě lineárně nezávislé. Matice  $\mathbf{\Lambda}$  je známá také pod názvem kanonická forma matice  $\mathbf{A}$ .

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & \dots & \dots & \dots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix}$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 & \dots & 0 \\ 0 & \lambda_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{pp} \end{bmatrix} \dots \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Vlastní hodnoty a vlastní vektory matice  $\mathbf{A}$  nalezneme pomocí rovnice

$$\mathbf{A}\mathbf{u}_j = \lambda_j\mathbf{u}_j, \quad (4.7)$$

pomocí které jsou vypočítány různé vlastní hodnoty  $\lambda_j$  a příslušné vlastní vektory  $\mathbf{u}_j$ . Počet vlastních hodnot a vlastních vektorů je stejný.

Výše uvedenou rovnici můžeme zapsat jako rozdíl dvou vektorů:

$$\mathbf{A}\mathbf{u}_j - \lambda_j\mathbf{u}_j = 0, \text{ dále pak } (\mathbf{A} - \lambda_j\mathbf{I})\mathbf{u}_j = 0 \quad (4.8)$$

Kromě triviálního řešení rovnice, kdy  $\mathbf{u}_j$  je nulový vektor, má tato rovnice následující řešení:

$$|\mathbf{A} - \lambda_j\mathbf{I}| = 0, \quad (4.9)$$

tj. determinant rozdílu mezi maticemi  $\mathbf{A}$  a  $\lambda_j \mathbf{I}$  musí být roven nule pro každé  $\lambda_j$ . Tuto rovnici nazýváme charakteristická rovnice.

Pro matici  $\mathbf{A}$  řádu  $p$  je charakteristická rovnice polynomem  $\lambda$  stupně  $p$ , jehož řešením jsou různé hodnoty  $\lambda_j$ . Na základě vypočítaných vlastních čísel lze jednoduše určit příslušné vlastní vektory.

Příklad:

Symetrická matice  $A = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$  má charakteristickou rovnici  $\begin{vmatrix} 2 & 2 \\ 2 & 5 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 0$ ,

$$\text{tj. } \begin{vmatrix} 2 & 2 \\ 2 & 5 \end{vmatrix} - \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} = 0 \text{ a } \begin{vmatrix} 2-\lambda & 2 \\ 2 & 5-\lambda \end{vmatrix} = 0$$

Charakteristický polynom můžeme najít rozvojem determinantu:  $(2-\lambda)(5-\lambda)-4=0$ , což dává:  $\lambda^2 - 7\lambda + 6 = 0$ . Rovnice má dvě řešení:  $\lambda_1 = 6$ ,  $\lambda_2 = 1$ .

Řazení vlastních hodnot je úplně náhodné, můžeme stejně správně uvádět  $\lambda_1 = 1$ ,  $\lambda_2 = 6$ .

Pomocí rovnice  $(\mathbf{A} - \lambda_j \mathbf{I})\mathbf{u}_j = 0$  můžeme najít vlastní vektory příslušející daným vlastním hodnotám.

Pro  $\lambda_1 = 6$

$$\left( \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - 6 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = 0$$

$$\begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = 0$$

Pro  $\lambda_2 = 1$

$$\left( \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = 0$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = 0$$

což je ekvivalentní páru lineárních rovnic:

$$-4u_{11} + 2u_{21} = 0$$

$$2u_{11} - 1u_{21} = 0$$

$$1u_{12} + 2u_{22} = 0$$

$$2u_{12} + 4u_{22} = 0$$

Tyto systémy lineárních rovnic vždy zahrnují jistou neurčitost. Jejich řešení totiž představuje jakýkoliv bod (vektor) ve stejném směru jako nalezený vlastní vektor.

K odstranění neurčitosti je určena libovolná hodnota pro jeden prvek vektoru  $\mathbf{u}$ , např. 1.

$$u_{11} = 1$$

$$\text{pak podle } -4u_{11} + 2u_{21} = 0$$

$$\text{dostáváme } -4 + 2u_{21} = 0$$

$$\text{a } u_{21} = 2$$

$$u_{12} = 1$$

$$\text{pak podle } 1u_{12} + 2u_{22} = 0$$

$$\text{dostáváme } 1 + 2u_{22} = 0$$

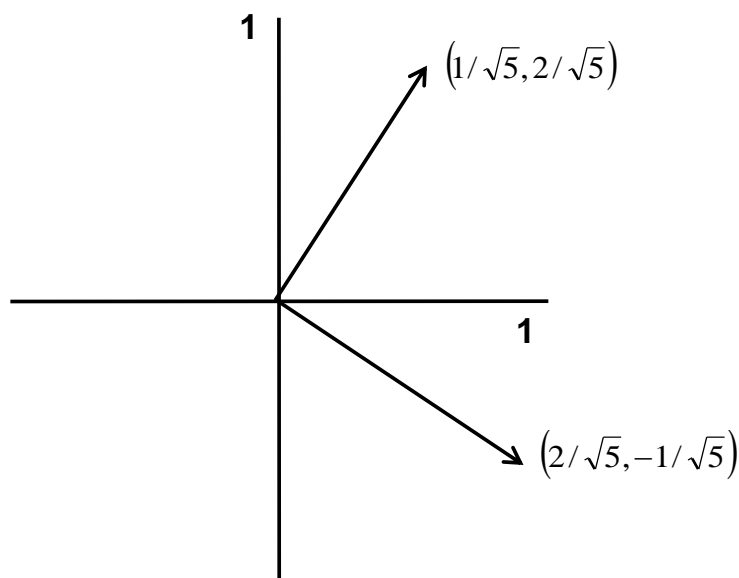
$$\text{a } u_{22} = -\frac{1}{2}$$

Vlastní vektory jsou tedy:  $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$  a  $\begin{bmatrix} 1 \\ -\frac{1}{2} \end{bmatrix}$ .

Zde je nutno poznamenat, že i jiné hodnoty  $u_{11}$  a  $u_{12}$  by byly rovněž vhodné; např. vektory  $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$  a  $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$  také vyhovují lineárním rovnicím. Tyto vlastní vektory jsou identické s výše uvedenými, liší se pouze v násobku skalárem. Proto je zvykem vlastní vektory standardizovat, resp. normalizovat. Jednou z běžných metod je normalizace vektorů tak, aby jejich délka byla rovna jedné (každý prvek vektoru je podělen délkou vektoru).

V našem příkladě jsou vektory  $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$  a  $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$  normalizovány na  $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$  a  $\begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$ .

Jelikož matice **A** byla symetrická, její vlastní vektory jsou ortogonální (na sebe kolmé; Obrázek 4.4). Vlastní vektory nesymetrické matice nejsou na sebe kolmé (ortogonální).



**Obrázek 4.4** Vlastní vektory symetrické matice jsou ortogonální.

Závěrem je nutno připomenout, že hledání vlastních hodnot a vlastních vektorů matice je základním principem některých mnohorozměrných statistických metod, kterými se budeme zabývat v dalších kapitolách.

## 5 Asociační koeficienty

V předchozí části jsem si představili pojem matice a asociační matice. Zopakujme si, že vícerozměrná data jsou typicky uchovávána a zpracovávána v maticové formě a všechny vícerozměrné metody jsou založeny na maticové algebře. Základním vstupem vícerozměrných analýz je matice  $n$  objektů (odběry, vzorky, profily, pacienti apod.) popsaná  $p$  proměnnými (chemické parametry, abundance jednotlivých druhů atd.). Na základě této matice je počítána **asociační matice**, tj. **matice vztahů**. Vztahy mohou být počítány jak mezi proměnnými (R mode analýza), tak mezi objekty (Q mode analýza).

Dříve, než budeme představovat jednotlivé vícerozměrné metody, musíme zmínit asociační koeficienty. Jako měřítko vazby **parametrů** je nejčastěji využívána **korelace** a **kovariance**. Vzniklá tzv. asociační matice parametrů je podkladem pro faktorovou analýzu a analýzu hlavních komponent. Pro **objekty** lze jako měřítko vztahu použít **metriky vzdálenosti** nebo **koeficienty podobnosti**. Míry podobnosti nabývají své maximální hodnoty v případě identických objektů a minimální hodnoty, když dva objekty jsou zcela odlišné. U vzdáleností je tomu obráceně. V případě potřeby lze podobnost převést na vzdálenost.

### 5.1 Asociační koeficienty mezi proměnnými

Vztah dvou proměnných  $x$  a  $y$  můžeme hodnotit pomocí **Pearsonova korelačního koeficientu**  $r$ .

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (5.1)$$

kde  $x_i$  je hodnota  $i$ -tého objektu proměnné  $x$  a  $\bar{x}$  je průměr dané proměnné,  $y_i$  je hodnota  $i$ -tého objektu proměnné  $y$  a  $\bar{y}$  je průměr dané proměnné.

Hodnoty tohoto koeficientu se pohybují v intervalu  $<-1, 1>$ . Čím je hodnota Pearsonova korelačního koeficientu blíží jedné, tím je silnější přímá lineární závislost mezi proměnnými  $x$  a  $y$ . Čím je blíží mínus jedné, tím je silnější nepřímá lineární závislost mezi těmito proměnnými.

Pearsonův korelační koeficient se používá tehdy, když předpokládáme normální rozdělení hodnot proměnných.

V případě, že proměnné nevyhovují podmínce normality rozložení (např. když hodnoty proměnných jsou měřeny na ordinální škále), můžeme použít **Spearmanův korelační koeficient**  $r^s$ .

$$r_{xy}^s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - Q_i)^2, \quad (5.2)$$

kde  $R_1, \dots, R_n$  jsou pořadí prvků proměnné  $x$  a podobně  $Q_1, \dots, Q_n$  jsou pořadí prvků proměnné  $y$ ,  $n$  je počet objektů. Hodnoty tohoto koeficientu se také pohybují v intervalu  $(-1, 1)$  a jeho interpretace je stejná jako u Pearsonova korelačního koeficientu.

Intenzitu vztahu dvou proměnných  $x$  a  $y$  můžeme hodnotit také pomocí **kovariance**. Kovariance není na rozdíl od korelačního koeficientu standardizovaná vzhledem k rozdílným měřítkům proměnných.

$$s_{xy} = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \quad (5.3)$$

Kovariance může nabývat hodnot z intervalu  $(-\infty, \infty)$ .

## 5.2 Asociační koeficienty mezi objekty – metriky vzdálenosti

Vztahy mezi objekty lze vyjádřit pomocí metrik vzdálenosti. Jejich společnou vlastností je, že maximální hodnotu dosahují dva objekty, které jsou úplně odlišné a objekty identické mají vzdálenost nulovou.

Vzdálenost budeme dále označovat symbolem  $D$ .

Metriky (*metrics*) musí splňovat následující kritéria:

- Když jsou objekty shodné, jejich vzdálenost je 0. Když  $a = b$  tak  $D(a, b) = 0$ .
- Když objekty nejsou shodné, jejich vzdálenost je kladné číslo. Když  $a \neq b$  tak  $D(a, b) > 0$ .
- Platí symetrie, vzdálenost objektu  $a$  od  $b$  je stejná jak vzdálenost objektu  $b$  od  $a$ .  $D(a, b) = D(b, a)$
- Platí trojúhelníková nerovnost, tj. součet dvou stran trojúhelníka je vždy roven nebo větší než strana třetí.  $D(a, b) + D(b, c) \geq D(a, c)$

Semimetriky (pseudometriky, *semimetrics*) nevyhovují poslední podmínce trojúhelníkové nerovnosti a neumožňují náležité uspořádání objektů v metrickém prostoru (v „normálním“ systému souřadnic).

Mnohé koeficienty podobnosti ( $S$ ) lze převést na vzdálenosti pomocí transformace  $D = 1 - S$  nebo  $D = \sqrt{1 - S}$  a výsledkem jsou často semimetrické nebo nemetrické koeficienty vzdálenosti.

Následující text shrnuje základní metriky vzdálenosti.

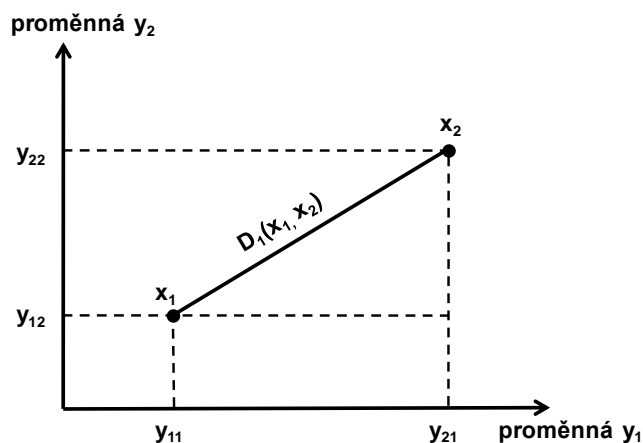
### Euklidovská vzdálenost (*Euclidean distance*)

Jde o nejpoužívanější míru vzdálenosti. Je založená na Pythagorově větě. Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a problém dvou nul. Nemá horní hranici hodnot.



$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2} \quad (5.4)$$

Obrázek 5.1 znázorňuje Euklidovskou vzdálenost dvou objektů v prostoru dvou proměnných.



**Obrázek 5.1** Výpočet Euklidovské vzdálenosti mezi objekty  $x_1$  a  $x_2$ .

Jako další měřítko se používá také čtverec této vzdálenosti. Jeho nevýhodou jsou semimetrické vlastnosti.

$$D_1(x_1, x_2)^2 = \sum_{j=1}^p (y_{1j} - y_{2j})^2 \quad (5.5)$$

### **Průměrná Euklidovská vzdálenost (*average distance*)**

Euklidovská vzdálenost nemá horní hranici. Vzdálenost se zvětšuje s počtem proměnných. Aby mohly být zahrnuty proměnné s různým rozsahem hodnot, je vhodné je před výpočtem standardizovat nebo transformovat. V případě hodnocení vzdálenosti společenstev na základě abundancí druhů bylo navrženo několik modifikací Euklidovské vzdálenosti tak, aby odstranily nedostatky této metriky. Vliv počtu proměnných (v tomto případě druhů) je minimalizovaný tak, že Euklidovská vzdálenost je přepočtena na počet proměnných.

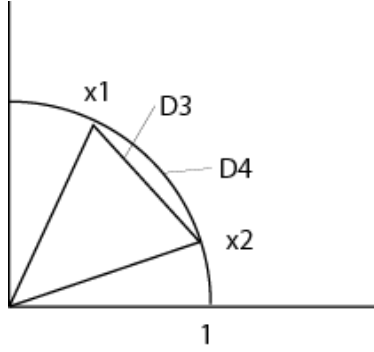
$$D_2^2(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p (y_{1j} - y_{2j})^2 \quad (5.6)$$

nebo

$$D_2(x_1, x_2) = \sqrt{D_2^2} \quad (5.7)$$

### **Tětivová vzdálenost (*chord distance*)**

Tětivová vzdálenost je Euklidovská vzdálenost po normalizaci. Její hodnoty se pohybují od nula po druhou odmocninu z počtu proměnných. Při výpočtu počítá pouze s poměry proměnných v rámci jednotlivých objektů (vzorků). Jde vlastně o Euklidovskou vzdálenost počítanou pro vektory objektů standardizované na délku jedna (Obrázek 5.2), nebo je možný přímý výpočet už zahrnující standardizaci. Odstraňuje problém dvou nul a vliv rozdílného rozpětí proměnných v objektech při výpočtu Euklidovské vzdálenosti.



$$D_3(x_1, x_2) = \sqrt{2 \left[ 1 - \frac{\sum_{j=1}^p y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^p y_{1j}^2} \sqrt{\sum_{j=1}^p y_{2j}^2}} \right]} \quad (5.8)$$

**Obrázek 5.2** Ukázka výpočtu tětivové vzdálenosti a geodetické metriky v prostoru dvou proměnných.

### Geodetická metrika (*geodesic metric*)

Transformace tětivové vzdálenosti je známá jako geodetická metrika. Počítá délku výseče jednotkové kružnice mezi normalizovanými vektory (viz. tětivová vzdálenost, Obrázek 5.2).

$$D_4(x_1, x_2) = \arccos \left[ 1 - \frac{D_3^2(x_1, x_2)}{2} \right] \quad (5.9)$$

### Mahalanobisova vzdálenost

Jde o obecné měřítko vzdálenosti beroucí v úvahu korelaci mezi proměnnými a je nezávislá na rozsahu hodnot proměnných. Respektuje rozdílnou variabilitu a také korelační strukturu v datech. Počítá vzdálenost mezi objekty v systému souřadnic, jehož osy nemusí být na sebe kolmé. V praxi se používá pro zjištění vzdálenosti mezi skupinami objektů. Jsou dány dvě skupiny objektů  $w_1$  a  $w_2$  o  $n_1$  a  $n_2$  počtu objektů a popsané  $p$  parametry:

$$D_5^2(w_1, w_2) = \overline{d}_{12} V^{-1} \overline{d}_{12}', \quad (5.10)$$

kde  $\overline{d}_{12}$  je vektor rozdílů mezi průměry  $p$  proměnných ve dvou skupinách objektů.  $V$  je vážená disperzní matice (matice kovariancí proměnných) uvnitř skupin objektů.

$$V = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)V_1 + (n_2 - 1)V_2] \quad (5.11)$$

kde  $V_1$  a  $V_2$  jsou disperzní matice jednotlivých skupin. Vektor  $\overline{d_{12}}$  měří rozdíl mezi  $p$ -rozměrnými průměry skupin v  $p$ -rozměrném prostoru a  $V$  vkládá do rovnice kovarianci mezi proměnnými.

### **Manhattanská vzdálenost (*Manhattan metric, city-block metric*)**

Základní forma Minkowského metriky, při  $\lambda = 1$  je známá jako Manhattanská vzdálenost. Jde vlastně o součet rozdílů jednotlivých proměnných popisujících objekty.

$$D_6(x_1, x_2) = \sum_{j=1}^p |y_{1j} - y_{2j}| \quad (5.12)$$

### **Průměrná Manhattanská vzdálenost (*mean character difference*)**

Podobně, jak jsme to viděli u Euklidovské vzdálenosti, máme i u Manhattanské vzdálenosti možnost minimalizovat vliv počtu proměnných a přepočítat Manhattanskou vzdálenost na počet proměnných. Její výhodou je, že hodnota se nezvyšuje s rostoucím počtem proměnných.

$$D_7(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p |y_{1j} - y_{2j}| \quad (5.13)$$

### **Minkowského metrika (*Minkowski's metric*)**

Je obecnou formou výpočtu vzdálenosti. Zahrnuje v sobě několik měr jako speciální případy. Podle zadaného koeficientu může odpovídat např. Euklidovské nebo Manhattanské metrice. Se stoupajícím koeficientem umocňování stoupá významnost větších rozdílů. Existuje ještě obecnější forma, kdy koeficient umocňování a odmocňování je zadáván zvlášť.

$$D_8(x_1, x_2) = \left[ \sum_{j=1}^p |y_{1j} - y_{2j}|^\lambda \right]^{\frac{1}{\lambda}} \quad (5.14)$$

$\lambda$  je celé číslo. V případě, že  $\lambda = 2$ , jde o Euklidovskou vzdálenost. V ekologii se nepoužívá číslo  $\lambda$  větší než 2, protože mocniny větší než 2 dávají příliš velkou důležitost největší odchylce  $|y_{1j} - y_{2j}|$ .

### **Vážená euklidovská vzdálenost**

Všechny míry odvozené od Minkowského metriky mají společné nevýhody, jde o již představenou závislost na použitých jednotkách měření, které někdy brání smysluplnému získání jakéhokoliv součtu pro různé proměnné, ale také o to, že když jsou proměnné uvažovány v součtu se stejnými váhami, silně korelované proměnné mají nepřiměřeně velký vliv na výsledek.

Právě proto se někdy používá vážená euklidovská vzdálenost

$$D_9(x_1, x_2) = \sqrt{\sum_{j=1}^p w_j^2 (y_{1j} - y_{2j})^2}, \quad (5.15)$$

kde  $w_j$  je váha proměnné  $j$ .

### **Whittakerův asociační index (*Whittaker's index of association*)**

Je dobře použitelný pro data abundancí. Každý druh (proměnná) je nejprve transformován ve svůj podíl ve společenstvu (v tomto případě společenstvo druhů tvoří součet hodnot všech proměnných ve vzorku – objektu). Následující výpočet je opět obdobou Manhattané vzdálenosti. Doplnkem asociačního indexu je následující vzdálenost:

$$D_{10}(x_1, x_2) = \frac{1}{2} \sum_{j=1}^p \left| \frac{y_{1j}}{\sum_{j=1}^p y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^p y_{2j}} \right| \quad (5.16)$$

Její hodnota je 0 v případě identických proporcí druhů (proměnných).

### ***Canberra metric***

Varianta Manhattané vzdálenosti používaná v ekologických studiích. Před výpočtem musí být odstraněny dvojité nuly a metrika jimi tedy není ovlivněna. Zajímavé je, že stejný rozdíl mezi početnými druhy ovlivňuje tuto vzdálenost méně než ten stejný rozdíl mezi druhy vzácnějšími. Ani tato vzdálenost nemá horní hranici.

$$D_{11}(x_1, x_2) = \sum_{j=1}^p \left( \frac{|y_{1j} - y_{2j}|}{(y_{1j} + y_{2j})} \right) \quad (5.17)$$

### **Koeficient divergence (*coefficient of divergence*)**

Koeficient divergence je obdobná metrika jako  $D_{11}$ , ale je založena na Euklidovské vzdálenosti a vztažena na počet proměnných. Také se používá na ekologická data druhových abundancí po odstranění dvojích nul z výpočtu (a tedy i z hodnoty počtu proměnných  $p$ ).

$$D_{12}(x_1, x_2) = \sqrt{\frac{1}{p} \sum_{j=1}^p \left( \frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}} \right)^2} \quad (5.18)$$

### ***Coefficient of racial likeness***

Umožňuje srovnávat skupiny objektů podobně jako Mahalanobisova vzdálenost, ale na rozdíl od ní neeliminuje vliv korelace proměnných. Dvě skupiny objektů  $w_1$  a  $w_2$  s počtem objektů  $n_1$  a  $n_2$  jsou charakterizovány průměrem proměnných ve skupinách  $\bar{y}_{ij}$  a rozptylem proměnných ve skupinách  $s_{ij}^2$ . Tento koeficient byl vyvinut pro potřeby antropologických studií.

$$D_{13}(w_1, w_2) = \sqrt{\frac{1}{p} \sum_{j=1}^p \frac{(\bar{y}_{1j} - \bar{y}_{2j})^2}{\left(\frac{s_{1j}^2}{n_1}\right) + \left(\frac{s_{2j}^2}{n_2}\right)}} - \frac{2}{p} \quad (5.19)$$

## $\chi^2$ metrika

První ze skupiny metrik založených na  $\chi^2$  využíváném pro výpočet vzdáleností kontingenčních tabulek, a tedy frekvenčních dat. Příkladem takových dat může být matice lokalit (objekty) charakterizovaná abundancemi nebo frekvencemi druhů (proměnné). V matici nejsou přípustné žádné záporné hodnoty. Data původní matice abundancí/frekvencí  $y$  jsou nejprve přepočítána do matice poměrných frekvencí tak, že řádkové součty jsou rovny jedné (druhy jsou na lokalitě vyjádřeny svým poměrným zastoupením, tedy relativní frekvencí). Jako dodatečné charakteristiky uplatňované při výpočtu jsou spočteny součty

$\sum_{j=1}^p y_{ij}$  a sloupců  $\sum_{i=1}^n y_{ij}$  celé matice  $n_{(i)}$  lokalit x  $p_{(j)}$  druhů.

Výpočet odstraňuje problém dvou nul. Nejjednodušším výpočtem je obdoba Euklidovské vzdálenosti

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^p \left( \frac{y_{1j}}{\sum_{j=1}^p y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^p y_{2j}} \right)^2}, \quad (5.20)$$

která je dále vážena součty jednotlivých druhů

$$D_{14}(x_1, x_2) = \sqrt{\sum_{j=1}^p \frac{1}{\sum_{i=1}^n y_{ij}} \left( \frac{y_{1j}}{\sum_{j=1}^p y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^p y_{2j}} \right)^2}. \quad (5.21)$$

Tuto metriku je možné využít i pro měření vzdáleností mezi druhy na základě jejich rozložení na lokalitách.

## $\chi^2$ vzdálenost

Výpočet je podobný  $\chi^2$  metrice, ale vážení je prováděno relativní četností řádku v matici místo jeho absolutního součtu. Při výpočtu se užívá hodnota  $\sum_{j=1}^p \sum_{i=1}^n y_{ij}$  (celkový součet matice).  $\chi^2$  vzdálenost je využívána také při výpočtu vztahů řádků a sloupců kontingenční tabulky.

$$D_{15}(x_1, x_2) = \sqrt{\frac{\sum_{j=1}^p \frac{1}{\sum_{i=1}^n y_{ij}} \left( \frac{y_{1j}}{\sum_{j=1}^p y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^p y_{2j}} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n y_{ij}}} \quad (5.22)$$

$$= \sqrt{\sum_{j=1}^p \sum_{i=1}^n y_{ij}} \sqrt{\sum_{j=1}^p \frac{1}{\sum_{i=1}^n y_{ij}} \left( \frac{y_{1j}}{\sum_{j=1}^p y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^p y_{2j}} \right)^2}$$

### 5.3 Asociační koeficienty mezi objekty – koeficienty podobnosti

Koeficienty podobnosti jsou používány k měření asociací mezi objekty. Oproti většině koeficientů vzdálenosti nejsou nikdy metrické, díky čemuž je vždy možno nalézt dva objekty, A a B, které jsou více podobné než suma jejich podobností s jiným, více vzdáleným objektem C. Z toho vyplývá, že podobnosti nemohou být přímo využity k umístění objektů v metrickém prostoru; musí být převedeny na vzdálenosti. Matice podobností často tvoří základ shlukovacích metod.

Koeficienty podobnosti byly nejprve vyvinuté pro binární data (data typu prezence/absence; ano/ne). S pozdějším rozvojem počítačů byly generalizovány i pro vícečetné proměnné.

Další rozdělení koeficientů podobnosti je určeno ošetřením tzv. problému dvou nul (*double zero problem*).

- **Symetrické koeficienty** podobnosti dávají proměnné, která nabývá u dvou srovnávaných objektů hodnotu nula, stejnou srovnávací hodnotu jako každé jiné proměnné. Tyto koeficienty se používají v případě, že nulový stav reprezentuje stejný druh informace jako kterákoliv jiná hodnota a tedy není jen označením chybějících údajů. Proto tyto koeficienty není vhodné používat v ekologických studiích k hodnocení proměnných, které představují např. přítomnost/nepřítomnost druhů.
- **Asymetrické koeficienty** podobnosti neuvažují duplicitní nulové hodnoty u srovnávaných objektů jako informaci o podobnosti. Uplatnění asymetrických koeficientů je zejména v ekologických studiích, kde proměnné představují druhy a hodnocení společné prezence a absence není symetrické. Na druhé straně, přítomnost druhu pouze v jednom ze dvou objektů naznačuje rozdíl mezi těmito objekty.

Nejdříve se budeme věnovat binárním koeficientům, tj. těm, které pracují s binárními proměnnými (data typu prezence/absence, ano/ne, atd.). U binárních dat dochází k následujícím případům u dvou srovnávaných objektů (Tabulka 5.1).

**Tabulka 5.1** Hodnoty šesti binárních proměnných (pr. 1 až pr. 6) u dvou objektů  $x_1$  a  $x_2$ .

	pr. 1	pr. 2	pr. 3	pr. 4	pr. 5	pr. 6
objekt 1 ( $x_1$ )	1	0	1	1	1	0
objekt 2 ( $x_2$ )	0	1	1	0	1	0
označení stavu	b	c	a	b	a	d

Pozorované stavy můžeme sumarizovat ve frekvenční tabulce (Tabulka 5.2) rozměru 2 x 2 se čtyřmi póly obsahující tyto početnosti:

- a počet proměnných, které nabývají pro oba objekty hodnotu 1
  - b počet proměnných, které nabývají u  $i$ -tého objektu 1 a u  $j$ -tého objektu 0
  - c počet proměnných, které nabývají u  $i$ -tého objektu 0, u  $j$ -tého objektu 1
  - d počet proměnných, které nabývají pro oba objekty hodnoty 0
- Platí  $a + b + c + d = p$ .

**Tabulka 5.2** Sumarizace Tabulky 5.1 ve frekvenční tabulce.

		objekt $x_2$		
		1	0	
objekt $x_1$	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	p

V našem příkladě z Tabulky 5.2 jsou tyto početnosti:  $a = 2$ ,  $b = 2$ ,  $c = 1$ ,  $d = 1$ .

### 5.3.1 Symetrické binární koeficienty

Základem všech indexů podobnosti pro kvalitativní binární data je, že dva objekty jsou vzájemně více podobné, když mají více souhlasných binárních proměnných a méně podobné, když je více proměnných unikátních pro jeden objekt. Při určení podobnosti dvou objektů budeme tedy pozorovat u  $p$  proměnných jejich společnou přítomnost resp. absenci v objektech.

**Jednoduchý srovnávací koeficient** (*simple matching coefficient*) je obvyklou metodou pro výpočet podobnosti mezi dvěma objekty. Jde o podíl počtu proměnných, které kódují objekt stejně, a celkového počtu proměnných.

$$S_1(x_1, x_2) = \frac{a + d}{p} \quad (5.23)$$

Koeficient patří do skupiny **symetrických binárních koeficientů**. Koeficienty této skupiny dávají stejnou váhu pozitivní shodě ( $1 - 1$ ) i negativní shodě ( $0 - 0$ ).

Další variantou tohoto koeficientu je jeho alternativa, která přiřazuje větší důležitost rozdílům než shodám (**Rogers a Tanimoto**).

$$S_2(x_1, x_2) = \frac{a + d}{a + 2b + 2c + d} \quad (5.24)$$

Další čtyři navržené koeficienty berou v úvahu dvojí nuly, ale jsou navrženy tak, aby se snížil vliv problému dvou nul (**Sokal a Sneath**):

$$S_3(x_1, x_2) = \frac{2a + 2d}{2a + b + c + 2d} \quad (5.25)$$

tento koeficient dává dvakrát větší váhu shodným proměnným než rozdílným;

$$S_4(x_1, x_2) = \frac{a + d}{b + c} \quad (5.26)$$

porovnává shody a rozdíly prostým podílem v měřítku, které nabývá hodnot od nula do nekonečna;

$$S_5(x_1, x_2) = \frac{1}{4} \left[ \frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right] \quad (5.27)$$

porovnává shodné deskriptory se součty okrajů tabulky;

$$S_6(x_1, x_2) = \frac{a}{\sqrt{(a+b)(a+c)}} \frac{d}{\sqrt{(b+d)(c+d)}} \quad (5.28)$$

je vytvořen z geometrických průměrů členů vztahujících se k  $a$  a  $d$ , podle koeficientu  $S_5$ .

### 5.3.2 Asymetrické binární koeficienty

V některých případech nelze dávat stejnou váhu pro společnou prezenci (1-1) a absenci (0-0) proměnných (např. druhů) v objektech. Pro tyto případy byly vyvinuty **asymetrické binární koeficienty**.

Ty se stejně jako předchozí symetrické koeficienty používají ke srovnání objektů, v ekologii běžně ke srovnání vzorků nebo lokalit na základě druhového složení. Používají se zde pro data prezence/absence druhů. Ve výpočtu nejsou zahrnuty proměnné, které u obou srovnávaných objektů nabývají nulové hodnoty.

Nejznámější z asymetrických koeficientů jsou Jaccardův a Sørensenův koeficient.

#### Jaccardův koeficient (*Jaccard's coefficient*)

$$S_7(x_1, x_2) = \frac{a}{a + b + c} \quad (5.29)$$

dává všem členům stejnou váhu.

#### Sørensenův koeficient (*Sørensen's coefficient*)



Sorensenův koeficient je variantou Jaccardova koeficientu, dává ovšem dvojnásobnou váhu dvojitému výskytu. Přítomnost druhů je více informativní než jejich nepřítomnost, která může být způsobena různými faktory a nemusí nutně odrážet rozdílnost prostředí. Výskyt druhu na obou lokalitách je silným ukazatelem jejich podobnosti. Jaccardův koeficient je monotónní k Sorensenovu koeficientu, proto podobnost pro dvě dvojice objektů vypočítaná podle  $S_7$  bude podobná stejnému výpočtu  $S_8$ . Oba koeficienty se liší pouze v měřítku. Jiná varianta tohoto koeficientu dává společným výskytům trojnásobnou váhu.

$$S_8(x_1, x_2) = \frac{2a}{2a + b + c} \quad (5.30)$$

$$S_9(x_1, x_2) = \frac{3a}{3a + b + c} \quad (5.31)$$

Řada dalších koeficientů dává různou váhu jednotlivým kombinacím proměnných.

Jako doplněk koeficientu  $S_2$  byl navrhnut koeficient, který dává dvojnásobnou váhu rozdílu ve jmenovateli (**Sokal a Sneath**).

$$S_{10}(x_1, x_2) = \frac{a}{a + 2b + 2c} \quad (5.32)$$

Další koeficient umožňuje porovnat počet společných prezencí proti celkovému počtu proměnných (druhů) ve všech objektech, včetně proměnných (druhů), které nabývají nulové hodnoty v obou uvažovaných objektech (d). (**Russel a Rao**)

$$S_{11}(x_1, x_2) = \frac{a}{p} \quad (5.33)$$

Další koeficient porovnává duplicitní prezence s diferencemi (**Kulczynski**).

$$S_{12}(x_1, x_2) = \frac{a}{b + c} \quad (5.34)$$

Úpravou kvantitativního koeficientu  $S_{18}$  pro binární data byl vytvořen následující koeficient (**Sokal a Sneath**):

$$S_{13}(x_1, x_2) = \frac{1}{2} \left[ \frac{a}{a + b} + \frac{a}{a + c} \right], \quad (5.35)$$

kde jsou duplicitní prezence srovnávány se součty okrajů tabulky  $(a+b)$  a  $(a+c)$ .

Obdobou symetrického koeficientu  $S_6$  tak, aby byl odstaněn problém dvou nul, je koeficient, který jako míru podobnosti používá geometrický průměr poměrů  $a$  k počtu druhů v každém objektu, tj. se součty okrajů tabulky  $(a+b)$  a  $(a+c)$  (**Ochiachi**).

$$S_{14}(x_1, x_2) = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (5.36)$$

### 5.3.3 Symetrické kvantitativní koeficienty

V biologii se můžeme kromě binárních proměnných setkat i s multistavovými kvalitativními nebo kvantitativními proměnnými. Pro takové případy mohou být využity koeficienty, které vznikly rozšířením binárních koeficientů tak, aby se přizpůsobily multistavovým proměnným.

#### Modifikovaný jednoduchý srovnávací koeficient (*simple matching coefficient*)

Modifikovaný jednoduchý srovnávací koeficient může být použit pro multistavové proměnné. Čitatel obsahuje počet proměnných, pro které jsou dva objekty ve stejném stavu.

$$S_1(x_1, x_2) = \frac{\text{shoda}}{p}. \quad (5.37)$$

Např. je-li dvojice objektů popsána následujícími deseti multistavovými proměnnými (Tabulka 5.3), potom hodnota koeficientu  $S_1$ , vypočítaná pro 10 multistavových proměnných bude  $S_1(x_1, x_2) = 4 \text{ shody} / 10 \text{ proměnných} = 0,4$ .

**Tabulka 5.3** Ukázka výpočtu jednoduchého srovnávacího koeficientu pro multistavové proměnné.

	proměnné										$\Sigma$
objekt $x_1$	9	3	7	3	4	9	5	4	0	6	
objekt $x_2$	2	3	2	1	2	9	3	2	0	6	
shoda	0	1	0	0	0	1	0	0	1	1	4

Podobným způsobem je možné rozšířit všechny binární koeficienty pro multistavové proměnné.

#### Gowerův obecný koeficient podobnosti

V případě, že máme objekty popsány několika kvantitativními a několika kvalitativními proměnnými, lze použít Gowerův koeficient podobnosti, který zahrnuje podobnost podle různých typů proměnných – binárních, kvalitativních a semikvantitativních i kvantitativních.

Podobnost mezi dvěma objekty je vypočítána jako průměr podobností vypočítaných pro všechny proměnné (těmito proměnnými mohou být např. druhy, nebo i environmentální proměnné).

$$S_{15}(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p s_{12j} \quad (5.38)$$

Pro každou proměnnou  $j$  je hodnota parciální podobnosti  $s_{12j}$  mezi objekty  $x_1$  a  $x_2$  vypočítána následovně:

- Pro binární proměnné  $s_j = 1$  (shoda) nebo 0 (neshoda). Gower navrhl dvě formy tohoto koeficientu, symetrickou i asymetrickou. Následující forma je symetrická, dává  $s_j = 1$  případům nepřítomnosti binární charakteristiky dvou objektů (0-0). Druhá forma, Gowerův asymetrický koeficient dává případům 0-0  $s_j = 0$ .

- Kvalitativní a semikvantitativní proměnné jsou upraveny podle jednoduchého srovnávacího pravidla zmíněného výše:  $s_j = 1$  při souhlasu a  $s_j = 0$  při nesouhlasu proměnných. Případy shodné nepřítomnosti binární charakteristiky dvou objektů (problém dvou nul) jsou ošetřeny stejně jako v předchozím případě.
- Kvantitativní deskriptory (reálná čísla) jsou zpracovány následovně: pro každou proměnnou se nejprve vypočte rozdíl mezi stavy obou objektů  $|y_{1j} - y_{2j}|$  stejně jako v případě koeficientu vzdálenosti patřícího do skupiny Minkowského metrik. Tento rozdíl je poté vydělen největším rozdílem  $R_j$  nalezeným pro danou proměnnou mezi všemi objekty ve studii (nebo v referenční populaci – doporučuje se vypočítat největší rozdíl  $R_j$  každé proměnné  $j$  pro celou populaci, aby byla zajištěna konzistence výsledků pro všechny parciální studie). Z tohoto podílu je normalizovaná vzdálenost odečtena od jedné, aby byla transformována na podobnost.

$$s_{12j} = 1 - \left[ \frac{|y_{1j} - y_{2j}|}{R_j} \right] \quad (5.39)$$

Gowerův koeficient může být nastaven tak, aby zahrnoval vážení významu proměnných. Žádné porovnání není vypočítáno u proměnných, u nichž chybí informace buď u jednoho, nebo u druhého objektu. Toto zajišťuje člen  $w_j$ , nazývaný Kroneckerovo delta, popisující přítomnost/nepřítomnost informace v obou objektech: je-li informace o proměnné  $y_j$  přítomna u obou objektů  $w_j = 1$ , jinak  $w_j = 0$ .

Konečná forma Gowerova koeficientu pak vypadá takto:

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^p w_{12j} s_{12j}}{\sum_{j=1}^p w_{12j}} \quad (5.40)$$

Další přiblížení ke komplexnosti umožňuje vážení různých proměnných, tj. přiřazení čísla z intervalu  $<0,1>$  parametru  $w_j$ .

Při výpočtu Gowerova koeficientu musíme dobře zvážit, které semikvantitativní proměnné zpracujeme jako kvantitativní a které nikoliv.

Gowerův koeficient nabývá hodnot podobnosti od nula do jedné, kde jedna značí největší podobnost objektů.

Pro ilustraci výpočtu koeficientu uvádíme dva objekty (plochy  $x_1$  a  $x_2$ ) popsány osmi kvantitativními chemickými proměnnými  $p$ , pro které je známý maximální rozdíl  $R_j$  z celé vzorkované plochy (tabulka 5.4).

**Tabulka 5.4.** Ukázka výpočtu Gowerova koeficientu.

	Proměnné $j$								$\Sigma$
objekt $x_1$	2	2	-	2	2	4	2	6	
objekt $x_2$	1	3	3	1	2	2	2	5	
$R_j$	1	4	2	4	1	3	2	5	
$w_{12j}$	1	1	0	1	1	1	1	1	7
$ y_{1j} - y_{2j} /R_j$	1	0.25	-	0.25	0	0.67	0	0.20	

$w_{12j}s_{12j}$	0	0.75	0	0.75	1	0.33	1	0.80	4.63
------------------	---	------	---	------	---	------	---	------	------

$S_{15}(x_1, x_2) = 4.63/7 = 0.66$  (podle Legendre, Legendre 1998).

Další obecný koeficient podobnosti, stejně jako Gowerův koeficient, počítá podobnost dvou objektů jako podíl sumy parciálních podobností proměnných a počtu těchto proměnných (**Estabrook a Rogers**). Obecný zápis tohoto koeficientu je proto stejný jako  $S_{15}$ :

$$S_{16}(x_1, x_2) = \frac{\sum_{j=1}^p w_{12j} s'_{12j}}{\sum_{j=1}^p w_{12j}} \quad (5.41)$$

a stejně jako u  $S_{15}$  mohou být parametry  $w_j$  (mezi 0 a 1) opět využity jako váhy místo toho, aby pouze hrály roly Kroneckerovy delta. Koeficient se liší výpočtem parciálních podobností  $s'_j$ . V původní podobě byly stavové hodnoty kladná celá čísla a proměnné byly buď uspořádané, nebo neseřazené. U tohoto koeficientu je parciální podobnost dvou objektů pro danou proměnnou  $j$  vypočítána použitím monotónní klesající funkce částečné podobnosti. Na základě zkušeností autoři navrhli použít funkci dvou čísel  $d$  a  $k$ :

$$\begin{aligned} s'_{12j} &= f(d_{12j}, k_j) = \frac{2(k+1-d)}{2k+2+dk} \text{ pro } d \leq k \\ s'_{12j} &= f(d_{12j}, k_j) = 0 \text{ pro } d > k, \end{aligned} \quad (5.42)$$

kde  $d$  je vzdálenost mezi dvěma stavy objektů  $x_1$  a  $x_2$  pro proměnnou  $j$ , tj. stejně jako v Gowerově koeficientu  $|y_{1j} - y_{2j}|$  a  $k$  je parametr určený a priori uživatelem pro každou proměnnou, který popisuje jaká maximální velikost nenulové parciální podobnosti je dovolena. Parametr  $k$  (obvykle malé číslo) je roven největšímu rozdílu  $d$ , pro který parciální podobnost  $s'_{12j}$  proměnné  $j$  může být nenulová.

Autoři vytvořili i další míru parciální podobnosti  $s_{12j}$  pro funkci  $S_{16}$ , pro případ, že by funkce  $f(d, k)$  nepopisovala správně vztahy mezi objekty proměnné  $j$ . Tato modifikace poskytuje výhodný nástroj zvláště při použití kvalitativních nebo semikvantitativních proměnných.

### 5.3.4 Asymetrické kvantitativní koeficienty

Stejně jako v předchozí části se nejprve zmíníme o možnostech rozšíření binárních koeficientů na multistavové.

#### Jaccardův koeficient

$$S_7(x_1, x_2) = \frac{\text{shoda}}{p - d}, \quad (5.43)$$

kde v čitateli je počet proměnných se stejnou hodnotou v porovnávaných objektech.

Tento koeficient můžeme použít v případě, že proměnné jsou kódovány malým počtem tříd a my chceme získat velké kontrasty v rozdílech v hodnotách. V jiných případech

samozřejmě použitím takového koeficientu dojde ke ztrátě části informace nesené hodnotami jednotlivých proměnných.

V ekologických studiích, kde jsou proměnné reprezentovány abundancemi druhů, je často nutná odmocninová nebo logaritmická transformace proměnných, protože distribuce druhových abundancí v ekologickém gradientu je často velmi nerovnoměrná. Další možností je použití stupnice relativních abundancí s hranicemi vytvořenými v geometrické řadě např. od 0 (absence) do 7 (velmi četné zastoupení). Normalizované abundance lépe vyjadřují roli jednotlivých druhů v ekosystému, než surová data abundancí.

Některé koeficienty snižují vliv velkých rozdílů a mohou proto být použity na původní data druhových abundancí, zatímco ostatní – porovnávající rozdíl v abundancích více lineárně – je lépe aplikovat na normalizovaná data.

### Sørensenův kvantitativní koeficient (Bray-Curtis; Steinhaus by Motyka)

Sørensenův kvantitativní koeficient (známý také pod názvem Bray-Curtis koeficient) se používá na data abundancí druhů. Patří mezi „klasické“ kvantitativní koeficienty.

$$S_{17}(x_1, x_2) = \frac{W}{(A+B)/2} = \frac{2W}{A+B} \quad (5.44)$$

$W$  je součet minimálních abundancí jednotlivých druhů,  $A$  a  $B$  jsou součty abundancí všech druhů ve dvou srovnávaných objektech, tj. celkový počet jedinců v každém vzorku (tabulka 5.5).

**Tabulka 5.5** Ukázka výpočtu Sørensenova kvantitativního koeficientu.

	Abundance druhů					A	B	W
vzorek $x_1$	7	3	4	5	1	20		
vzorek $x_2$	2	4	7	6	3		22	
minimum	2	3	4	5	1			15

$$S_{17}(x_1, x_2) = \frac{2 \cdot 15}{20 + 22} = 0.714$$

Tento koeficient je příbuzný se Sørensenovým koeficientem ( $S_8$ ). Nahradíme-li četnosti druhů daty prezence/absence, změní se  $S_{17}$  na  $S_8$ .

### Kulczynskeho koeficient

Tento koeficient porovnává součet minim k celkovému počtu jedinců ve vzorku a pak je vypočítán průměr ze získaných dvou hodnot.

$$S_{18}(x_1, x_2) = \frac{1}{2} \left( \frac{W}{A} + \frac{W}{B} \right) \quad (5.45)$$

Pro příklad z Tabulky 5.5 určíme tento koeficient:  $S_{18}(x_1, x_2) = \frac{1}{2} \left( \frac{15}{20} + \frac{15}{22} \right) = 0.716$

Nahradíme-li počty druhů daty prezence/absence, změní se  $S_{18}$  na  $S_{13}$ .

### Morisita-Horn koeficient

Dalším oblíbeným koeficientem je Morisita-Horn koeficient:

$$S_{19} = \frac{2 \sum n_{1i} n_{2i}}{(d_1 + d_2) N_1 N_2}, \quad (5.46)$$

kde  $n_{1i}$  a  $n_{2i}$  je počet jedinců  $i$ -tého druhu v prvním a druhém objektu,  $N_1$  a  $N_2$  jsou součty abundancí všech druhů ve srovnávaných objektech, a  $d_1 = \frac{\sum n_{1i}^2}{N_1^2}$  a  $d_2 = \frac{\sum n_{2i}^2}{N_2^2}$ .

Následující koeficienty jsou přizpůsobeny pro normalizovaná data abundancí; tj. adaptovány na vyrovnané rozložení frekvencí. Jsou podobné koeficientům  $S_{15}$  a  $S_{16}$ .

**Gover** navrhl, že jeho obecný koeficient podobnosti může vyloučit problém dvou nul z porovnání (viz výše) a je tak dobře uplatnitelný pro kvantitativní data abundancí druhů. Protože rozdíly mezi stavy abundancí jsou vypočteny jako  $|y_{1j} - y_{2j}|$  a jsou proto lineárně závislé na měřítku, měl by být tento koeficient používán na normalizovaná data.

$$S_{20}(x_1, x_2) = \frac{\sum_{j=1}^p w_{12j} s_{12j}}{\sum_{j=1}^p w_{12j}}, \quad (5.47)$$

kde

$$s_{12j} = 1 - \left[ \frac{|y_{1j} - y_{2j}|}{R_j} \right] \quad (5.48)$$

jako v  $S_{15}$  a  $w_{12j} = 0$  když  $y_{1j}$  nebo  $y_{2j}$  je chybějící informace, nebo když  $y_{1j}$  a  $y_{2j}$  je nepřítomný druh ( $y_{1j} + y_{2j} = 0$ ).

$w_{12j} = 1$  ve všech ostatních případech.

U dat abundance druhů může opět  $w_j$  stejně jako u  $S_{15}$  nabývat hodnot od 0 do 1, aby ve formě váhy výpočtu pomohlo vyjádřit biomasu, nebo biologický objem různých druhů, nebo kompenzovalo účinnost odběru daného druhu.

Další obecný koeficient podobnosti vychází z koeficientu  $S_{16}$ , byl navržen **Legendrem a Chodorowskim**. Používá modifikovanou verzi funkce částečné podobnosti  $f(d, k)$  nebo matici částečné podobnosti jako u  $S_{16}$ . Protože  $S_{21}$  zpracovává všechny rozdíly  $d$  stejným způsobem, bez ohledu na to, zda odpovídají vysokým, nebo nízkým hodnotám v měřítku abundancí, je lepší používat ho s vyrovnanými daty abundancí. Jediný rozdíl mezi  $S_{16}$  a  $S_{21}$  je v ošetření problému dvou nul. Koeficient ve své obecné formě představuje součet částečných podobností všech druhů vydělený celkovým počtem druhů nalezených v obou objektech.

$$S_{21}(x_1, x_2) = \frac{\sum_{j=1}^p w_{12j} s'_{12j}}{\sum_{j=1}^p w_{12j}} \quad (5.49)$$

kde

$$s'_{12j} = f(d_{12j}, k_j) = \frac{2(k+1-d)}{2k+2+dk} \quad \text{pro } d \leq k$$

$$s'_{12j} = f(d_{12j}, k_j) = 0 \quad \text{pro } d > k,$$

$$s'_{12j} = f(d_{12j}, k_j) = 0 \quad \text{když } y_{1j} \text{ nebo } y_{2j} = 0 \text{ (tj. } y_{1j} \times y_{2j} = 0)$$

anebo  $s'_{12j} = f(y_{1j}, y_{2j})$  danou parciální maticí podobnosti ve které je  $s'_{ij} = 0$  když  $y_{1j}$  nebo  $y_{2j} = 0$

$w_{12j} = 0$  když  $y_{1j}$  nebo  $y_{2j}$  je chybějící informace, nebo

když  $y_{1j}$  a  $y_{2j}$  je absence druhu ( $y_{1j} + y_{2j} = 0$ )

$w_{12j} = 1$  ve všech ostatních případech.

$w_{12j}$  může nabývat hodnot od 0 do 1 jak bylo vysvětleno výše pro koeficient  $S_{20}$ .

### **$\chi^2$ podobnost ( $\chi^2$ similarity)**

Je posledním kvantitativním koeficientem, jenž eliminuje problém dvou nul (*double zero problem*). Je to doplněk  $\chi^2$  metriky ( $D_{14}$ ).

$$S_{22}(x_1, x_2) = 1 - D_{14}(x_1, x_2) \quad (5.50)$$

Výběr správné metriky vzdálenosti silně ovlivňuje výsledky všech vícerozměrných analýz.

## 6 Shluková analýza

Jednou z možností využití informace obsažené ve vícerozměrných pozorováních je roztrídění objektů do několika poměrně homogenních skupin – shluků tak, aby objekty patřící do stejné skupiny si byly podobnější než objekty z různých skupin. Různými možnostmi a aspekty tvorby homogenních skupin objektů se zabývá shluková analýza (*cluster analysis*). Shlukovou analýzou se sníží počet dimenzí objektů tak, že řadu uvažovaných proměnných zastoupí jediná proměnná, vyjadřující příslušnost objektu k definované skupině.

Shluková analýza identifikuje skupiny v datech a pomáhá tak najít skrytou strukturu v datech. Ovšem i když data tvoří souvislou strukturu, shluková analýza v nich hledá strukturu skupin, to znamená, že kontinuum je rozděleno do systému skupin.

Použití metod shlukové analýzy je prospěšné zejména tam, kde se studovaný soubor reálně rozpadá do tříd, tj. objekty mají tendenci se seskupovat do přirozených shluků. Použitím vhodných algoritmů se pak podaří odhalit strukturu studované množiny objektů a jednotlivé objekty klasifikovat. Zbývá pak již pouze najít vhodnou interpretaci pro popsání rozklad, tj. charakterizovat vzniklé třídy (shluky, skupiny).

Shlukovou analýzu můžeme použít i v případech, kdy objekty nejeví tendenci k tvoření přirozených skupin, ale spíše připomínají víceméně homogenní chaos. V takovém případě je ovšem na místě vyšší opatrnost při interpretaci výsledků.

Formálně může být cíl shlukové analýzy popsán následovně: máme k dispozici datovou matici  $\mathbf{X}$  typu  $n \times p$ , kde  $n$  je počet objektů (v ekologii nejčastěji vzorky, odběry, případně lokality) a  $p$  je počet proměnných (v ekologii nejčastěji environmentální charakteristiky, taxony, ale také např. ekologické skupiny – gildy). Uvažujeme různé rozklady  $S^{(k)}$  množiny  $n$  objektů do  $k$  shluků a hledáme takový rozklad, který by byl z určitého hlediska nejvýhodnější. Zde připouštíme pouze rozklady s disjunktími shluky, tj. jeden objekt patří pouze jednomu shluku. Cílem je dosáhnout, aby si objekty uvnitř shluku byly co nejvíce podobné a s objekty z různých shluků co nejméně.

Shluková analýza pracuje s asociační maticí podobností, resp. vzdáleností objektů. Problematické asociačních koeficientů jsme se věnovali v předchozí kapitole. Při výběru asociačního koeficientu je třeba brát v úvahu metodu shlukování a charakter souboru dat. V některých případech je způsob výpočtu podobnosti/vzdálenosti objektů dán již konkrétní shlukovací metodou.

Cílem shlukování je zejména:

- popsat strukturu dat;
- nalézt určité skupiny podobných objektů, tj. shluky.

Existuje několik typů shlukové analýzy, které se liší postupem shlukování. Shlukování může být **hierarchické** nebo **nehierarchické**.

- **Hierarchická shluková analýza** vytváří systém skupin a podskupin tak, že každá skupina může obsahovat několik podskupin nižšího řádu a sama může být součástí skupiny vyššího řádu. Výsledek se dá graficky znázornit stromem – dendrogramem.
- **Nehierarchická shluková analýza** (*partitioning methods*) rozdělí objekty do několika shluků stejného řádu.

V ekologii bývá shluková analýza používána ke klasifikaci vzorků (lokalit), ale v některých případech i na klasifikaci druhů, resp. taxonů, nebo environmentálních proměnných.



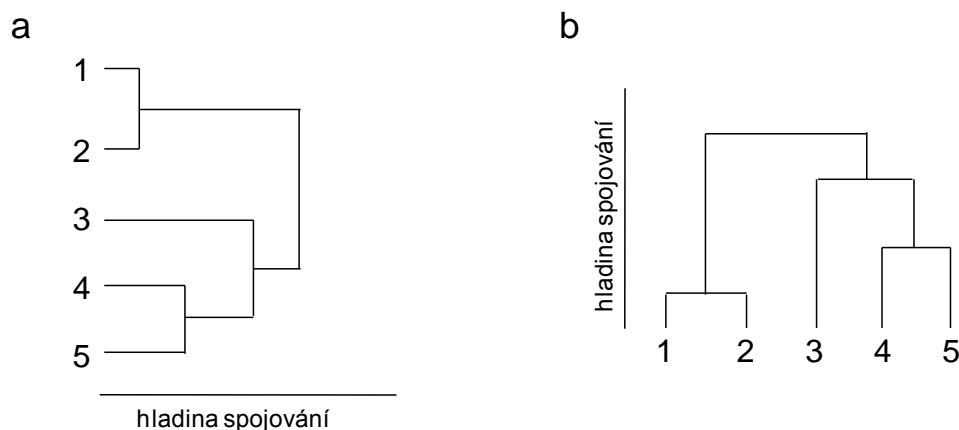
## 6.1 Hierarchické shlukování

Hierarchické shlukovací metody uspořádají skupiny do hierarchické struktury. Jsou dvě možnosti k vytvoření hierarchického shlukování: **aglomerativní** a **divizivní**.

- **Aglomerativní metody.** Při aglomerativních metodách spojujeme objekty navzájem nejpodobnější a poté s každou skupinou pracujeme jako se samostatným objektem až do okamžiku, kdy zůstane pouze jedna skupina. Tento postup není vhodný pro velmi objemná data.
- **Divizivní metody.** Celý soubor se dělí nejčastěji na dvě části – každou z nich lze potom považovat za samostatný soubor, který se znovu dělí.

Metody jsou konstituovány tak, aby podobnost uvnitř skupin a rozdíl mezi skupinami byly co největší.

Výsledky hierarchických shlukovacích metod lze graficky znázornit v podobě **stromu** – **dendrogramu** (Obrázek 6.1). Představíme si jej na příkladu aglomerativního shlukování. Na vodorovné ose je stupnice pro hladinu spojování. Vlevo začíná strom  $n$  větvemi – objekty (v příkladu na obrázku je jich pět). V každém kroku se spájí dvě větve v bodě, který odpovídá příslušné hladině spojení (*linkage distance*). V příkladu na obrázku jsou si nejpodobnější objekty 1 a 2, jsou spojeny na nejnižší hladině. Dendrogram lze zobrazit nejen horizontálně (Obrázek 6.1a), ale i vertikálně (Obrázek 6.1b).



**Obrázek 6.1** Ukázka dendrogramu (stromu) pěti objektů. Strom lze zobrazit horizontálně (a) i vertikálně (b).

### 6.1.1 Hierarchické aglomerativní shlukování

Agglomerativní shluková analýza pracuje se samostatnými objekty, které jsou shlukovány do větších shluků. V mnohých vědních disciplínách jsou aglomerativní techniky používány častěji než divizivní metody. Existuje mnoho aglomerativních metod, každá z nich využívá jiný pohled na data.

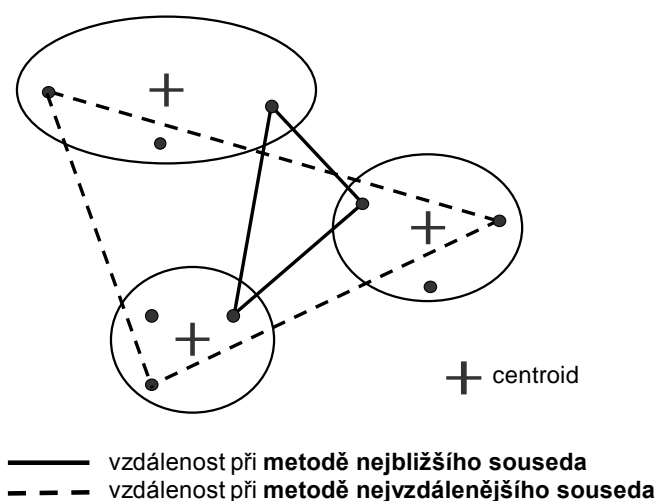
Základním krokem tohoto shlukování je výpočet podobností/vzdáleností mezi všemi dvojicemi objektů, tj. vytvoření asociační matice. V různých etapách algoritmu posuzujeme podobnost/vzdálenost dvou objektů, podobnost/vzdálenost objektu a shluku a podobnost/vzdálenost dvou shluků. Způsob výpočtu podobnosti/vzdálenosti zásadním způsobem ovlivňuje výsledek shlukování.

V předchozí kapitole jsou uvedeny různé míry podobnosti a metriky vzdálenosti. Většinou požadujeme, aby podobnost nabývala hodnot od nuly pro maximální rozdílnost po jedničku pro totožnost. Často se však z praktických důvodů používají různé míry vzdálenosti, tentýž je tedy měřen v opačném směru. Nevyplývají z toho žádné problémy; ostatně každou míru vzdálenosti  $D$  ( $D \geq 0$ ) lze převést na míru podobnosti  $S$ ,  $0 \leq S \leq 1$ , např.  $S = e^{-D}$  a naopak.

V dalším textu stručně představíme několik způsobů stanovení podobnosti/vzdálenosti mezi shluky. S tímto procesem se můžeme setkat také pod názvem aglomerativní metoda, aglomerativní postup, nebo shlukovací algoritmus.

### Vzdálenost mezi shluky (aglomerativní metody)

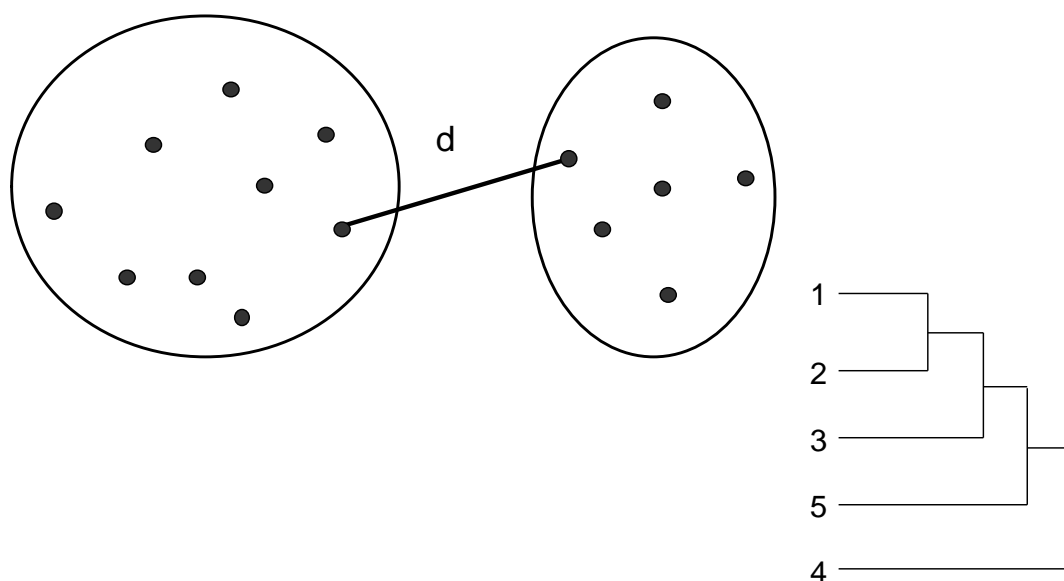
Všechny aglomerativní metody jsou založeny na shlukování jednotlivých objektů nebo shluků do větších skupin. Skupiny, které jsou si nejvíc podobné, jsou sloučeny. Definice vzdálenosti mezi shluky se u jednotlivých metod liší. Metody se navzájem liší chápáním této vzdálenosti (Obrázek 6.2).



**Obrázek 6.2** Vnímání vzdálenosti při metodě nejbližšího a nejvzdálenějšího souseda.

- Metoda nejbližšího souseda (jednospojná metoda, metoda jediné vazby, *single-linkage clustering, the nearest neighbor method*)

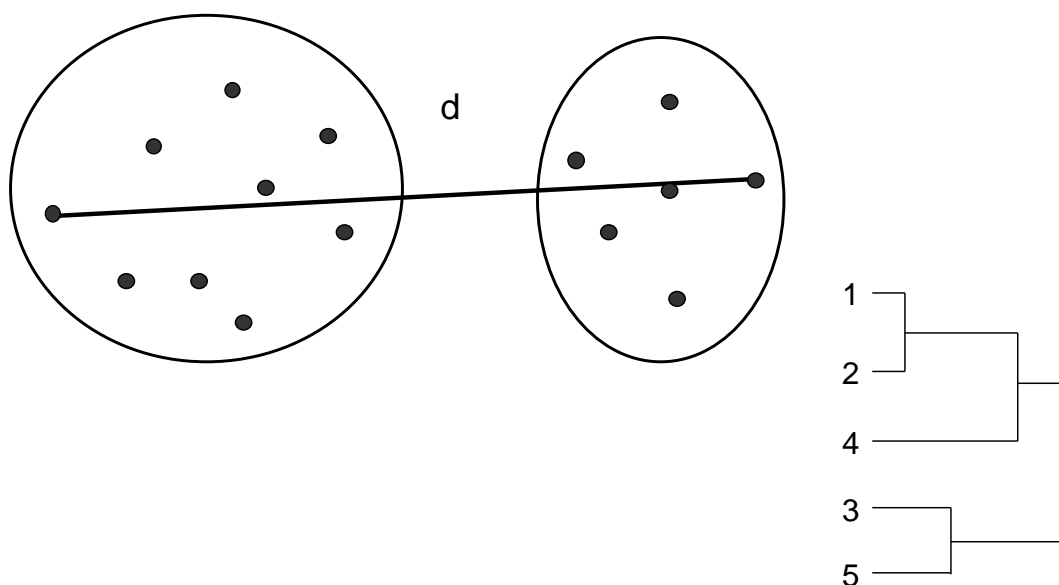
Historicky nejstarší metoda. Vzdálenost mezi dvěma shluky (na počátku analýzy reprezentované jednotlivými objekty) je daná jako minimální vzdálenost mezi všemi možnými zástupci shluků (Obrázek 6.3). To znamená, že ve dvou shlucích, o jejichž spojení uvažujeme, nás zajímají pouze ty dva objekty, které jsou k sobě nejbližší. Při použití této metody se často i značně vzdálené objekty mohou sejít ve stejném shluku, pokud větší počet dalších objektů mezi nimi vytvoří jakýsi most. Toto charakteristické řetězení objektů se považuje za nevýhodu, zvláště když máme důvod požadovat, aby shluky měly obvyklý eliptický tvar se zhuštěným jádrem.



**Obrázek 6.3** *Vzdálenost u metody nejbližšího souseda a ukázka dendrogramu vzniklého touto metodou (podle Marhold, Suda 2002).*

- Metoda nejvzdálenějšího souseda (všespojná metoda, *complete-linkage clustering, the furthest neighbor method*)

Tato metoda je založena na opačném principu než jednospojná metoda. Vzdařenost mezi dvěma shluky je daná maximální vzdáleností mezi všemi možnými zástupci obou shluků (Obrázek 6.4). Tato metoda produkuje shluky, které jsou mezi sebou dobře odděleny. Nežádoucí řetězový efekt zde odpadá, naopak je tu tendence ke tvorbě kompaktních shluků, nikoli mimořádně velkých.



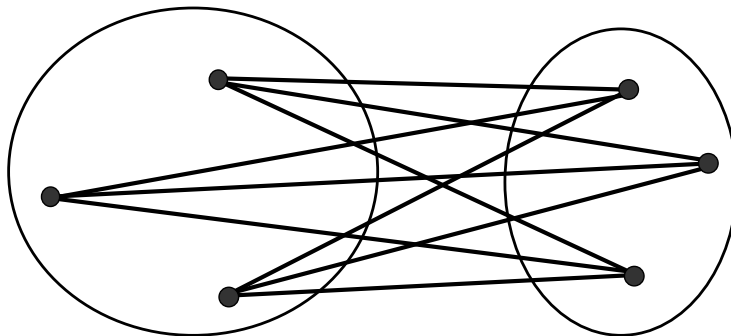
**Obrázek 6.4** *Vzdálenost u metody nejvzdálenějšího souseda a ukázka dendrogramu vzniklého touto metodou (podle Marhold, Suda 2002).*

- Metoda průměrné vazby (středospojná metoda, *average-linkage clustering*)  
Existují čtyři metody průměrného shlukování. První dvě metody UPGMA a WPGMA používají průměrnou vzdálenost mezi všemi členy shluků jako kritérium vzdálenosti mezi

shluky. Metody UPGMC a WPGMC počítají mezishlukovou vzdálenost jako vzdálenost mezi centroidy (těžišti) shluků. Dalším rozdílem u těchto metod je vážení velikosti shluků. Metody UPGMA a UPGMC dávají stejné váhy původním podobnostem a zároveň váhy shluků jsou proporcionální k velikosti shluků, u metod WPGMA a WPGMC jsou váhy shluků stejné bez ohledu na velikost shluku.

- *UPGMA (unweighted pair-group method using arithmetic averages)*

Při této metodě shlukování je vzdálenost mezi shluky definována jako průměr ze všech možných mezishlukových vzdáleností objektů (Obrázek 6.5). Metoda vede často k podobným výsledkům jako metoda nejvzdálenějšího souseda.



**Obrázek 6.5** Vzdálenost u metody průměrné vazby (podle Marhold, Suda 2002).

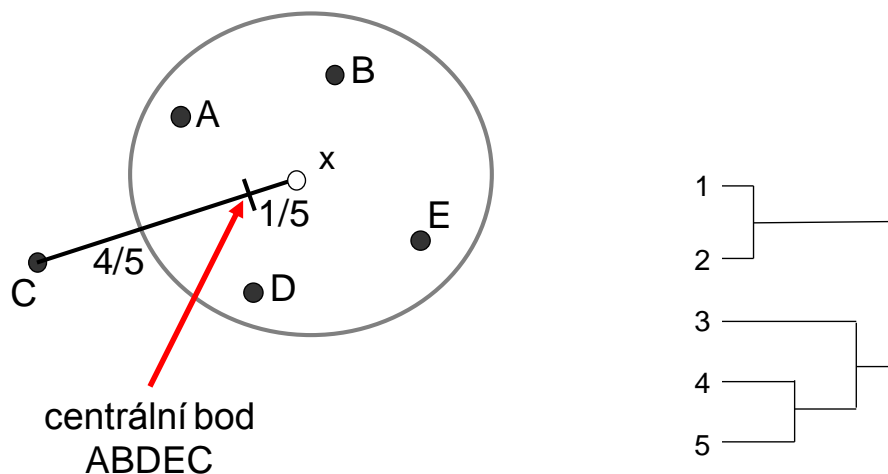
- *WPGMA (weighted pair-group method using arithmetic averages)*

Tato metoda je obdobou předchozí metody ovšem doplněna o vážení shluků jejich velikostí (pod velikostí shluků rozumíme počet jejich objektů). Proto by se tato metoda měla používat v případech, když očekáváme různě velké shluky.

- *UPGMC (unweighted pair-group method using centroids, unweighted centroid clustering, Gowerova metoda)*

Tato metoda nevychází již ze shrnování informací o mezishlukových vzdálenostech objektů. Kritérium je vzdálenost centroidů (těžišť). Při této metodě je vzdálenost mezi shluky počítána jako vzdálenost mezi centroidy těchto shluků. Při shlukování se tedy spojují shluky, jejichž centroidy leží nejbližše. Centroid nového shluku je definován podle polohy původních objektů, nikoliv jako centroid vypočtený z centroidů spojených shluků (Obrázek 6.6). To znamená, že nový centroid získáme jako nevážený průměr ze všech bodů nového shluku a jako vážený průměr původních centroidů.

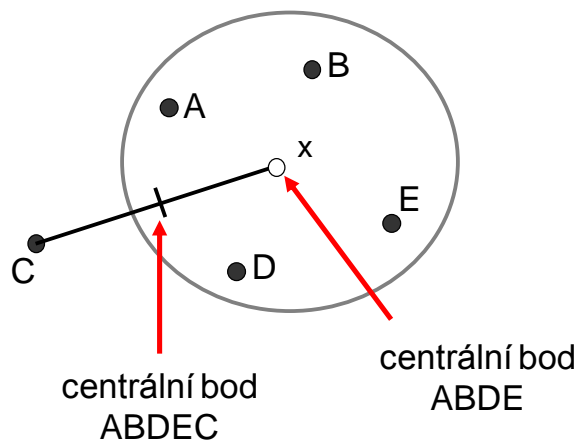
Nevýhodou centroidní metody je skutečnost, že v případě spojování dvou shluků velmi rozdílné velikosti bude centroid (těžiště) nového shluku velmi blízko většího shluku (nebo dokonce uvnitř). Vlastnosti menšího shluku se tak do jisté míry ztrácí.



**Obrázek 6.6** Vzdálenost u centroidní metody a ukázka dendrogramu vzniklého touto metodou (podle Marhold, Suda 2002).

- WPGMC (weighted pair-group method using centroids, weighted centroid clustering, median method, mediánová metoda)

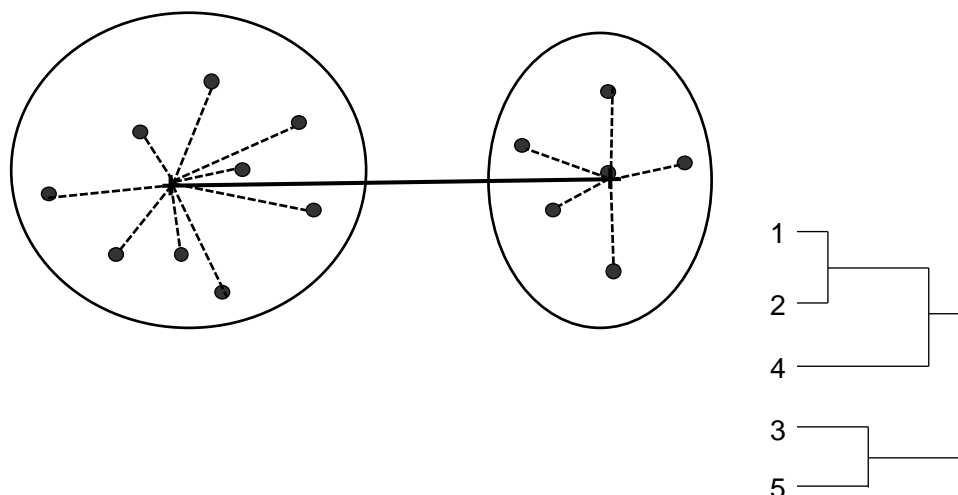
Mediánová metoda odstraňuje problém daný rozdílnou velikostí spojovaných shluků. Analyzované shluky se považují za stejně velké, centroid nového shluku je proto vždy v polovině vzdálenosti mezi centroidy spojovaných shluků. To znamená, že nový centroid získáme jako nevážený průměr původních centroidů (Obrázek 6.7). Jde ovšem o vážený průměr ze všech bodů nového shluku. Tato metoda je preferována tehdy, když očekáváme velké rozdíly ve velikosti shluků.



**Obrázek 6.7** Vzdálenost u mediánové metody (podle Marhold, Suda 2002).

- Wardova metoda (*minimum variance clustering, Ward's method*)

Wardova metoda je podobná středospojné a centroidní metodě. Kritérium pro spojování shluků je přírůstek celkového vnitroskupinového součtu čtverců odchylek pozorování od shlukového průměru (Obrázek 6.8). Přírůstek je vyjádřený jako součet čtverců v nově vznikajícím shluku, zmenšený o součty čtverců v obou zanikajících shlucích. Wardova metoda má tendenci odstraňovat malé shluky, tedy tvořit shluky zhruba shodné velikosti, což je často vítaná vlastnost.

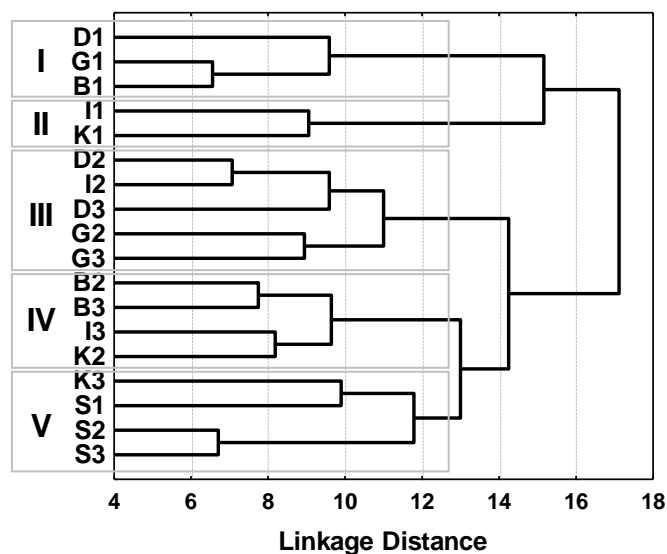


**Obrázek 6.8** Vzdálenost u Wardovy metody a ukázka dendrogramu vzniklého Wardovou metodou (podle Meloun, Militký 2004 a Marhold, Suda 2002).

Aglomerativní hierarchický algoritmus můžeme definovat následovně:

1. Vypočteme asociační matici vhodných měr vzdálenosti.
2. Proces začneme od rozkladu  $S^{(n)}$ , tj. od  $n$  shluků, z nichž každý obsahuje jeden objekt.
3. V asociační matici najdeme dva objekty/shluky ( $g$ -tý a  $h$ -tý), jejichž vzdálenost je minimální.
4. Spojíme dva shluky nalezené v bodě 3 ( $g$ -tý a  $h$ -tý) do nového shluku ( $i$ -tý). V původní matici vymažeme  $g$ -tý a  $h$ -tý řádek i sloupec a nahradíme je řádkem i sloupcem pro nový shluk. Řád matice se sníží o jednu.
5. Zaznamenejme pořadí cyklu rozkladu  $I = 1, 2, \dots, n-1$ , dále identifikaci spojených objektů/shluků a hladinu pro spojení.
6. Když proces vytváření rozkladů ještě neskončil spojením všech objektů do jediného shluku  $S^{(1)}$ , pokračujeme znovu bodem 3.

Interpretaci výsledku hierarchického aglomerativního shlukování si představíme na konkrétním příkladu. Cílem bylo zjistit podobnost šesti lokalit ve třech časových obdobích z hlediska výskytu korýšů. Zajímalo nás, zda si jsou lokality podobnější v čase nebo v prostoru. Vstupní matici tvořilo 64 taxonů korýšů vyskytujících se v 18 objektech. Objekty představovalo šest lokalit v záplavové oblasti Dunaje ve třech obdobích (1: 1991-1992 před přehrazením Dunaje, 2: 1993-1997 prvních 5 let po přehrazení, 3: 1999-2004 dalších 6 let po přehrazení). Sledovanými lokalitami byly: D: Dobrohošť, G: Gabčíkovo, B: Bodíky, I: Istragov, K: Královská lúka, S: Sporná sihoť. Použitá byla všespojná shlukovací metoda (*complete linkage*) a jako míra vzdálenosti Euklidovská vzdálenost.

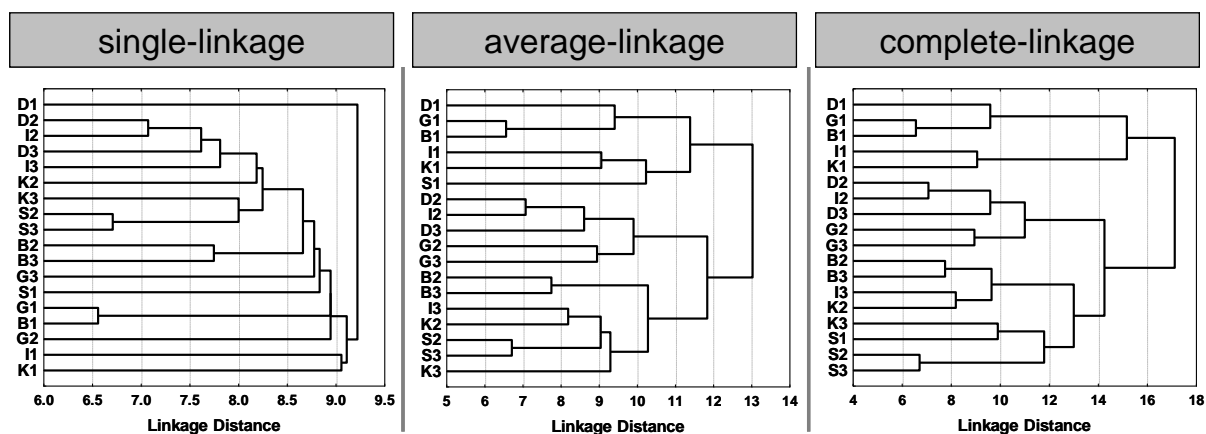


**Obrázek 6.9** Ukázka výsledku shlukové analýzy společenstev koryšů (Illyová, Némethová 2005).

Interpretace dendrogramu je následovná (Obrázek 6.9): na určené hladině spojování (*linkage distance*) se vytvořilo pět shluků lokalit. První shluk (I) obsahuje lokality D1, G1, B1 – lokality Dobrohošť, Bodíky a Gabčíkovo před přehrazením Dunaje. V tomto shluku jsou si nejpodobnější lokality Gabčíkovo a Bodíky (jsou sloučeny na nižší hladině spojování). Druhý shluk (II) obsahuje lokality I1, K1 - Istragov a Královská lúka v období před přehrazením. Třetí shluk obsahuje lokality D2, D3, G2, G3, I2: Dobrohošť a Gabčíkovo ve druhém a třetím období (po přehrazení) společně s lokalitou Istragov ve druhém období. V tomto shluku jsou si nejpodobnější lokality Dobrohošť ve druhém období a Istragov taky ve druhém období. Čtvrtý shluk je tvořen lokalitami B2, B3, I3, K2: Bodíky (druhé a třetí období), Istragov (třetí období) a Královská lúka (druhé období). Poslední pátý shluk je tvořen lokalitami K3, S1, S2, S3: Sporná síhoť (všechna období) a Královská lúka ve třetím období.

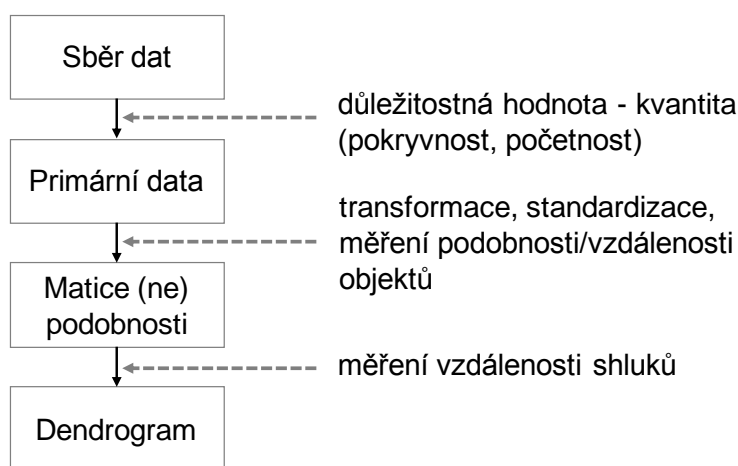
Je velmi žádoucí doplnit takové zhodnocení dendrogramu o popis, co mají dané objekty (v tomto případě lokality v časových obdobích) v jednotlivých shlucích společné (výskyt konkrétních taxonů) a čím se shluky lokalit mezi sebou liší.

Na Obrázku 6.10 lze vidět, jak různé jsou výsledné dendrogramy při použití různých shlukovacích algoritmů.



**Obrázek 6.10** Dendrogramy vytvořené pomocí stejné metriky vzdálenosti (Euklidovská vzdálenost) a tří různých shlukovacích algoritmů: jednospojné (single), středospojné (average) a všespojné (complete linkage) metody. V případě jednospojné metody je zjevné silné řetězení objektů. (Společenstva koryšů šesti lokalit ve třech časových obdobích; Illyová, Némethová 2005.)

Výsledek hierarchického aglomerativního shlukování je ovlivněn na několika úrovních (Obrázek 6.11). Jde nejenom o typ vstupních dat, ale také o jejich případnou transformaci a standardizaci, dále o měření vzdálenosti/podobnosti mezi objekty a následně o měření vzdálenosti mezi shluky (shlukovací algoritmus). Podle Kováře a Lepše (1986) mají transformace dat větší vliv na výsledek shlukování než metoda shlukování (měření vzdálenosti mezi shluky).



**Obrázek 6.11** Výsledek hierarchického aglomerativního shlukování je ovlivněn na několika úrovních (podle Lepš, Šmilauer 2000).

Závěrem můžeme definovat tyto hlavní kritické problémy hierarchické aglomerativní analýzy:

- Velké množství proměnných nebo objektů v dendrogramu je obtížné interpretovat.
- Analýza je silně závislá na zvolení vhodné metriky vzdálenosti/koefficientu podobnosti.
- Analýza je silně závislá na shlukovacím algoritmu (způsobu měření vzdálenosti mezi shluky).
- Může nastat situace, kdy se v asociační matici vyskytnou tzv. shody (*ties*) – stejné hodnoty u různých skupin objektů, případně shluků. Dochází k tomu zejména při analýze binárních dat. Existuje několik možností řešení těchto shod v závislosti od typu vazeb mezi objekty (např. spojení všech objektů najednou, paralelní vytvoření skupin tzv. *multiple fusion*, náhodné spojení tzv. *silent mode*, *single linkage*, *suboptimal fusions*). Různé způsoby vypořádání se se shodami ovšem ovlivňují výsledný dendrogram.

Hierarchické aglomerativní metody jsou velice populární a jejich výhody jsou následovné:

- Jsou vhodné pro méně objemná data.
- Výsledný dendrogram je jednoduše interpretovatelný.



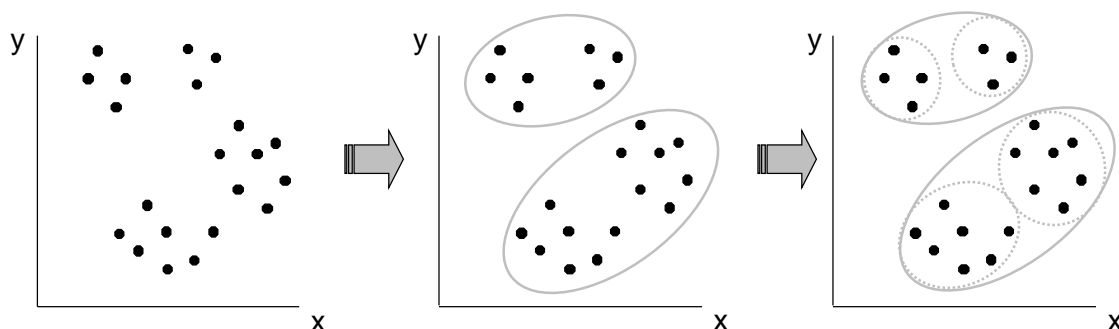
### 6.1.2 Hierarchické divizivní shlukování

Divizivní metody pracují ze začátku se všemi objekty jako s jednou skupinou. Nejdříve je tato skupina rozdělena do dvou menších skupin. Dále se s každou skupinou pracuje samostatně, dochází k její rozdělení na dvě podskupiny. Dělení podskupin pokračuje dále až dokud není splněno kritérium, které ukončí analýzu (např. předem definovaný počet kroků, případně rozklad na samostatné objekty; Obrázek 6.12). Principem tohoto způsobu shlukování je, že větší rozdíly přetrvávají nad méně důležitými rozdíly: celková struktura shluku determinuje podskupiny.

Divizivní hierarchický postup můžeme tedy formalizovat následovně: vycházíme od jediného shluku  $S^{(1)}$  a v každém kroku jeden ze shluků rozštěpíme na dva, takže na konci procesu dostáváme rozklad  $S^{(n)}$ .

Divizivní metody mohou být

- **monotetické** – dělení souboru probíhá podle jediné proměnné;
- **polytetické** – dělení probíhá podle komplexní charakteristiky získané na základě všech proměnných v rámci souboru.



**Obrázek 6.12** Princip divizivního shlukování.

Divizivní metody jsou často používány v ekologii, konkrétně ke klasifikaci biologických společenstev. Jejich výhody jsou následovné:

- divizivní metody jsou vhodné pro objemné datové soubory;
- ke každému dělení je připojeno kritérium, podle kterého dělení proběhlo.

#### Monotetické metody

Význam monotetických metod je hlavně historický. Jednou z nich, která se osvědčila, je **asociační analýza** (*association analysis*). V současnosti se již nepoužívá, my ji zde ovšem uvádíme zejména kvůli vysvětlení principu divizivního shlukování.

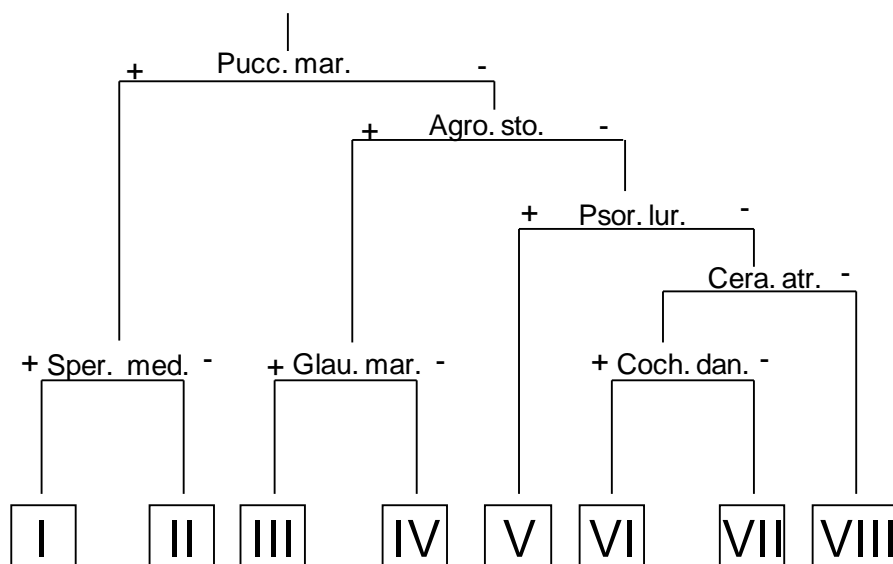
Asociační analýza byla používána v ekologii ke klasifikaci společenstev. Použitelná je pro binární kvalitativní data (v ekologii jde o data prezence-absence druhů). Shluk se dělí na základě jedné proměnné (prezence-absence jednoho tzv. kritického druhu). Na začátku asociační analýzy se určí proměnná, která je maximálně asociovaná s ostatními proměnnými: asociace mezi proměnnými je odhadována jako kvalitativní korelační koeficient pro binární data, bez ohledu na jeho znaménko. Pro každou proměnnou je spočtena suma všech asociací. Proměnná, která má nejvyšší sumární hodnotu asociací určuje dělení shluku na dvě skupiny. Jedna skupina je skupina objektů (vzorky, odběry, nebo lokality), ve kterých je proměnná kódovaná jednotkou, druhá skupina je skupina objektů, ve kterých je tato proměnná kódovaná

nulou (Obrázek 6.13). Tato proměnná je vyřazena z dalšího výpočtu a postup se opakuje pro každý z obou vytvořených shluků samostatně.

Metoda je citlivá na přítomnost vzácných druhů a nepřítomnost běžnějších druhů. Proto se již nepoužívá ve svojí původní formě. Z ekologické zkušenosti je zřejmé, že přítomnost a zvláště nepřítomnost určitého jediného druhu je velmi slabou indikací pro zařazení lokality nebo společenstva k určité skupině. Divizivní monotetické shlukování tedy není robustní.

Výhodou monotetické metody je jednoduchý klíč, který může být použit ke klasifikaci dalších objektů podle prezenze a absence druhů.

Zřejmou nevýhodou metody je její monotetická povaha.



**Obrázek 6.13** Ukázka výsledku asociační analýzy. Binární klíč k identifikaci typů slanisk západního Irsku (Ivimey-Cook, Proctor 1966 v Digby, Kempton 1987).

## Polytetické metody

U polytetických metod probíhá dělení souboru na základě všech proměnných. Skupiny vytvořené polytetickou metodou jsou homogennější než skupiny vytvořené monotetickou metodou.

Mezi ekology je velice oblíbená metoda **two way indicator species analysis** a program **TWINSPAN**. Jde o polytetickou metodu, která dělí objekty (vzorky, odběry, lokality) podle výsledků ordinace korespondenční analýzou. Toto rozdělení je tedy založeno na všech proměnných (v ekologii druzích).

TWINSPAN pracuje pouze s kvalitativními daty. Aby mohla být zahrnuta informace o kvantitě druhů, byl vyvinut kvalitativní ekvivalent druhové abundance, tzv. pseudo-druh (*pseudo-species*). Každá abundance druhu je nahrazena přítomností jednoho nebo více pseudo-druhů. Čím víc početnější je druh, tím víc pseudo-druhů je definováno. Každý pseudo-druh je definován minimální abundancí korespondujícího druhu, tzv. hraniční hodnotou (*cut level*, *cut-off level*). Pseudo-druh je tedy přítomen, pokud zastoupení druhu přesáhne hraniční hodnotu (tabulka 6.1).

**Tabulka 6.1** Ukázka tvorby pseudo-druhů pro TWINSPAN při použití hraničních hodnot 0, 1, 5, 20 (podle Lepš, Šmilauer 2000, 2003).

Druh	Vzorek 1	Vzorek 2
------	----------	----------

Původní tabulka	<i>Cirsium oleraceum</i>	0	1
	<i>Glechoma hederacea</i>	6	0
	<i>Juncus tenuis</i>	15	25
Tabulka s pseudo-druhy použitá v TWINSpan	Cirsoler1	0	1
	Glechede1	1	0
	Glechede2	1	0
	Junctenu1	1	1
	Junctenu2	1	1
	Junctenu3	1	1
	Junctenu4	0	1

Výhodou nahrazení kvantitativní proměnné několika kvalitativními proměnnými je, že když abundance druhu vykazuje unimodální odezvu podél gradientu, každý pseudo-druh také vykazuje unimodální křivku odezvy, a když je křivka odezvy pro abundanci zešíkmená, pak se křivky odezvy pseudo-druhů liší ve svých optimech.

Proces dělení (*dichotomy, division*) objektů do skupin probíhá pomocí korespondenční analýzy. Objekty se rozdělí do dvou skupin: na levou – zápornou a pravou – kladnou stranu dichotomie podle jejich skóre na první ose korespondenční analýzy. Osa je rozdělena v centroidu (těžišti). Ordinance se zopakuje s přiřazením větší váhy druhům, které upřednostňují jednu nebo druhou stranu dichotomie. Algoritmus je komplikovaný, jde o výpočet polarizovaných ordinací a získání většiny vzorků mimo těžiště. Pak je klasifikace založena hlavně na druzích typických pro levou nebo pravou stranu dichotomie. Po rozdělení souboru objektů na dvě části je každá část dále podrobena další ordinaci, vzniknou čtyři skupiny, atd.

Výhody TWINSpan-u jsou následovné:

- TWINSpan nejenom klasifikuje objekty (lokality), ale poskytuje i kritérium použité pro to které dělení. Klasifikace vzorků je doplněna klasifikací druhů.
- TWINSpan je užitečný hlavně při analýze velkých datových souborů.

Nevýhodou této metody, tak často používané v ekologii společenstev, je nutnost zvolit hraniční hodnoty pro tvorbu pseudo-druhů. Výsledek analýzy je těmito hraničními hodnotami silně ovlivněn.

## 6.2 Nehierarchické shlukování

Často existují případy, kdy není výhodné používat hierarchickou shlukovou analýzu, protože data nevykazují hierarchickou strukturu. V takových případech může být vhodnější použití nehierarchického shlukování, při kterém jsou vytvořeny skupiny stejného řádu. Skupiny by měly být uvnitř co nejvíc homogenní a mezi sebou odlišné.

Nehierarchické metody shlukování jsou vhodné pro velmi objemná data.

### 6.2.1 Metoda *K*-průměrů (*K-means clustering*)

Nejběžnější nehierarchickou metodou je **metoda *K*-průměru**. Hlavním cílem metody je nalezení takových skupin v mnohorozměrném prostoru, kdy vnitroskupinová podobnost je co největší. Princip vytvoření shluků je stejný jako při Wardově metodě: minimalizace celkové

sumy čtverců vzdáleností uvnitř skupin. Výsledkem je vytvoření  $K$  skupin, které jsou co nejvíce odděleny od sebe.

Algoritmus metody je následovný:

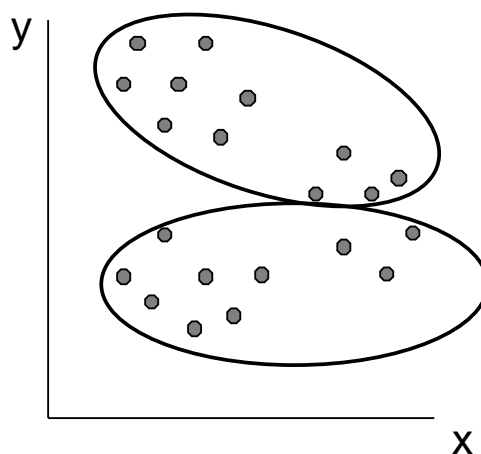
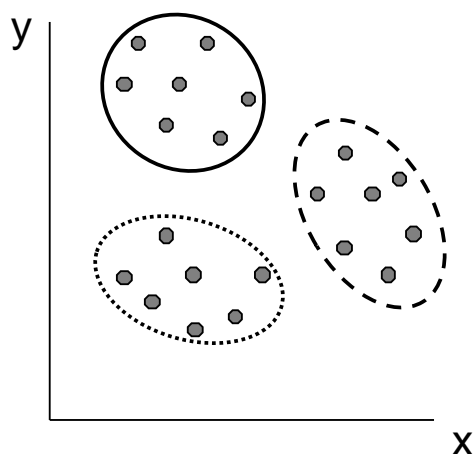
1. Zvolíme počátečný rozklad do  $K$  shluků, nejčastěji náhodně (podkladem ovšem může být také např. výsledek již provedeného shlukování, který chceme zlepšit).
2. Určíme centroidy pro všechny shluky v aktuálním rozkladu.
3. Postupně zhodnotíme pozici všech objektů. Pokud má objekt nejbližší k vlastnímu centroidu, ponecháme jej na místě, jinak jej přesuneme do shluku, k jehož centroidu má nejbližší.
4. Centroidy každého z  $K$  shluků jsou přepočítány.
5. Body 3 a 4 se opakují, až když už žádný další přesun nezlepší kritéria. Tímto způsobem se v  $K$  skupinách objekty přesouvají tak, aby se minimalizovala variabilita uvnitř skupin a maximalizovala variabilita mezi skupinami (jde o relokační proceduru). Proces je tedy iterativní.

Tento algoritmus je základní, existuje ovšem i několik modifikací:

- Proces lze zahájit s  $K$  vybranými objekty místo počátečního rozkladu. Pak se dostáváme hned ke kroku č. 3. Další postup je již stejný.
- Přepočet centroidů lze provést po každém přesunu objektu (nikoli tedy jen po každém cyklu). Průběh shlukování a výsledek je pak závislý také na pořadí objektů, ve kterém vstupují do 3. kroku.

Nevýhodou metody  $K$ -průměrů je, že pracuje se čtverci Euklidovských vzdáleností. To může být v některých případech problém, zejména když se vyskytují odlehle objekty. Metoda  $K$ -průměrů je citlivá na odlehle hodnoty.

Další nevýhodou metody je nutnost definovat počet skupin  $K$  předem. Je potřebné si uvědomit, že takto můžeme získat pouze lokální extrém, o kterém nemáme jistotu, že je zároveň extrémem globálním (Obrázek 6.14). Proto je vhodné provést analýzu pro několik různých počátečních  $K$  skupin a následně určit poměr vnitroskupinové a meziskupinové variability pro všechny analýzy (všechny  $K$ ). Nakonec bude jako nejlepší určen takový počet shluků  $K$ , při kterém je poměr vnitroskupinové a meziskupinové variability nejmenší.



**Obrázek 6.14** Ukázka rozdělení objektů do shluků nehierarchickou metodou  $K$ -průměrů. Výsledek je ovlivněn volbou počtu shluků. Vlevo: počet shluků tři je dobrá volba; vpravo: počet shluků dva je špatná volba.

### 6.2.2 Metoda $X$ -průměrů ( $X$ -means clustering)

Pro nejrozšířenější nehierarchickou shlukovací metodu  $K$ -průměrů můžeme definovat dva hlavní problémy: 1. počet shluků  $K$  musí být definován uživatelem a 2. hledání  $K$  shluků podléhá lokálnímu minimu. Řešení prvního problému a částečně i druhého problému nabízí **metoda  $X$ -průměrů**.

V algoritmu metody  $X$ -průměrů se počet shluků vypočítá dynamicky, přičemž je uživatelem zadávána pouze dolní a horní hranice pro  $K$ .

Algoritmus je tvořen dvěma kroky, které se opakují.

1. V prvním kroku je aplikována tradiční metoda  $K$ -průměrů pro  $K$  shluků ( $K$  je nejprve rovno dolní hranici určené uživatelem).
2. V druhém kroku se zjišťuje, zda a kde se má objevit nový centroid, nový shluk. Toto je dosaženo tím, že se některé shluky nechají rozpadnout na dva. Proces začíná tím, že se každý centroid shluku (nazveme jej rodičovský centroid) rozdělí na dva centroidy (dcerské centroidy) v opačném směru podél náhodně zvoleného vektoru. Pak se pro každou rodičovskou oblast, čili pro každý pár dcerských centroidů, vypočítá lokální metoda  $K$ -průměru pro dva shluky. Hranice rodičovských oblastí se nemění. Srovnáním Bayesovského informačního kritéria (BIC) pro model s dceřinými centroidy a model s rodičovským centroidem se rozhodne o výsledné struktuře. Podle výsledku testu je buď zachován rodičovský centroid (a tedy rodičovský shluk), nebo je nahrazen dceřinými centroidy (tj. dvěma dceřinými shluky).
3. Když  $K \geq K_{max}$  (horní hranice určena uživatelem) proces se ukončí a vyhodnotí se nejlepší model v průběhu hledání, tj. sada centroidů s nejlepší hodnotou testového kritéria. Jinak se pokračuje znovu krokem 1.

Jako kritérium pro dělení shluku na dva dcerské shluky může být kromě BIC použito i jiné, např. Akaikovo informační kritérium (AIC).

Výhodou tohoto postupu je také fakt, že regionální metoda  $K$ -průměrů s pouze dvěma shluky je méně citlivá na lokální minima.

### 6.2.3 Metoda $K$ -medoidů: PAM ( $K$ -medoids method: partitioning around medoids)

**Metoda  $K$ -medoidů** je velice podobná metodě  $K$ -průměrů, s tím rozdílem, že zástupcem středu shluku není centroid ale tzv. reprezentativní objekt – **medoid**.

Další rozdíl mezi metodami  $K$ -průměrů a  $K$ -medoidů je v míře, kterou se hodnotí vzdálenost objektů od středu shluku (centroidů v metodě  $K$ -průměrů, medoidů v metodě  $K$ -medoidů).

Princip metody  $K$ -medoidů je v hledání  $K$  reprezentativních objektů, které nazýváme medoidy. Medoid je definován jako objekt shluku, jehož průměrná nepodobnost ke všem objektům v shluku je minimální, tj. je to nejcentrálněji umístěný bod v daném datovém souboru. Shluk je pak definován jako soubor objektů, které jsou přiřazeny ke stejnému medoidu.

Metodu  $K$ -medoidů můžeme považovat za robustnější obdobu metody  $K$ -průměrů.

Nejčastější realizací shlukování  $K$ -medoidů je algoritmus **PAM** *Partitioning around medoids*:

1. Postupně je selektováno  $K$  reprezentativních objektů. První objekt je ten, pro který je suma nepodobností ke všem dalším objektům co nejmenší. Tento objekt je umístěn nejvíce centrálně v sadě objektů. Postupně je v každé iteraci vybrán další objekt, který snižuje sumu (přes všechny objekty) nepodobností k nejpodobnějšímu vybranému objektu co nejvíce. Proces pokračuje, až dokud není nalezeno  $K$  reprezentativních objektů – medoidů.
2. Všechny objekty jsou spojeny s nejbližším medoidem. Míra nepodobnosti/vzdálenosti je definována jakoukoliv platnou metrikou vzdálenosti, nejčastěji Euklidovskou vzdáleností, Manhattanskou vzdáleností, Minkowského vzdáleností,  $1 - \text{korelace}$ .
3. V druhé fázi algoritmu se zlepšuje sada medoidů a tedy shlukování. To se děje srovnáním všech párů objektů, kde jeden z nich je medoidem a druhý ne. Pro každý medoid  $m$  a postupně pro každý objekt  $o$ , který není medoidem, se vymění pozice  $m$  a  $o$  a zjišťuje se hodnota kritéria shlukování pro tuto konfiguraci. Když se zlepší kritérium shlukování, testovaný objekt se stane medoidem místo původního medoidu. Tato procedura se opakuje, až dokud již nedochází k žádnému dalšímu zlepšení.

Výhody metody  $K$ -medoidů:

- Metoda nevyžaduje původní data, může být aplikována také přímo na matici nepodobností.
- Shlukování je možné na základě jakékoliv míry vzdálenosti (důležité např. v biologických aplikacích, kdy se může jednat např. o shlukování korelovaných prvků).
- Medoidy jsou robustními představiteli středů shluků, jsou méně citlivé k odlehlým pozorováním než centroidy v metodě  $K$ -průměrů (tato robustnost je důležitá, když objekty nepatří jasně k žádnému shluku).
- Shlukování není závislé na pořadí objektů v datové matici (s výjimkou když existují ekvivalentní řešení, co je velice zřídka).

Nevýhodou metody  $K$ -medoidů je stejně jak tomu bylo v metodě  $K$ -průměrů potřeba definovat počet shluků  $K$  předem. Tento problém lze řešit pomocí koeficientu siluety (*silhouette coefficient*).

## 6.3 Určení optimálního počtu shluků

Validací shlukové analýzy se rozumí měření kvality shlukování pro jednotlivé algoritmy nebo stejný algoritmus, který počítal několikrát s jinými proměnnými. Validace shlukové analýzy je velmi důležitý krok, protože výsledek shlukování musí být ověřen ve většině aplikací. Ve většině případů musí být počet výsledných shluků nastaven uživatelem. Existuje několik přístupů, jak určit správný počet shluků.

### 6.3.1 Analýza rozptylu (ANOVA)

Velmi snadným a dobře pochopitelným způsobem určení počtu shluků může být analýza rozptylu (ANOVA), popřípadě její neparametrická obdoba Kruskal-Wallisova analýza

rozptylu. Při použití této metody jako validační techniky sledujeme vliv rozdělení datového souboru do shluků na jednotlivé proměnné. Sledujeme, zda proměnné mají v jednotlivých shlucích rozdílné hodnoty. Vybíráme takový počet shluků, který nám nejlépe odděluje požadované proměnné. Jedná se o jednorozměrnou metodu, která pracuje přímo s datovou maticí oproti ostatním metodám, které pracují s asociačními maticemi.

### 6.3.2 Dunnův validační index (*Dunn's validity index*)

Tento index je založen na předpokladu, že nalezené shluky jsou kompaktní a dobře oddělené. Pro všechny oddělené shluky, kde  $c_i$  představuje  $i$  – tý shluk, je Dunnův validační index počítán podle vzorce:

$$D = \min_{1 \leq i \leq n} \left\{ \min_{\substack{1 \leq i \leq n \\ i \neq j}} \left\{ \frac{d(c_i, c_j)}{\max_{1 \leq i \leq n} \{d'(c_k)\}} \right\} \right\} \quad (6.1)$$

kde  $d(c_i, c_j)$  představuje vzdálenost mezi shluky  $c_i$  a  $c_j$  (mezishluková vzdálenost),  $d'(c_k)$  je vzdálenost uvnitř shluků,  $n$  je počet shluků. Minimum je počítáno pro všechny shluky, které byly získány. Hlavním cílem tohoto indexu je maximalizovat vzdálenost mezi shluky a minimalizovat vzdálenost uvnitř shluků. Z toho vyplývá, že vysoké hodnoty indexu indikují optimální počet shluků.

### 6.3.3 Daviesův-Bouldinův validační index (*Davies-Bouldin validity index*)

Daviesův-Bouldinův validační index je podíl sumy vnitro–shlukového rozložení a mezi–shlukového rozložení. Hodnoty tohoto indexu získáme ze vzorce:

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{i \neq j} \left\{ \frac{S_n(Q_i) + S_n(Q_j)}{S_n(Q_i, Q_j)} \right\} \quad (6.2)$$

kde  $n$  je počet shluků,  $S_n(Q_i)$  je průměrná vzdálenost objektů ve shluku od středu shluku a  $S_n(Q_i, Q_j)$  je vzdálenost mezi středy shluků. Nízký podíl získáme, když jsou shluky kompaktní a daleko od sebe. Nízké hodnoty tohoto indexu indikují optimální počet shluků.

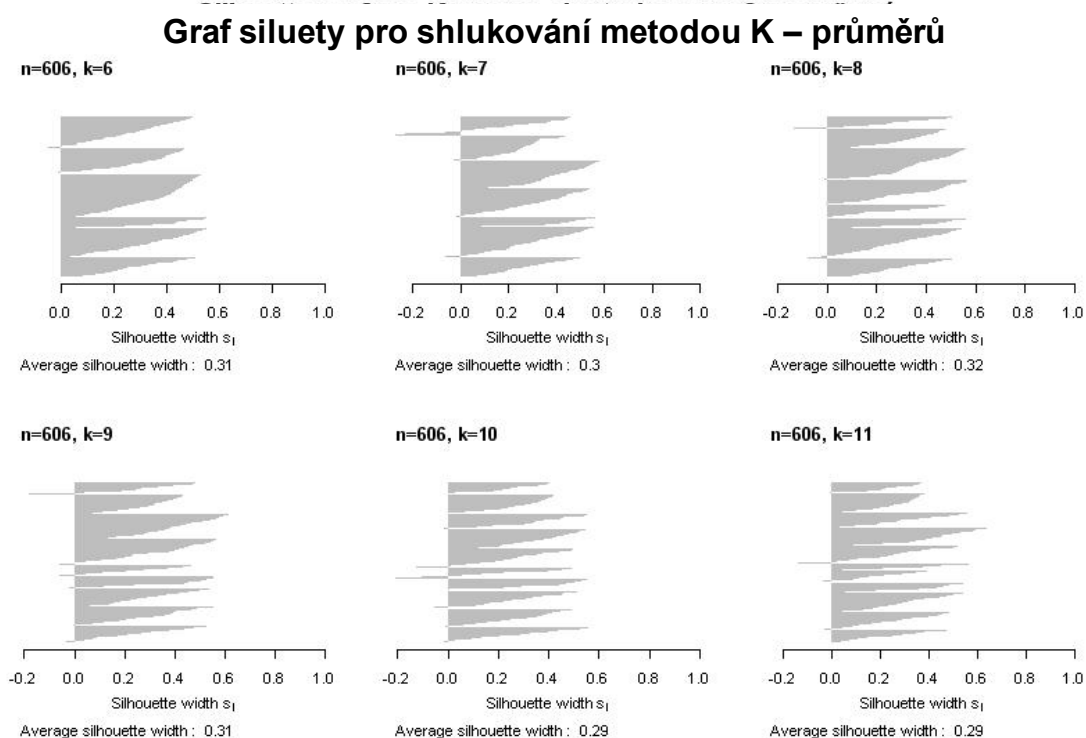
### 6.3.4 Validační metoda siluety

Validační metoda siluety, počítá hodnotu šířky siluety pro každý objekt, průměrnou hodnotu šířky siluety pro každý shluk a průměrnou hodnotu šířky siluety pro celý soubor. Tento přístup je založen na porovnání průměrné šířky siluety pro daný shluk. Silueta zde reprezentuje poměr podobnosti a odlišnosti od ostatních shluků. Průměrná šířka siluety může být použita k validaci shlukové analýzy a k rozhodnutí o vhodnosti zvoleného počtu shluků. K získání hodnoty  $S(i)$  použijeme vzorec

$$S(i) = \frac{(b(i) - a(i))}{\max\{b(i), a(i)\}} \quad (6.3)$$

kde  $a(i)$  je průměrná odlišnost  $i$  – tého objektu od všech ostatních vzorků ve stejném shluku,  $b(i)$  je minimum z průměrů odlišnosti  $i$  – tého objektu ke všem vzorkům v ostatních shlucích.  $S(i)$  může nabývat hodnot  $\langle -1, 1 \rangle$ . Když je hodnota siluety blízká jedné, znamená to, že objekt je zařazen do správného shluku, je-li hodnota siluety blízká nule, znamená to, že objekt můžeme zařadit také do jiného shluku, vzorek leží stejně daleko od obou shluků. Hodnota mínus jedna nám indikuje špatně zařazený objekt, nachází se někde mezi shluky. Celková průměrná hodnota pro celý datový soubor je jednoduše průměr ze všech získaných  $S(i)$ .

Největší hodnota celkové průměrné siluety indikuje nejlepší shlukování (počet shluků). Proto počet shluků s největší průměrnou hodnotou šířky siluety je optimální řešení. Výstupem této metody bývá sada grafů, kde jsou vyznačeny hodnoty siluety pro všechny objekty ve shlucích pro více variant shlukování (Obrázek 6.15).



**Obrázek 6.15** Graf siluety. Bylo zde shlukováno 606 lokalit do 6-ti až 11-ti shluků. Optimální počet shluků je 8, kde je nejvyšší hodnota průměrné siluety. Také si můžeme všimnout záporných hodnot siluety, které nám indikují špatně zařazené shluky.

### 6.3.5 Izolační index (*Isolation index*)

Tento index je založen na tvrzení, že sousední vzorky (v prostoru) patří do stejného shluku. Izolace každého shluku je měřena pomocí pravidla  $k$ –nejbližšího souseda, kde pravidlo pro každý případ  $a$  je definováno jako procento  $k$ –nejbližších sousedů, které byly zařazený do stejného shluku jako  $a$ . Průměrováním přes všechny případy v datech můžeme homogenitu rozdělení spočítat podle vzorce:

$$I_k = \frac{1}{n} \sum_{i=1}^n v_k(x_i) \quad (6.4)$$



Vysoké hodnoty tohoto indexu znamenají dobře oddělené shluky. Autoři uvádí, že index odměňuje rozklad na kompaktní a dobře oddělené shluky, avšak nedokáže penalizovat případy, kdy se shluky překrývají, protože každý objekt je limitován okolím.

### 6.3.6 C-index

Tento index je definován vztahem:

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}}. \quad (6.5)$$

Kde  $S$  je suma vzdáleností mezi všemi páry objektů ve shluku. Necht'  $p$  je počet takovýchto párů objektů patřících do jednoho shluku a  $P$  je počet takovýchto párů objektů v celém datovém souboru. Všechny páry v datovém souboru seřadíme podle jejich vzdálenosti, a vybereme  $p$  nejmenších vzdáleností a  $p$  největších vzdáleností. Takto získáme  $S_{\min}$ , což je suma nejmenších  $p$  vzdáleností v datovém souboru, a  $S_{\max}$ , sumu  $p$  největších vzdáleností. Nízké hodnoty čitatele ve vzorci znamenají, že v daném shluku se vyskytují páry objektů s malou vzdáleností. Minimální hodnoty C-indexu indikují dobře oddělené shluky. Počet shluků, který minimalizuje hodnotu C-indexu, je optimální.

### 6.3.7 Goodmanův-Kruskalův index (*Goodman-Kruskal index*)

Pro daný datový soubor Goodmanův-Kruskalův index hodnotí všechny možné čtveřice objektů ( $a, b, c, d$ ). Necht'  $d$  je vzdálenost mezi dvěma objekty ( $a$  a  $b$  nebo  $c$  a  $d$ ).

Pak se čtveřice nazývá shoda (*concordant*), když platí  $d(a,b) < d(c,d)$ , přičemž  $a$  a  $b$  jsou ve stejném shluku a  $c$  a  $d$  nejsou ve stejném shluku nebo  $d(a,b) > d(c,d)$ , přičemž  $c$  a  $d$  jsou ve stejném shluku a  $a$  a  $b$  jsou ve shlucích odlišných.

Naopak se čtveřice nazývá neshoda (*discordant*), když platí  $d(a,b) < d(c,d)$  a  $a$  a  $b$  nejsou ve stejném shluku, zatím co  $c$  a  $d$  ve stejném shluku jsou. Nebo také  $d(a,b) > d(c,d)$  přičemž  $a$  a  $b$  jsou ve stejném shluku a  $c$  a  $d$  nejsou ve stejném shluku. Dobré rozdělení datového souboru by mělo obsahovat hodně shod a málo neshod těchto čtveřic. Označme počet shod  $N_c$  a neshod  $N_d$ . Goodmanův-Kruskalův index dále získáme podle vzorce

$$GK = \frac{N_c - N_d}{N_c + N_d}. \quad (6.6)$$

Vysoké hodnoty GK indexu znamenají dobře vytvořené shluky a počet shluků, který maximalizuje hodnoty indexu, dává optimální počet shluků.

### 6.3.8 Meansim (MSA)

Nejedná se přímo o validační metodu pro určení správného počtu shluků. Její výsledky nám pomohou pouze vybrat optimální řešení ze shluků již vytvořených nezávisle na asociační matici, která byla použita při shlukové analýze. Tuto metodu můžeme použít například v případě kdy máme datový soubor obsahující data jak o složení společenstva, tak o parametrech prostředí. Objekty (lokality) zde shlukujeme na základě proměnných prostředí a následně nás zajímá, jak dobře nám tyto shluky oddělují společenstva na vzorcích.

Tato metoda hodnotí sílu klasifikace (*Classification strength* – CS). Byla speciálně navržena pro mnoho vzorků a relativně málo shluků. Klasifikační síla shlukování je stanovena tím, do jaké míry si jsou objekty ve stejném shluku průměrně navzájem podobné oproti podobnosti objektů s objekty z jiných shluků.

Analýza je založena na matici podobnosti mezi vzorky. CS je počítána jako rozdíl mezi průměrem všech podobností uvnitř shluků (W) a průměrem všech podobností mezi shluky (B) podle vzorce:

$$CS = W - B. \quad (6.5)$$

Hodnoty CS se pohybují mezi nulou a jedničkou. Hodnoty blízké jedné indikují dobrou klasifikaci mezi skupinami (tj. uvnitř skupin je vysoká podobnost a mezi skupinami nízká).

## 6.4 Shluková analýza: shrnutí

**Vstupem shlukové analýzy je:**

- matice podobnosti nebo vzdáleností objektů nebo
- tabulka objektů charakterizovaných několika proměnnými.

**Výstupem shlukové analýzy je:**

- strom (dendrogram) – při hierarchické shlukové analýze;
- zařazení objektů do předem definovaného počtu shluků – při nehierarchické shlukové analýze.

**Při použití shlukové analýzy je nutno pamatovat na níže uvedené problémy:**

- hierarchické aglomerativní shlukování není efektivní pro velmi velká data;
- při hierarchické aglomerativní analýze je výsledek silně ovlivněn výběrem indexu podobnosti, resp. metrikou vzdálenosti a shlukovacím algoritmem;
- při hierarchické divizivní analýze TWINSpan je výsledek silně ovlivněn nastavením hraničních hodnot;
- při nehierarchickém shlukování je nutné určit počet předpokládaných shluků předem.

## Seznam použité literatury

- Davies, D. L., Bouldin, D. W. 1979. A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell.* 1 (4): 224-227.
- Digby, P.G.N., Kempton, R.A. 1987. *Multivariate analysis of ecological communities.* Chapman and Hall, London – New York.
- Dunn, J. C., 1974. Well separated clusters and optimal fuzzy partitions. *J.Cybern.* 4: 95-104.
- Gnanadesikan, R. 1977. *Methods for statistical data analysis of multivariate observations.* John Wiley & Sons, New York – London – Sydney – Toronto.
- Goodman, L., Kruskal, W. 1954. Measures of associations for cross-validations. *J. Am. Stat. Assoc.* 49: 732-764.
- Hebák, P., Hustopecký, J. 1987. *Vícerozměrné statistické metody s aplikacemi.* SNTL, Alfa, Praha.
- Hebák, P., Hustopecký, J., Jarošová, E., Pecáková, I. 2007. *Vícerozměrné statistické metody (1). 2. přepracované vydání,* Informatorium, Praha, ISBN 9788073330569
- Hubert, L., Schultz, J. 1976. Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie.* 29: 190-241.
- Jongman, R.H., ter Braak, C.J.F., van Tongeren, O.F.R. 1987. *Data analysis in community and landscape ecology.* Pudoc, Wageningen.
- Latka, F. 1981. *Minilexikon matematiky.* Alfa, Bratislava, 158pp.
- Legendre, P., Legendre, L. 1998, *Numerical Ecology*, 2nd Engl. Ed., Elsevier, Amsterdam, ISBN 0444892494.
- Lepš, J., Šmilauer, P. 1994. *Metody mnohorozměrné statistiky v analýze ekologických dat. Studijní materiál ke kursu.* Biologická fakulta Jihočeské university, České Budějovice.
- Lepš, J., Šmilauer, P. 2000. *Mnohorozměrná analýza ekologických dat.* Biologická fakulta Jihočeské univerzity v Českých Budějovicích. České Budějovice.
- Lepš, J., Šmilauer, P. 2003. *Multivariate Analysis of Ecological Data using CANOCO.* Cambridge University Press. ISBN 0 521 81409 X hardback, ISBN 0 521 89108 6 paperback.
- Marhold, K., Suda, J. 2002. *Statistické zpracování mnohorozměrných dat v taxonomii (Fenologické metody).* Učební texty Univerzity Karlovy v Praze. Univerzita Karlova v Praze, Nakladatelství Karolinum. 160pp. ISBN 80-246-0438-8.
- Palmer, M.W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology* 74: 2215-2230.
- Pauwels, E. J., Frederix, G. 1999 Finding salient regions in images: nonparametric clustering for image segmentation and grouping. *Computer Vision and Image Understanding*, 75: 73-85.
- Podani, J. 2001. *SYN-TAX 2000. Computer program for data analysis in ecology and systematics. User's Manual.* Scientia Publishing, Budapest.
- Rousseeuw, P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* 20: 53-65.
- StatSoft, Inc. 2005. *STATISTICA (data analysis software system), version 7.1.* www.statsoft.com.
- ter Braak, C. J. F., Šmilauer, P. 1998. *CANOCO References Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4).* Ithaca, NY, USA: Microcomputer Power.

- Tvrđík, J. 2003. Analýza vícerozměrná dat. Ostravská univerzita, Přírodovědecká fakulta, Ostrava [Online pdf] 18.10.2010 přístupný na: [http://prf.osu.cz/doktorske\\_studium/dokumenty/Multivariable\\_Data\\_Analysis.pdf](http://prf.osu.cz/doktorske_studium/dokumenty/Multivariable_Data_Analysis.pdf)
- Urban, D.L. (2000) Multivariate analysis in ecology. Principal Components Analysis. [http://www.env.duke.edu/lel/env358/mv\\_pca.pdf](http://www.env.duke.edu/lel/env358/mv_pca.pdf)
- Urban, D.L. 2000. Multivariate analysis in ecology. Nonhierarchical agglomeration. [http://www.env.duke.edu/lel/env358/mv\\_kmeans.pdf](http://www.env.duke.edu/lel/env358/mv_kmeans.pdf)
- van der Lann, M. J., Pollard, K. S., Bryan, J. 2002. A New Partitioning Around Medoids Algorithm. *Journal of Statistical Computation and Simulation* 73: 575–584.
- Van Sickle, J. 1997. Using Mean Similarity Dendrograms to Evaluate Classifications, *Journal of Agricultural, Biological and Environmental Statistics* 2: 370 – 388.
- Wolda, H. 1981. Similarity Indices, Sample Size and Diversity. *Oecologia (Berlin)* 50: 296-302.
- Zvára, K. 2001. Biostatistika. Učební texty Univerzity Karlovy v Praze. Univerzita Karlova v Praze – Nakladatelství Karolinum. 212 pp. ISBN 80-7184-773-9.