INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Vícerozměrné statistické metody v biologii

Danka Haruštiaková, Jiří Jarkovský, Simona Littnerová, Ladislav Dušek

Březen 2011

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.

Předmluva

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, medicíně, 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 potřebnou 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 matematické biologie i dalších přírodovědných oborů.

Česká a ani anglická terminologie používána v dostupné literatuře není zcela stabilizovaná a často se stává, že totožné metody jsou v různých učebnicích a statistických programech uváděny pod různými názvy. Z tohoto důvodu uvádíme jak anglické názvy metod, tak i české alternativní názvy.

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, březen 2011

Danka Haruštiaková
Jiří Jarkovský
Simona Littnerová
Ladislav Dušek

ISBN: XXX-XX-XXXX-XXX-X

Obsah

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 popsaný ve formě datové tabulky nebo jej dokonce nějakým způsobem popsat jinému člověku – nastává zde tedy místo pro speciální typ analýzy, 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 jednolitého 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 tím zjednodušení jejich struktury;
- jednoduchá vizualizace dat, kdy se v jediném grafu skrývá informace např. z 20 proměnných;
- umožnění a/nebo zjednodušení interpretace dat na základě jejich zjednodušení a vizualizace.

Ačkoliv je v případě vícerozměrných analýz používána celá řada matematických postupů, jedno mají všechny tyto analýzy společné – hledání souvislostí a jejich výklad.

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 a medicínských oborech, ale také v technice, kybernetice, 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 Statistický 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.

Software **R** (*The R Project for Statistical Computing*) 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 softwaru nabízí např. hodnocení výsledků shlukování ve formě tzv. Silhouette plot.

SPSS (*Statistical Package for the Social Sciences*) je běžný komerční software s rozšířenými možnostmi zpracování dat a jejich analýzy. Vícerozměrné metody jsou součásti 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. Obsahuje metody hierarchického shlukování, nehierarchického shlukování a ordinace. Výhodou tohoto softwarového nástroje jsou široké možnosti uživatelského přizpůsobení analýz, které nejsou v běžných komerčních softwarech 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 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 (*PAlaeontological STatistics*) 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 ze 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 odlehlých hodnot apod.

Klasickým příkladem je provázanost analýzy hlavních komponent s parametrickou kovariancí nebo korelací, kdy přítomnost odlehlé 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 odlehlé hodnoty v datech a nijak nepřispívá k pochopení zdrojů

variability dat. Naproti tomu některé vícerozměrné metody lze považovat za velmi robustní a analogické k 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 2-1) 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 2-1 Ukázka datové tabulky

Vzorek	Půdní typ	*Quercus* (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 přiřazeny 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-Blanquetové stupnici, která pokryvnost rostlinných taxonů hodnotí na 7stupňové 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*): Na intervalové stupnici je kromě vlastností předchozích dvou stupnic možné také sčítání a odečítání. Na rozdíl od pořadové stupnice zde lze 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. teplota ve stupních Kelvina, hodnoty délky, plochy nebo objemu).

Z hlediska statistického zpracování dat můžeme proměnné rozdělit na:

- ***kvalitativní*** (*qualitative*)
  - binární (*binary*, dvoustavové, alternativní) – nabývají pouze dvou hodnot, většinou je kódujeme 0 a 1 (např. přítomnost nebo nepřítomnost určitého živočišného druhu)
  - vícestavové (*multistate*) – nabývají vícero hodnot, např. výše uvedené typy půd
- ***semikvantitativní*** (*semiquantitative*) – do této skupiny patří proměnné, jejichž hodnoty jsou vyjádřeny pomocí pořadové stupnice, která nemá konstantní rozdíly mezi sousedícími hodnotami (např. Braun-Blanquetova stupnice pokryvnosti)
- ***kvantitativní*** (*quantitative*) – proměnné lze vyjádřit měřitelnou stupnicí, na níž jsou konstantní rozdíly mezi jednotkami
  - nespojité, diskrétní (*discontinuous*, *discrete*) – proměnné, které nabývají pouze určité reálné hodnoty (např. počet květů)
  - spojité, 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. indikátorových proměnných (*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 první 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 máme několik možností:

- Objekty, ve kterých hodnoty chybí, můžeme vypustit. Toto řešení je vhodné tehdy, když jsou chybějící data pouze v několika málo objektech.
- Proměnné, u kterých hodnoty chybí, můžeme vypustit, pokud jich není mnoho a nejde o klíčové proměnné.
- Chybějící hodnoty můžeme doplnit, a to různými metodami:
  - doplnění průměru z hodnot, které jsou k dispozici;
  - 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

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

Tato transformace se často používá ze čtyř různých důvodů:

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

Odmocninová transformace

$$y_{ij} = \sqrt{x_{ij}} \tag{2.2}$$

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

$$y_{ij} = \sqrt{x_{ij} + 0{,}5} \tag{2.3}$$

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

$$y_{ij} = \arcsin \sqrt{x_{ij}} \tag{2.4}$$

- Používá se v kombinaci s odmocninovou transformací a předpokládá, že data jsou měřena v intervalu $\langle 0;1 \rangle$.
- Používá se na úpravu relativních hodnot vyjádřených v intervalu $\langle 0;1 \rangle$ (např. vegetační pokryvnosti druhů).

Exponenciální transformace

$$y_{ij} = a^{x_{ij}} \tag{2.5}$$

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 a intervaly tříd nemusí být stejné.

Typickou transformací na ordinální škálu je použití Braun-Blanquetovy stupnice při kvantifikaci pokryvnosti vegetace (Tabulka 2-2).

Tabulka 2-2 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.

$$y_{ij} = 0 \quad \text{když } x_{ij} = 0 \qquad y_{ij} = 1 \quad \text{když} \quad x_{ij} > 0 \tag{2.6}$$

Transformací na ordinální škálu se vždy ztrácí 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 běžné např. 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; přestává tedy záležet na skutečném rozměru příslušné proměnné. K nejčastějším úpravám patří centrování a standardizace směrodatnou odchylkou.

Standardizace rozpětím

$$y_{ij} = \frac{x_{ij} - \min_j \{x_{ij}\}}{\max_j \{x_{ij}\} - \min_j \{x_{ij}\}} \tag{2.7}$$

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.

$$y_{ij} = x_{ij} - \bar{x}_j \tag{2.8}$$

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.

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

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

$$y_{ij} = \frac{x_{ij}}{\sum_i x_{ij}} \tag{2.10}$$

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 výrazně neliší. Tato standardizace bývá používána v případech, kdy se v seznamu druhů vyskytují různé trofické úrovně, protože vyšší trofické úrovně jsou méně zastoupeny (a proto může vyhovovat jejich nadvážení).

$$y_{ij} = \frac{x_{ij}}{\sum_j x_{ij}} \tag{2.11}$$

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 těch případech, kdy jsou veliké rozdíly ve vyrovnanosti vzorků.

$$y_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}} \tag{2.12}$$

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.

$$y_{ij} = \frac{x_{ij}}{\max_j \{x_{ij}\}} \tag{2.13}$$

Standardizace na jednotkovou délku vektoru řádku

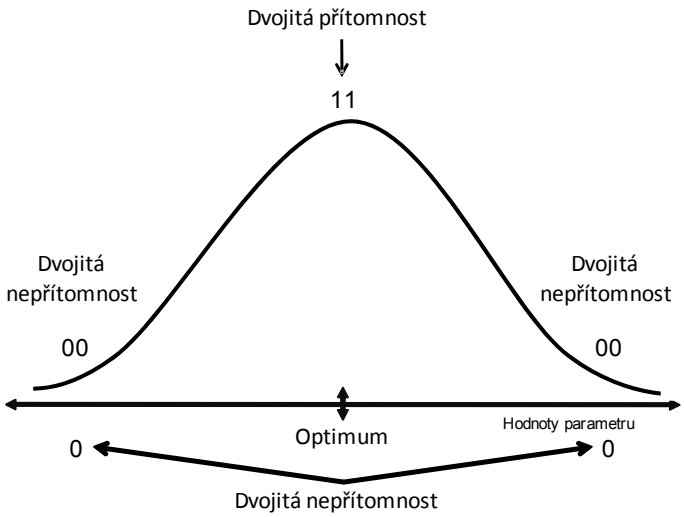
Podělením hodnot proměnných u objektu odmocninou sumy čtverců hodnot se všechny vektory objektů zobrazí na jednotkové kružnici prostoru tvořeného proměnnými (v ekologických studiích jde o druhy). Euklidovské vzdálenosti se touto standardizací redukují na tětivové vzdálenosti (*chord distance*).

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^{\,2}}} \tag{2.14}$$

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

Problém dvou nul je v ekologických studiích častým problémem. Vyskytuje se u proměnných, kde nula znamená nepřítomnost a ne hodnotu stupnice. Typickým příkladem jsou početnosti (abundance) druhů. Druhy jsou známy unimodální distribucí niky podél environmentálního 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 vlastnosti nik obou lokalit jsou buď „vyšší“ než na optimální nice, anebo má jedna z nich „vyšší“ a druhá „nižší“ vlastnosti, než jsou 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 přírodních a mnohdy i technických oborech 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

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

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

Vícerozměrné normální rozložení má tyto vlastnosti:

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

Pro jednorozměrné normální rozložení má předešlý vzorec tvar

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

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

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

$$C^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \tag{3.3}$$

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

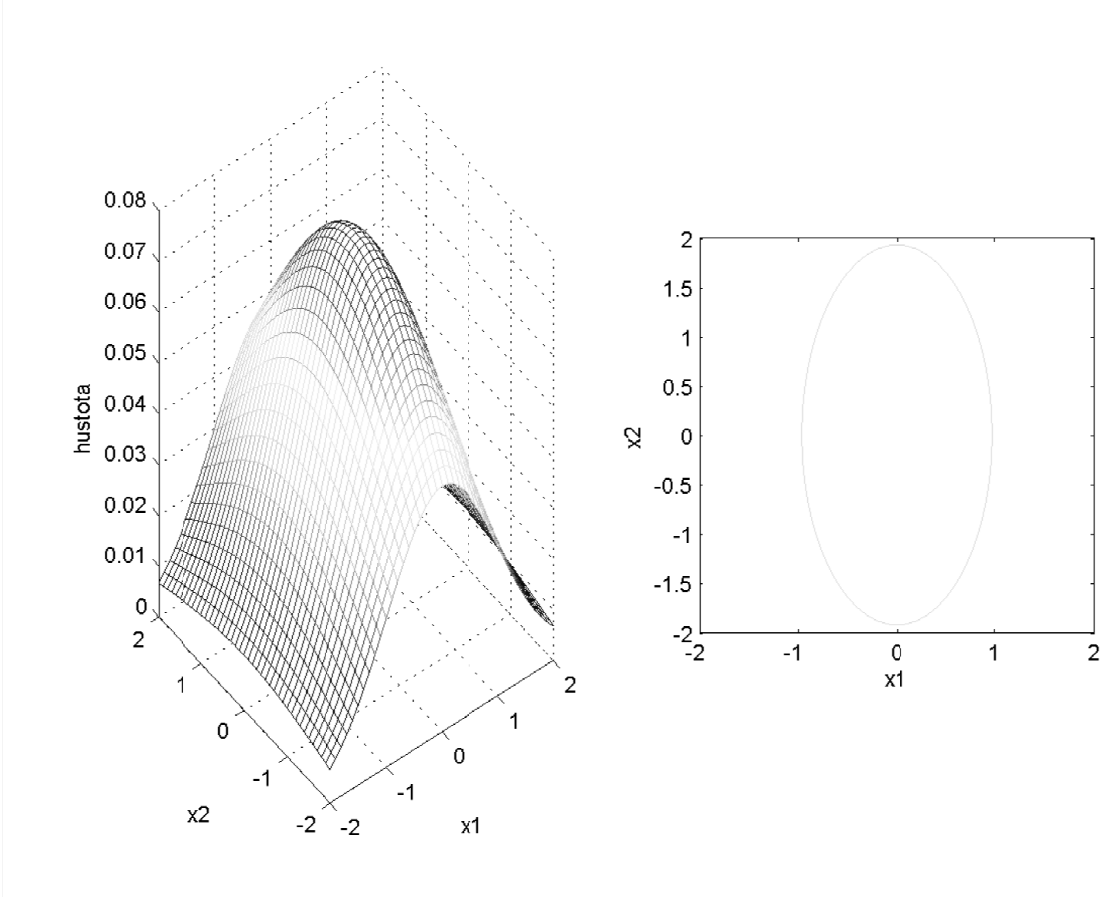
$$C^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2(p)$$

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 μ_1 a μ_2, s rozptyly σ_1^2, σ_2^2 a s kovariancí σ_{12}, pak je možné determinant kovarianční matice **Σ** možné vyjádřit jako $\sigma_1^2\sigma_2^2(1-\rho^2)$, kde ρ je korelační koeficient definovaný jako $\frac{\sigma_{12}}{\sigma_1\sigma_1}$. 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 $\sigma_1^2(1-\rho^2)$

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

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).

$$f(x_1, x_2) = konst.$$



Obrázek 3.1 Hustota dvourozměrného normálního rozdělení a elipsy konstantní hustoty, $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = 0$.

3.1 Vícerozměrné charakteristiky rozdělení

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

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(\mathrm{X}_1) \\ \mathrm{E}(\mathrm{X}_2) \\ \vdots \\ \mathrm{E}(\mathrm{X}_p) \end{bmatrix}$$

a kovariační matice

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

kde σ_{ij} je kovariance dvou náhodných veličin, tj.

$$\sigma_{ij} = \mathrm{cov}(\mathrm{X}_i, \mathrm{X}_j) = \mathrm{E}(\mathrm{X}_i - \mathrm{E}(\mathrm{X}_i))(\mathrm{X}_j - \mathrm{E}(\mathrm{X}_j)) \tag{3.4}$$

a $\sigma_{ii} = \sigma_i^2$ je rozptyl $\mathrm{var}(\mathrm{X}_i)$. 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 vzdálenosti 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. tří- a vícerozměrný prostor). Tento termín se používá při shlukové analýze.

3.2 Wishartovo rozdělení

Uvažujeme ν nezávislých náhodných vektorů $\mathbf{u_i}$, i = 1, 2, … ν, vesměs s rozdělením $\mathrm{N_p}(\mathbf{o_p}, \boldsymbol{\Sigma})$. Potom náhodná matice $\mathbf{A} = \sum_{\mathbf{i=1}}^{\nu} \mathbf{u_i}\mathbf{u_i}^{\mathrm{T}}$ má *p*-rozměrné Wishartovo rozdělení se ν stupni volnosti, tedy $\mathbf{A} \sim \mathrm{W_p}(\nu, \boldsymbol{\Sigma})$.

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.

$$\left.\begin{array}{c}\mathbf{A}=\mathbf{A_1}+\mathbf{A_2}+...+\mathbf{A}_H \\ \mathbf{A}_h \sim W_p(\nu_h,\boldsymbol{\Sigma}), h=1,2,...,H\end{array}\right\}\longrightarrow \mathbf{A}_h \sim W_p\left(\sum_{h=1}^{H}\nu_h,\boldsymbol{\Sigma}\right) \tag{3.5}$$

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 $W_p(\nu,\boldsymbol{\Sigma})$ a na A nezávislý *p*-položkový vektor a s rozdělením $N_p\left(\mathbf{o}_p,\boldsymbol{\Sigma}/c\right)$. Potom kvadratická forma

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

má Hotellingovo rozdělení T^2 (p, ν – p+1).

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

$$X \sim N(\mu,\sigma^2)\longrightarrow \frac{\bar{x}-\mu}{\sqrt{\frac{s^2(x)}{n}}} \sim t(n-1) \tag{3.6}$$

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 μ, 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

$$x \sim N_p(\mu,\Sigma)\longrightarrow n(x-\mu)^T S^{-1} \sim T^2(p,n-p). \tag{3.7}$$

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

$$T^2 = \frac{n_1 n_2}{n}\left(\overline{x_1}-\overline{x_2}-\delta\right)^T \bar{S}^{-1}\left(\overline{x_1}-\overline{x_2}-\delta\right), \tag{3.8}$$

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

4 Asociační koeficienty

V předchozí části jsme 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 nabývají tehdy, když jsou dva objekty zcela odlišné. U vzdáleností je tomu obráceně. V případě potřeby lze podobnost převést na vzdálenost.

4.1 Asociační koeficienty mezi proměnnými

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

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} , \qquad (4.1)$$

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 bližší jedné, tím je silnější přímá lineární závislost mezi proměnnými *x* a *y*. Čím je bližší 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 jsou hodnoty proměnných měřeny na ordinální škále), můžeme použít **Spearmanův korelační koeficient r^s**.

$$r_{xy}{}^{s} = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n}(R_i - Q_i)^2 , \qquad (4.2)$$

kde $R_1, \ldots R_n$ jsou pořadí prvků proměnné *x* a podobně $Q_1, \ldots Q_n$ 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. Kovariance může nabývat hodnot z intervalu $(-\infty, \infty)$.

$$s_{xy} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{4.3}$$

4.2 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 kriteria:

- 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).

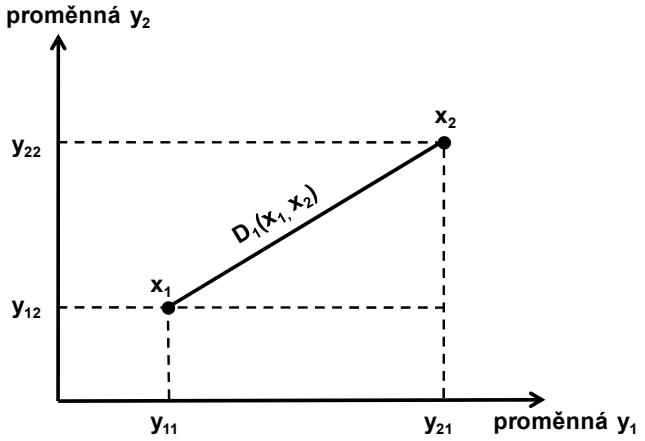
Mnohé koeficienty podobnosti (*S*) lze převést na vzdálenosti pomocí transformace $D = 1 - S$ nebo $D = \sqrt{1 - S}$ a výsledkem jsou často semimetrické nebo nemetrické koeficienty vzdáleností. Následující text shrnuje základní metriky vzdálenosti.

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.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \tag{4.4}$$

Obrázek 4.1 znázorňuje Euklidovskou vzdálenost dvou objektů v prostoru dvou proměnných.



Obrázek 4.1 Výpočet Euklidovské vzdálenosti mezi objekty x1 a x2.

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

$$D_1(x_1, x_2)^2 = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2 \tag{4.5}$$

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 navrhnuto 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.

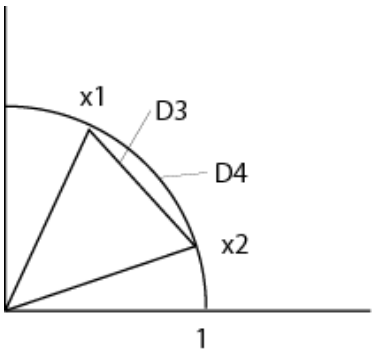
$$D_2^2(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} (y_{1j} - y_{2j})^2 \tag{4.6}$$

nebo

$$D_2(x_1, x_2) = \sqrt{D_2^2} \tag{4.7}$$

Tětivová vzdálenost (chord distance)

Tětivová vzdálenost je Euklidovská vzdálenost po normalizaci. Její hodnoty se pohybují od nuly 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 4.2), nebo je možný přímý výpočet, který již zahrnuje standardizaci. Odstraňuje problém dvou nul a vliv rozdílného rozpětí proměnných v objektech při výpočtu Euklidovské vzdálenosti.



$$D_3(x_1, x_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)} \qquad (4.8)$$

Obrázek 4.2 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 4.2).

$$D_4(x_1, x_2) = \arccos\left[1 - \frac{D_3^2(x_1, x_2)}{2}\right] \qquad (4.9)$$

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:

$$D_5^2(w_1, w_2) = \overline{d_{12}} V^{-1} \overline{d_{12}'}, \qquad (4.10)$$

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.

$$V = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)V_1 + (n_2 - 1)V_2\right] \qquad (4.11)$$

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 λ =1 je známá jako Manhattanská vzdálenost. Jde vlastně o součet rozdílů jednotlivých proměnných, které objekty popisují.

$$D_6(x_1, x_2) = \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \tag{4.12}$$

Průměrná Manhattanská vzdálenost (mean character difference)

Podobně, jako 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 se hodnota nezvyšuje s rostoucím počtem proměnných.

$$D_7(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \tag{4.13}$$

Minkowského metrika (Minkowski´s metric)

Je obecnou formou výpočtu vzdálenosti. Zahrnuje v sobě několik metrik 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 je koeficient umocňování a odmocňování zadáván zvlášť.

$$D_8(x_1, x_2) = \left[\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|^{\lambda} \right]^{\frac{1}{\lambda}} \tag{4.14}$$

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

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.

$$D_9(x_1, x_2) = \sqrt{\sum_{j=1}^{p} w_j^2 (y_{1j} - y_{2j})^2} \tag{4.15}$$

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 na svůj podíl ve společenstvu (v tomto případě společenstvo druhů tvoří součet hodnot všech proměnných ve vzorku – objektu). Následující výpočet je opět obdobou Manhattanské vzdálenosti. Doplňkem asociačního indexu je následující vzdálenost:

$$D_{10}(x_1, x_2) = \frac{1}{2}\sum_{j=1}^{p}\left|\frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}}\right| \tag{4.16}$$

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

Canberra metric

Varianta Manhattanské 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.

$$D_{11}(x_1, x_2) = \sum_{j=1}^{p}\left(\frac{\left|y_{1j} - y_{2j}\right|}{\left(y_{1j} + y_{2j}\right)}\right) \tag{4.17}$$

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*).

$$D_{12}(x_1, x_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}}\right)^2} \tag{4.18}$$

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 $\overline{y_{ij}}$ a rozptylem proměnných ve skupinách s_{ij}^2. Tento koeficient byl vyvinut pro potřeby antropologických studií.

$$D_{13}(w_1, w_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\frac{\left(\overline{y}_{1j} - \overline{y}_{2j}\right)^2}{\left(\frac{s_{1j}^2}{n_1}\right) + \left(\frac{s_{2j}^2}{n_2}\right)}\right)} - \frac{2}{p} \tag{4.19}$$

χ^2 metrika

První ze skupiny metrik založených na χ^2 využívaném 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 frekvenci). 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

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}} \right)^2} , \tag{4.20}$$

která je dále vážena součty jednotlivých druhů

$$D_{14}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{\sum_{i=1}^{n} y_{ij}} \left(\frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}} \right)^2} . \tag{4.21}$$

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

χ^2 vzdálenost

Výpočet je podobný χ^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). χ^2 vzdálenost je využívána také při výpočtu vztahů řádků a sloupců kontingenční tabulky.

$$D_{15}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{\frac{\sum_{i=1}^{n} y_{ij}}{\sum_{j=1}^{p} \sum_{i=1}^{n} y_{ij}}} \left(\frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}} \right)^2} \tag{4.22}$$

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

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

- **Symetrické koeficienty** podobnosti 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 4-1).

Tabulka 4-1 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 4-2) 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 a u *j*-tého objektu 1
d počet proměnných, které nabývají pro oba objekty hodnoty 0

Platí a + b + c + d = p.

Tabulka 4-2 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 (Tabulka 4-2) jsou tyto početnosti: a = 2, b = 2, c = 1, d = 1.

4.3.1 Symetrické binární koeficienty

Základem všech indexů podobnosti pro kvalitativní binární data je, že dva objekty jsou si 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.

$$S_1(x_1, x_2) = \frac{a+d}{p} \tag{4.23}$$

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**).

$$S_2(x_1, x_2) = \frac{a+d}{a+2b+2c+d} \tag{4.24}$$

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

$$S_3(x_1, x_2) = \frac{2a+2d}{2a+b+c+2d} \tag{4.25}$$

tento koeficient dává dvakrát větší váhu shodným proměnným než rozdílným;

$$S_4(x_1, x_2) = \frac{a+d}{b+c} \tag{4.26}$$

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

$$S_5(x_1, x_2) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right] \tag{4.27}$$

porovnává shodné deskriptory se součty okrajů tabulky;

$$S_6(x_1, x_2) = \frac{a}{\sqrt{(a+b)(a+c)}} \frac{d}{\sqrt{(b+d)(c+d)}} \tag{4.28}$$

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

4.3.2 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)

$$S_7(x_1, x_2) = \frac{a}{a+b+c} \tag{4.29}$$

dává všem členům stejnou váhu.

Sørensenův koeficient (Sørensen´s coefficient)

Sorensenův koeficient je variantou Jaccardova koeficientu, dává ovšem dvojnásobnou váhu dvojitým výskytům. 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.

$$S_8(x_1, x_2) = \frac{2a}{2a+b+c} \tag{4.30}$$

$$S_9(x_1, x_2) = \frac{3a}{3a+b+c} \tag{4.31}$$

Řada dalších koeficientů dává různou váhu jednotlivým kombinacím proměnných. Jako doplněk koeficientu S_2 byl navrhnut koeficient, který dává dvojnásobnou váhu rozdílům ve jmenovateli (**Sokal a Sneath**).

$$S_{10}(x_1, x_2) = \frac{a}{a + 2b + 2c} \tag{4.32}$$

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**)

$$S_{11}(x_1, x_2) = \frac{a}{p} \tag{4.33}$$

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

$$S_{12}(x_1, x_2) = \frac{a}{b + c} \tag{4.34}$$

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

$$S_{13}(x_1, x_2) = \frac{1}{2}\left[\frac{a}{a + b} + \frac{a}{a + c}\right], \tag{4.35}$$

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

Obdobou symetrického koeficientu S_6 tak, aby byl odstraněn 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**).

$$S_{14}(x_1, x_2) = \frac{a}{\sqrt{(a + b)(a + c)}} \tag{4.36}$$

4.3.3 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, 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.

$$S_1\,(x_1, x_2) = \frac{shoda}{p}. \tag{4.37}$$

Např. je-li dvojice objektů popsána následujícími deseti multistavovými proměnnými (Tabulka 4-3), potom hodnota koeficientu S_1, vypočítaná pro 10 multistavových proměnných bude $S_1(x_1,x_2)$ = 4 shody/10 proměnných = 0,4.

Tabulka 4-3 Ukázka výpočtu jednoduchého srovnávacího koeficientu pro multistavové proměnné.

	proměnné										Σ
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).

$$S_{15}(x_1,x_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j} \tag{4.38}$$

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 $\left|y_{1j} - y_{2j}\right|$, 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.

$$s_{12j} = 1 - \left[\frac{\left|y_{1j} - y_{2j}\right|}{R_j}\right] \tag{4.39}$$

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

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \tag{4.40}$$

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 nuly 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 4-4).

Tabulka 4-4 Ukázka výpočtu Gowerova koeficientu.

	Proměnné j								Σ
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
$\lvert y_{1j} - y_{2j}\rvert / R_j$	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}:

$$S_{16}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12j} s'_{12j}}{\sum_{j=1}^{p} w_{12j}} \tag{4.41}$$

a stejně jako u S_{15} mohou být parametry w_j (mezi 0 a 1) opět využity jako váhy místo toho, aby pouze hrály roli Kroneckerova 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:

$$s'_{12j} = f\left(d_{12j}, k_j\right) = \frac{2(k+1-d)}{2k+2+dk} \text{ pro } d \le k$$
$$s'_{12j} = f\left(d_{12j}, k_j\right) = 0 \text{ pro } d > k, \tag{4.42}$$

kde d je vzdálenost mezi dvěma stavy objektů x_1 a x_2 pro proměnnou j, tj. stejně jako v Gowerově koeficientu $\left|y_{1j} - y_{2j}\right|$ 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.

4.3.4 Asymetrické kvantitativní koeficienty

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

Jaccardův koeficient

$$S_7(x_1, x_2) = \frac{\text{shoda}}{p-d}, \tag{4.43}$$

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 takovéhoto 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.

$$S_{17}(x_1, x_2) = \frac{W}{(A+B)/2} = \frac{2W}{A+B} \tag{4.44}$$

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 4-5).

Tabulka 4-5 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

$$S_{17}(x_1, x_2) = \frac{2.15}{20 + 22} = 0.714$$

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 následně je vypočítán průměr ze dvou získaných hodnot.

$$S_{18}(x_1, x_2) = \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right) \tag{4.45}$$

Pro příklad z Tabulka 4-5 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:

$$S_{19} = \frac{2\sum n_{1i} n_{2i}}{(d_1 + d_2) N_1 N_2}, \tag{4.46}$$

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}.

Gower 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 $\left|y_{1j} - y_{2j}\right|$ a jsou proto lineárně závislé na měřítku, měl by být tento koeficient používán na normalizovaná data.

$$S_{20}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}, \tag{4.47}$$

kde

$$s_{12j} = 1 - \left[\frac{|y_{1j} - y_{2j}|}{R_j} \right] \tag{4.48}$$

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, biologický objem různých druhů, nebo kompenzovalo účinnost odběru daného druhu.

Další obecný koeficient podobnosti vychází z koeficientu S_{16} a 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.

$$S_{21}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12j} s'_{12j}}{\sum_{j=1}^{p} w_{12j}} \tag{4.49}$$

kde

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

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

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

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

χ2 podobnost (χ2 similarity)

Je posledním kvantitativním koeficientem, jenž eliminuje problém dvou nul (*double zero problem*). Jedná se o doplněk χ^2 metriky (D_{14}).

$$S_{22}(x_1, x_2) = 1 - D_{14}(x_1, x_2) \tag{4.50}$$

5 Shluková analýza

Jednou z možností, jak využít informace obsažené ve vícerozměrných pozorováních, je roztřídění objektů do několika poměrně homogenních skupin – shluků tak, aby si objekty patřící do stejné skupiny 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 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 je následně možné odhalit strukturu studované množiny objektů a jednotlivé objekty klasifikovat. Pak již zbývá pouze najít vhodnou interpretaci pro popsaný 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 **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 disjunktními shluky, tj. jeden objekt patří pouze jednomu shluku. Cílem je dosáhnout toho, aby si objekty uvnitř shluku byly co nejvíce podobné a od objektů z ostatních shluků se co nejvíce lišily.

Shluková analýza pracuje s asociační maticí podobností, resp. vzdáleností objektů. Problematice 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.

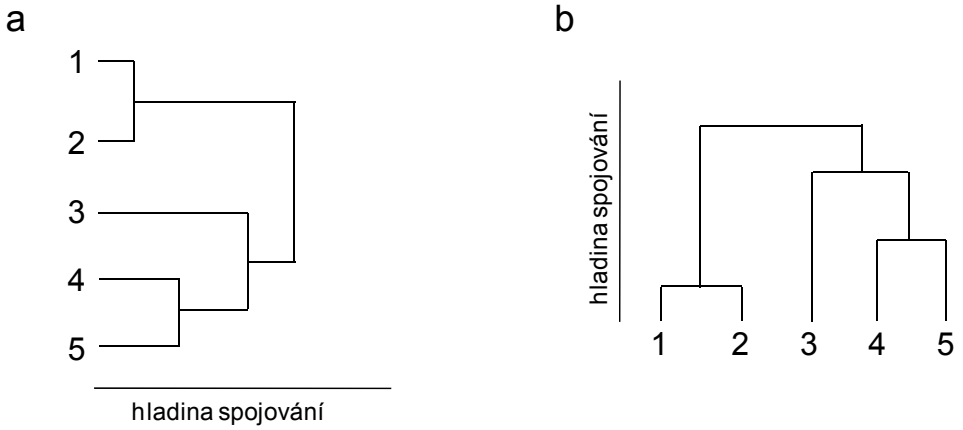
5.1 Hierarchické shlukování

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 5.1). 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 5.1a), ale i vertikálně (Obrázek 5.1b).



Obrázek 5.1 Ukázka dendrogramu (stromu) pěti objektů. Strom lze zobrazit horizontálně (a) i vertikálně (b).

5.1.1 Hierarchické aglomerativní shlukování

Aglomerativní 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, přičemž 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.

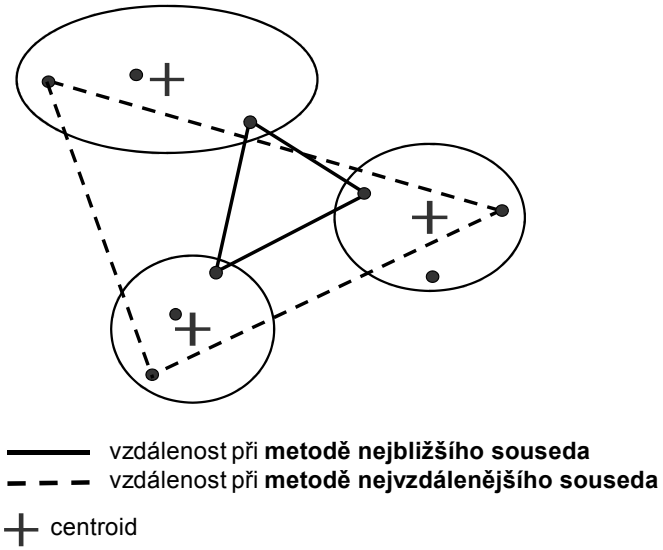
V předchozí kapitole jsou uvedeny různé míry podobnosti a metriky vzdálenosti. Většinou požadujeme, aby podobnost nabývala hodnot od nuly pro maximální rozdílnost po jedničku pro totožnost. Často se však z praktických důvodů používají různé míry vzdálenosti, tentýž jev je

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 postupem se můžeme setkat také pod názvem aglomerativní metoda, aglomerativní postup, nebo shlukovací algoritmus.

Vzdálenost mezi shluky (aglomerativní metody)

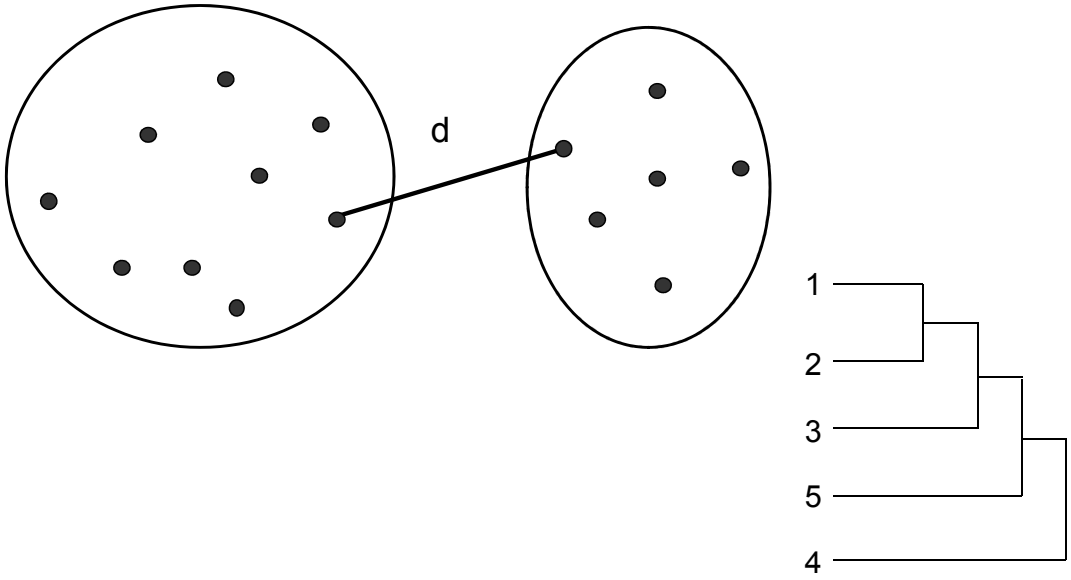
Všechny aglomerativní metody jsou založeny na shlukování jednotlivých objektů nebo shluků do větších skupin. Skupiny, které jsou si nejvíc podobné, jsou sloučeny. Definice vzdálenosti mezi shluky se u jednotlivých metod liší. Metody se navzájem liší chápáním této vzdálenosti (Obrázek 5.2).



Obrázek 5.2 Vnímání vzdálenosti při metodě nejbližšího a nejvzdálenějšího souseda.

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

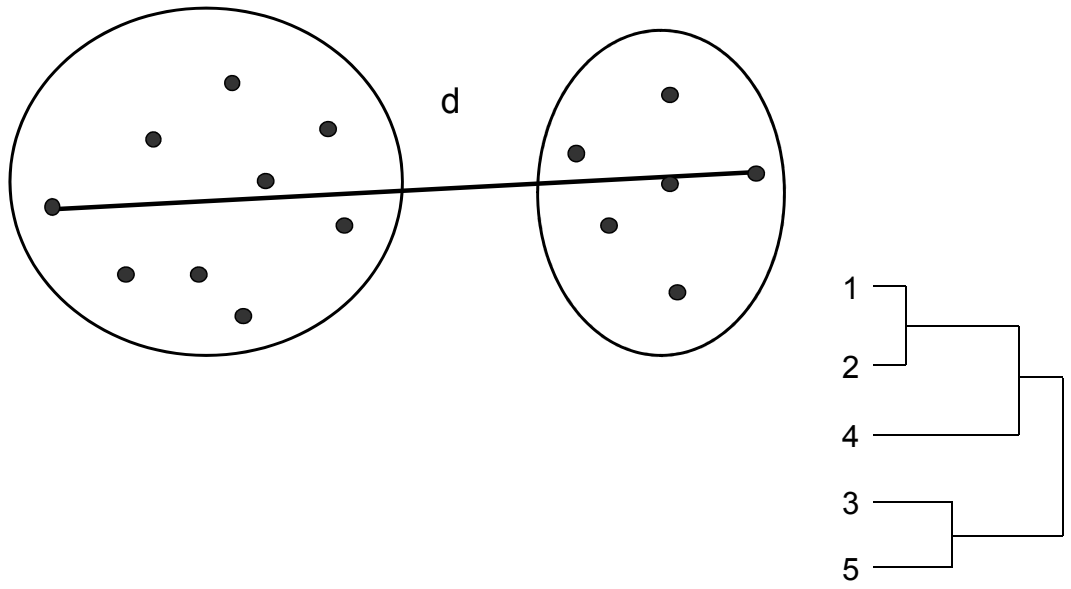
Historicky nejstarší metoda. Vzdálenost mezi dvěma shluky (na počátku analýzy reprezentované jednotlivými objekty) je daná jako minimální vzdálenost mezi všemi možnými zástupci shluků (Obrázek 5.3). To znamená, že ve dvou shlucích, o jejichž spojení uvažujeme, nás zajímají pouze ty dva objekty, které jsou k sobě nejblíže. 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 zhutněným jádrem.



Obrázek 5.3 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. Vzdálenost mezi dvěma shluky je daná maximální vzdáleností mezi všemi možnými zástupci obou shluků (Obrázek 5.4). Tato metoda produkuje shluky, které jsou mezi sebou dobře odděleny. Nežádoucí řetězový efekt zde odpadá, naopak je tu tendence ke tvorbě kompaktních shluků, které nebývají velké.



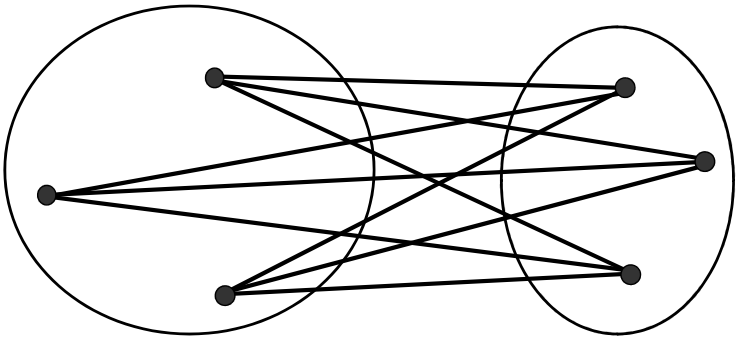
Obrázek 5.4 Vzdálenost u metody nejvzdálenějšího souseda a ukázka dendrogramu vzniklého touto metodou (podle Marhold, Suda 2002).

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 5.5). Metoda vede často k podobným výsledkům jako metoda nejvzdálenějšího souseda.



Obrázek 5.5 Vzdálenost u metody průměrné vazby (podle Marhold, Suda 2002).

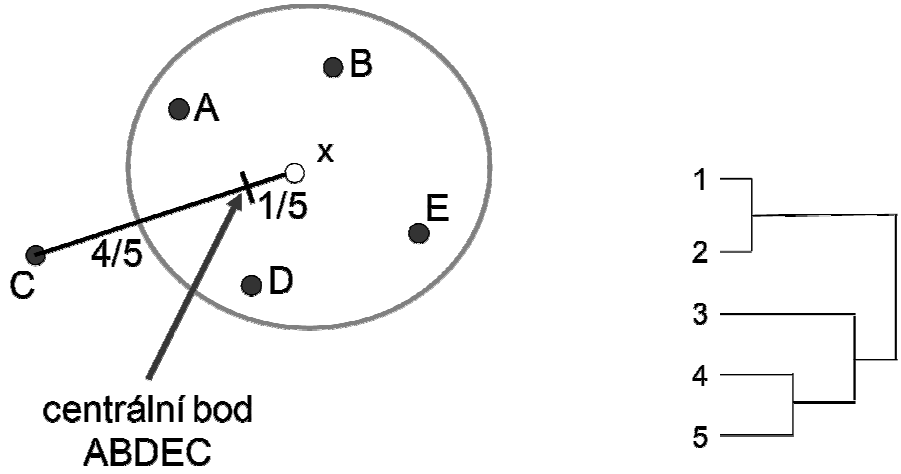
WPGMA (weighted pair-group method using arithmetic averages)

Tato metoda je obdobou předchozí metody ovšem doplněna o vážení shluků jejich velikostí (pod velikostí shluků rozumíme počet jejich objektů), tak aby různě velké shluky měly při výpočtu stejnou váhu. 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 již nevychází 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 nejblíže. Centroid nového shluku je definován podle polohy původních objektů, nikoliv jako centroid vypočtený z centroidů spojených shluků (Obrázek 5.6). To znamená, že nový centroid získáme jako průměr ze všech bodů nového shluku.

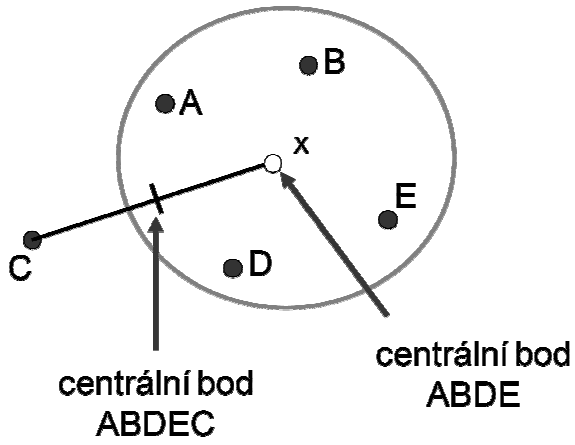
Nevýhodou centroidní metody je skutečnost, že v případě spojování dvou shluků velmi rozdílné velikosti bude centroid (těžiště) nového shluku velmi blízko většího shluku (nebo dokonce uvnitř). Vlastnosti menšího shluku se tak do jisté míry ztrácejí.



Obrázek 5.6 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)

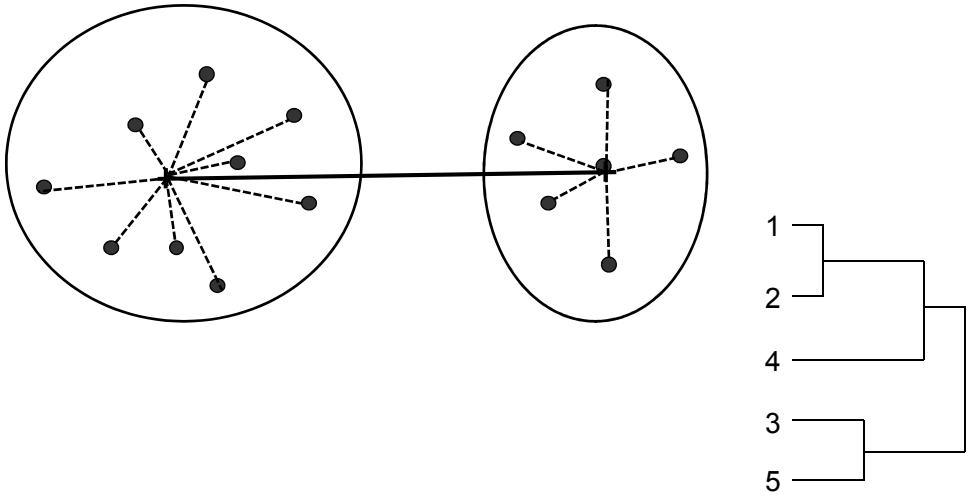
Mediánová metoda odstraňuje problém daný rozdílnou velikostí spojovaných shluků. Analyzované shluky se považují za stejně velké a tedy se stejnou vahou při výpočtu, centroid nového shluku je proto vždy v polovině vzdálenosti mezi centroidy spojovaných shluků. To znamená, že nový centroid získáme jako nevážený průměr původních centroidů (Obrázek 5.7). Jde ovšem o vážený průměr ze všech bodů nového shluku. Tato metoda je preferována tehdy, když očekáváme velké rozdíly ve velikosti shluků.



Obrázek 5.7 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 5.8). 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.



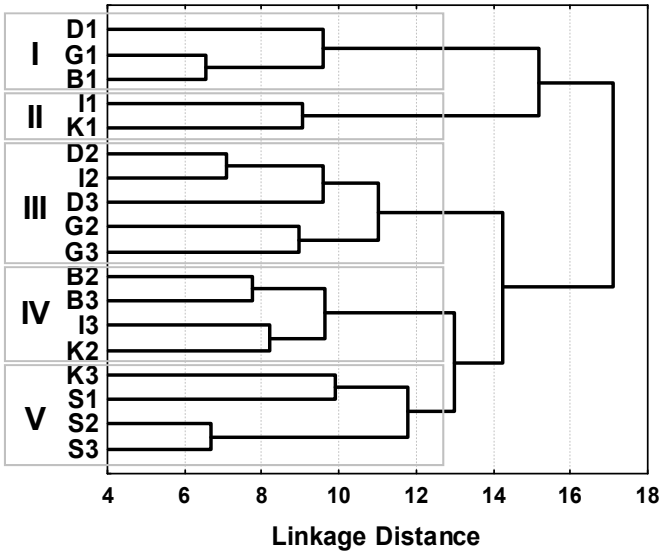
Obrázek 5.8 Vzdálenost u Wardovy metody a ukázka dendrogramu vzniklého Wardovou metodou (podle Meloun, Militký 2004 a Marhold, Suda 2002).

Obecný postup aglomerativního hierarchického shlukování

Aglomerativní hierarchický algoritmus můžeme definovat následovně:

- Vypočteme asociační matici vhodných měr vzdálenosti.
- Proces začneme od rozkladu $S^{(n)}$, tj. od *n* shluků, z nichž každý obsahuje jeden objekt.
- V asociační matici najdeme dva objekty/shluky (*g*-tý a *h*-tý), jejichž vzdálenost je minimální.
- 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.
- Zaznamenáme pořadí cyklu rozkladu I = 1, 2, … n-1, dále identifikaci spojených objektů/shluků a hladinu pro spojení.
- Pokud proces vytváření rozkladů spojením všech objektů do jediného shluku $S^{(1)}$ neskončil, 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, jestli 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: Kráľovská lúka, S: Sporná sihoť. Použita byla všespojná shlukovací metoda (*complete linkage*) a jako míra vzdálenosti Euklidovská vzdálenost.

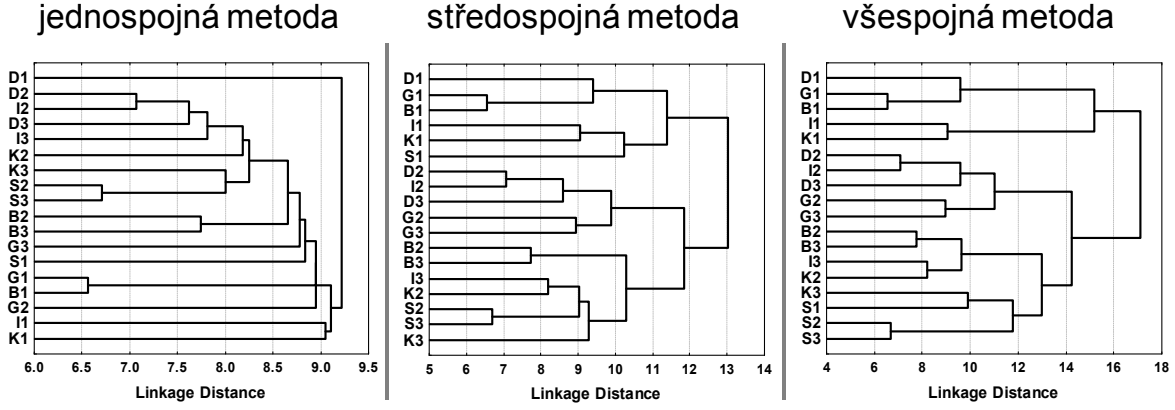


Obrázek 5.9 Ukázka výsledku shlukové analýzy společenstev korýšů (Illyová, Némethová 2005).

Interpretace dendrogramu je následovná (Obrázek 5.9): 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 Kráľovská 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 tvořený lokalitami B2, B3, I3, K2: Bodíky (druhé a třetí období), Istragov (třetí období) a Kráľovská lúka (druhé období). Poslední pátý shluk je tvořen lokality K3, S1, S2, S3: Sporná sihoť (všechna období) a Kráľovská lúka ve třetím období.

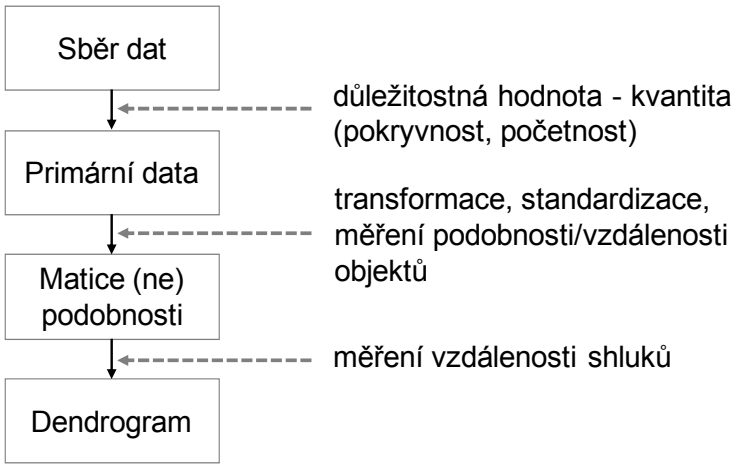
Je velmi žádoucí doplnit takové zhodnocení dendrogramu o popis, co mají dané objekty (v tomto případě lokality v časových obdobích) v jednotlivých shlucích společné (výskyt konkrétních taxonů) a čím se shluky lokalit mezi sebou liší.

Na Obrázek 5.10 lze vidět, jak různé jsou výsledné dendrogramy při použití různých shlukovacích algoritmů.



Obrázek 5.10 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 korýšů šesti lokalit ve třech časových obdobích; Illyová, Némethová 2005.)

Výsledek hierarchického aglomerativního shlukování je ovlivněn na několika úrovních (Obrázek 5.11). 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 5.11 Výsledek hierarchického aglomerativního shlukování je ovlivněn na několika úrovních (podle Lepš, Šmilauer 2000).

Aglomerativní hierarchické shlukování: shrnutí

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/koeficientu 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 následující:

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

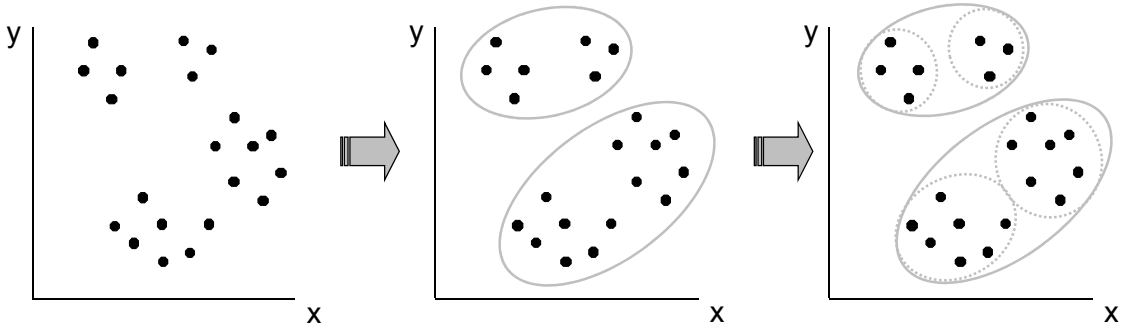
5.1.2 Hierarchické divizivní shlukování

Divizivní metody pracují ze začátku se všemi objekty jako s jednou skupinou. Nejdříve je tato skupina rozdělena do dvou menších skupin. Dělení podskupin pokračuje dále, dokud není splněno kritérium, které ukončí analýzu (např. předem definovaný počet kroků, případně rozklad na samostatné objekty; Obrázek 5.12). 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 5.12 Princip divizivního shlukování.

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

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

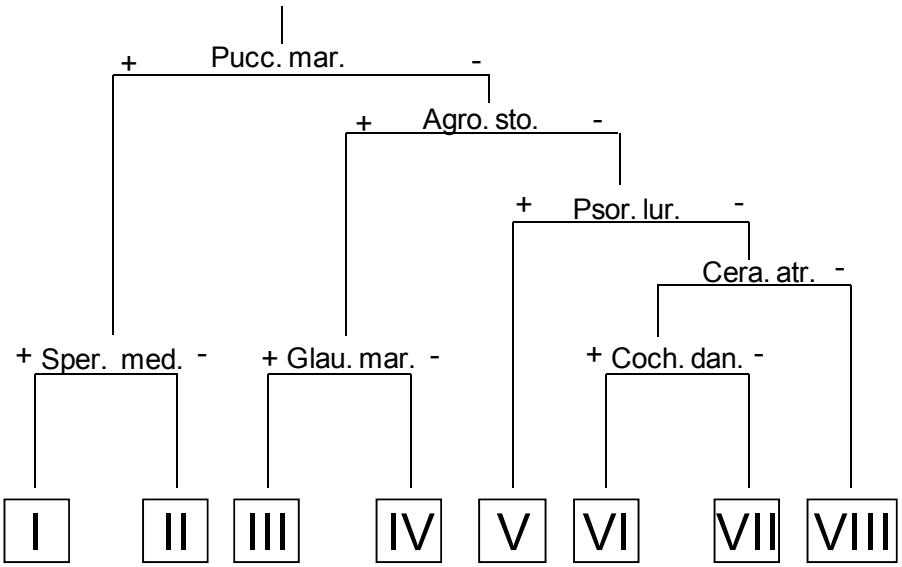
Monotetické metody

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 jedničkou, druhá skupina je skupina objektů, ve kterých je tato proměnná kódovaná nulou (Obrázek 5.13). 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 prezence a absence druhů. Zřejmou nevýhodou metody je její monotetická povaha.



Obrázek 5.13 Ukázka výsledku asociační analýzy. Binární klíč k identifikaci typů slanisk západního Irska (Ivemey-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 přítomnosti jednoho nebo více pseudo-druhů. Čím víc je druh početnější, 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 5-1).

Tabulka 5-1 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	*Cirsium oleraceum*	0	1
	Glechoma hederacea	6	0
	Juncus tenuis	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

Výhoda nahrazení kvantitativní proměnné několika kvalitativními proměnnými spočívá v tom, že když abundance druhu vykazuje unimodální odezvu podél gradientu, každý pseudo-druh také vykazuje unimodální křivku odezvy, a když je křivka odezvy pro abundanci zešikmená, pak se křivky odezev 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). Ordinace 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 TWINSPANu jsou následující:

- 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.

5.2 Nehierarchické shlukování

Často se setkáme s případy, kdy není výhodné používat hierarchickou shlukovou analýzu, protože data nevykazují hierarchickou strukturu. V takových případech může být vhodnější použití nehierarchického shlukování, při němž 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.

5.2.1 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 od sebe co nejvíce odděleny.

Algoritmus metody je následovný:

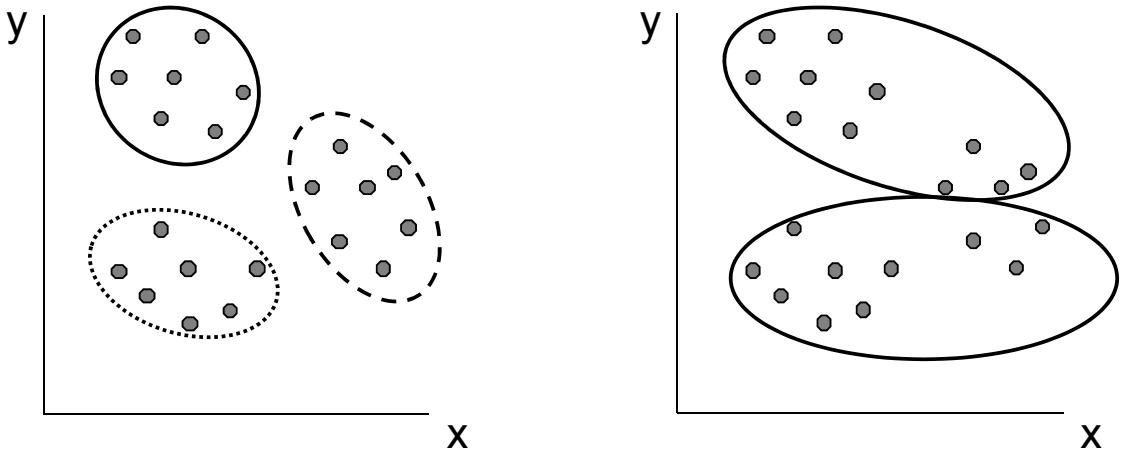
- Zvolíme počáteční rozklad do *K* shluků, nejčastěji náhodně (podkladem ovšem může být také např. výsledek již provedeného shlukování, který chceme zlepšit).
- Určíme centroidy pro všechny shluky v aktuálním rozkladu.
- Postupně zhodnotíme pozici všech objektů. Pokud má objekt nejblíže k vlastnímu centroidu, ponecháme jej na místě, jinak jej přesuneme do shluku, k jehož centroidu má nejblíže.
- Centroidy každého z *K* shluků jsou přepočítány.
- Body 3 a 4 se opakují do té doby, kdy 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 rovnou 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 při výskytu odlehlých objektů. Metoda *K*-průměrů je citlivá na odlehlé 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 5.14). Proto je vhodné provést analýzu pro několik různých počátečných *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 5.14 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.

5.2.2 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.

- V prvním kroku je aplikována tradiční metoda *K*-průměrů pro *K* shluků (*K* je nejprve rovno dolní hranici určené uživatelem).
- V druhém kroku se zjišťuje, zda a kde se má objevit nový centroid, nový shluk. Tohoo je dosaženo tím, že se některé shluky nechají rozpadnout na dva. Proces začíná tak, že se každý centroid shluku (nazveme jej rodičovský centroid) rozdělí na dva centroidy (dceřiné centroidy) v opačném směru podél náhodně zvoleného vektoru. Poté se pro každou rodičovskou oblast, čili pro každý pár dceřiných

centroidů, vypočítá lokální metoda *K*-průměru pro dva shluky. Hranice rodičovských oblastí se nemění. Srovnáním Bayesovského informačního kriteria (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).

- 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 kriteria. Jinak se pokračuje znovu krokem 1.

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

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

5.2.3 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*:

- 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.
- Všechny objekty jsou spojeny s nejbližším medoidem. Míra nepodobnosti/vzdálenosti je definována jakoukoliv platnou metrikou vzdálenosti, nejčastěji Euklidovskou vzdáleností, Manhattanskou vzdáleností, Minkowského vzdáleností, 1 – korelace.
- V druhé fáze 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 kriteria shlukování pro tuto konfiguraci. Když se zlepší kriterium shlukování, testovaný objekt se stane medoidem místo původního medoidu. Tato procedura se opakuje, dokud již nedochází k žádnému dalšímu zlepšení.

Výhody metody *K*-medoidů:

- Metoda nevyžaduje původní data, může být aplikována také přímo na matici nepodobnosti.
- 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 případů. kdy existují ekvivalentní řešení, což je velice zřídka).

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

5.3 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.

5.3.1 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, na rozdíl od ostatních metod, které pracují s asociačními maticemi.

5.3.2 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:

$$D = \min_{1 \le i \le n} \left\{ \min_{\substack{1 \le i \le n \\ i \ne j}} \left\{ \frac{d(c_i, c_j)}{\max_{1 \le i \le n} \{d'(c_k)\}} \right\} \right\} \tag{5.1}$$

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.

5.3.3 Daviesův-Bouldinův validační index (*Davies-Bouldin validity index*)

Daviesův-Bouldinův validační index je podíl sumy vnitroshlukového rozložení a mezishlukového rozložení. Hodnoty tohoto indexu získáme ze vzorce:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S_n(Q_i,Q_j)}\right\} \tag{5.2}$$

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.

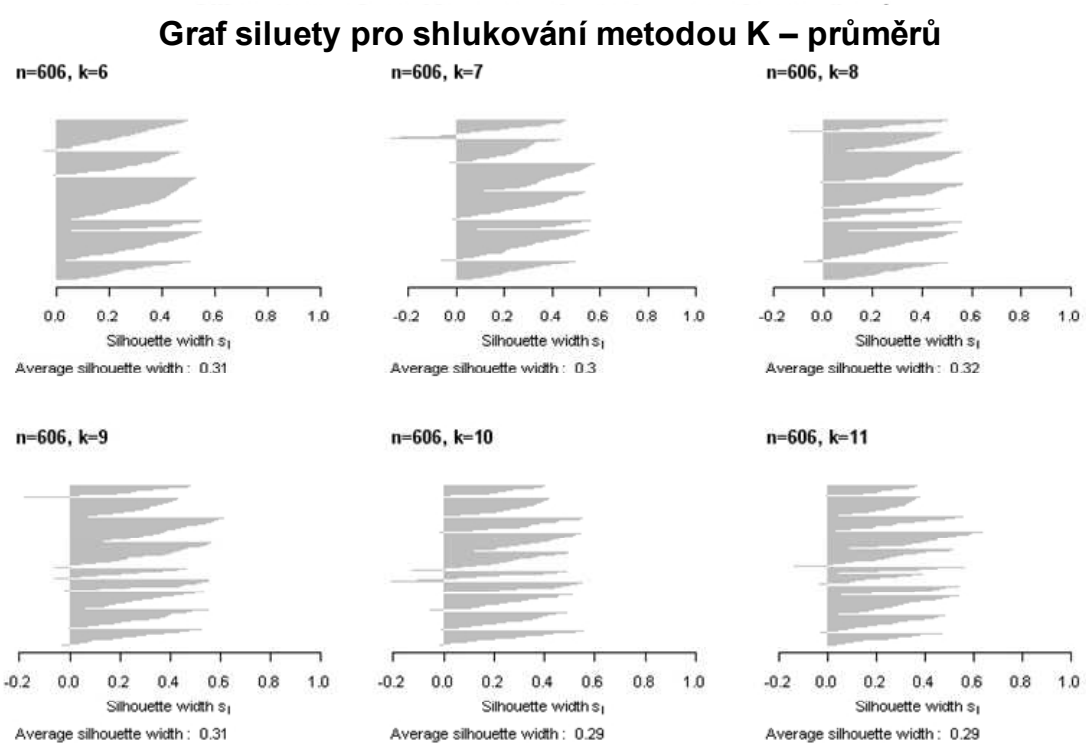
5.3.4 Validační metoda siluety

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:

$$S(i) = \frac{(b(i)-a(i))}{\max\{b(i),a(i)\}} \tag{5.3}$$

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 je počet shluků s největší průměrnou hodnotou šířky siluety optimální řešení. Výstupem této metody bývá sada grafů, kde jsou vyznačeny hodnoty siluety pro všechny objekty ve shlucích pro více variant shlukování (Obrázek 5.15).



Obrázek 5.15 Graf siluety. Shlukováno bylo 606 lokalit do 6 až 11 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.

5.3.5 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 zařazeny do stejného shluku jako *a*. Průměrováním přes všechny případy v datech můžeme homogenitu rozdělení spočítat podle vzorce:

$$I_k = \frac{1}{n}\sum_{i=1}^{n} v_k(x_i) \tag{5.4}$$

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.

5.3.6 C-index

Tento index je definován vzorcem:

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}}, \tag{5.5}$$

kde S je suma vzdáleností mezi všemi páry objektů ve shluku. Nechť *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.

5.3.7 Goodmanův-Kruskalův index (*Goodman-Kruskal index*)

Pro daný datový soubor Goodmanův-Kruskalův index hodnotí všechny možné čtveřice objektů (*a, b, c, d*). Nechť *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 (*disconcordant*), 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

$$GK = \frac{N_c - N_d}{N_c + N_d}. \tag{5.6}$$

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

5.3.8 Meansim (MSA)

Nejedná se přímo o validační metodu pro určení správného počtu shluků. Její výsledky nám pouze pomohou 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 ve 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:

$$CS = W - B. \tag{5.7}$$

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).

5.4 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.

6 Ordinační analýza

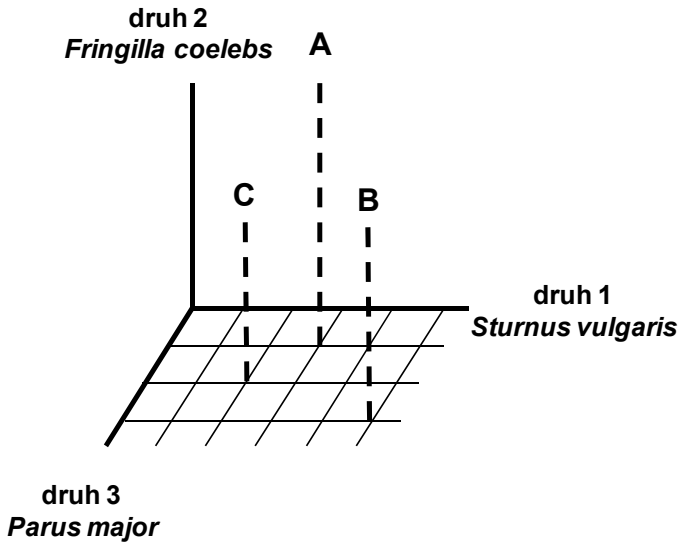
6.1 Úvod

Při řešení množství problémů se setkáváme se situací, kdy je počet sledovaných proměnných velmi rozsáhlý, nepřehledný a vztahy mezi nimi jsou velmi těžko interpretovatelné.

Objekty charakterizované *p* proměnnými si můžeme představit jako body v *p*-rozměrném prostoru. V tomto prostoru každý z rozměrů představuje hodnoty jedné proměnné. Když pracujeme pouze se dvěma nebo třemi proměnnými, situaci si můžeme zobrazit v dvoj- nebo trojrozměrném grafu. Tady lze bez problémů sledovat vztahy mezi objekty, jejich vzdálenost a seskupení. Situaci si můžeme představit na příkladě ekologických společenstev, kde objekty jsou lokality a proměnné druhy (Tabulka 6-1, Obrázek 6.1).

Tabulka 6-1 Zastoupení třech druhů ptáků na třech lokalitách.

	Druh 1 *Sturnus vulgaris*	Druh 2 *Fringilla coelebs*	Druh 3 *Parus major*
Lokalita A	3	5	1
Lokalita B	5	4	3
Lokalita C	2	3	2



Obrázek 6.1 Umístění lokalit (A, B, C) v prostoru vytvořeném třemi ptačími druhy.

Když je počet proměnných větší, nemůžeme je jednoduše prozkoumat v trojrozměrném grafu. Lze se podívat na umístění objektů v prostoru definovaném dvojicemi proměnných, ovšem prozkoumat tímto způsobem všechny možné páry proměnných by bylo velice pracné; kromě toho některé problémy a vztahy dat nejsou v kombinaci pouze dvou proměnných pozorovatelné. Mnohorozměrná data se tedy snažíme zjednodušit tak, že odhalíme hlavní trendy variability v celém souboru proměnných. Mnohorozměrné řešení spočívá v zobrazení objektů v mnohorozměrném grafu s tolika osami, kolik je původních proměnných. Takový diagram je ovšem možné zobrazit ve dvou- nebo trojrozměrném prostoru. Proto se používá projekce takového mnohorozměrného diagramu do roviny nesoucí nejvíce variability. Při tomto procesu

nedochází k větší ztrátě informace. Tomuto procesu říkáme ***redukce dimenzionality dat*** a je základním ***principem ordinačních metod***.

Ordinace je obecné označení pro skupinu metod, které slouží k seřazení objektů podél tzv. ***ordinační osy*** (teoretického gradientu, resp. hypotetické – latentní proměnné) tak, aby byl zachován trend a struktura v datech. Ordinační metody umožňují odhalit vztahy mezi proměnnými stejně jako vztahy mezi objekty. Úspěšnost ordinačních metod závisí na struktuře obsažené v datech. Dobře strukturovaná data, tedy data, v nichž existují vztahy mezi proměnnými, umožňují koncentraci podstatné části variability do několika málo ordinačních os.

Všechny ordinační techniky redukující dimenzionalitu jsou založeny na ***vlastní analýze*** (*eigenanalysis*), tj. hledání ***vlastních vektorů*** (*eigenvectors*) asociační matice. Výpočet má dvě nejběžnější řešení, obě z nich se ve vícerozměrné analýze používají v různých oborech ponejvíce z historických a interpretačních důvodů:

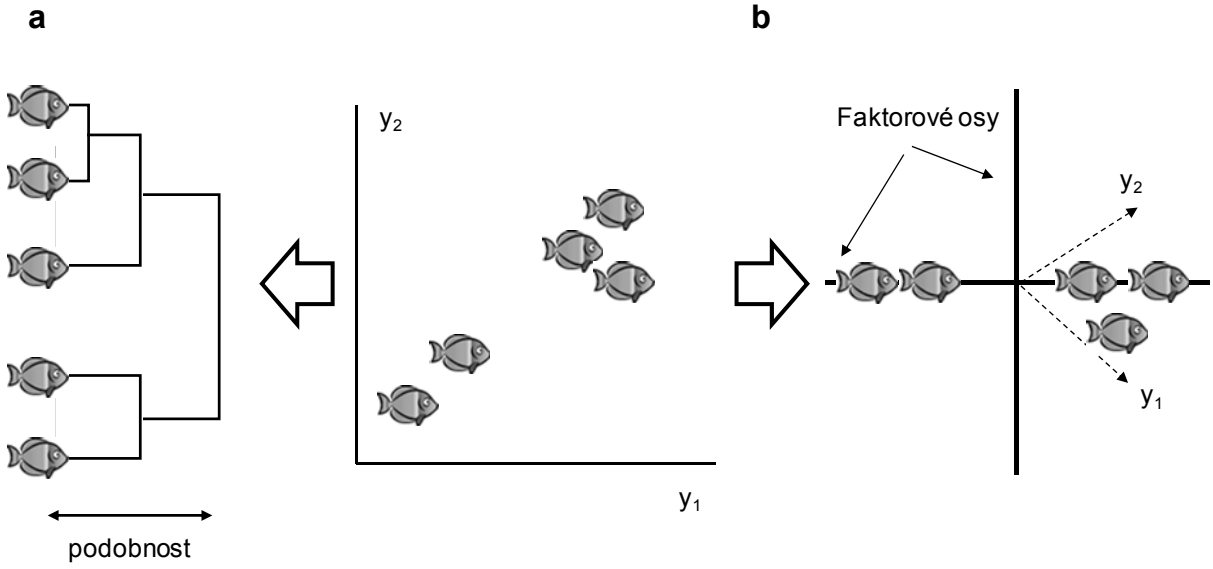
- výpočetní přístup maticové algebry (viz příloha Základy maticové algebry):
  - výsledek je získán jednoznačným matematickým postupem pomocí výpočtu vlastních čísel a vlastních vektorů čtvercové asociační matice;
  - vlastní vektory matice představují řešení definující směr ortogonálních os ordinační analýzy ve vícerozměrném prostoru;
  - míra variability nesené na osách ordinační analýzy je popsána vlastními čísly matice; v případě existence vztahů mezi původními proměnnými nesou první z těchto os ordinační analýzy větší podíl variability dat, než připadá na původní proměnné;
  - *nevýhoda:* metoda není pro nematematické obory intuitivní; výpočetně náročné pro velké datové matice a proto řada softwaru používá iterativní postup;
- geometrický iterativní přístup (detailněji popsán v kapitole o korespondenční analýze):
  - metoda je intuitivní z pohledu významu ordinačních analýz pro praktickou analýzu dat;
  - pracuje s představou rotace pohledu kolem objektů ve vícerozměrném prostoru, která nalézá optimální pohled nesoucí největší množství variability;
  - interpretačně jde v analýze ekologických dat o analogii postupu váženého průměrování používané v analýze valenčních charakteristik taxonů;
  - jde o iterační algoritmus:
    - v prvním kroku je náhodně zvolena osa (vektor) ve vícerozměrném prostoru;
    - je vyhodnocena variabilita spjatá s touto osou a ověřeno, zda jiné proložení osy (nyní již jde o systematický postup, náhodné je pouze stanovení osy v prvním kroku výpočtu) nevysvětluje variabilitu dat lépe;
    - postup je opakován, dokud není možno dosáhnout v dané ose vyšší vyčerpané variability;
    - osa je zafixována a algoritmus pokračuje hledáním další osy, ortogonální k již nalezené ose; celý postup je opakován až do dosažení maximálního počtu os pro daný typ ordinační analýzy;
  - *nevýhoda:* metoda nemusí nalézt nejlepší řešení, pokud při iteračním procesu nalezne lokální minimum dostatečné pro zastavení algoritmu (nicméně v praxi tento problém nastává vzácně);
  - *nevýhoda:* různé softwarové implementace mohou dávat mírně odlišné výsledky.

Úvod do vlastní analýzy vlastních čísel a vlastních vektorů je v Příloze: Základy maticové algebry. V kontextu ordinačních metod je ovšem nutno zdůraznit několik bodů.

- Vlastní analýza probíhá na čtvercové symetrické matici odvozené z datové matice.
- Vlastní analýza má jediné řešení a není závislá na řádu matice.
- Směr ordinační osy je určen vlastním vektorem (*eigenvector*) a je k ní přiřazena vlastní hodnota (*eigenvalue*).
- Osy jsou seřazeny podle jejich vlastních hodnot, tudíž první osa má nejvyšší vlastní hodnotu, druhá osa druhou nejvyšší atd.
- Vlastní hodnoty mají matematický význam, který může pomoci interpretaci. V analýze hlavních komponent (PCA) vlastní hodnoty představují vysvětlenou variabilitu, rozptyl (variance extracted). V korespondenční analýze a metodách od ní odvozených představují vlastní hodnoty vysvětlenou inercii (inertia extracted).
- Osy jsou navzájem kolmé (ortogonální).
- Počet ordinačních os je obvykle roven počtu proměnných, resp. počtu proměnných minus jedna (nebo počtu objektů minus jedna, když je tato hodnota menší). Počet ordinačních os, které se vyplatí interpretovat, by měl být co nejmenší při zachování maximální interpretovatelnosti.
- Pozice objektů a proměnných v ordinačním diagramu jsou vypočítány zároveň a proto můžou být zobrazeny v tom samém ordinačním diagramu (takový diagram se nazývá *biplot*).

Ordinační metody jsou především průzkumné metody, které se používají ke tvorbě hypotéz a primárně neslouží k jejich testování. Ordinační metody se nezabývají příčinnými vztahy.

Rozdíl mezi ***shlukovou analýzou*** a ***ordinací*** je znázorněn na Obrázek 6.2. Shluková analýza nachází v datech skupiny; klasifikuje objekty nebo proměnné do skupin. Ordinační metody seřazují objekty a/nebo proměnné podél ordinačních os.



Obrázek 6.2 Dvě možnosti zpracování mnohorozměrných dat: a shluková analýza, b ordinace. y1, y2 – původní proměnné.

6.1.1 Interpretace výsledků ordinační analýzy

Matematický význam výsledků ordinační analýzy je jasný – jde o sadu vzájemně ortogonálních (nezávislých) vektorů (ordinačních os, dimenzí ordinačního prostoru) přeskupujících variabilitu spjatou s jednotlivými proměnnými tak, aby se co nejvíce z této

variability dalo vyjádřit v co nejmenším počtu dimenzí. Nad touto matematickou interpretací je nicméně ještě interpretace daná účelem analýzy:

- zjednodušení vícerozměrného souboru do co nejmenšího počtu dimenzí bez ohledu na jejich interpretovatelnost (používá se pro zmenšení objemu vysoce dimenzionálních dat);
- zjednodušení vícerozměrného souboru do malého počtu dimenzí interpretovatelných za pomoci původních proměnných (používá se pro zjednodušení analýzy, kdy interpretovatelné vícerozměrné osy vstupují do dalších výpočtů);
- identifikace korelační struktury dat (cílem je zjištění a vizualizace vzájemných vztahů proměnných);
- identifikace přirozeně existujících shluků objektů (v případě, že rozdělení objektů do shluků souvisí s variabilitou dat, je možné tyto shluky identifikovat v prostoru prvních ordinačních os);
- v případě analýzy ekologických dat jsou ordinační osy často interpretovány jako tzv. environmentální gradienty. Protože tato interpretace má přímou vazbu na řadu ekologických teorií vysvětlujících utváření biologických společenstev, je blíže rozebrána v následujícím textu.

Výsledkem ordinace je ordinační diagram (graf). Pro interpretaci ordinačního diagramu platí:

- Směr osy (např. vlevo versus vpravo, nahoře, dole) je náhodný a neovlivňuje interpretaci; díky tomu jej můžeme změnit, když je to interpretačně výhodné.
- Numerická škála osy není potřebná k interpretaci (s výjimkou DCA, kde je škála jednotkou beta diverzity).
- Pořadí os je důležité (s výjimkou NMDS). První osa je důležitější než druhá osa atd.
- Třetí a další osy můžou být sestrojeny; rozhodnutí, kde skončit interpretaci dalších os, je zejména záležitostí kvality a kvantity dat a schopnosti interpretovat výsledky.
- Je vhodné, aby osy nebyly korelované, aby představovaly různé latentní proměnné. Většina technik automaticky spěje k nekorelovaným (ortogonálním) osám.
- Nejdůležitějšími nástroji pro interpretaci výsledků ordinačních analýz jsou odborné zkušenosti biologa z terénu/laboratoře a znalosti z literatury.

6.1.2 Interpretace os ordinační analýzy jako environmentálních gradientů

Výskyt organismů a potažmo celých biologických společenstev je ovlivněn podmínkami prostředí prostřednictvím zákonů ekologického optima a minima; tedy organismus se vyskytuje pouze tam, kde jsou splněny jeho minimální nároky na podmínky prostředí, a záznam o výskytu organismu v datech je dokladem toho, že minimální nároky organismu byly splněny.

V ordinační analýze ekologických dat je tento pohled často interpretačně obrácen a předpokládá se, že výsledkem analýzy je skrytý environmentální gradient, který jako kombinace reálných faktorů prostředí ovlivňuje výskyt organismů. V tomto kontextu je vlastně pozice organismu na ose ordinační analýzy jejím optimem pro tento skrytý gradient a pozice společenstva (lokality) na ose je průměrným váženým optimem organismů tvořících toto společenstvo; čím dále od středu ordinačního prostoru, tím extrémnější podmínky na gradientu jsou pro společenstvo nebo organismus optimální.

Jde o velmi podobný přístup uplatňovaný v analýze valenčních charakteristik taxonů pod názvem vážené průměrování a například korespondenční analýza, ale i jiné ordinační metody byly do ekologických věd uvedeny s touto interpretací na pozadí a jsou zde počítány pomocí iteračního algoritmu váženého průměrování (Hill, 1974).

Tento přístup je z hlediska interpretace ekologických dat přínosný, nicméně je při něm třeba mít na paměti i některá úskalí:

- skrytý gradient nemusí být přímo spjat s abiotickými charakteristikami prostředí (nadmořská výška, pH, kontaminace apod.), ale i vzájemnými interakcemi organismů, které v datových souborech často nejsou podchyceny a u některých typů organismů hrají podstatnou roli;
- skrytý gradient bez interpretace vůči původním proměnným nemusí oproti jednorozměrné analýze těchto proměnných přinášet interpretovatelné a relevantní informace o studovaném problému;
- gradient musí být vždy interpretován v kontextu použité statistické metody (předpoklady metod, velikost vzorku, poměr počtu lokalit vůči počtu proměnných apod.);
- reprezentativnost vzorkování vůči realitě determinuje rozhodujícím způsobem výsledky analýzy a vytvořené skryté gradienty musí být vždy interpretovány v tomto kontextu.

6.1.3 Typy ordinačních metod

Podobně jako do skupiny shlukových analýz patří několik různých metod, i ordinačních metod je několik.

- Analýza hlavních komponent (PCA, *principal component analysis*) je limitovaná kvantitativními proměnnými.
- Faktorová analýza (FA, *factor analysis*) zaštiťuje analýzu hlavních komponent a bývá široce používána zejména v sociologii a psychologii, ale i v jiných oborech.
- Korespondenční analýza (CA, *correspondence analysis, reciprocal averaging*) umožňuje současné zobrazení řádků a sloupců kontingenční tabulky.
- Detrendovaná korespondenční analýza (DCA, *detrended correspondence analysis*) je detrendovaná forma korespondenční analýzy a je oblíbenou metodou mezi ekology.
- Analýza hlavních koordinát (PCoA, *principal coordinate analysis, metric multidimensional scaling*) umožňuje zachovat v redukovaném prostoru vzdálenosti mezi objekty na základě metrické asociační matice.
- Nemetrické mnohorozměrné škálování (NMDS, *nonmetric multidimensional scaling*) pracuje s jakoukoliv metrickou nebo semimetrickou asociační maticí. Je velice populární mezi ekology.

Podrobně se těmto technikám budeme věnovat v následujícím textu. Samostatnou kapitolu představují kanonické ordinační metody (*canonical ordination*), které spojují ordinaci s regresí a umožňují testovat hypotézy.

6.2 Analýza hlavních komponent a faktorová analýza

K řešení problému redukce dimenzionality dat byly vytvořeny dvě příbuzné vícerozměrné metody, a to analýza hlavních komponent (PCA, principal component analysis) a faktorová analýza (factor analysis). Obě tyto metody vycházejí z analýzy kovarianční, případně korelační matice výchozích proměnných. Pokoušejí se najít skryté (neměřitelné, latentní) proměnné, označované jako hlavní komponenty nebo faktory, vysvětlující variabilitu a závislost původních proměnných. Analýza hlavních komponent i faktorová analýza se tedy snaží o vyjádření původních proměnných pomocí latentních proměnných, které se nedají přímo měřit, můžou

ovšem mít určitou věcnou interpretaci. Cílem je zjednodušení původního systému proměnných a zároveň zjištění struktury jejich závislostí.

V analýze hlavních komponent i faktorové analýze je závislost výchozích proměnných zkoumána symetricky. Proměnné tu nejsou apriorně členěny podle směru závislosti na vysvětlující a vysvětlované; jejich vzájemná závislost není vysvětlovaná příčinnými vztahy mezi těmito proměnnými, ale působením skrytých proměnných – hlavních komponent, či faktorů. Od hlavních komponent, resp. faktorů se v obou metodách požaduje, aby maximálně vysvětlovaly původní proměnné. Způsob, jakým tyto dvě metody reprezentují původní proměnné, je odlišný.

- při analýze hlavních komponent (PCA) nové (latentní) proměnné (hlavní komponenty, *principal components*) vysvětlují maximum celkového rozptylu původních proměnných, případně maximálně reprodukují celkovou kovarianční (nebo korelační) matici výchozích proměnných,
- u faktorové analýzy soubor latentních proměnných (společných faktorů, *common factors*, faktorů, *factors*) maximálně reprodukuje nediagonální prvky kovarianční (korelační) matice původních proměnných, tedy vysvětluje především vzájemné závislosti mezi pozorovanými proměnnými. Metodu faktorové analýzy možno považovat za zobecnění analýzy hlavních komponent.

6.2.1 Analýza hlavních komponent (PCA, principal component analysis)

PCA nahrazuje původní soubor proměnných souborem nových (hypotetických), proměnných sumarizujících rozptyly původních proměnných. Tyto nové proměnné nazýváme hlavní komponenty (*principal components*) a jsou lineární kombinací původních proměnných. Hlavní komponenty jsou na sobě nezávislé, čili kolmé (ortogonální). Zda jsou tyto nové proměnné umělými charakteristikami, či zda skutečně odrážejí určité reálné faktory, tj. mají určitý předmětný obsah, je otázkou interpretace, kterou je třeba provádět na základě věcných znalostí zkoumaných proměnných.

Proces hledání hlavních komponent je postupný. Nejdříve se vytvoří první hlavní komponenta, která je vedena ve směru největší variability mezi objekty a tedy vysvětluje největší část rozptylu původních dat. Po nalezení první hlavní komponenty je nalezena druhá hlavní komponenta, která vysvětluje největší část zbytkového rozptylu a zároveň je nezávislá (ortogonální) na první hlavní komponentě. Podobně jsou nalezeny další komponenty. Výsledkem jsou nekorelované ortogonální faktory.

Hlavní komponenty jsou uspořádány podle jejich klesajícího rozptylu. Proto několik prvních hlavních komponent v sobě zahrnuje podstatnou část rozptylu sledovaného souboru objektů.

Algebraicky PCA hledá vlastní hodnoty (*eigenvalues*) a vlastní vektory (*eigenvectors*) asociační matice. Prvky vlastních vektorů jsou váhy původních proměnných. Tyto udávají pozici objektů vzhledem k novému systému vytvořenému hlavními komponentami. Analýza hlavních komponent vychází ze symetrické matice založené na původních proměnných. Touto maticí může být ***korelační matice*** nebo ***kovarianční matice***. Hlavní komponenty korelační nebo kovarianční matice jsou určeny nalezením vlastních hodnot a s nimi souvisejících vlastních vektorů matice.

Výpočetní algoritmus je následující:

Hlavní komponenty asociační matice, kterou označíme **A,** získáme řešením vztahu:

$$(\mathbf{A} - \lambda_k \mathbf{I})\mathbf{u}_k = \mathbf{0} \qquad (6.1)$$

kde $(\mathbf{A} - \lambda_k \mathbf{I})$ je charakteristická rovnice, která se používá k výpočtu vlastních hodnot λ_k. Ty získáme z rovnice:

$$\left|\mathbf{A} - \lambda_k \mathbf{I}\right| = 0 \qquad (6.2)$$

kde $\left|\mathbf{A} - \lambda_k \mathbf{I}\right|$ je determinant charakteristické rovnice. Vlastní vektory $\mathbf{u}_k$ souvisí s vlastními hodnotami λ_k, jak lze vidět v rovnici (6.1). Vlastní hodnoty představují rozptyl odpovídající hlavním komponentám. Každá vlastní hodnota odpovídá jedné komponentě. Tím dostáváme tolik vlastních hodnot, kolik máme proměnných.

Všechny vlastní hodnoty jsou kladné nebo rovné nule, jsou seřazeny od největší po nejmenší. Největší vlastní hodnota a k ní příslušný vlastní vektor odpovídá první komponentě, která vysvětluje největší podíl variability v datech.

Tedy pro PCA platí:

- Vlastní hodnoty určují množství rozptylu vysvětlené příslušnou komponentou.
- Vlastní vektory jsou hlavními komponentami.
- K symetrické matici řádu *p* je možné přiřadit *p* vlastních hodnot. Počet vlastních hodnot a vlastních vektorů (hlavních komponent) je tedy stejný jako počet původních proměnných.
- Nezávislost hlavních komponent vyplývá ze symetrie korelační/kovarianční matice.
- Vlastní komponenty jsou seřazeny podle vlastních hodnot, tj. množství vysvětleného rozptylu. Proto velká část rozptylu původní datové matice může být zachycena několika prvními hlavními komponentami.

Typy PCA

Analýza hlavních komponent (PCA) je definována pro korelační a kovarianční matici. Korelace jsou standardizované kovariance. Hlavní komponenty, které získáme z korelační matice, neodpovídají komponentám získaným z kovarianční matice. Vzdálenosti mezi objekty v těchto dvou případech nejsou stejné. To znamená, že výsledek PCA závisí na tom, zda se rozhodneme pracovat na korelační nebo kovarianční matici.

PCA na kovarianční matici – centrovaná PCA

Původní proměnné jsou centrovány. Součet vlastních hodnot kovarianční matice je roven součtu rozptylů proměnných. Počáteční bod nové souřadnicové soustavy je posunut z původního počátečního bodu do centroidu ordinovaných objektů. Vzdálenost mezi objekty v nové souřadnicové soustavě zůstávají stejné jako v původní soustavě.

Tento typ PCA volíme tehdy, jsou-li jednotlivé proměnné vyjádřeny v příbuzných jednotkách. Příkladem může být analýza ekologických společenstev, kde jsou všechny druhy (proměnné) měřeny ve stejných jednotkách.

PCA na korelační matici – standardizovaná PCA

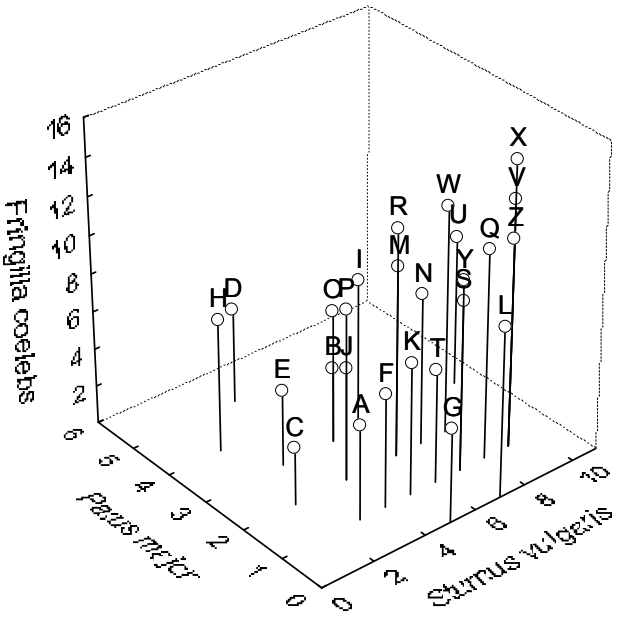
Původní proměnné jsou standardizovány na nulový průměr a jednotkový rozptyl. Součet vlastních hodnot korelační matice je roven řádu korelační matice, tj. počtu proměnných. Počáteční bod nové souřadnicové soustavy je posunut z původního počátečního bodu do centroidu ordinovaných objektů a zároveň jsou původní proměnné přeškálovány tak, aby měly jednotkový rozptyl. Vzdálenosti mezi objekty pak nejsou závislé na jednotkách měření proměnných.

Tento typ PCA volíme tehdy, jsou-li jednotlivé proměnné vyjádřeny ve zcela rozdílných jednotkách měření. Hlavní komponenty jsou totiž lineární kombinací původních proměnných a v případě použití různých měřítek u původních proměnných jejich lineární kombinace nemají význam. A proto je v takém případě vhodné založit PCA na normovaných proměnných, čili na korelační matici.

Geometrický význam hlavních komponent

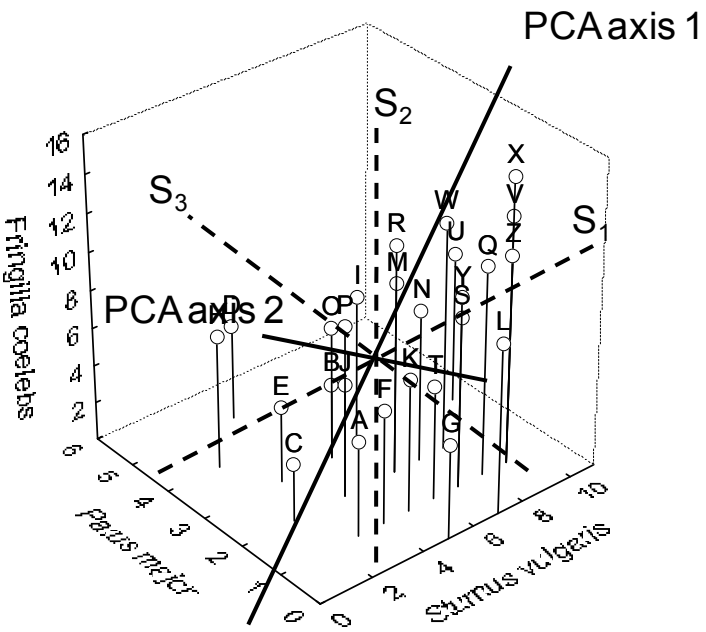
Geometricky je PCA rotací původní datové matice a může být definována jako projekce objektů do nového systému hlavních komponent tak, že maximum rozptylu je promítnuto neboli extrahováno podél první hlavní komponenty, maximum rozptylu nekorelovaného s první hlavní komponentou je promítnuto na druhé hlavní komponentě, atd. Tato rotace souřadnicového systému tedy umožňuje zachytit na několika prvních komponentách maximum informace o prostorové struktuře souboru vícerozměrných pozorování. První dvě hlavní komponenty popisují rovinu s největším rozptylem. Formálně lze tento princip představit jako zobrazení shluku *n* bodů (objektů) v *p*-rozměrném euklidovském prostoru, jehož osy odpovídají jednotlivým proměnným $X_1, X_2, \ldots, X_p$. Relativní pozice objektů v původním prostoru *p* proměnných a v prostoru určeném hlavními komponentami je stejná. Původní systém se tedy natáčí do směru maximální variability mezi objekty, přičemž se zachovávají euklidovské vzdálenosti mezi objekty. Střed souřadnicového systému je bod se souřadnicemi danými výběrovými průměry proměnných. Principem PCA je nalezení lineárních kombinací proměnných, což geometricky odpovídá rotaci původní souřadnicové soustavy provedené tak, že nové osy procházejí směry maximálního rozptylu shluku bodů.

Analýzu hlavních komponent si představíme na konkrétním příkladě. Na Obrázek 6.1 je zobrazen jednoduchý příklad tří lokalit v prostoru tří proměnných, kterými jsou početnosti ptačích druhů. Na Obrázek 6.3 je ve stejném prostoru zobrazeno celkem 26 lokalit označených A až Z. Tento jednoduchý příklad jsme zvolili kvůli názornosti, protože na vztah tří parametrů se dovedeme podívat v třírozměrném prostoru. Pro více rozměrů již nejsme schopni vytvořit odpovídající zobrazení a potřebujeme vícerozměrnou analýzu, kterou naše data zjednodušíme a zobrazíme. Pomocí PCA situaci z Obrázek 6.3 také zobrazíme v redukovaném prostoru.



Obrázek 6.3 Zobrazení 26 lokalit v prostoru vytvořeném třemi proměnnými – početnostmi třech ptačích druhů.

Jelikož vícerozměrná analýza zjednodušuje naměřená data na základě analýzy jejich vzájemných vazeb, v dalším kroku je nevyhnutné vytvořit měřítko této vazby. Jako měřítko vazby proměnných v PCA se používá korelace nebo kovariance. V našem příkladě jsme jako vstup do PCA použily matici kovariancí, protože jednotlivé proměnné (v našem případě početnosti tří ptačích druhů) byly měřeny ve stejném měřítku. V průběhu analýzy proběhne tedy centrování a následná rotace souřadnicové soustavy (Obrázek 6.4).



Obrázek 6.4 Princip rotace prostoru prostoru tří proměnných a jejich zobrazení v prostoru prvních dvou hlavních komponent PCA (PCA axis 1, 2). S_1, S_2, S_3 – centrované původní proměnné.

Kritéria pro určení počtu komponent, které interpretujeme

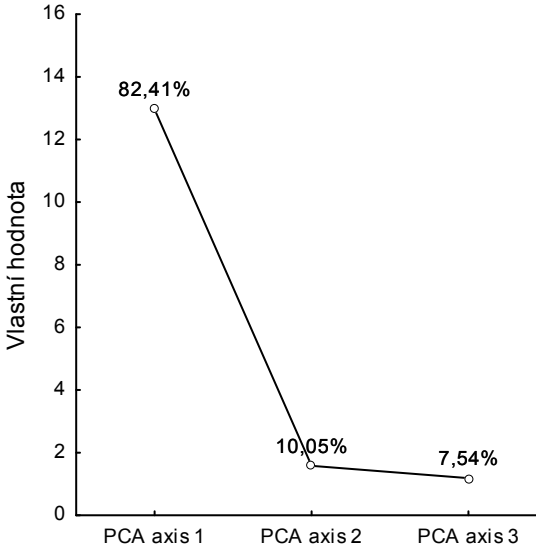
Hlavní komponenty postupně vysvětlují stále menší a menší část celkového rozptylu. Je proto potřeba určit kolik, komponent je rozumné interpretovat. Při interpretaci hlavních komponent je vhodné se omezit především na první komponenty s vysokými vlastními hodnotami. Interpretace komponent s vyššími pořadovými čísly bývá nezřídka obtížná a problematická. Rozumné je brát v úvahu hlavní komponenty, jejichž vlastní hodnoty jsou větší než průměr všech vlastních hodnot.

Ve většině případů pracujeme s korelační maticí, kde je rozptyl (variabilita) všech proměnných roven 1.0. Pak je celkový rozptyl datové matice rovný počtu proměnných. Například když máme 10 proměnných, každou s rozptylem 1, pak je celkový rozptyl, který může být vysvětlený, roven 10. Jedním z nejčastěji používaných kriterií k volbě počtu hlavních komponent, které zachováme, je tzv. Kaiserovo kriterium, které navrhuje ponechat pouze komponenty s vlastní hodnotou větší než 1. V takovém případě totiž hlavní komponenta vysvětluje větší část rozptylu než jedna původní proměnná.

U PCA založené na kovarianční matici je suma vlastních hodnot rovna součtu rozptylů proměnných. I zde je ovšem velice jednoduché určit, kolik komponent budeme interpretovat. Interpretujeme pouze ty komponenty, jejichž vlastní hodnoty jsou nadprůměrné.

Poznámka: Použití Kaiserova kriteria v uvedené formě je možné při interpretaci výsledků PCA založené na korelační matici ze softwaru Statistica. V jiných programech může být součet všech vlastních hodnot normalizovaný. Tak je tomu např. v software Canoco, kde je součet všech vlastních hodnot PCA roven jedné, a to jak u PCA na korelační matici, tak i u PCA na kovarianční matici. Všechny vlastní hodnoty hlavních komponent pak mají hodnoty nižší než jedna. Pak je vhodné zachovat a interpretovat ty hlavní komponenty, jejichž vlastní hodnota je větší než podíl jedné a počtu původních proměnných.

Další možností určení počtu komponent, které budeme interpretovat, je graf – tzv. *scree plot* vlastních hodnot. Hlavní komponenty jsou postupně vyneseny na ose x, příslušné vlastní hodnoty na ose y. Když se díváme doprava na komponenty s vyšším pořadovým číslem, vlastní hodnoty klesají. Počet hlavních komponent, které vysvětlují podstatnou část rozptylu, určíme z grafu podle tvaru křivky tak, že sledujeme, kdy pokles vlastních hodnot u následných komponent ustane a křivka se ohne k menšímu ubývání hodnot (Obrázek 6.5).



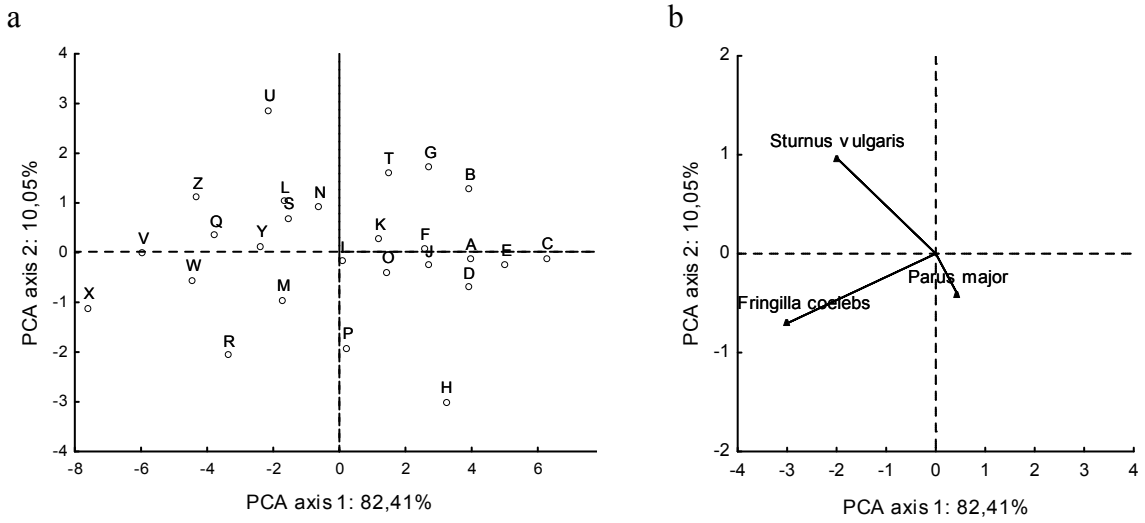
Obrázek 6.5 Ukázka tzv. scree plot. Zobrazení vlastních hodnot kovarianční matice; PCA tří ptačích druhů a 26 lokalit. (PCA axis 1 – 3: první až třetí hlavní komponenta).

Výsledky důležité pro interpretaci PCA zahrnují:

- Vlastní hodnoty, vlastní čísla (*eigenvalues*) vyjadřují podíl rozptylu původního datového souboru vyjádřeného příslušnou hlavní komponentou. Pro interpretaci nejsou důležité konkrétní hodnoty vlastních čísel, ale jejich procentuální podíl ze součtu všech vlastních hodnot.
- Korelace proměnných s hlavními komponenty vyjadřují vztah původních proměnných a hlavních komponent. Čím je absolutní hodnota této korelace vyšší, tím vyšší vliv má příslušná původní proměnná na danou hlavní komponentu. Při interpretaci hlavních komponent se tedy zaměříme na proměnné, které s komponentami korelují.
- Ordinační diagram objektů zobrazuje objekty v ordinačním prostoru (Obrázek 6.6a). Objekty jsou znázorněny body. Pozice těchto bodů v prostoru hlavních komponent jsou dány tzv. komponentním skóre. Je možné je interpretovat přímo prostřednictvím ordinačního diagramu nebo použít pro další analýzy, např. shlukové analýzy.
- Ordinační diagram proměnných zobrazuje proměnné v ordinačním prostoru (Obrázek 6.6b). Proměnné jsou znázorněny vektory s počátkem ve středu souřadnicové soustavy. Hodnoty proměnné kontinuálně rostou ve směru vektoru a klesají v opačném směru. Čím je vektor proměnné delší, tím větší je její vliv. Je důležité brát ohled pouze na proměnné, které jsou dobře reprezentovány v rovině prvních dvou hlavních komponent. Úhel mezi proměnnou a hlavní komponentou je úměrný korelaci mezi touto proměnnou a hlavní komponentou. Čím menší je úhel mezi vektorem proměnné a příslušnou komponentou, tím silněji proměnná ovlivňuje příslušnou komponentu. Vztahy mezi proměnnými se v ordinačním diagramu interpretují na základě úhlů mezi vektory. Kosinus úhlu mezi proměnnými je úměrný jejich korelaci. Tento úhel je stejný jako při kovarianci, protože při použití korelace a kovariance nedochází ke změně pozice proměnné v mnohorozměrném prostoru, ale pouze ke změně v délce její osy, tedy délce vektoru.
- Biplot je graf, který zobrazuje objekty i proměnné ve společném ordinačním diagramu. Při interpretaci je nutno pamatovat na způsob hodnocení vztahů mezi objekty a proměnnými, kdy je chybou interpretace vztahů objektů a proměnných podle jejich blízkosti v redukovaném prostoru. Správná interpretace vychází

z projekce objektů na vektor proměnné nebo na jeho prodloužení a rovněž je důležitý výběr typu biplotu – buď biplot vzdáleností nebo biplot korelací.

Jako výsledek našeho konkrétního příkladu uvádíme ordinační diagram objektů (Obrázek 6.6a) a původních proměnných (Obrázek 6.6b).



Obrázek 6.6 Zobrazení 26 lokalit (a) a tří proměnných (b) v prostoru prvních dvou hlavních komponent (PCA axis 1 a PCA axis 2; podle Palmera 2010).

Tabulka 6-2 Výsledek analýzy hlavních komponent založené na kovarianční matici (vlastní hodnoty, procento vysvětleného rozptylu, korelace původních proměnných s hlavními komponentami).

	Hlavní komponenty		
	1.hl.komp.	2.hl.komp.	3.hl.komp.
Vlastní hodnota	12,964	1,580	1,186
% celkového rozptylu	82,4	10,0	7,5
Kumulativní %	82,4	92,5	100,0
Korelace původních proměnných s hlavními komponentami			
Sturnus vulgaris	-0,89	0,43	-0,17
Fringilla coelebs	-0,97	-0,23	0,03
Parus major	0,38	-0,35	-0,86

Analýza hlavních komponent byla spočítána v programu Statistica. Vlastní hodnoty hlavních komponent byly: $\lambda_1 = 12{,}964$, $\lambda_2 = 1{,}580$, $\lambda_3 = 1{,}186$. Celkový počet komponent je stejný jako počet původních proměnných, a to tři. Součet všech vlastních hodnot je tedy 15,731. Nadprůměrnou vlastní hodnotu má pouze první komponenta (12,964 > 5.244). Je postačující, když se při interpretaci našeho výsledku zaměříme pouze na první komponentu. Tato vysvětluje 82,4% rozptylu původní datové matice (Tabulka 6-2).

Z Tabulka 6-2 a grafu na Obrázek 6.6b je zřejmá silná záporná korelace proměnných *Sturnus vulgaris* a *Fringilla coelebs* s první komponentou, a tak můžeme první komponentu interpretovat na základě těchto dvou původních proměnných.

V našem příkladě budeme interpretovat první hlavní komponentu. Původní proměnné jsou v prostoru hlavních komponent znázorněny na základě jejich korelace s nimi. První hlavní

komponenta souvisí s početností druhů *Fringilla coelebs* a *Sturnus vulgaris* (Obrázek 6.6b, Tabulka 6-2).

Biplot a jeho typy

Biplot je graf, který zobrazuje objekty i proměnné ve společném ordinačním diagramu. V závislosti na použité standardizaci vlastních vektorů (eigenvektorů) existují dva typy biplotů (Tabulka 6-3):

- Biplot vzdáleností (distance biplot)
  - Standardizace délky vlastních vektorů (eigenvektorů) na jednotkovou délku
  - Pozice objektů na faktorových osách mají rozptyl rovný vlastnímu číslu (*eigenvalue*)
  - Interpretace biplotu
    - Umožňuje interpretovat euklidovské vzdálenosti objektů v prostoru PCA (jsou aproximací euklidovských vzdáleností v původním prostoru).
    - Projekce objektu v pravém uhlu na původní proměnnou aproximuje pozici objektu na této původní proměnné.
    - Délka projekce jednotlivých původních proměnných v prostoru faktorových os popisuje jejich příspěvek k definici daného faktorového prostoru.
    - Úhly mezi původními proměnnými ve faktorovém prostoru nemají žádnou intepretaci.
- Biplot korelací (correlation biplot)
  - Standardizace délky vlastních vektorů (eigenvektorů) na druhou odmocninu z vlastních čísel (*eigenvalue*)
  - Pozice objektů na faktorových osách mají jednotkový rozptyl
  - Interpretace biplotu
    - Euklidovské vzdálenosti objektů v prostoru PCA nejsou aproximací euklidovských vzdáleností v původním prostoru.
    - Projekce objektu v pravém uhlu na původní proměnnou aproximuje pozici objektu na této původní proměnné.
    - Délka projekce jednotlivých původních proměnných v prostoru faktorových os popisuje jejich směrodatnou odchylku.
    - Úhly mezi původními proměnnými ve faktorovém prostoru souvisí s jejich korelací.
    - Není vhodný, pokud má smysl interpretovat vzdálenosti (vzájemné vztahy) mezi objekty.

Tabulka 6-3 Standardizace vlastních vektorů a její vliv na projekci proměnných a objektů v biplotu

Původní proměnná (centrovaná)	Standardizace eigenvektoru			
	$\sqrt{\lambda_k}$	1	$\sqrt{\lambda_k}$	1
Celková délka	s_j	1	1	1
Úhly proměnných v redukovaném prostoru	Projekce kovariancí (korelací)	90° rotace systému os	Projekce korelací	90° rotace systému os
Hranice příspěvku k definici faktorové osy	$s_j\sqrt{d/p}$	$\sqrt{d/p}$	$\sqrt{d/p}$	$\sqrt{d/p}$
Projekce na faktorovou osu k	Kovariance s k	Proporcionální kovarianci s k	Korelace s k	Proporcionální korelaci s k
Korelace s faktorovou osou k	$\frac{u_{jk}\sqrt{\lambda_k}}{s_j}$	$\frac{u_{jk}\sqrt{\lambda_k}}{s_j}$	$u_{jk}\sqrt{\lambda_k}$	$u_{jk}\sqrt{\lambda_k}$

Kde λ_k je vlastní číslo (*eigenvalue*) faktorové osy k, s_j je Směrodatná odchylka původní proměnné *j*, *d* je počet původních proměnných, *p* je počet faktorových os a u_{jk} je hodnota eigenvektoru faktorové osy k pro původní proměnnou j

Předpoklady a omezení PCA.

Předpoklady PCA jsou:

- mnohorozměrné normální rozdělení proměnných (pokud cílem není identifikace shluků spjatých s variabilitou dat nebo vícerozměrně odlehlých hodnot);
- proměnné jsou kvantitativní a je možné pro ně vypočítat kovarianci nebo korelaci;
- nezávislost pozorování (objektů).

K těmto bodům je vhodné doplnit dále:

- PCA byla původně navržena pro data s mnohorozměrným normálním rozdělením. Na menší odchylky od mnohorozměrného normálního rozdělení je PCA dostatečně robustní.
- Původně byla PCA navržena pro kvantitativní data. PCA je ovšem částečně robustní i pro zpracování semikvantitativních a binárních proměnných. PCA není vhodná pro vícestavové kvalitativní proměnné, na které nelze použít euklidovskou metriku. V těchto případech se používají jiné metody, např. PCoA.
- Když data obsahují mnoho nul (*double zero problem*), není pro jejich zpracování PCA vhodná. V takovémto případě je vhodné použít jinou metodu, např. PCoA, NMDS nebo korespondenční analýzu.
- Počet proměnných *p* by měl být menší, než je počet objektů *n*. Obecně se doporučuje, aby se počet objektů blížil druhé mocnině počtu proměnných. Analýzu lze spočítat i v případě většího počtu proměnných, než je počet objektů, je ovšem potřeba se zaměřit pouze na několik prvních vlastních vektorů, které jsou málo ovlivněny tím, zda je matice singulární. Proto např. při hodnocení molekulárních dat, kdy počet proměnných převyšuje počet objektů, je výhodnější použít např. PCoA.

V případě příliš složité struktury v datech může být interpretace ordinace složitá. Představme si soubor objektů, který se podél první hlavní komponenty rozdělí na dvě základní skupiny. Pokud jsou tyto dvě skupiny uvnitř členěné komplikovaným způsobem, druhá a další komponenty bývají jistým kompromisem mezi strukturou v obou základních skupinách. V takových případech je vhodné každou skupinu analyzovat samostatně.

PCA se často používá v ekologii biologických společenstev, kdy jsou objekty – většinou lokality nebo snímky – charakterizovány hodnotami několika živočišných nebo rostlinných druhů (počet jedinců, dominance, u rostlin např. pokryvnost). Pomocí PCA hledáme takové latentní proměnné (hlavní komponenty), ke kterým je vztah všech druhů co nejtěsnější. PCA se dá použít pouze v případech, kdy předpokládáme lineární vztah druhů k hlavním komponentám.

6.2.2 Faktorová analýza (Factor analysis)

Faktorová analýza je vícerozměrná statistická metoda, jejíž podstatou je rozbor struktury vzájemných závislostí proměnných na základě předpokladu, že tyto závislosti jsou důsledkem působení určitého menšího počtu v pozadí stojících nezměřitelných faktorů, které jsou nazývány společné faktory (nebo faktory, *common factors, factors*).

Cílem faktorové analýzy je:

- redukce počtu proměnných (charakterizování sady p proměnných menším počtem společných faktorů),
- odhalení struktury vztahů mezi proměnnými.

Faktorová analýza vznikla v oblasti psychologie a byla po dlouhou dobu používána téměř výhradně v tomto oboru. V posledních desetiletích ovšem pronikla i do dalších vědních oborů a našla uplatnění i v biologii a medicíně.

Faktorovou analýzu lze považovat za rozšíření metody hlavních komponent. Na rozdíl od PCA vychází ze snahy vysvětlit závislosti proměnných. Mezi nedostatky PCA (na kovarianční matici) patří zejména fakt, že není invariantní vůči změnám měřítka proměnných. Přístup faktorové analýzy umožňuje tento nedostatek odstranit.

Předpokladem faktorové analýzy je stejně jako u PCA vícerozměrné normální rozdělení proměnných. Problémy ve faktorové analýze můžou spočívat v:

- nejednoznačnosti odhadů faktorových parametrů (problém rotace);
- nutnosti specifikovat počet společných faktorů (*common factors*) před provedením analýzy.

Ve faktorové analýze se vysvětluje vzájemná lineární závislost pozorovaných proměnných $X_1, X_2, \ldots, X_p$ existencí menšího počtu nepozorovatelných faktorů $f_1, f_2, \ldots, f_m$ (zvaných společné faktory, *common factors*) a p dalších zdrojů variability $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ (zvaných chybové či specifické faktory nebo též rušivé či reziduální složky). Společné faktory vyvolávají korelace mezi proměnnými, zatímco chybové faktory pouze přispívají k rozptylu jednotlivých pozorovaných proměnných. Předmětem zájmu faktorové analýzy jsou především společné faktory.

Na tomto místě je potřeba představit dva pojmy související s faktorovou analýzou, a to: faktorové váhy nebo zátěže (*factor loadings*) a komunalita (*communality*).

- Faktorové váhy (*factor loadings*) jsou korelační koeficienty (nebo kovarianční koeficienty v případě použití kovarianční matice) proměnných se společnými faktory.
- Komunality (*communality*) proměnných jsou diagonální prvky redukované kovarianční/korelační matice. Komunalita i-té proměnné udává část jejího rozptylu, která je vysvětlena působením společných faktorů. Zbývající část rozptylu proměnné se nazývá specifickým, či chybovým rozptylem proměnné.

Faktorová analýza pracuje podobně jako PCA. Rovněž jako PCA pracuje s korelační nebo kovarianční maticí a nalézá první hlavní faktor tak, aby vysvětloval největší část rozptylu datové matice. Další faktory jsou konstruovány tak, aby byly nezávislé, čili nekorelované, a vyčerpávaly sestupně maximum celkového rozptylu. Na rozdíl od PCA faktorová analýza odhaduje, kolik rozptylu je vysvětleno komunalitou (*communality*). Rozdíl mezi faktorovou analýzou a analýzou hlavních komponent je i v dalším kroku analýzy. Tady jsou hlavní faktory rotovány tak, aby co nejjednodušeji popisovaly proměnné, tj. aby byly co nejblíže situovány co nejvíce původním proměnným. To je dosaženo v situacích, kdy hlavní faktory jsou nejblíže skupině silně korelovaných proměnných. V těchto situacích můžou být hlavní faktory do určité míry korelovány (viz níže neortogonální rotace faktorů).

Při specifikaci rotace je potřeba určit počet faktorů, které chceme rotovat, tj. zachovat a interpretovat. Postup analýzy je pak následující:

- Spočítáme analýzu pro stejný počet faktorů, jako je počet proměnných (*p*). Tato první fáze analýzy probíhá tedy stejně jako PCA. Získáme provizorní váhy faktorů (*factor loadings*).
- Podle vlastních hodnot faktorů (případně podle *scree plot*-u) určíme počet faktorů *m*, které zachováme a budeme interpretovat, tedy i rotovat.
- Pro stanovený počet faktorů určíme rotaci faktorů a znovu spočítáme analýzu.

I když první fáze faktorové analýzy probíhá stejně jako PCA, interpretace výsledků je jiná než při PCA, což je způsobeno právě rotací faktorů ve druhé fázi analýzy.

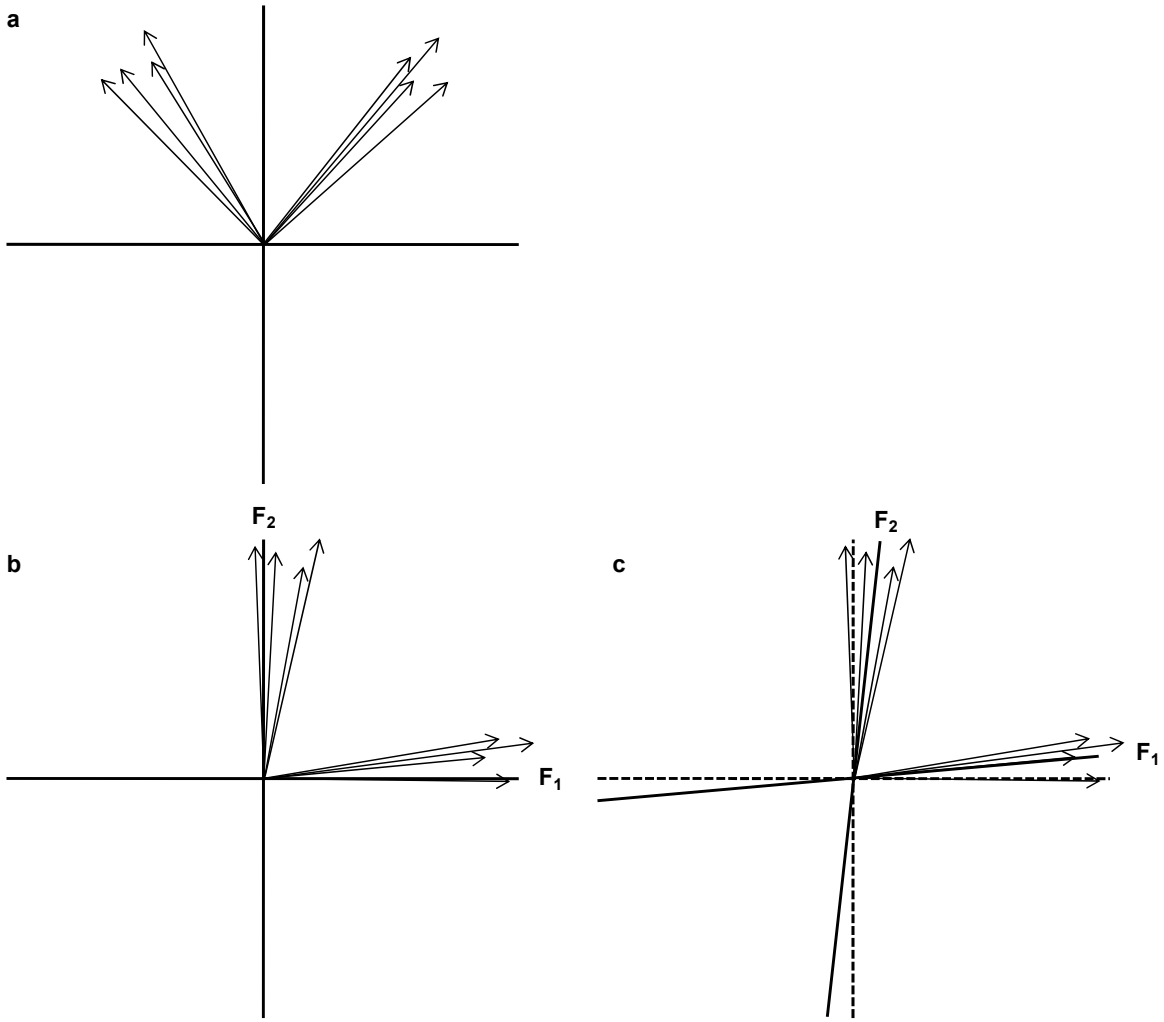
Rotace faktorů slouží k usnadnění jejich interpretace. Cílem je lokalizace souřadnicové soustavy do prostoru společných faktorů tak, aby co nejjednodušeji popisovala proměnné. Každá proměnná by proto měla mít vysoké faktorové váhy (*factor loadings*) u co nejmenšího počtu společných faktorů a nízké či středně vysoké váhy u zbývajících faktorů.

Metody rotace

Pro rotaci faktorů existuje několik možností (Obrázek 6.7). Rotace faktorů může být:

- ortogonální (orthogonal) – zachovává nezávislost faktorů, tyto jsou tedy nekorelované;
- neortogonální (non-orthogonal, oblique) – nové faktory se stávají do určité míry korelované.

Nejznámější metody ortogonální rotace jsou varimax (*variance maximazing*) a quartimax. Rotace varimax je nejběžnější možností rotace. Maximalizuje sumu rozptylů všech faktorů. Quartimax rotace minimalizuje počet faktorů potřebných k vysvětlení všech proměnných. Obě tyto rotace můžou být použity s normalizací vah faktorů nebo bez této normalizace. Všechny možnosti jsou součástí software Statistica.



Obrázek 6.7 a. Nerotovaný prostor, b. Ortogonální rotace v prostoru dvou faktorů F1 a F2, c. Neortogonální rotace stejné situace.

6.2.3 Analýza hlavních komponent a faktorová analýza: shrnutí

- Vstupem analýzy hlavních komponent a faktorové analýzy je:
  - Matice korelací nebo kovariancí původních proměnných.

- Výstupem analýzy hlavních komponent a faktorové analýzy je:
  - Ordinační diagram.
  - Vlastní hodnoty hlavních komponent, resp. faktorových os.
  - Procento vysvětleného rozptylu hlavními komponentami, resp. faktorovými osami.
  - Korelace původních proměnných s hlavními komponentami, resp. s faktorovými osami.

- Při použití analýzy hlavních komponent a faktorové analýzy je nutno pamatovat na níže uvedená omezení:
  - Parametrická metoda.
  - Problém odlehlých hodnot.
  - Závislé na rozdělení proměnných.
  - Nelze použít, když jsou faktory úplně nezávislé (jejich korelace je nulová).

6.3 Korespondenční analýza (CA, correspondence analysis) a detrendovaná korespondenční analýza (DCA, detrended correspondence analysis)

6.3.1 Korespondenční analýza (CA, correspondence analysis)

Korespondenční analýza (CA, *correspondence analysis*) je nástrojem pro analýzu vztahů mezi řádky a sloupci kontingenčních tabulek. Umožňuje tak společné zobrazení řádků a sloupců kontingenční tabulky. Kontingenční tabulky jsou základním nástrojem pro zkoumání vztahů mezi dvěma proměnnými. Jde o frekvenční tabulku, která zaznamenává kumulativní četnosti dvou nominálních (kategoriálních) proměnných. Každý sloupec a každý řádek tabulky pak reprezentuje jednu kategorii dané proměnné (Obrázek 6.8).

	1	…	j	…	p
1	y_{11}	…	y_{1j}	…	y_{1p}
⋮	⋮		⋮		⋮
i	y_{i1}	…	y_{ij}	…	y_{ip}
⋮	⋮		⋮		⋮
n	y_{n1}	…	y_{nj}	…	y_{np}

Obrázek 6.8 Ukázka kontingenční tabulky.

Hodnota y_{ij} v tabulce o rozměrech $n \times p$ označuje počet pozorování neboli frekvenci, které současně náleží do i-té řádkové kategorie a j-té sloupcové kategorie pro $i = 1, ..., n$ a $j = 1, ..., p$ (Obrázek 6.8).

Základní myšlenkou korespondenční analýzy je odvodit tzv. indexy, tj. ordinační osy, které budou kvantifikovat vztahy mezi řádkovými a sloupcovými kategoriemi. Z těchto indexů je pak možné odvodit, která sloupcová kategorie má větší či menší váhu v daném řádku a naopak.

Korespondenční analýza se také vztahuje k otázce snížení dimezionality dat podobně jako např. analýza hlavních komponent a ke snaze o rozklad frekvenční tabulky na faktory. Hledá vlastně podprostor, který zachová největší část tzv. inerce. Celková inerce tabulky je definována jako celková χ^2 statistika frekvenční tabulky podělena celkovým součtem pozorování v tabulce. Korespondenční analýza rozkládá celkovou inerci na sadu vlastních hodnot, resp. na ortogonální faktory.

Podobně jako u dalších ordinačních metod se i v případě korespondenční analýzy snažíme získat ordinační osy v klesajícím stupni důležitosti tak, aby se hlavní informace obsažená v tabulce dala shrnout do podprostoru s co možná nejmenším počtem dimenzí. První osa prochází směrem maximální inerce shluku řádkových (resp. sloupcových) bodů v prostoru sloupcových (resp. řádkových) kategorií. Druhá osa je ze všech kolmých směrů na první osu taková, která prochází směrem maximální inerce shluků bodů, atd. Počet ordinačních os, a tedy vlastních vektorů a vlastních hodnot je minimum z počtu řádků a počtu sloupců snížený o jednu.

Většinu celkové inerce původní tabulky vysvětluje zpravidla několik málo prvních os. Proto většinou postačuje znázornit výsledek do prostoru prvních dvou nebo tří ordinačních os. Je ovšem možné určit přesněji, kolik ordinačních os je rozumné interpretovat. Můžeme rozhodnout dvěma způsoby.

- Zvolíme hraniční hodnotu (např. 80 %) a zjistíme, kolik os má sumární inerci větší než námi zvolená hraniční hodnota.
- Interpretujeme ordinační osy, jejichž vlastní hodnota je nadprůměrná, tj. větší než průměr všech vlastních hodnot.

Algoritmus korespondenční analýzy je jednoduchý, podobně jako u PCA jde o vlastní analýzu, a tedy o hledání vlastních hodnot a vlastních vektorů matice. Na rozdíl od PCA, kde vlastní hodnoty představují vysvětlený rozptyl příslušnou komponentou, v případě CA vlastní hodnoty extrahují inerci, neboli vztah mezi sloupcovými a řádkovými kategoriemi. Rozdílem oproti PCA je, že k získání vlastních čísel datové matice se používá rozklad na singulární hodnoty. Výpočtu vlastních hodnot a vlastních vektorů předchází několik kroků.

Nejdříve je původní datová matice převedena na příspěvek standardizovaných reziduií, které získáme podle vzorce

$$\mathbf{Z} = \mathbf{D}_\mathbf{r}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathbf{T})\mathbf{D}_\mathbf{c}^{-1/2} \tag{6.3}$$

kde matice **P** a **rc** pocházejí z původní datové matice. Původní datová matice rozměru *n x p* (Obrázek 10.1), je následně převedena na matici relativních hodnot, kde $p_{ij} = \frac{y_{ij}}{y}$, $r_i = \frac{y_{i+}}{y}$ a $c_i = \frac{y_{+i}}{y}$.

$$\begin{bmatrix} \mathbf{P} & \mathbf{r} \\ \mathbf{c}^\mathbf{T} & \mathbf{1} \end{bmatrix} = \left[\begin{array}{cccc|c} p_{11} & p_{12} & \cdots & p_{1s} & r_1 \\ p_{21} & p_{22} & \cdots & p_{2s} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rs} & r_r \\ \hline c_1 & c_2 & \cdots & c_s & 1 \end{array}\right] \tag{6.4}$$

dále matice $\mathbf{D_r}$ a $\mathbf{D_c}$ jsou diagonální matice, které mají na diagonále prvky vektoru r a c. Tedy prvky matice **Z** nabývají hodnot podle vzorce

$$z_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \tag{6.5}$$

Rozklad matice Z na singulární hodnoty je následující

$$\mathbf{Z} = \mathbf{U}\Gamma\mathbf{V} \tag{6.6}$$

kde matice **U** je typu *r x k* a její sloupce jsou tvořeny levými zobecněnými singulárními vektory. Matice **V** je typu *s x k* a je složena ze sloupců tvořených z pravých zobecněných singulárních vektorů. Matice **Γ** je typu *k x k* a její diagonála je tvořena singulárními hodnotami (rozklad matice na singulární hodnoty viz Příloha: Základy maticové algebry).

Vektory matice **U** jsou rovny normalizovaným vlastním (charakteristickým) vektorům matice $\mathbf{ZZ}^\mathrm{T}$, a vektory matice **V** jsou rovny normalizovaným vektorům matice $\mathbf{Z}^\mathrm{T}\mathbf{Z}$. Singulární hodnoty matice **Γ** jsou rovny odmocninám vlastních čísel matice $\mathbf{ZZ}^\mathrm{T}$ tedy $\mathbf{Z}^\mathrm{T}\mathbf{Z}$**.**

Následný výpočet souřadnic bodů, které představují buď řádky, nebo sloupce původní datové matice, je závislý na vazbě, kterou sledujeme:

Pokud nás zajímají pouze řádky matice, je výpočet souřadnic řádků (případů) následující:

$$\mathbf{G} = \mathbf{D}_\mathbf{c}^{-1/2}\mathbf{V}\Gamma \tag{6.7}$$

K souřadnicím řádků se souřadnice sloupců dopočítají podle vzorce:

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U} \tag{6.8}$$

Analogicky je tomu u sloupců, pokud nás zajímají pouze vazby mezi sloupci původní datové matice. Souřadnice sloupců získáme podle vzorce:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Gamma} \tag{6.9}$$

K těmto souřadnicím sloupců získáme souřadnice řádků podle vzorce:

$$\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V} \tag{6.10}$$

Další možnost je zobrazení řádkových a sloupcových kategorií v jednom grafu, kde souřadnice řádků ani sloupců nejsou váženými průměry druhé sady kategorií. Souřadnice řádků získáme podle vztahu:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Gamma} \tag{6.11}$$

Matici souřadnic sloupců získáme ze vztahu:

$$\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Gamma} \tag{6.12}$$

Výpočetně může být korespondenční analýza řešena také jednoduchou procedurou váženého průměrování (*weighted averaging*). Oba způsoby řešení vedou samozřejmě k obdobnému výsledku. Jako příklad tohoto výpočtu můžeme uvést matici p druhů vyskytujících se v n vzorcích. Korespondenční analýza je jednou z oblíbených metod mezi ekology, kteří se zabývají výskytem rostlinných nebo živočišných druhů ve vzorcích (na lokalitách). Korespondenční analýza seřazuje jednotlivé vzorky na osách, které jsou určeny pouze druhovým složením společenstev. Každý vzorek můžeme považovat za bod v p-rozměrném prostoru, kde p je celkový počet druhů a jednotlivé osy odpovídají druhům, tj. skóre, neboli souřadnice vzorku na ose je definována zastoupením odpovídajícího druhu ve vzorku. Úkolem korespondenční analýzy je zobrazit množinu bodů, představujících jednotlivé objekty (vzorky, lokality), do redukovaného prostoru tak, aby nové osy zachytávaly co nejvíce inerce a aby docházelo k minimálnímu zkreslení prostorových vztahů. Jinými slovy, aby podobné objekty (vzorky, lokality) byly ve výsledné projekci blízko sebe a nepodobné daleko od sebe. Výsledek je podobný výsledku analýzy hlavních komponent na korelační matici. V případě tabulek obsahujících mnoho nulových hodnot (v ekologii to naznačuje silný environmentální gradient) je použití korespondenční analýzy vhodnější, protože předpokládá unimodální odezvu druhů na gradient ordinační osy. Korespondenční analýza byla v ekologii velmi často používána v 80. letech minulého století. V současnosti je významněji používána její detrendovaná forma, detrendovaná korespondenční analýza.

Princip váženého průměrování vysvětlíme na příkladě společenstev ptačích druhů na třech lokalitách. Kvůli zjednodušení předpokládejme, že společenstva tvořily pouze tři druhy (Tabulka 6-4). Procedura váženého průměrování obsahuje proces opakované křížové kalibrace mezi skóre řádků a sloupců, jehož výsledkem je společná ordinace řádků i sloupců kontingenční tabulky. Skóre řádků jsou váženými průměry skóre sloupců a skóre sloupců jsou váženými průměry skóre řádků. Ordinační skóre řádků a sloupců jsou odvozeny tak, aby byla maximalizována korelace mezi skóre řádků a skóre sloupců.

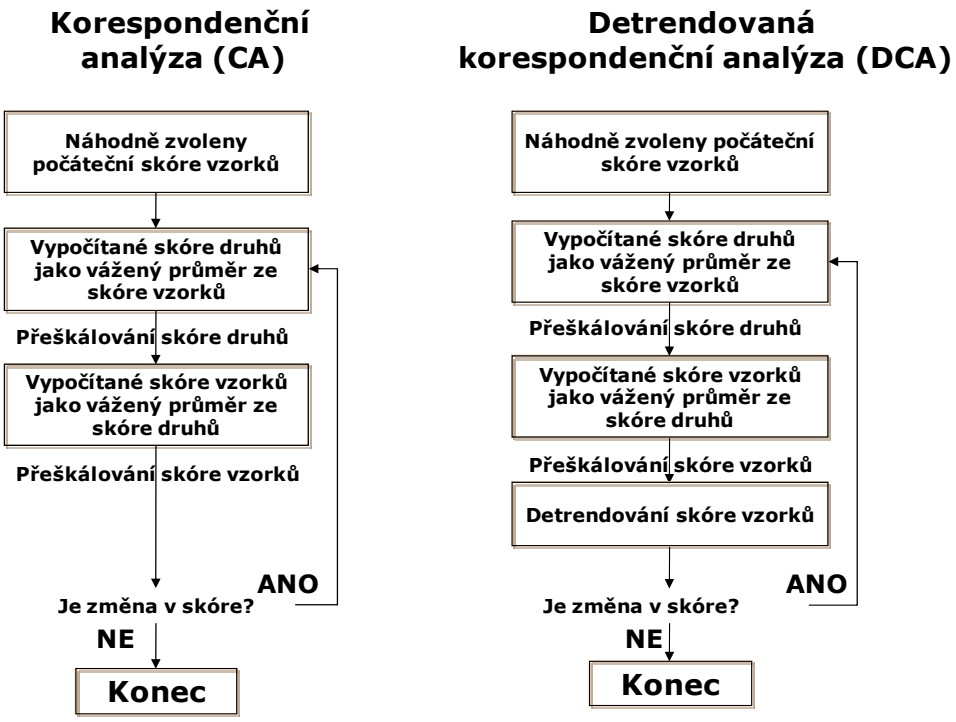
Tabulka 6-4 Ukázka výpočtu první osy korespondenční analýzy metodou váženého průměrování na příkladě tří společenstev (A, B, C). WA 1 – WA 5 – skóre druhů/vzorků vypočítané jako vážený průměr ze skóre vzorků/druhů. Resc. – přeškálování na rozpětí 0–100.

	Sturnus vulgaris	Fringilla coelebs	Parus major	počáteční skóre	WA 1	resc.	WA 2	resc.	WA 3	resc.	WA 4	resc.	WA 5	resc.
A	3	5	1	1	17.8	0.0	26.0	0.0	28.3	0.0	28.9	0.0	29.0	0.0
B	5	4	3	2	33.3	94.2	43.6	100.0	46.4	100.0	47.2	100.0	47.4	100.0
C	2	3	2	3	34.3	100.0	41.3	87.2	43.3	82.6	43.8	81.4	43.9	81.2
WA 1	1.9	1.8	2.2											
resc.	20.0	0.0	100.0											
WA 2	67.1	56.4	80.4											
resc.	44.5	0.0	100.0											
WA 3	67.4	55.1	79.1											
resc.	51.4	0.0	100.0											
WA 4	66.5	54.0	77.5											
resc.	53.2	0.0	100.0											
WA 5	66.3	53.7	77.1											
resc.	53.7	0.0	100.0											

Metoda váženého průměrování vychází z náhodně zvolených čísel přiřazených ke každému vzorku. Výsledek není ovlivněn volbou počátečných hodnot, je možné zvolit libovolné nenulové číslo, pro každý vzorek však rozdílné. Tyto hodnoty budeme označovat jako počáteční skóre vzorků (*site scores*). Další kroky výpočtu jsou:

- Výpočet skóre druhů (*species scores*) jako vážené průměry skóre vzorků, přičemž váhami jsou y_{ij}, tj. početnosti druhů ve vzorcích.
- Restandardizace skóre druhů. V tomto kroku je možné použít libovolné lineární přeškálování, např. převedení na škálu od 0 do 100. Toto zajišťuje, aby se rozpětí hodnot při iterativním procesu nezmenšovalo.
- Výpočet nových skóre vzorků jako vážené průměry ze skóre druhů všech druhů vyskytujících se v daném vzorku. I zde platí, že váhy druhů jsou jejich početnosti.
- Restandardizace skóre vzorků.

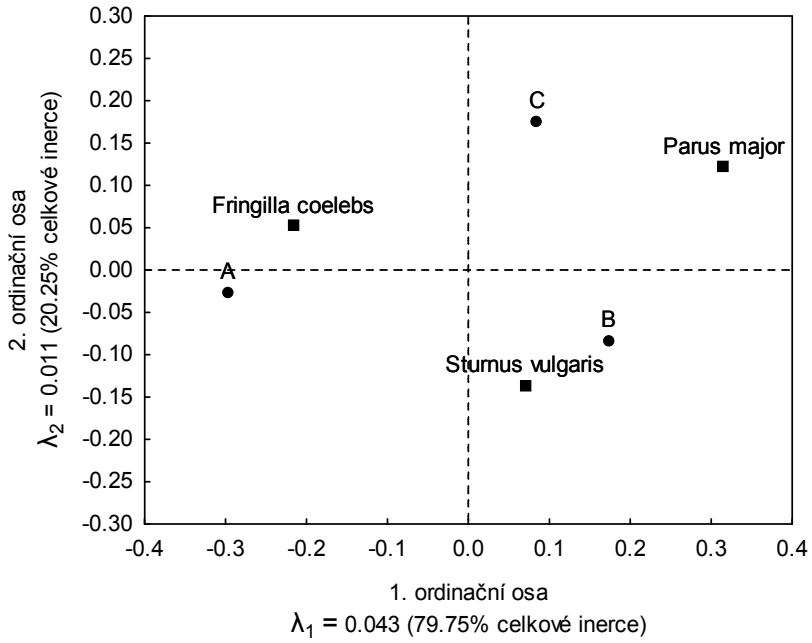
Algoritmus pokračuje recipročním průměrováním a restandardizací skóre vzorků a druhů, dokud mezi dvěma iteracemi již nedojde k žádné zjevné změně ve skóre druhů a vzorků (Tabulka 6-4, Obrázek 6.9). Při procesu váženého průměrování platí, že výpočet konverguje ke stejnému výsledku bez ohledu na zvolené počáteční hodnoty. Výsledkem je první osa korespondenční analýzy: skóre, čili souřadnice všech vzorků a druhů na první ose korespondenční analýzy.



Obrázek 6.9 Algoritmus korespondenční analýzy (CA) a detrendované korespondenční analýzy (DCA) (podle Palmera 1993).

Výpočet druhé a dalších os je složitější, ovšem principiálně stejný jako je uvedeno výše. Algoritmus výpočtu druhé osy je doplněný o krok, který zajistí lineární nezávislost první a druhé osy, podobně je výpočet třetí osy doplněný o krok zajišťující její lineární nezávislost s prvními dvěma osami atd.

Ordinační diagram na Obrázek 6.10 znázorňuje výsledek korespondenční analýzy tří ptačích druhů na třech lokalitách (data z Tabulka 6-4) v prostoru prvních dvou os. Z ordinačního diagramu je zřejmé, že pozice druhů i lokalit na první ordinační ose odpovídá vypočítaným skóre z Tabulka 6-4. V ordinačním diagramu je zřejmá vazba druhu *Fringilla coelebs* a lokality A a také druhu *Sturnus vulgaris* a lokality B; uvedené druhy byly na těchto lokalitách nejpočetnější. Různé software škálují skóre řádkových a sloupcových kategorií různě, toto škálování ovšem neovlivňuje interpretaci výsledku.



Obrázek 6.10 Ukázka ordinačního diagramu. Pozice druhů a vzorků v prostoru prvních dvou os korespondenční analýzy.

Grafické znázornění vztahů, které získáme z korespondenční analýzy, je založeno na myšlence reprezentovat všechny sloupce a řádky a interpretovat relativní pozice bodů jako váhy příslušné danému sloupci a řádku. Systém nalezených ordinačních os tedy bude poskytovat souřadnice každého sloupce a řádku, které můžeme zobrazit v jednom grafu, ordinačním diagramu – biplotu. Z biplotu můžeme poznat, které sloupcové kategorie jsou více důležité v řádkových kategoriích a naopak. V takovém grafu můžeme interpretovat vzdálenosti mezi řádkovými kategoriemi a vzdálenosti mezi sloupcovými kategoriemi, ne ovšem vzdálenosti mezi řádkovými body a sloupcovými body. Můžeme ale interpretovat relativní pozici bodu z jedné sady s ohledem ke všem bodům druhé sady. Pro ***ordinační diagram*** obecně platí, že:

- blízkost dvou řádků (sloupců) značí podobný profil v těchto dvou řádcích (pojmem profil označujeme distribuci podmíněné četnosti);
- pokud jsou od sebe řádky či sloupce vzdáleny, jejich profil je značně odlišný;
- blízkost určitého řádku a určitého sloupce znamená, že tento řádek má důležitou váhu v daném sloupci;
- pokud jsou od sebe určitý řádek a sloupec daleko, nejsou v daném sloupci téměř žádná pozorování, která přísluší danému řádku;
- body poblíž středu ordinačního diagramu nemají výrazný profil; střed ordinačního diagramu je těžištěm bodů jak řádkových, tak sloupcových kategorií.

V ordinačním diagramu jsou řádky i sloupce původní matice (v našem případě druhy a vzorky) znázorněny body. Pozice druhů v ordinačním prostoru představuje jeho optimum vzhledem k ordinačním osám. Ordinační osy představují teoretické gradienty. Korespondenční analýza předpokládá unimodální závislost druhů na gradientu tvořeném ordinačními osami.

V ordinačním diagramu z výše uvedeného vyplývá:

- vzorky, které mají podobné druhové složení, budou v ordinačním diagramu umístěny poblíž sebe;
- vzorky, které nemají společné druhy, budou v ordinačním diagramu umístěny dále od sebe;
- druhy, které se vyskytovaly spolu ve vzorcích, budou v ordinačním diagramu umístěny poblíž sebe;

- druhy, které se vyskytovaly v jiných vzorcích, budou v ordinačním diagramu umístěny dále od sebe;
- druhy umístěny poblíž vzorků byly pro tyto vzorky typické, resp. se vyskytovaly pouze v nich;
- když se druh v daném vzorku nevyskytoval, budou od sebe v ordinačním diagramu vzdáleny.

Požadavky na data a omezení korespondenční analýzy:

- Korespondenční analýza se používá ke zpracování kontingenčních tabulek, které obsahují pouze pozitivní hodnoty nebo nuly. Pouze pro takovou kontingenční tabulku lze určit podmíněné pravděpodobnosti. CA nemůže být použita na data obsahující negativní hodnoty. Data proto nesmí být centrována nebo standardizována.
- Kontingenční tabulka nesmí obsahovat řádek s celkovým součtem nula ani sloupec s celkovým součtem nula.
- CA je citlivá na odlehlé hodnoty.
- Data by měla být dimenzionálně homogenní, měřeny ve stejných jednotkách. Pouze v takovém případě je smysluplné hodnotit vzdálenosti mezi řádky a mezi sloupci matice. Při řádových rozdílech hodnot vstupní matice se doporučuje logaritmická transformace.

6.3.2 Detrendovaná korespondenční analýza (detrended correspondence analysis, DCA)

Při zpracování ekologických dat korespondenční analýzou často dochází ke dvěma problémům.

1. Vzorky nacházející se na koncích první osy jsou si svojí pozicí navzájem bližší než vzorky, které se nacházejí ve střední části osy.
2. Druhové složení je výborně vysvětlené seřazením vzorků a druhů podél první osy a důležitost druhé osy by měla být minimální. Tak tomu ovšem není a skóre vzorků na druhé osy vykazují kvadratický vztah s jejich skóre na první ose. Tento nedostatek označujeme jako obloukový efekt (*arch effect*). Označení podkovový efekt (horseshoe effect) je běžnější u PCA a ne u CA, kde koncové body nemají tendenci ohýbat se dovnitř. (Obrázek 6.11). K obloukovému efektu dochází v případech, když máme velké množství dvojic vzorků, které nemají společný žádný druh.

Obloukový efekt je matematický artefakt metody a nesouvisí s reálnou strukturou dat. Dochází k němu v případech, když první osa celkem vysvětluje druhová data. Pak je možné získat druhou osu přeložením první osy ve středu a složením jejích konců k sobě. Takto poskládaná osa není lineárně závislá s první osou. I když je v datech skutečný druhý gradient, korespondenční analýza jej neodhalí jako druhou osu když je rozptyl menší než rozptyl upravené přeložené první osy.

V takových případech se doporučuje použít detrendovanou formu korespondenční analýzy – detrendovanou korespondenční analýzu (*detrended correspondence analysis*, DCA).

Detrendovaná korespondenční analýza je odvozena od korespondenční analýzy a liší se od ní pouze v jednom kroku, kdy probíhá detrendování (Obrázek 6.9). Detrendování se týká druhé, třetí a dalších ordinačních os. Detrendování odstraňuje obloukový efekt (Obrázek 6.11, Obrázek 6.12).

Detrendování je možné dvěma různými metodami.

- Detrendování segmenty. K detrendování druhé osy metodou segmentace je první osa rozdělena na segmenty a vzorky uvnitř každého segmentu jsou centrovány tak, aby měly nulový průměr na druhé ose. Postup je opakován pro různé „startovací body" segmentů. Výsledky jsou v některých případech citlivé na počet segmentů. Detrendování dalších os se děje podobným procesem.
- Detrendování polynomem. Jde o nalezení polynomické rovnice, kterou vysvětlujeme vztahy objektů, a odčítání jejího vlivu. Je to elegantnější forma detrendování než metoda segmentace. Nejdříve je provedena regrese tak, aby druhá osa byla polynomickou funkcí první osy. Pak je druhá osa nahrazena rezidui z této regrese. Podobný postup je použit pro třetí a další osy. Bohužel, výsledky detrendování polynomy nemusí být vyhovující, a proto bývá preferované detrendování segmenty.

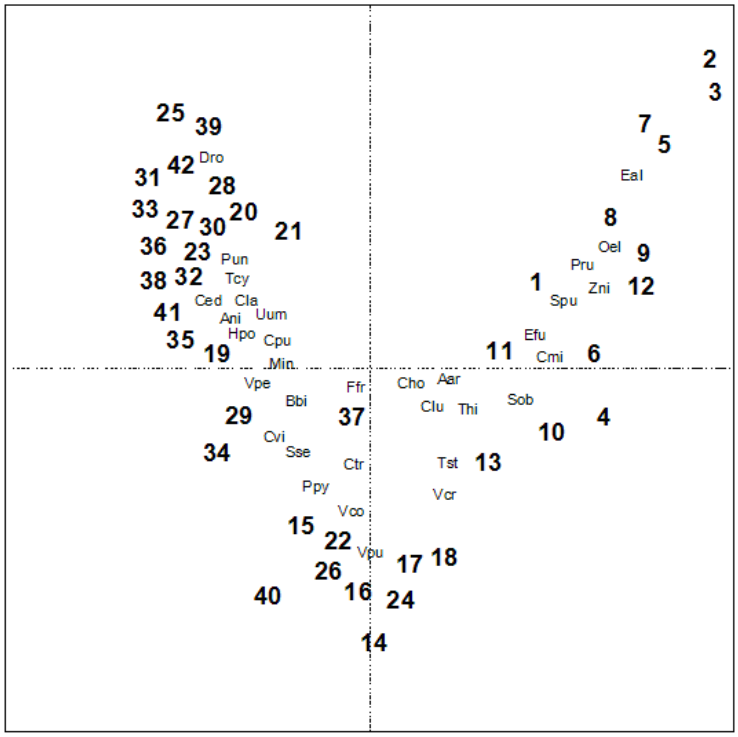
Přeškálování osy segmentací má následující důsledky:

- body na konci osy si již nejsou navzájem bližší než body uprostřed osy;
- unimodální křivky všech druhů mají standardizovanou toleranci 1 s.d., měřenou v násobcích směrodatné odchylky, tj. většina křivky prochází přes 4 s.d.;
- délka ordinační osy je měřena v násobcích směrodatné odchylky (s.d.).
- Při detrendování segmenty je tedy možné odměřit délku gradientu os. Je užitečné vědět, že když je délka první osy blízká 4 s.d., můžeme předpokládat, že vzorky na opačných koncích první osy nemají společný žádný druh.
- Požadavky na data vstupující do DCA jsou stejné jako u CA, jelikož jde o modifikaci CA.
- Kromě ekologických studií se detrendovaná korespondenční analýza uplatnila např. při analýze behaviorálních dat.

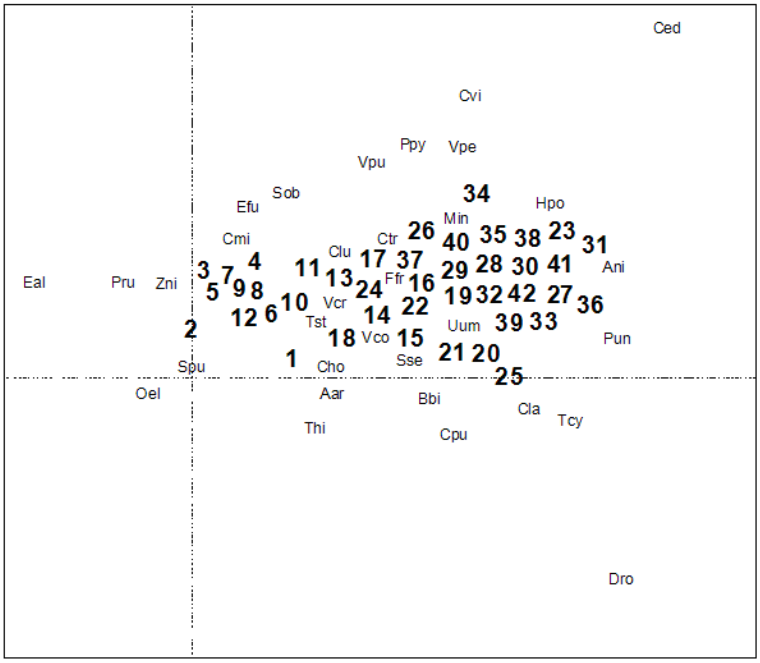
Detrendování by se nemělo používat automaticky, ale pouze při prokazatelném obloukovém efektu v CA. I přesto je DCA jednou z nejpopulárnějších metod analýzy ekologických dat.

Jako příklad uvádíme analýzu 33 druhů měkkýšů ze 42 lesních lokalit od měkkých lužních lesů přes přechodné lužní lesy až po tvrdé lužní lesy (lokality jsou v Obrázek 6.11 a Obrázek 6.12 označeny čísly 1 až 42). Početnost druhů byla vyjádřena čtyřstupňovou škálou podle hodnot dominance. V souboru dat se vyskytovaly skupiny lokalit, které neměly společný žádný druh. Gradient první osy byl dlouhý (3,485 s.d.), což se projevilo ve výsledku korespondenční analýzy jako obloukový efekt (arch effect; Obrázek 6.11). Proto byla k analýze těchto dat použita detrendovaná korespondenční analýza (Obrázek 6.12), která je vhodnější pro komplexní ekologická data reprezentující celou délku gradientu.

Výpočet první osy detrendované korespondenční analýzy je naprosto stejný jako u korespondenční analýzy. Detrendování se týká až druhé, třetí a dalších os. Interpretace první osy CA a DCA je proto naprosto stejná. První ordinační osa jak u CA, tak u DCA představuje vlhkostní gradient (lze si všimnout postupnosti čísel lokalit od nízkých hodnot – měkký lužní les až po vysoké hodnoty – tvrdý lužní les, Obrázek 6.11, Obrázek 6.12).



Obrázek 6.11 Výsledek korespondenční analýzy 33 druhů měkkýšů (označeny zkratkou názvu) na 42 lokalitách (označeny číslem 1-42). $\lambda_1 = 0.547$, $\lambda_2 = 0.174$, první a druhá osa CA vysvětlují 38.8 % celkové inerce. V ordinačním diagramu je zřetelně vidět obloukový efekt (arch effect). (Čejka et al. 2008)



Obrázek 6.12 Výsledek detrendované korespondenční analýzy 33 druhů měkkýšů (označeny zkratkou názvu) na 42 lokalitách (označeny číslem 1-42). $\lambda_1 = 0.547$, $\lambda_2 = 0.122$, první a druhá osa CA vysvětlují 36.0 % celkové inerce. Délka gradientu první osy je 3.485 s.d., druhé osy 1.943 s.d. (Čejka et al. 2008)

6.3.3 Korespondenční analýza a detrendovaná korespondenční analýza: shrnutí

- Vstup korespondenční analýzy:
  - Kontingenční tabulka.

- Výstup korespondenční analýzy:
  - Vlastní hodnoty matice.
  - Procento vysvětlené inerce ordinačními osami.
  - Skóre (souřadnice) řádků a sloupců na ordinačních osách.
  - Ordinační diagram kombinující jak skóre řádků, tak i skóre sloupců v ordinačním diagramu – tzv. biplotu.

- Při použití korespondenční analýzy je nutno pamatovat na níže uvedená omezení:
  - Velký počet malých skupin vzorků může způsobit problematickou interpretaci výsledků a nestabilitu výpočtu.
  - Problémem korespondenční analýzy může být tzv. obloukový efekt, který je možné odstranit pomocí detrendované korespondenční analýzy.

6.4 Analýza hlavních koordinát (PCoA, principal coordinate analysis, metric multidimensional scaling)

Mnohorozměrné škálování (MDS, *multidimensional scaling*) se používá jako průzkumná metoda. Cílem analýzy je najít smysluplné dimenze, které umožňují vysvětlit pozorované vzdálenosti (nepodobnosti) nebo podobnosti mezi objekty.

Jednoduchou metrickou technikou mnohorozměrného škálování je analýza hlavních koordinát (PCoA, *principal coordinate analysis*), nazývaná i klasické škálování. PCoA pracuje s maticí vzdáleností a výsledkem je rozmístění objektů v novém prostoru definovaném ordinačními osami. Podobně jako se ordinační osy u PCA nazývají hlavní komponenty, u PCoA je označujeme hlavní koordináty (*principal coordinates*). PCoA je podobná analýze hlavních komponent (PCA), umožňuje ovšem použití i jiných měr vzdáleností než euklidovské vzdálenosti. Při použití euklidovské vzdálenosti je PCoA ekvivalentní k PCA.

PCoA je vhodná pro zpracování všech typů proměnných – binárních proměnných, vícestavových kvalitativních proměnných nebo smíšených dat. Analýza hlavních koordinát zahrnuje dva základní kroky.

- V prvním kroku se z primární matice dat vypočítá asociační matice vzdáleností objektů, která je symetrická (je ekvivalentní korelační nebo kovarianční matici v PCA).
- Ve druhém kroku se podobně jako u PCA vypočítají vlastní čísla, vlastní vektory asociační matice a komponentní skóre.

Interpretace výsledků PCoA je podobná jako u PCA. Výrazným rozdílem je ovšem skutečnost, že hlavní koordináty nejsou lineární kombinací původních proměnných. Proto není možné určit vliv původních proměnných na jednotlivé hlavní koordináty. Je ovšem možné vypočítat korelace nebo kovariance mezi hlavními koordináty a proměnnými a pomocí nich interpretovat hlavní koordináty.

Euklidovská vzdálenost objektů v prostoru hlavních koordinát (v ordinačním diagramu) je u PCoA aproximací vzdálenosti objektů v asociační matici, která může být založena na libovolném koeficientu vyjadřujícím vztah mezi objekty. PCoA tedy vytváří takové rozmístění objektů v euklidovském prostoru (na ordinačním diagramu), které co nejlépe odráží vztahy mezi objekty v asociační matici; jde tedy o nejlepší euklidovskou aproximaci neeuklidovské matice. PCoA je

ovšem citlivá na použité metrice vzdálenosti, tj. při různých koeficientech vzdálenosti budou výsledky analýzy různé. V případě použití pseudometrických nebo nemetrických vzdáleností může nastat, že je jedna nebo více vlastních hodnot matice negativní. V takovém případu můžou nastat problémy s interpretací výsledku.

PCoA se velmi často používá v biologii, kdy charakter dat vyžaduje použití jiné míry vzdálenosti než je euklidovská vzdálenost. Jde zejména o analýzu binárních nebo smíšených dat.

PCoA se často používá v analýze molekulárních dat, kde nezřídka dochází k tomu, že počet proměnných převyšuje počet objektů. Analýzu je možné použít i v takovém případě.

Nevýhodou PCoA je, že hlavní koordináty nelze jednoduše interpretovat pomocí původních proměnných.

6.5 Nemetrické mnohorozměrné škálování (NMDS, nonmetric multidimensional scaling)

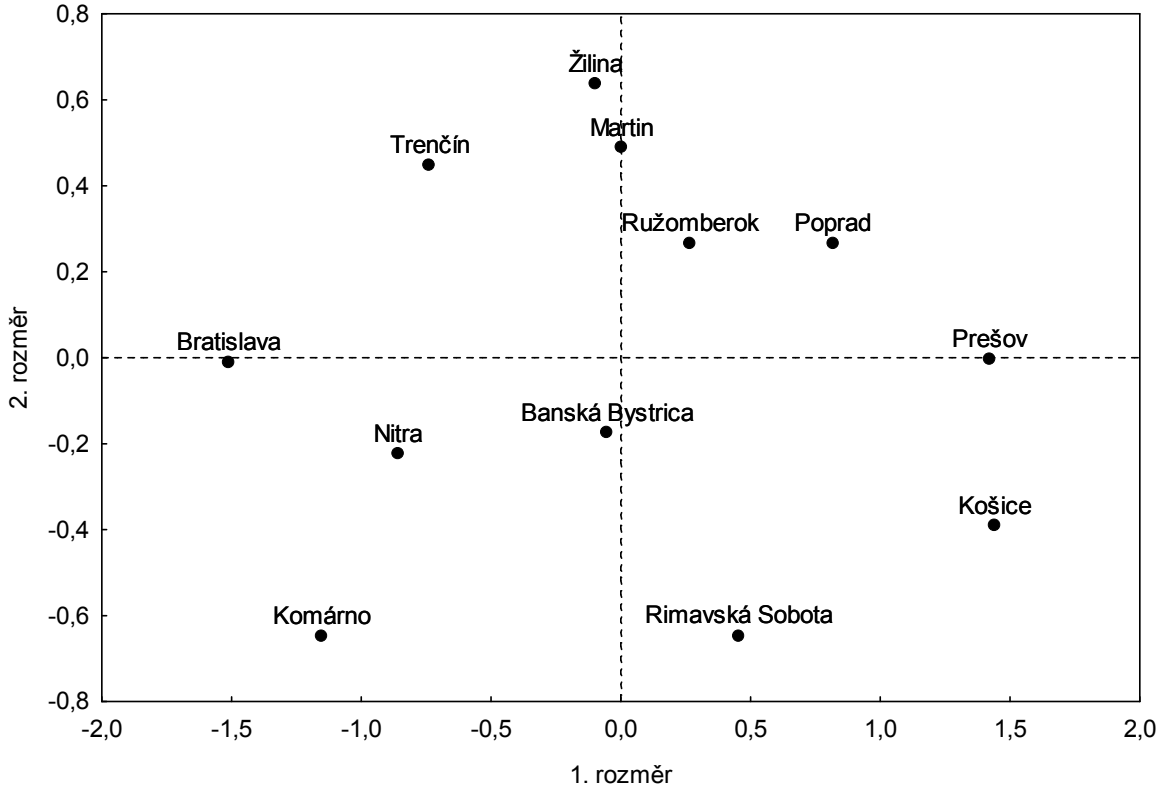
Nemetrické mnohorozměrné škálování (NMDS, *nonmetric multidimensional scaling*) analyzuje libovolnou metrickou nebo semimetrickou matici vzdálenosti nebo podobnosti. NMDS zobrazí pozorované vzdálenosti nebo podobnosti mezi objekty v euklidovském prostoru.

Pomocí následujícího příkladu demonstrujeme princip mnohorozměrného škálování. Předpokládejme, že máme k dispozici matici vzdáleností měst Slovenska z mapy. Naším cílem bude reprodukovat vzdálenosti mezi městy v dvourozměrném prostoru (Tabulka 6-5). Obecně, NMDS seřadí objekty (města Slovenska) v prostoru s určitým počtem rozměrů – dimenzí tak, aby byly zachovány pozorované vzdálenosti (Obrázek 6.13). Z výsledků NMDS budeme schopni vysvětlit vzdálenosti ve smyslu ordinačních os, v našem případě můžeme vysvětlit vzdálenosti pomocí dvou geografických rozměrů: sever/jih a východ/západ. Aktuální orientace os je náhodná.

Vraťme se k našemu příkladu. Otáčením mapy libovolným směrem se vzdálenosti mezi městy nemění. Výsledná orientace os v rovině nebo prostoru je většinou výsledkem subjektivního rozhodnutí tak, aby byl výsledek co nejjednodušeji interpretovatelný.

Tabulka 6-5 Ukázka asociační matice – vzdálenosti měst Slovenska v km.

	Banská Bystrica	Bratislava	Komárno	Košice	Martin	Nitra	Poprad	Prešov	Rimavská Sobota	Ružomberok	Trenčín	Žilina
B. Bystrica	0	204	188	214	92	119	124	208	105	53	139	117
Bratislava	204	0	100	402	227	85	328	412	273	257	124	202
Komárno	188	100	0	342	214	69	312	396	213	241	160	238
Košice	214	402	342	0	234	317	120	36	129	195	337	259
Martin	92	227	214	234	0	145	114	198	171	39	103	25
Nitra	119	85	69	317	145	0	243	327	188	172	91	169
Poprad	124	328	312	120	114	243	0	84	133	75	217	139
Prešov	208	412	396	36	198	327	84	0	165	159	301	223
R. Sobota	105	273	213	129	171	188	133	165	0	140	208	196
Ružomberok	53	257	241	195	39	172	75	159	140	0	142	64
Trenčín	139	124	160	337	103	91	217	301	208	142	0	78
Žilina	117	202	238	259	25	169	139	223	196	64	78	0



Obrázek 6.13 Výsledek mnohorozměrného škálování (příklad z Tabulka 6-5)

Cílem nemetrického mnohorozměrného škálování (NMDS) je stejně jako v případě metrického mnohorozměrného škálování (PCoA) vytvořit na základě asociační matice s libovolnou metrikou její euklidovskou reprezentaci. NMDS se na rozdíl od PCoA neomezuje na euklidovskou geometrii, pracuje s jakoukoliv maticí podobnosti nebo vzdálenosti, buď symetrickou, nebo nikoliv. Může to být i přímé hodnocení podobnosti nebo vzdálenosti na semikvantitativní škále. Vzdálenosti mezi objekty nemusí být metrické ani spojité. Metoda si dokáže poradit i s vyšším počtem chybějících hodnot v asociační matici, pokud zůstává dostatek informací k umístění každého objektu s ohledem na několik dalších objektů.

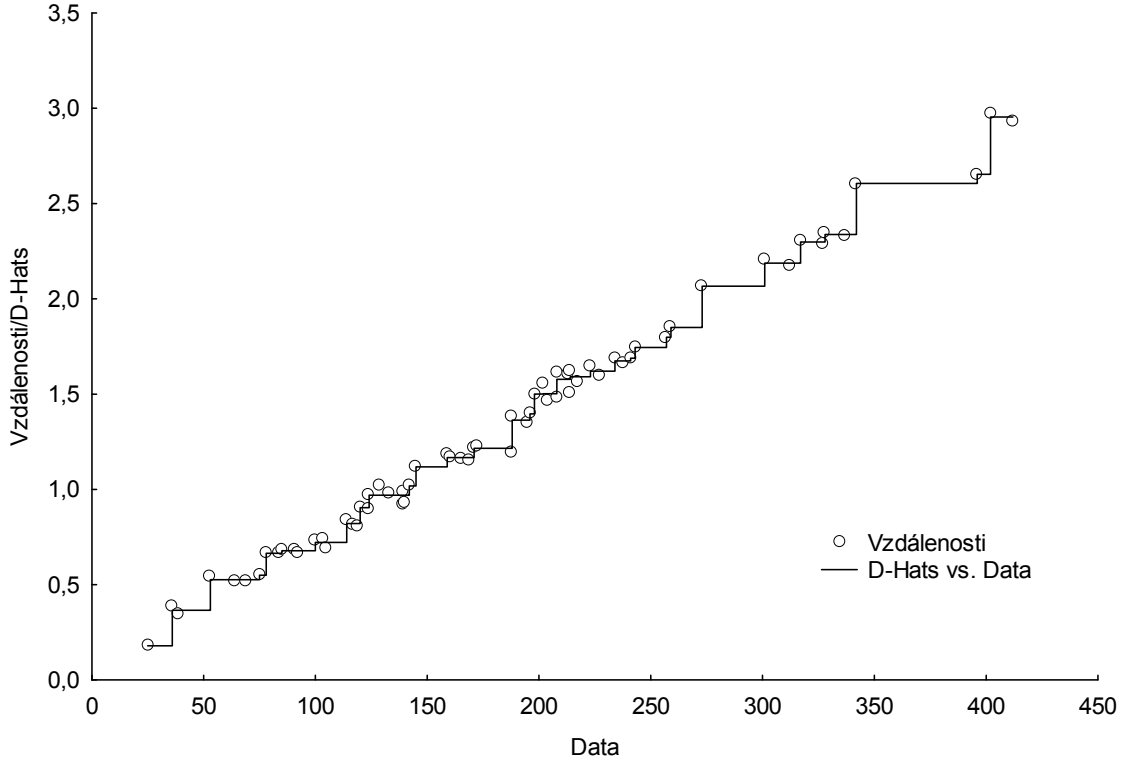
Hlavní odlišnost NMDS oproti PCoA je v tom, že tato technika se nesnaží o zachování přesných vzdáleností mezi objekty v původním prostoru proměnných, ale o prezentaci objektů v malém počtu rozměrů (dvou nebo třech). Namísto zachování přesných vzdáleností mezi objekty zachovává jen pořadí vzdáleností mezi objekty. Problémem metody je nutnost specifikovat počet ordinačních os (dimenzí) předem. NMDS je citlivé vůči nesprávnému stanovení dimenzionality. Výsledkem je výpočet souřadnic všech objektů pro tyto osy. Jde o iterativní proces.

Obecně platí, že se objekty nedají seřadit tak, aby byly v redukovaném prostoru vzájemné vzdálenosti mezi nimi stejné, jako jsou spočítány hodnoty vzdálenosti/nepodobnosti. Proto je zavedená míra, která jednoduchým číslem vyjadřuje, jak dobře nebo jak špatně korespondují vzdálenosti v redukovaném prostoru s hodnotami vzdáleností/nepodobností. Tato míra se nazývá funkce stresu (*loss function* nebo *stress function*). Nabývá hodnot od nuly do jedné; čím je hodnota nižší, tím je výsledek lepší. V průběhu iterativního algoritmu je analýza modifikovaná tak, aby byl minimalizovaný stres. V NMDS různé počáteční nastavení vedou k různým výsledkům vzhledem k lokálním minimům ve funkci stresu (*stress function*). Proto se doporučuje provést více analýz s různým nastavením dimenzí, které chceme extrahovat. Z těchto pokusů pak zvolíme analýzu s minimální hodnotou stresu.

Algoritmus NMDS je následující:

- Uživatel specifikuje počet dimenzí (*t*) a volí přiměřenou míru vzdálenosti/nepodobnosti.
- Vypočítá se matice vzdáleností.
- Je určeno počáteční uspořádání objektů v *t*-rozměrném prostoru.
- Vypočítá se míra stresu (stres – záměna mezi pořadím vzdáleností v asociační matici a pořadím vzdáleností v ordinaci NMDS).
- Objekty jsou mírně posunuty ve směru snížení stresu.
- Předchozí dva kroky se opakují, až dokud hodnota stresu nedosáhne minimum. Finální uspořádání objektů může být rotováno.

Vztah původního a redukovaného prostoru je možné sledovat pomocí Shepardova diagramu. Je to grafické znázornění reprodukovaných vzdáleností pro určitý počet rozměrů vůči pozorovaným vstupním datům (vzdálenostem). Na vertikální ose (y) jsou znázorněny vzdálenosti v ordinačním prostoru, na horizontální ose (x) původní vzdálenosti, případně podobnosti (Obrázek 6.14). V diagramu jsou znázorněny i tzv. D-hat hodnoty, které jsou výsledkem monotónní transformace vstupních dat. V případě, že všechny reprodukované vzdálenosti spadají na linii D-hat hodnot, řazení vzdáleností (nebo podobností) perfektně reprodukuje dané řešení. Tak je tomu v našem příkladě (Obrázek 6.14).



Obrázek 6.14 Ukázka Shepardova diagramu (příklad měst Slovenska, Tabulka 6-5, Obrázek 6.13)

Obecně platí, že čím víc dimenzí používáme k reprodukci matice vzdáleností, tím lépe reprodukovaná matice vysvětluje pozorované vzdálenosti v původních datech (tj. tím menší je stres). Skutečně, když použijeme tolik rozměrů, kolik je proměnných, perfektně reprodukujeme pozorovanou matici vzdáleností. Samozřejmě naším cílem je redukce pozorovaných dat, tj. vysvětlit matici vzdáleností pomocí menšího počtu dimenzí. Vraťme se k příkladu vzdáleností mezi městy. Když máme dvourozměrnou mapu, jsou vizualizované vzdálenosti mezi městy o mnoho informativnější, než je samotná matice vzdáleností.

Výsledkem NMDS je finální uspořádání objektů, tj. určení skóre všech objektů pro *t* dimenzí. Uspořádání je závislé na počtu zvolených dimenzí (*t*). První dvě osy z třírozměrného řešení nemusí být nutně podobné dvourozměrnému řešení.

Připomeňme si, že v NMDS je pořadí os náhodné: první osa není nutně důležitější než druhá osa, atd. Proto je někdy užitečné výsledek zrotovat (např. metodou varimax), ačkoliv není možné tvrdit, že výsledné řešení představuje nějaký „gradient".

Výhodou NMDS je:

- Možné použití nemetrické vzdálenosti,
- Možné použití nesymetrické matice.
- V případě metrických vzdáleností NMDS sumarizuje vzdálenosti v méně dimenzích než škálování v PCoA.

Mnohorozměrné škálování může sloužit pro přípravu podkladů pro shlukovací metodu *k*-průměrů (*k-means clustering*) pokud není možné na data použít euklidovskou vzdálenost.

Metoda nemetrického mnohorozměrného škálování je nejenom praktická metoda, ale v současnosti do jisté míry i módní záležitost.

6.5.1 Mnohorozměrné škálování: shrnutí

- Vstup mnohorozměrného škálování:
  - Matice vzdáleností/podobnosti objektů.

- Výstup mnohorozměrného škálování:
  - Ordinační diagram.
  - Skóre (souřadnice) objektů na ordinačních osách.

- Při použití mnohorozměrného škálování je nutno pamatovat na níže uvedená omezení:
  - Latentní dimenze (NMDS) stejně jako hlavní koordináty (PCoA) nejsou lineárně závislé na hodnotách původních proměnných.
  - NMDS je velice citlivá na výběr metriky vzdálenosti.

7 Kanonická ordinační analýza

7.1 Úvod

V kapitole 6 jsme představili ordinační metody, které slouží zejména jako průzkumné metody odhalující trendy v datech. Kapitola 7 je věnována technikám kanonické ordinační analýzy, které vyhodnocují vztah mezi dvěma sadami proměnných. Postupně si představíme několik kanonických ordinačních metod:

- Kanonická korespondenční analýza (CCA, *canonical correspondence analysis*) je asymetrická metoda, která pomocí mnohorozměrné regrese zjišťuje, do jaké míry je skupina závislých proměnných vysvětlená skupinou nezávislých proměnných. Používá se k modelování vztahu mezi nelineárními závislými proměnnými a sadou nezávislých proměnných.
- Redundanční analýza (RDA, *redundancy analysis*) podobně jako CCA zjišťuje závislost jedné skupiny proměnných od druhé skupiny proměnných. Je vhodná v takových případech, kdy dvě sady proměnných mají lineární vztah.
- Kanonická korelační analýza (CCorA, *canonical correlation analysis*) je symetrická metoda, která hledá maximální lineární korelaci mezi dvěma sadami proměnných.
- Diskriminační analýza (DFA, *discriminant analysis*) se zabývá diskriminací skupin.

7.2 Kanonická korespondenční analýza (CCA, canonical correspondence analysis)

Kanonická korespondenční analýza (CCA) je kanonickou, čili omezenou formou korespondenční analýzy (CA). Vstupními daty pro CCA jsou dvě sady proměnných: matice nezávislých proměnných **X** (v ekologii např. environmentální data měřena ve vzorcích) a matice závislých proměnných **Y** (v ekologii např. zastoupení druhů ve vzorcích). CCA používá mnohorozměrnou regresi k určení lineární kombinace proměnných, která nejlépe vysvětluje inerci ordinačních skóre získaných ze závislých proměnných. Podobně jako tomu bylo u CA, i CCA maximalizuje inerci skóre závislých proměnných, ovšem tak, aby ordinační osy byly lineární kombinací nezávislých proměnných. Proto ordinačním osám říkáme kanonické, nebo omezené. Právě kvůli tomuto omezení jsou vlastní hodnoty kanonických os CCA menší než v CA.

Nezávislé, vysvětlující proměnné nemusí nutně splňovat předpoklady rozdělení, neměly by ovšem obsahovat odlehlé hodnoty ani být výrazně asymetrické.

V některých případech je ovšem vhodné transformovat nezávislé proměnné. CCA není ovlivněna lineární transformací nezávislých proměnných, nelineární transformace dat ovšem již ovlivňuje výsledek analýzy. Nezávislé proměnné jsou před vstupem do analýzy standardizovány.

Do CCA můžeme zařadit vysvětlující proměnné více typů:

- kvantitativní (spojité)
- semikvantitativní
- nominální (kategoriální).

V případě, že do CCA vstupují nominální vysvětlující proměnné, je potřeba uvádět je ve formě tzv. indikátorových (*dummy*) proměnných (Tabulka 7-1), tj. každá kategorie bude zastoupena jednou proměnnou nabývající hodnoty 0 (ne – vlastnost nepřítomna) a 1 (ano – vlastnost přítomna).

Tabulka 7-1 Ukázka přepisu kategoriální proměnné na indikátorové proměnné pro použití v CCA.

	Původní proměnná		Indikátorové proměnné (dummy variables)		
Vzorek	Rybí pásmo	Kód	Lipanové pásmo	Parmové pásmo	Cejnové pásmo
1	lipanové	1	1	0	0
2	lipanové	1	1	0	0
3	parmové	2	0	1	0
4	cejnové	3	0	0	1
5	cejnové	3	0	0	1
6	parmové	2	0	1	0
7	lipanové	3	1	0	0

Pro každou kategoriální proměnnou s *K* kategoriemi můžeme do analýzy zařadit pouze *K - 1* indikátorových proměnných. Problém totiž nastává při lineární závislosti skupiny proměnných. Je zřejmé, že součet hodnot všech indikátorových proměnných pro každý vzorek je rovný jedné. Proto jedna z indikátorových proměnných nebude do analýzy zařazena. Přitom ovšem nedochází k žádné ztrátě informace; když odstraníme proměnnou cejnové pásmo, informace o něm zůstává, protože cejnové pásmo se vyskytne v každém vzorku, kde není lipanové a parmové pásmo. Některé software (např. Canoco) odstraní nadbytečnou indikátorovou proměnnou automaticky.

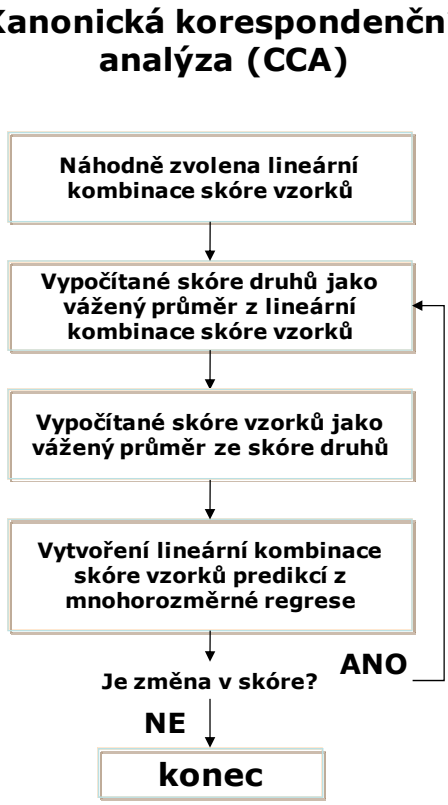
CCA je široce používaná v ekologii k modelování kanonických vztahů mezi druhovým složením a měřenými proměnnými prostředí.

Algoritmus výpočtu CCA si představíme jako rozšíření algoritmu váženého průměrování na příkladu ekologických dat. Závislé proměnné tvoří početnosti nebo frekvence druhů ve vzorcích a nezávislé proměnné jsou proměnné prostředí měřeny v těch samých vzorcích.

Podobně jako u detrendované korespondenční analýzy (DCA), i pro kanonickou korespondenční analýzu (CCA) platí, že se od korespondenční analýzy (CA) liší pouze v jednom kroku (Obrázek 7.1). Tento nový krok je ovšem do algoritmu přidaný ne kvůli odstranění nežádoucího efektu, ale proto, aby bylo možné vysvětlit kanonické osy pomocí konkrétních nezávislých proměnných.

V CCA jsou skóre vzorků, které jsou determinovány váženým průměrováním druhů, dále podrobeny mnohorozměrné lineární regresi, do které tyto skóre vzorků vstupují jako závislá proměnná a environmentální proměnné jako nezávislé proměnné. Nové skóre vzorků jsou predikované regresní rovnicí. Tato regresní rovnice je lineární kombinací proměnných. Nové skóre vzorků budeme označovat jako LC skóre, na rozdíl od skóre určeném váženým průměrováním, které označíme WA (Obrázek 7.1).

Řešení CCA je nejběžněji získáno algoritmem váženého průměrování, což je ekvivalentní řešení vlastní analýzy. Nicméně je algoritmus váženého průměrování intuitivně lépe pochopitelný, a proto ho zde uvádíme.



Obrázek 7.1 Algoritmus kanonické korespondenční analýzy (CCA) (podle Palmera 1993).

Statistický model, na kterém je založena CCA, předpokládá unimodální odezvu druhů na gradient prostředí. CCA je aproximací Gausovské regrese za určitých předpokladů.

CCA, stejně jako CA, není vhodná pro extrémně krátké gradienty, na kterých mají abundance druhů nebo jejich frekvence lineární nebo monotónní vztah ke gradientu.

Výsledkem CCA jsou dvě sady skóre vzorků. Není jednoznačné, které jsou vhodnější pro použití v ordinačním diagramu, zdali WA skóre, nebo LC skóre. Ve většině situací se doporučuje použití WA skóre. LC skóre jsou získány přímo z mnohorozměrné regrese nezávislých proměnných, a tak mají dva vzorky identické LC skóre, když mají stejné hodnoty nezávislých proměnných, a to i tehdy, když nemají společný žádný druh.

Kanonických ordinačních os je tolik, kolik je nezávislých proměnných. Další osy jsou konstruovány jako neomezené. Může nastat situace, že první neomezená osa má vyšší vlastní hodnotu než první kanonická – omezená osa. Neomezené osy jsou velice užitečné v explorativní analýze, můžou naznačit, které důležité proměnné pravděpodobně chybí.

Celková vysvětlená inerce je suma vlastních hodnot kanonických os. Celková inerce závislých proměnných (druhových dat) je suma vlastních hodnot kanonických a ordinačních os (omezených a neomezených os) a je ekvivalentní sumě vlastních hodnot, nebo celkové inerci v CA. Proto můžeme vysvětlenou inerci ve srovnání s celkovou inerci použít jako míru, která hodnotí, jak dobře jsou závislé proměnné vysvětleny nezávislými proměnnými.

Když se počet proměnných blíží počtu objektů (vzorků), vysvětlená inerce se blíží celkové inerci a výsledek CCA se blíží výsledku CA. V takovém případě již ordinace není omezená proměnnými a může např. dojít k obloukovému efektu, jak tomu bývá u CA. Obloukový efekt je možné v CCA elegantně odstranit vyloučením proměnných, které korelují s druhou osou. Existuje i další varianta CCA známá jako detrendovaná kanonická korespondenční analýza (DCCA), která ve svém algoritmu zahrnuje detrendování i lineární regresi. Detrendování by ovšem nemělo být v CCA potřebné.

Pro ***interpretaci*** ordinačního diagramu CCA platí stejná pravidla jako při CA, co se týče rozmístění objektů a závislých proměnných (např. vzorků a druhů). Na rozdíl od CA nejsou

ovšem kanonické osy v CCA teoretickými gradienty, ale lineární kombinací nezávislých proměnných. Ordinační diagram, který nazýváme triplot, zobrazuje vzorky jako body, druhy jako body a nezávislé proměnné jako vektory (nebo body). Kvantitativní nezávislé proměnné jsou v ordinačním diagramu znázorněny vektory s počátkem ve středu souřadnicové soustavy. Směřování vektoru proměnné udává směr nárůstu hodnot této proměnné, opačný směr udává směr poklesu hodnot dané proměnné. Pozice vektoru nezávislé proměnné vzhledem ke kanonické ose je dána jejich vzájemnou korelací. Podobně i vzájemná pozice nezávislých proměnných v ordinačním diagramu odráží korelační koeficienty mezi těmito proměnnými. Nezávislé proměnné s delším vektorem jsou silněji korelované s ordinačními osami než proměnné s krátkými vektory. V případě kategoriálních proměnných jsou kategorie znázorněny body umístěnými v centroidu vzorků patřících k dané kategorii.

Nezávislé proměnné jsou mezi sebou porovnatelné, protože byly standardizovány. Kanonické osy můžeme interpretovat buď na základě kanonických koeficientů, nebo na základě korelací nezávislých proměnných s osami. Oba přístupy poskytují stejné informace v případě, že proměnné nejsou korelované. Když jsou nezávislé proměnné silně korelovány mezi sebou (např. proto, že se počet proměnných blíží k počtu objektů), kanonické koeficienty nejsou stabilní. Korelace ovšem netrpí problémem multikolinearity. V případě silně korelovaných nezávislých proměnných se doporučuje ponechat v analýze pouze jednu proměnnou ze skupiny. Vlastní hodnoty tímto klesnou pouze nepatrně. Při výraznějším poklesu vlastních hodnot došlo pravděpodobně k vyloučení příliš mnoha proměnných, případně k vyloučení nesprávných proměnných.

V mnohých případech bývá triplot CCA příliš přeplněný. V takových případech jsou následovné možnosti:

- Rozdělit triplot na biploty nebo diagramy zobrazující pouze jeden typ informace (druhy, vzorky, nebo nezávislé proměnné).
- Přeškálování vektorů tak, aby pozice druhů a vzorků byly více rozestoupené.
- Zobrazení pouze nejpočetnějších druhů (je ovšem vhodné zachovat vzácné druhy v analýze).
- Nezobrazovat skóre vzorků. Jsou pouze lineární kombinací vysvětlujících proměnných. Zobrazení pozic vzorků je užitečné pro identifikaci odlehlých hodnot.

Testování hypotéz

Výhodou CCA je možnost testovat hypotézy. Testování hypotéz v CCA je možné pomocí permutačního testu. První vlastní hodnota (případně také suma všech vlastních hodnot) je porovnána s příslušnou statistikou získanou z náhodných permutací dat. Tyto permutace nemění aktuální data, pouze náhodně přiřadí data vysvětlující proměnné k hodnotám vysvětlované proměnné. Když je příslušná statistika větší nebo rovna 95 % statistik z permutovaných dat, můžeme zamítnout nulovou hypotézu, že závislé proměnné nemají vztah k nezávislým proměnným. Testování první vlastní hodnoty určuje, zda je první osa CCA silnější než náhodně vytvořená osa. Podobně testování sumy všech kanonických os hodnotí, zda existuje celkový vztah mezi závislými a nezávislými proměnnými.

Součástí některých software, např. Canoco for Windows, je možnost výběru statisticky významných proměnných ze skupiny nezávislých proměnných. Takové proměnné mají statisticky významný vztah k matici závislých proměnných. Postup výběru statisticky významných proměnných se označuje termínem „forward selection“ a pracuje s použitím Monte-Carlo permutačního testu. Jelikož CCA je omezená ordinace a výsledek silně závisí na tom, zda máme k dispozici správné nezávislé proměnné; doporučuje se vždy otestovat statistickou významnost první kanonické osy. K tomu rovněž slouží Monte-Carlo permutační test, jenž je součástí software Canoco.

Software Canoco umožňuje i další sofistikované metody, jako je např. parciální kanonická analýza, při níž lze odčítat vliv určitých nezávislých proměnných a hodnotit pouze vliv ostatních nezávislých proměnných na matici závislých proměnných.

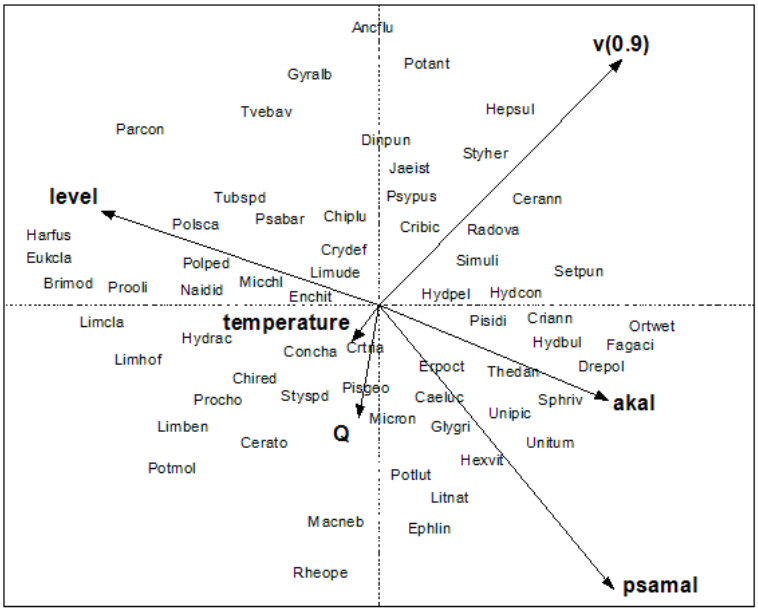
Většina omezení CCA je stejných, jako je tomu u mnohorozměrné regrese. Proto je nutné si uvědomit, že:

- korelace neznamená kauzální vztah, a proměnná, která se zdá být silná může mít jen vztah k neměřenému ale „skutečnému“ gradientu;
- vysoce korelované proměnné by neměly být do analýzy zařazeny; jejich vliv je velice náročné interpretovat;
- když se počet proměnných blíží počtu objektů, řešení analýzy není již omezeno proměnnými a analýza je neomezená;
- interpretovatelnost výsledků je přímo závislá na volbě a kvalitě vysvětlujících proměnných;
- přestože mnohorozměrná regrese i CCA hledají lineární kombinaci vysvětlujících proměnných, nemáme záruku nalézt skutečný gradient, který může být vztažen k neměřené nebo neměřitelné proměnné.

Při použití CCA se vždy doporučuje provést také CA na matici závislých proměnných (matice Y). Když jsou skóre řádků a sloupců matice v CA podobně umístěny v ordinačním diagramu, jako je tomu v CCA, můžeme být spokojeni, že měřené nezávislé proměnné vysvětlují podstatnou část inerce závislých proměnných.

Jako příklad kanonické korespondenční analýzy uvádíme analýzu společenstva makrozoobentosu makrozoobentosu zahrnující 63 taxonů a 60 vzorků. Jako vysvětlujících proměnných bylo použito 13 použito 13 proměnných charakterizujících vodní režim a základní fyzikálně-chemické podmínky. podmínky. V průběhu analýzy bylo permutačními testy vybráno šest proměnných, které měly statisticky statisticky významný vliv na druhová data (Obrázek 7.2,

Tabulka 7-2).



Obrázek 7.2 Výsledek kanonické korespondenční analýzy (CCA): 63 taxonů makrozoobentosu (označeny zkratkou názvu) v prostoru prvních dvou os CCA vytvořených jako lineární kombinace šesti environmentálních proměnných (plné názvy proměnných jsou uvedeny v

Tabulka 7-2).

Tabulka 7-2 Výsledek CCA 63 taxonů makrozoobentosu a šesti environmentálních proměnných.

	1. osa	2. osa	
Vlastní hodnota kanonické osy	0,367	0,267	
Kumulativní % vysvětlené inerce nezáv. proměnných	7,8	13,4	
Celková inerce			4,727
Suma kanonických vlastních hodnot			1,150
Korelace nezávislých proměnných s kanonickými osami			
temperature: teplota vody (°C)	-0,068	-0,096	
level: stav vodní hladiny (cm)	-0,738	0,250	
Q: průtok vody (m3 s-1)	-0,054	-0,300	
v(0.9): rychlost proudu 0.9 m pod hladinou (m s-1)	0,646	0,655	
akal: substrát dna – štěrk (%)	0,612	-0,250	
psamal: substrát dna – písek (%)	0,625	-0,755	

Výsledkem CCA je ordinační diagram, ve kterém jsou druhy i vzorky znázorněny body, kvantitativní proměnné vektory, kategoriální proměnné centroidy kategorií (Obrázek 7.2; vzorky nejsou znázorněny, žádné kategoriální proměnné nebyly použity).

Kanonické osy jsou lineární kombinací vybraných proměnných prostředí. Interpretace kanonických os je v případě CCA přímá a většinou se opírá o korelaci nezávislých proměnných s kanonickými osami. V našem příkladě lze první kanonickou osu interpretovat jako gradient velikosti toku: od toků s vysokou vodní hladinou a nízkou rychlostí proudu k tokům s nízkou vodní hladinou a vysokou rychlostí proudu (korelace proměnné stav vodní hladiny s první kanonickou osou: -0,738, rychlost proudu 0,9 m pod hladinou: 0,646). Gradient druhé kanonické osy nejlépe charakterizuje korelace s proměnnou psamal: písečný substrát dna (korelace proměnné psamal s druhou kanonickou osou: -0,755).

Vlastní hodnota první kanonické osy byla $\lambda_1 = 0,367$, vlastní hodnota druhé osy $\lambda_2 = 0,267$. První dvě osy vysvětlují 13,4 % inerce druhových dat. Vybrané proměnné prostředí vysvětlují celkem 24,3 % inerce druhových dat (1,150/4,727*100%).

Z grafu je velice dobře vidět vazba jednotlivých taxonů k vlastnostem prostředí charakterizovaném vybranými environmentálními proměnnými. Je potřebné si ovšem uvědomit, že vztahy jsou pouze popisné, ne kauzální.

7.3 Redundanční analýza (RDA, redundancy analysis)

Redundanční analýza (RDA) je kanonickou, neboli omezenou formou analýzy hlavních komponent (PCA). Vstupem do RDA jsou dvě sady proměnných: matice nezávislých proměnných **X** (např. environmentální data) a matice závislých proměnných **Y** (např. druhová data).

Cílem RDA je maximalizovat odpověď sady závislých proměnných **Y** na sadu nezávislých proměnných **X**. Metoda je v podstatě rozšířením PCA o krok, ve kterém jsou skóre objektů sady závislých proměnných omezeny tak, aby byly lineární kombinací sady nezávislých proměnných. RDA je proto úzce spjatá s mnohorozměrnou regresní analýzou a dává podobné výsledky jako kanonická korelační analýza. RDA je možné popsat jako mnohorozměrnou regresní analýzu následovanou analýzou hlavních komponent, a to v těchto krocích:

- Regrese každé závislé proměnné Y_i na sadě nezávislých proměnných **X** pomocí mnohorozměrné regrese a získání regresních koeficientů.

- PCA na sadě regresních koeficientů z mnohorozměrné regrese a získání matice kanonických vlastních vektorů.
- Použití kanonických vlastních vektorů k získání skóre objektů buď ve faktorovém prostoru **X** nebo prostoru závislých proměnných **Y**. Skóre v prostoru závislých proměnných jsou známé jako vážené průměry (WA), zatímco skóre ve faktorovém prostoru jsou známé jako lineární kombinace (LC). V mnohých aplikacích jsou WA skóre důležitější a lépe interpretovatelné.

Výsledek RDA je možné zobrazit v biplotu, který se skládá z bodů objektů a z vektorů obou sad proměnných. Kosinus úhlu mezi vektory proměnných je odhadem korelačního koeficientu mezi těmito proměnnými. V případě většího počtu proměnných nebo objektů je vhodné zobrazit dva ordinační diagramy, a to pro každou sadu proměnných samostatně.

Pro RDA platí stejné předpoklady a omezení jako pro PCA. Je nutno ovšem zdůraznit, že RDA je založena na lineární mnohorozměrné regresi a PCA, a proto by měla být použita na úplně lineární datové soubory.

Podobně jako předchozí metody i RDA bývá používána v ekologii společenstev. Na rozdíl od korespondenční analýzy a jejích odvozených forem se PCA a RDA používají v případech, kdy očekáváme lineární vztah mezi abundancemi nebo frekvencemi druhů k proměnným. V mnoha případech není možné tento předpoklad dodržet, dá se předpokládat pouze na krátkém ekologickém gradientu. Proto není použití RDA v ekologických studiích významně časté.

7.4 Kanonická korelační analýza (CCorA, canonical correlation analysis)

Kanonická korelační analýza (CCorA) hodnotí vztah mezi dvěma sadami kvantitativních proměnných. Zjišťuje, zda se jedna skupina proměnných chová stejně jako druhá skupina proměnných pro ty samé objekty a když ano, co je podstatou této shody. Vstupem do CCorA jsou dvě matice proměnných, které můžeme považovat za vzájemně závislé proměnné, nebo přistupujeme k jedné matici jako k vysvětlujícím, nezávislým proměnným a ke druhé matici jako k vysvětlovaným, závislým proměnným. V druhém případě je CCorA velice podobná RDA.

Podobně jako u analýzy hlavních komponent a ve faktorové analýze se i v kanonické analýze transformuje systém vzájemně korelovaných proměnných do systému nových hypotetických (skrytých) proměnných. V CCorA se vztah mezi dvěma skupinami vzájemně závislých proměnných vyjadřuje pomocí menšího počtu nově vytvořených proměnných. Tyto nové proměnné jsou lineárními funkcemi původních proměnných a jsou založeny na analýze kovariančních nebo korelačních matic výchozích proměnných.

CCorA hledá lineární kombinaci proměnných z první sady a lineární kombinaci proměnných z druhé sady, které mají maximální korelaci mezi sebou. CCorA měří tedy intenzitu lineární závislosti, tj. korelovanosti lineárních funkcí dvou skupin proměnných. Výsledná korelace lineárních kombinací dvou sad proměnných se nazývá kanonická korelace a je odmocninou vlastní hodnoty matice v CCorA.

CCorA vytváří kanonickou funkci, která maximalizuje kanonické korelační koeficienty mezi dvěma lineárními kombinacemi proměnných. Označme počet proměnných v první sadě k a počet proměnných ve druhé sadě n. Když k je větší než n, existuje k možných vlastních hodnot, přičemž $k - n$ z nich jsou nulové. První kanonická korelace je největší možná korelace mezi lineárními kombinacemi první sady proměnných a lineárními kombinacemi druhé sady proměnných. K ní přísluší kanonická funkce, která je první kanonickou osou. Další kanonické osy jsou nekorelované s předchozími kanonickými osami.

Kanonická korelační analýza je generalizací mnohorozměrné lineární regrese. CCorA na rozdíl od mnohorozměrné regrese nehledá závislost jedné závislé proměnné na sadě nezávislých proměnných, ale vztah dvou sad proměnných. Když je k rovno jedné, dostáváme pouze jednu

kladnou vlastní hodnotu a kanonické korelační rovnice jsou redukovány na problém mnohorozměrné regrese.

Kanonickou osu interpretujeme pomocí kanonických vah, tj. korelací jednotlivých proměnných a jejich příslušných lineárních kombinací. Kanonické váhy jsou podobné faktorovým váhám proměnných a faktorů ve faktorové analýze.

Výsledek CCorA je možné graficky zobrazit v biplotu, ze kterého je možné vidět přibližnou kovarianci, resp. korelaci mezi oběma skupinami proměnných podobně jako v RDA. Když je CCorA počítána z korelační matice a ne z kovarianční matice, interpretace musí brát v úvahu fakt, že lineární kombinace se vztahují k standardizovaným proměnným a ne k původním. Požadavky na data:

- data musí být kvantitativní;
- metoda je citlivá na odlehlé hodnoty; požadavek normality ovšem není silný;
- počet proměnných první sady plus počet proměnných druhé sady proměnných musí být menší, než je počet objektů;
- proměnné mají mít mezi sebou lineární vztah (což je zřídka možné předpokládat v ekologických studiích).

Přestože CCorA není tak populární jako jiné kanonické techniky, její použití může být užitečné např. při hodnocení změn stavu dvou skupin proměnných stejného typu (můžeme např. korelovat dvě taxocenózy s cílem zjistit, zda se mění stejným způsobem), nebo při zjišťování korelace mezi skupinou fyzikálních proměnných a skupinou druhů, apod. Uplatnění CCorA můžeme najít nejen v biologii, ale i v medicíně (např. hodnocení vztahu skupiny rizikových faktorů a skupiny symptomů nemoci), v psychologii, sociologii, atd.

7.4.1 Kanonická analýza: shrnutí

- Vstupem kanonické analýzy je:
  - Matice závislých proměnných (kontingenční tabulka např. druhy x vzorky).
  - Matice nezávislých proměnných měřených na stejných objektech.

- Výstupem kanonické analýzy je:
  - Ordinační diagram.
  - Vlastní hodnoty kanonických os.
  - Procento vysvětleného rozptylu kanonickými osami.
  - Skóre (souřadnice) řádků a sloupců matice závislých proměnných na kanonických osách.
  - Korelace vysvětlujících proměnných s kanonickými osami.

- Při použití kanonické analýzy je nutné pamatovat na níže uvedená omezení:
  - Velký počet malých skupin objektů může způsobit problematickou interpretaci výsledků a nestabilitu výpočtu.
  - Při použití nevhodných vysvětlujících proměnných nebudou výsledky analýzy relevantní.

7.5 Diskriminační analýza (Discriminant function analysis, Canonical variate analysis)

Častým cílem v přírodních i sociálních vědách je diskriminovat již známé skupiny objektů na základě několika kvantitativních proměnných. Důvodem pro to může být přiřazení nového

objektu do jedné ze skupin (identifikace), nebo interpretovat dané skupiny, tj. určit vlastnosti jednotlivých skupin (diskriminace).

Diskriminační analýza je velice užitečný nástroj:

- k určení proměnných, které diskriminují mezi dvěma nebo více skupinami (k tomu slouží kanonická diskriminační analýza),
- ke klasifikaci objektů do různých skupin (k tomu slouží klasifikační diskriminační analýza).

Zabývá se tedy závislostí jedné kvalitativní proměnné (určuje zařazení objektů do skupin) na několika kvantitativních proměnných. Vstupní matici tvoří objekty charakterizovány sadou kvantitativních proměnných a jednou kategoriální proměnnou, která určuje příslušnost objektů do jedné ze skupin (Tabulka 7-3).

Tabulka 7-3 Ukázka datové tabulky vstupující do diskriminační analýzy. Objekty příslušející dvou skupinám jsou charakterizovány dvěma různými kvantitativními proměnnými (podle Legendre, Legendre 1983).

ID	kvalitativní proměnná (skupina)	kvantitativní proměnná 1 y_1	kvantitativní proměnná 2 y_2
1	A	3	5
2	A	3	7
3	A	5	5
4	A	5	7
5	A	5	9
6	A	7	7
7	A	7	9
8	B	6	2
9	B	6	4
10	B	8	2
11	B	8	4
12	B	8	6
13	B	10	4
14	B	10	6

Diskriminační analýza je parametrická metoda lineárního modelování.

Jejími předpoklady jsou:

- proměnné charakterizující každou skupinu by měly splňovat požadavek mnohorozměrného normálního rozdělení (techniky diskriminační analýzy jsou vůči odchylkám od normality celkem robustní, jsou ovšem citlivé na odlehlé hodnoty; statistické testy normalitu předpokládají);
- shoda skupinových kovariančních matic;
- proměnné, které se použijí k diskriminaci skupin, nemohou být úplně redundantní, tj. žádná z proměnných nesmí být lineární kombinací jiných proměnných;
- pro počty skupin (g), počet proměnných (p), počty objektů ve skupinách a celkové počty objektů v analýze (n) musí platit:
  - musí být alespoň dvě skupiny objektů: $g \geq 2$;
  - v každé ze skupin musí být nejméně 2 objekty;
  - počet proměnných musí být menší než počet objektů zmenšený o počet skupin: $0 < p < (n\text{-}g)$; doporučuje se ovšem, aby počet objektů v kterékoliv skupině byl výrazně vyšší než počet znaků;
  - žádná proměnná by neměla být v některé skupině konstantní.

7.5.1 Kanonická diskriminační analýza

Pro diskriminační analýzu má význam uvažovat pouze ty kvantitativní proměnné, u kterých byla zjištěna souvislost s kategoriální proměnnou (tj. byly zjištěny rozdíly mezi vektory středních hodnot v různých skupinách). Následně se hledá lineární kombinace proměnných, které nejlépe diskriminují mezi jednotlivými skupinami. Výpočet tak směruje k nalezení ***diskriminačních funkcí*** (kanonických os, *discriminant function, canonical root*) a k zjištění relativního příspěvku jednotlivých proměnných k celkové diskriminaci skupin. Počet diskriminačních funkcí je rovný počtu skupin snížený o jednu, případně počtu proměnných (v případě, že počet proměnných je menší než počet skupin snížený o jedničku).

V případě dvou skupin je analýza analogická mnohonásobné regresi a výsledkem je jedna ***diskriminační funkce***:

$$d = a + u_1 y_1 + u_2 y_2 + ... + u_p y_p \qquad (7.1)$$

kde a je konstanta a $u_1, \ldots u_p$ jsou koeficienty diskriminační funkce.

Proměnné s největšími (standardizovanými) koeficienty přispívají nejvíce k diskriminaci skupin. V případě více skupin je výsledkem více diskriminačních funkcí. Koeficienty pro první funkce se odvodí tak, aby skupinová těžiště (centroidy, průměry) byla maximálně vzdálená. Koeficienty vypočtené pro druhou funkci musí dále maximalizovat rozdíly mezi skupinovými centroidy a současně hodnoty obou funkcí nesmí být korelovány. Další funkce se odvozují stejným způsobem. Výsledek diskriminační analýzy více než dvou skupin, a tedy s nejméně dvěma diskriminačními funkcemi (kanonickými osami), lze graficky znázornit v ordinačním diagramu.

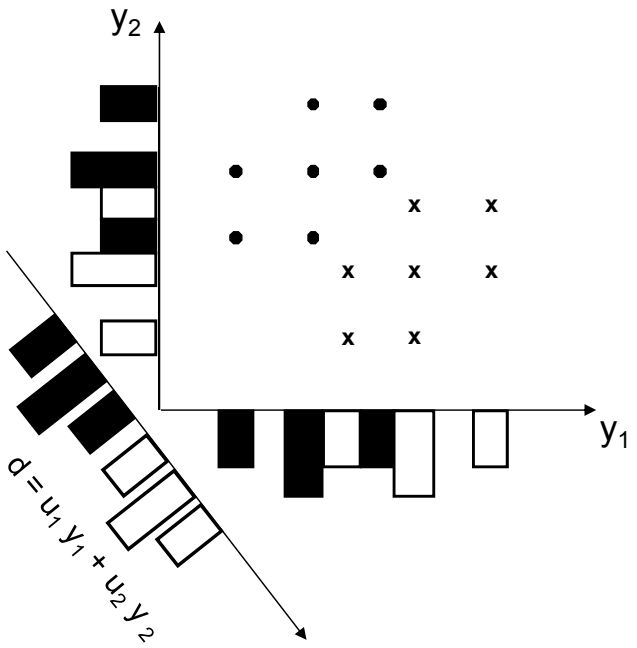
S diskriminačními funkcemi jsou spojeny vlastní čísla (*eigenvalues*), které určují míru rozptylu zachycenou těmito funkcemi. Diskriminační funkce (kanonické osy) bývají uspořádány podle klesajících vlastních čísel. Procentuální podíl vlastního čísla vzhledem k součtu všech vlastních čísel určuje důležitost diskriminační funkce (kanonické osy).

Výsledkem diskriminační analýzy dat v Tabulka 7-3 jsou tyto koeficienty (Tabulka 7-4):

Tabulka 7-4 Výsledky kanonické diskriminační analýzy dat z Tabulka 7-3.

	nestandardizované koeficienty	standardizované koeficienty	korelace proměnné s diskriminační funkcí (kanonickou osou)
y_1	-0,612	-1,0	-0,5
y_2	0,612	1,0	0,5
konstanta	0,612		
Vlastní hodnota	3,938	3,938	

Obě proměnné přispívají stejnou mírou k diskriminaci skupin, což je zřejmé také z následujícího obrázku (Obrázek 7.3).



Obrázek 7.3 Dvě skupiny, každá se sedmi objekty, nemůžeme oddělit pomocí proměnných y_1 nebo y_2 (histogramy na osách). Ovšem tyto skupiny lze ideálně oddělit pomocí diskriminační funkce **d** (podle Legendre, Legendre 1983).

Obrázek 7.3 a Tabulka 7-4 ukazuje idealizovaný příklad dvou skupin popsaných pouze dvěma kvantitativními proměnnými. Skupiny nemůžou být odděleny žádnou ze dvou proměnných. Řešením je nová diskriminační funkce **d**, která je lineární kombinací původních proměnných. Diskriminační funkce **d** přechází směrem největší meziskupinové variability.

Rozdíl mezi standardizovanými a nestandardizovanými koeficienty diskriminační funkce je následující:

- ***nestandardizované koeficienty*** diskriminační funkce (*unstandardized coefficients*) jsou závislé na použitém měřítku příslušné proměnné;
- ***standardizované koeficienty*** (*standardized coefficients*) vyjadřují jedinečný příspěvek proměnné pro diskriminační funkci, resp. jedinečný příspěvek pro oddělení skupin podél dané osy (diskriminační funkce).

K interpretaci diskriminačních funkcí (kanonických os) jsou nejdůležitější ***korelační koeficienty*** mezi jednotlivými proměnnými a diskriminačními funkcemi (kanonickými osami). Tyto korelační koeficienty pro příslušnou proměnnou se počítají bez ohledu na vliv ostatních proměnných na danou diskriminační funkci. Tím se tyto koeficienty významně liší od standardizovaných koeficientů diskriminační funkce. Pokud jsou původní proměnné vzájemně korelované, vyšší hodnota standardizovaného kanonického koeficientu bude přiřazena pouze jedné z dvojice nebo skupiny korelovaných proměnných.

Statistická významnost diskriminačních funkcí (os) bývá testována pomocí kritéria Wilks´ Lambda, chí kvadrát, případně poměr věrohodnosti (*likelihood ratio*), které testují hypotézu, že vektory středních hodnot jednotlivých skupin jsou totožné a že dané skupiny se podél příslušné osy a všech dalších os od sebe neliší.

7.5.2 Klasifikační diskriminační analýza

Klasifikační diskriminační analýza slouží k identifikaci objektů. Výsledkem jsou ***klasifikační funkce*** (*classification functions*), které můžou být použity k určení pravděpodobnosti příslušnosti objektů do skupin.

V tomto případě máme skupinu objektů se známým zařazením do skupin (trénovací soubor, informativní výběr) a skupinu objektů, které musíme zařadit do jedné ze skupin. Na základě trénovacího souboru sestavíme klasifikační funkce, pomocí kterých určíme pravděpodobnost zařazení neznámých objektů do skupin.

Jednou z možností odvození klasifikačního pravidla je výpočet lineární klasifikační funkce pro každou skupinu. Počet klasifikačních funkcí je tedy roven počtu skupin. Každá funkce umožní vypočítat klasifikační skóre pro každý objekt pro každou skupinu při použití vzorce:

$$s_i = c_i + w_{i1}y_1 + w_{i2}y_2 + ... + w_{ip}y_p\,, \qquad (7.2)$$

kde i určuje skupinu, *1,2, ... p* označují p proměnných, c_i je konstanta pro i-tou skupinu, w_{ij} je váha j-té proměnné ve výpočtu klasifikačního skóre pro i-tou skupinu; y_j je pozorovaná hodnota pro příslušný objekt a j-tou proměnnou, s_i je výsledné klasifikační skóre.

Objekt bude zařazen do skupiny, pro kterou klasifikační skóre dosáhne nejvyšší hodnoty. Klasifikační funkce můžou být použity přímo pro vypočítání klasifikačního skóre pro nové objekty. V našem příkladě (Tabulka 7-3) jsou klasifikační funkce následující:

$s_1 = -10.443 + 0.750y_1 + 2.250y_2$

$s_2 = -12.693 + 3.000y_1 - 0y_2$

Účinnost klasifikačního kriteria lze zjistit několika různými způsoby. V tomto případě je příslušnost všech studovaných objektů k jednotlivým skupinám známá.

- Resubstituce (*resubstitution*) – účinnost klasifikačního kriteria testujeme na stejném souboru dat, z něhož se toto klasifikační pravidlo odvozuje.
- Křížové ověření (*leave-one-out cross-validation*) – je vhodné v případě menšího počtu objektů. Ze souboru n objektů vybereme n – 1 objektů, které použijeme jako trénovací soubor, z něhož odvodíme klasifikační kritérium. Toto pak aplikujeme na jeden vypuštěný případ. Postup opakujeme n-krát.

Výsledkem obou způsobů je procentuální vyjádření úspěšnosti zařazení objektů do skupin sumarizované v tzv. klasifikační tabulce (*classification table*).

7.5.3 Diskriminační analýza: shrnutí

- Vstupem diskriminační analýzy je:
  - Tabulka objektů charakterizovaných několika kvantitativními proměnnými a jednou kategoriální proměnnou, která přiřazuje objektům příslušnost ke skupině.

- Výstupem diskriminační analýzy je:
  - Diskriminační funkce.
  - Klasifikační funkce.
  - Ordinační diagram (osy jsou kořeny, čili diskriminační funkce).

- Při použití diskriminační analýzy je nutno pamatovat na níže uvedená omezení:
  - Parametrická metoda, vyžaduje normální rozdělení proměnných v každé skupině.
  - Problém odlehlých hodnot.
  - Výsledky udává v pravděpodobnostech.
  - Není schopna zachytit nelineární vztahy mezi proměnnými.
  - Při použití silně korelovaných proměnných je nutné zvýšené opatrnosti při interpretaci koeficientů diskriminačních funkcí; silně redundantní proměnné mají vliv na stabilitu modelu a jeho koeficientů a pokud možno by v modelu neměly být používány společně.

8 Ordinační metody v ekologii společenstev

Ekosystémy jsou tvořeny mnoha biotickými a abiotickými složkami, které se navzájem ovlivňují. Způsob, jakým abiotické environmentální proměnné ovlivňují složení společenstev, je často zkoumán následujícím způsobem. Nejprve jsou vytipovány vzorky a zaznamenány vyskytující se druhy včetně jejich kvantity (abundance, frekvence). Jelikož počet druhů je zpravidla velký, používá se ordinační analýza na sumarizování a uspořádání dat v ordinačním diagramu. Ten je často interpretovaný podle toho, co je známo o prostředí ve vzorcích.

Když chybí jednoznačná environmentální data, k analýze dat používáme ordinační metody. Interpretace ordinačních os je nepřímá, proto tuto skupinu analýz můžeme označit jako ***nepřímá gradientová analýza*** (*indirect gradient analysis*). Když byla naměřena environmentální data, můžeme analyzovat vztah druhových dat a proměnných prostředí pomocí kanonických ordinačních metod. Interpretace výsledků je v tomto případě formální, přímá, proto kanonické ordinační analýzy označujeme jako ***přímá gradientová analýza*** (*direct gradient analysis*).

Ordinační, případně kanonické osy označujeme termínem ***gradienty***, nebo i trendy.

Cílem, kterého se snažíme v ekologických výzkumech dosáhnout pomocí ordinační analýzy, případně kanonické ordinační analýzy, je zformulovat hypotézy týkající se vztahů mezi druhovým složením společenstva a základními gradienty, které jsou buď teoretické, nebo určeny na základě environmentálních proměnných.

Vztah druhů k prostředí můžeme hodnotit na základě dvou modelů odpovědí druhu na gradient prostředí. Pod teoretickým gradientem si můžeme představit například vlhkost. Představme si vztah nějakého rostlinného druhu k vlhkosti prostředí. Z ekologie je známé, že druhy mají při určité hodnotě vlhkosti své optimum a při snižující či zvyšující se vlhkosti se jejich početnost, resp. pravděpodobnost výskytu snižuje. Při určitých hodnotách, které jsou pro daný druh suboptimální, se tento již nevyskytuje. Odpověď daného druhu na vlhkostní gradient je unimodální. V ordinačních metodách můžeme pracovat buď s lineární odpovědí druhu na gradient prostředí nebo s unimodální.

Lineární model předpokládá, že abundance nebo pravděpodobnost výskytu každého druhu buď roste, nebo klesá s hodnotami každé environmentální proměnné nebo gradientu.

Unimodální model předpokládá, že abundance nebo pravděpodobnost výskytu každého druhu má v rámci rozpětí hodnot každého gradientu optimum.

Přehled nejoblíbenějších metod používaných v analýze biologických společenstev je uveden v tabulce (Tabulka 8-1).

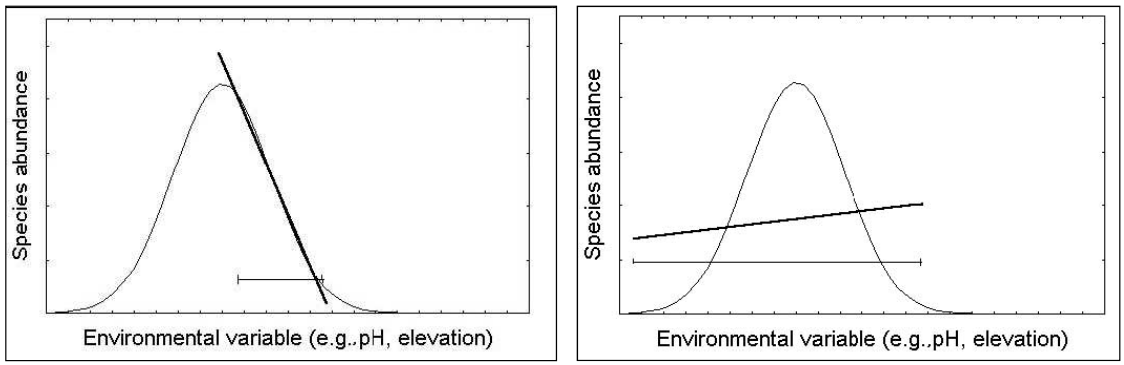
Tabulka 8-1 Rozdělení ordinačních metod nejčastěji používaných v ekologii společenstev.

	Model odezvy druhu na gradient	Metoda
Nepřímá gradientová analýza Ordinační osy (gradienty) jsou neomezené, jejich interpretace je nepřímá.	Lineární model	Analýza hlavních komponent (PCA)
	Unimodální model	Korespondenční analýza (CA) Detrendovaná korespondenční analýza (DCA)
	Nemetrická ordinace	Mnohorozměrné škálování (NMDS)
Přímá gradientová analýza Kanonické osy jsou lineární kombinací konkrétních environmentálních proměnných, jejich interpretace je přímá.	Lineární model	Redundanční analýza (RDA)
	Unimodální model	Kanonická korespondenční analýza (CCA) Detrendovaná kanonická korespondenční analýza (DCCA)

8.1 Unimodální a lineární model odezvy druhu na gradient prostředí

V této části představíme způsob, jak volit mezi unimodálním a lineárním modelem odezvy druhu na gradient prostředí.

Unimodální modely jsou o mnoho obecnější než modely monotónní (Obrázek 8.1), proto se doporučuje začít s unimodálním modelem a rozhodnout se později, zda si můžeme tento model zjednodušit na lineární.



a b

Obrázek 8.1 Unimodální křivka může být na krátkém gradientu dobře odhadnuta lineárním vztahem (a). Na delším gradientu lineární aproximace není účinná (b) (podle Lepš, Šmilauer 2000).

Abychom mohli rozhodnout o použití lineárního anebo unimodálního modelu, musíme odměřit délku nejdelšího gradientu. Nejdelší bývá gradient první ordinační osy. Délka gradientu se měří v násobcích směrodatné odchylky (s.d.). Druhová data jsou standardizována tak, že unimodální křivka probíhá přes 4 s.d. Proto u vzorků, které jsou od sebe vzdáleny 4 s.d., můžeme předpokládat, že nemají společný žádný druh. Doporučená volba mezi unimodálním a lineárním modelem je:

- když délka nejdelšího gradientu ≥ 4 s.d., volíme ***unimodální model***;
- když délka nejdelšího gradientu < 3 s.d., volíme ***lineární model*** (není ovšem nutnost použít lineární model).

Když je ovšem délka gradientu menší než 2 s.d., většina druhových křivek je monotónní a můžeme použít PCA nebo RDA. Výhodou použití PCA (případně její kanonické formy RDA) je, že zobrazení druhů a vzorků poskytuje víc kvantitativních informací než CA, DCA a (D)CCA. Nevýhodou PCA a RDA je předpoklad lineárních dat.

Zpravidla platí, že techniky váženého průměrování (CA, DCA, (D)CCA) jsou lepší pro heterogenní data, a techniky založené na modelu lineární odpovědi (PCA, RDA) jsou vhodné pro homogenní datové soubory.

8.2 Přímá a nepřímá gradientová analýza

V případě, že máme k dispozici pouze druhová data, pracujeme s metodami nepřímé gradientové analýzy. Přímou gradientovou analýzu můžeme použit až tehdy, když máme k dispozici environmentální proměnné. Není ovšem pravidlem, že když máme naměřeny environmentální proměnné, používáme vždy přímou gradientovou analýzu. I v tomto případě můžeme totiž použít nepřímou gradientovou analýzu a uplatnit v ní environmentální proměnné pouze externě k lepší interpretaci ordinační os. Tyto přístupy jsou komplementární a měly by se použít oba ke zhodnocení vzájemných pozic vzorků a druhů v přímé i nepřímé gradientové analýze. Když jsou si pozice vzorků a druhů podobné v obou výsledcích, environmentální proměnné spolehlivě vysvětlují druhová data.

8.3 Hybridní analýza

Jakýmsi „kříženc em“ mezi přímou a nepřímou ordinací je ***hybridní analýza***.

V případě, že máme k dispozici i druhová data i environmentální proměnné, můžeme použít přímou i nepřímou gradientovou analýzu. Za určitých podmínek je velice vhodné zkonstruovat několik os pomocí přímé gradientové analýzy. Tyto osy budou kanonické ordinační osy, čili omezené. Zbývající osy budou neomezené, vytvořeny pouze na základě druhových dat.

V přímé ordinaci je tolik omezených (kanonických) os, kolik je nezávislých vysvětlujících proměnných a až další ordinační osy jsou neomezené. V hybridní analýze předem definujeme počet kanonických ordinačních os (většinou to bývají dvě osy) a další ordinační osy jsou neomezené. Neomezené osy můžou naznačit další významné gradienty, které jsme environmentálními proměnnými nedokázali změřit. Je důležité porovnat vlastní hodnoty omezených a neomezených os. V hybridní analýze se může stát, že vlastní hodnota první neomezené osy je větší než vlastní hodnota první kanonické osy, co naznačuje silný gradient neomezený měřenými environmentálními proměnnými.

8.4 Parciální ordinační analýza

V případě, že nám je známý vliv nějaké proměnné, případně skupiny proměnných, na druhové společenstvo, a zajímá nás pouze variabilita, kterou touto skupinou proměnných neumíme vysvětlit, použijeme metody dílčí, tzv. ***parciální ordinace***. Skupinu proměnných, jejichž vliv na společenstvo v analýze oddělujeme, nazýváme ***kovariáty***. Parciální ordinace je možné použít na všechny metody, které jsme představili. Principem parciálních ordinací je oddělení vlivu kovariát a převedení analýzy pouze na zbývající, reziduální variabilitě. Vstupem do parciální ordinace je:

- matice druhů + matice kovariát (když používáme nepřímou ordinaci);
- matice druhů + matice kovariát + matice environmentálních proměnných (když používáme přímou ordinaci).

9 Seznam použité literatury

[1] Čejka, T., Horsák, M. & Némethová, D. The composition and richness of Danubian floodplain forest land snail faunas in relation to forest type and flood frequency. Journal of Molluscan Studies 74: 37-45. (2008)

[2] Davies, D. L., Bouldin, D. W. A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. 1 (4): 224-227. (1979)

[3] Digby, P.G.N., Kempton, R.A Multivariate analysis of ecological communities. Chapman and Hall, London – New York. (1987)

[4] Dunn, J. C. Well separated clusters and optimal fuzzy partitions. J.Cybern. 4: 95-104. (1974)

[5] Gnanadesikan, R. Methods for statistical data analysis of multivariate observations. John Wiley & Sons, New York – London – Sydney – Toronto. (1977)

[6] Goodman, L., Kruskal, W. Measures of associations for cross-validations. J. Am. Stat. Assoc. 49: 732-764. (1954)

[7] Hebák, P., Hustopecký, J. Vícerozměrné statistické metody s aplikacemi. SNTL, Alfa, Praha. (1987)

[8] Hebák, P., Hustopecký, J., Jarošová, E., Pecáková, I. Vícerozměrné statistické metody (1). 2. přepracované vydání, Informatorium, Praha, ISBN 9788073330569. (2007)

[9] Hill, M. O. Correspondence Analysis: A Neglected Multivariate Method. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 23, No. 3, pp. 340-354. (1974)

[10] Hubert, L., Schultz, J. Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie. 29: 190-241. (1976)

[11] Jongman, R.H., ter Braak, C.J.F., van Tongeren, O.F.R. Data analysis in community and landscape ecology. Pudoc, Wageningen. (1987)

[12] Kenkel, N. C., Derksen, D. A., Thomas, A. G., Watson, P. R. Multivariate analysis in weed science research. Weed Science, 50: 281–292. (2002)

[13] Latka, F. Minilexikon matematiky. Alfa, Bratislava, 158pp. (1981)

[14] Legendre, P., Legendre, L. Numerical Ecology, 2nd Engl. Ed., Elsevier, Amsterdam, ISBN 0444892494. (1998)

[15] Lepš, J., Šmilauer, P. Metody mnohorozměrné statistiky v analýze ekologických dat. Studijní materiál ke kursu. Biologická fakulta Jihočeské university, České Budějovice. (1994)

[16] Lepš, J., Šmilauer, P. Mnohorozměrná analýza ekologických dat. Biologická fakulta Jihočeské univerzity v Českých Budějovicích. České Budějovice. (2000)

[17] Lepš, J., Šmilauer, P. Multivariate Analysis of Ecological Data using CANOCO. Cambridge University Press. ISBN 0 521 81409 X hardback, ISBN 0 521 89108 6 paperback. (2003)

[18] Manly, B.F.J. Multivariate Statistical Methods. Second edition. Chapman & Hall. 232 pp. (1994)

[19] Marhold, K., Suda, J. 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. (2002)

[20] McGarigal, K., Cushman, S. & Stafford, S.G., Multivariate Statistics for Wildlife and Ecology Research, Springer, New York. (2000)

[21] Palmer, M. Ordination Methods for Ecologists. http://ordination.okstate.edu/ vstup 3.12.2010

[22] Palmer, M.W. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology 74: 2215-2230. (1993)
[23] Pauwels, E. J., Frederix, G. Finding salient regions in images: nonparametric clustering for image segmentation and grouping. Computer Vision and Image Understanding, 75: 73-85. (1999)
[24] Podani, J. 2001. SYN-TAX Computer program for data analysis in ecology and systematics. User's Manual. Scientia Publishing, Budapest. (2000)
[25] Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20: 53-65. (1987)
[26] StatSoft, Inc. STATISTICA (data analysis software system), version 7.1. www.statsoft.com. (2005)
[27] ter Braak, C. J. F., Šmilauer, P. CANOCO References Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4). Ithaca, NY, USA: Microcomputer Power. (1998)
[28] Tvrdík, J. 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 (2003)
[29] Urban, D.L. Multivariate analysis in ecology. Principal Components Analysis. http://www.env.duke.edu/lel/env358/mv_pca.pdf (2000)
[30] Urban, D.L. Multivariate analysis in ecology. Nonhierarchical agglomeration. http://www.env.duke.edu/lel/env358/mv_kmeans.pdf (2000)
[31] van der Lann, M. J., Pollard, K. S., Bryan, J. A New Partitioning Around Medoids Algorithm. Journal of Statistical Computation and Simulation 73: 575–584. (2002)
[32] Van Sickle, J. Using Mean Similarity Dendrograms to Evaluate Classifications, Journal of Agricultural. Biological and Environmental Statistics 2: 370 – 388. (1997)
[33] Wolda, H. Similarity Indices, Sample Size and Diversity. Oecologia (Berlin) 50: 296-302. (1981)
[34] Zvára, K. Biostatistika. Učební texty Univerzity Karlovy v Praze. Univerzita Karlova v Praze – Nakladatelství Karolinum. 212 pp. ISBN 80-7184-773-9. (2001)

10 Příloha – 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* x *p*. (Obrázek 10.1).

	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 10.1 Ukázka matice rozměru *n* x *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.

Mnohorozměrnými postupy lze analyzovat:

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

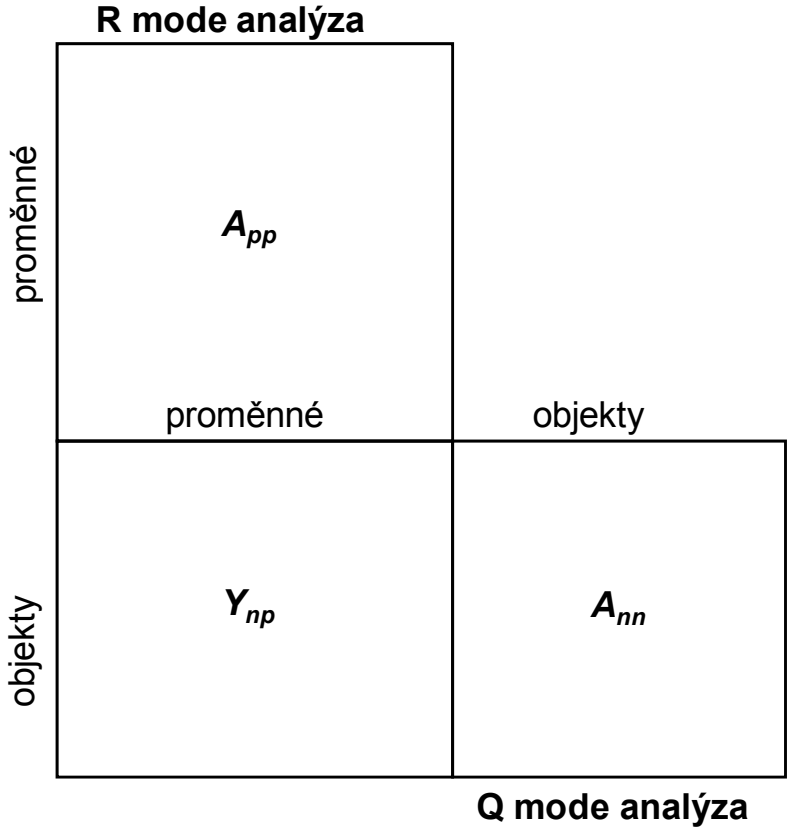
Matematické postupy aplikované při Q mode analýze jsou jiné než při R mode analýze. Např. korelační koeficient můžeme použít při sledování vztahů mezi proměnnými, nelze je ovšem použít pro vztah dvou objektů. Tady se používají jiné míry asociace, např. míry podobnosti. Výše uvedenou matici rozměru *n* x *p* můžeme zapsat ve tvaru

$$Y = \left[y_{ij}\right] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2P} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} \qquad (10.1)$$

.

10.1 Asociační matice

Asociační matice je v typickém případě čtvercová symetrická matice, kde sloupce a řádky odpovídají proměnným/objektům původní *n x 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 10.2). 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 10.2 Původní data tvořila matice Ynp rozměru *n* (objekty) x *p* (proměnné). Z této matice lze vytvořit dvě asociační matice App (proměnné x proměnné) a Ann (objekty x objekty) (podle Legendre, Legendre 1998).

Asociační matici mezi proměnnými označíme

$$A_{pp} = \left[a_{ij}\right] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2P} \\ \dots & \dots & \dots & \dots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix} \qquad (10.2)$$

asociační matici mezi objekty

$$A_{nn} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{10.3}$$

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

U asociační matice mezi objekty $\mathbf{A}_{nn}$ 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 $\mathbf{A}_{pp}$, kde je mírou asociace korelace, jsou hodnoty na diagonále a_{ii} rovny jedné.

10.2 Speciální matice

Matice se stejným počtem řádků a sloupců je **čtvercová**. Jak uvidíme dále, pouze pro takovou matici můžeme vypočítat determinant, inverzní matici, 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 (**I*B = B*I = B**).

Podobně **skalární matice** je diagonální matice formy

$\begin{bmatrix} k & 0 & ... & 0 \\ 0 & k & ... & 0 \\ ... & ... & ... & 0 \\ 0 & 0 & 0 & k \end{bmatrix} = kI$, 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 původní matice **B** rozměru *n* x *p* je označena $\mathbf{B}^T$. Její formát bude *p* x *n* a platí, že $b^T_{ij} = b_{ji}$. Jednoduše řečeno, řádky jedné matice jsou sloupci druhé matice. Např. transponovaná matice k matici

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

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

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

10.3 Vektory a normalizace

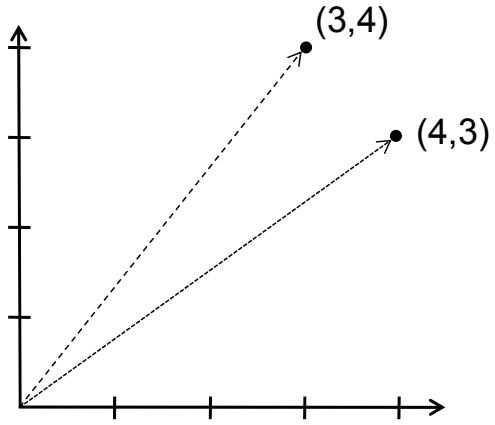
Sloupcová matice rozměru *n* x *1* se nazývá **vektor**.

Vektor zapíšeme následujícím způsobem:

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

Vektor je definován jako uspořádána *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 uspořádána dvojice reálných čísel (4, 3), kterou můžeme zakreslit do Euklidovského prostoru (Obrázek 10.3).



Obrázek 10.3 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 Pytagorovy 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

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

10.4 Sčítání a násobení matic

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

$$\mathbf{A} + \mathbf{B} = \mathbf{C}, \text{ kde } c_{ij} = a_{ij} + b_{ij} \qquad (10.4)$$

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: **A** + **B** = **B** + **A**
- Asociativnost: **A** + (**B** + **C**) = (**A** + **B**) + **C**
- Distributivnost: (*c* + *d*)**A** = *c***A** + *d***A**; *c*(**A** + **B**) = *c***A** + *c***B**
- Neutrálnost nuly – součet matice **A** a nulové matice (obě stejného rozměru) se rovná matici **A**: **A** + **0** = **0**+ **A** = **A**
- Opačná matice k matici **A** se značí **-A** a platí **A** + (**-A**) = **0**. Existence opačné matice umožňuje odčítání dvou matic: **A** – **B** = **A** + (**-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**A** = **A**
- Když *c*, *d* jsou reálná čísla, tak *c*(*d***A**) = (*c*.*d*)**A**

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

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

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

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

$$c_{ij} = a_i \cdot b_j = \begin{bmatrix} a_{i1} & a_{i2} & \dots & a_{ip} \end{bmatrix} \cdot \begin{bmatrix} b_{1j} \\ b_{2j} \\ \dots \\ b_{pj} \end{bmatrix} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{ip}b_{pj} \qquad (10.5)$$

Pro násobení matic platí následující:

- Dvě matice je možné spolu násobit pouze tehdy, když první matice má tolik sloupců, kolik má druhá matice řádků.
- 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. $\mathbf{B} \cdot \mathbf{B}^T$ a také $\mathbf{B}^T \cdot \mathbf{B}$ vždy existují.
- $\mathbf{B} \cdot \mathbf{B}$ (tedy druhá mocnina matice **B**) existuje, pouze když je matice **B** čtvercová.
- Násobení matic není kumulativní. $\mathbf{AB} \neq \mathbf{BA}$. 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**.
- $[\mathbf{AB}]^T = \mathbf{B}^T \cdot \mathbf{A}^T$ a $[\mathbf{ABCD\dots}]^T = \mathbf{\dots D}^T \cdot \mathbf{C}^T \cdot \mathbf{B}^T \cdot \mathbf{A}^T$.

10.5 Determinant matice

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

$$|A| = \sum (-1)^I a_{1j_1} \cdot a_{2j_2} \cdot \dots \cdot a_{nj_n} \ , \qquad (10.6)$$

kde počet sčítanců je $n!$ a I je počet inverzí v permutací $(j_1, j_2, \dots j_n)$ prvků 1, 2, … n.
Determinant matice druhého řádu se vypočítá jednoduše:

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \qquad (10.7)$$

Např. $\begin{vmatrix} 2 & 5 \\ 1 & 3 \end{vmatrix} = 2.3 - 5.1 = 6 - 5 = 1$.

Získáné 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 $n = 3$):

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \tag{10.8}$$

$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{23}a_{32}a_{11}$$

Např. $\begin{vmatrix} 1 & 3 & -2 \\ -3 & 0 & 1 \\ 2 & 5 & 6 \end{vmatrix} = 1.0.6 + 3.1.2 + (-3).5.(-2) - (-2).0.2 - 3.(-3).6 - 1.5.1 = 85$

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:

$$|A| = a_{11}|A_{11}| - a_{21}|A_{21}| + a_{31}|A_{31}| - a_{41}|A_{41}|, \tag{10.9}$$

kde determinant $|A_{11}|$ je determinantem submatice $\mathbf{A_{11}}$, 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 $\mathbf{A_{21}}$, atd.

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: $|\mathbf{A}| = |\mathbf{A}^T|$.
- 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.

10.6 Hodnost matice

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

Např. vektory $\begin{bmatrix} -4 \\ -6 \\ -8 \end{bmatrix}$ a $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$ 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 rovný 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 $h = 3$.

10.7 Inverzní matice

V maticové algebře neexistuje dělení matic. Lze jej ovšem nahradit násobením matice tzv. inverzní maticí. Inverzní matici matice **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 sestrojit inverzní matice. Pro obdélníkovou matici lze sestrojit 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 $\mathbf{A}^{-1} = \mathbf{A}^{T}$, **A** je ortogonální matice (matice, jejíž normalizované vektory jsou ortogonální, tj. vzájemně kolmé) a $\mathbf{A}\mathbf{A}^{T} = \mathbf{I}$

Inverzní matici $\mathbf{A}^{-1}$ k dané čtvercové matici **A** lze vypočítat pomocí Gauss-Jordanovy eliminační metody. Postup je následující:

Sestavíme matici **B** složenou z původní matice **A** a jednotkové matice **I**.

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

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

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

$$[AI] = \begin{bmatrix} 1 & 3 & 0 & 1 & 0 & 0 \\ -1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 3 & 3 & 0 & 0 & 1 \end{bmatrix}$$

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.

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

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

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

10.8 Vlastní hodnoty a vlastní vektory matice

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

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 rozptyl dat než původní proměnné.

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

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

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

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

Vlastní hodnoty a vlastní vektory matice **A** nalezneme pomocí rovnice

$$\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_j, \tag{10.10}$$

pomocí které jsou vypočítány různé vlastní hodnoty λ_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:

$$\mathbf{A}\mathbf{u}_j - \lambda_j \mathbf{u}_j = 0, \text{ dále pak } (\mathbf{A} - \lambda_j \mathbf{I})\mathbf{u}_j = 0 \tag{10.11}$$

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

$$|\mathbf{A} - \lambda_j \mathbf{I}| = 0, \tag{10.12}$$

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

Pro matici **A** řádu p je charakteristická rovnice polynomem λ stupně p, jehož řešením jsou různé hodnoty λ_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

$$\left| \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0, \text{ tj. } \left| \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0 \text{ a } \begin{vmatrix} 2-\lambda & 2 \\ 2 & 5-\lambda \end{vmatrix} = 0$$

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 *u*, např. 1.

$u_{11} = 1$
pak podle $-4u_{11} + 2u_{21} = 0$
dostáváme $-4 + 2u_{21} = 0$
a $u_{21} = 2$

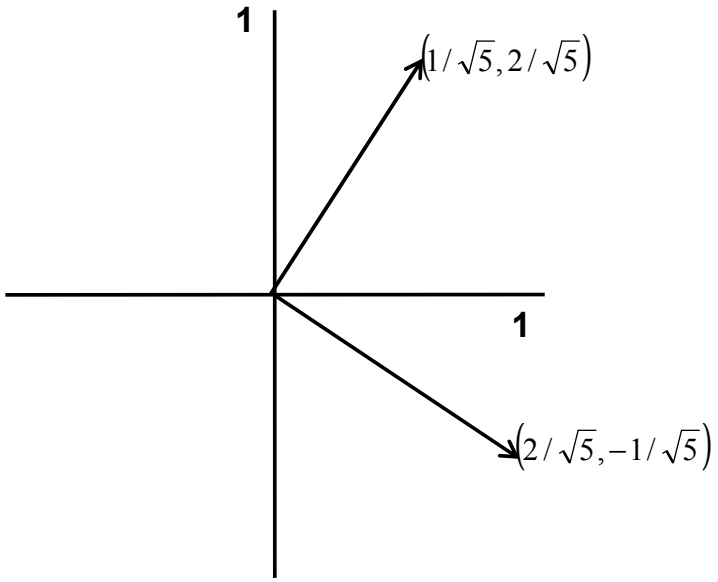
$u_{12} = 1$
pak podle $1u_{12} + 2u_{22} = 0$
dostáváme $1 + 2u_{22} = 0$
a $u_{22} = -\frac{1}{2}$

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 10.4). Vlastní vektory nesymetrické matice nejsou na sebe kolmé (ortogonální).



Obrázek 10.4 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.

10.9 Rozklad na singulární hodnoty (SVD)

Každou matici složenou z reálných dat lze rozdělit na součin tří matic speciálních vlastností. Tento postup se nazývá rozklad na singulární hodnoty (singular value decomposition-SVD), datovou matice lze rozdělit podle vztahu:

$$\mathbf{D}_{(r,s)} = \mathbf{U}_{(r,k)}\Gamma_{(k,k)}\mathbf{V}^{\mathrm{T}}_{(s,k)} \qquad r>s \tag{10.13}$$

Matice **U** a **V** jsou ortogonální a normované (ortonormální). To znamená, že když matici **U** nebo **V** vynásobíme danou transponovanou maticí, získáme matici jednotkovou. Dále matice **U** je složena z vlastních (charakteristických) vektorů čtvercové matice $\mathbf{DD}^{\mathrm{T}}$ a matice **V** z vlastních vektorů matice $\mathbf{D}^{\mathrm{T}}\mathbf{D}$.

$$\mathbf{U}^{\mathrm{T}}\mathbf{U}=\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{I} \tag{10.14}$$

Matice **Γ** je typu k x k a její diagonála je tvořena singulárními hodnotami, které jsou na hlavní diagonále uspořádány podle klesající velikosti.

$$\Gamma_{11} > \Gamma_{22} > \Gamma_{33} > \ \ldots\ldots\ \Gamma_{k,k} \tag{10.15}$$

Singulární hodnoty nesou informaci o významnosti jednotlivých sloupců matice **U** (skórů – scores) a odpovídajících sloupců matice **V** (zátěží – loadings). Singulární hodnoty matice **Γ** jsou rovny odmocninám vlastních čísel matice $\mathbf{DD}^{\mathrm{T}}$ tedy $\mathbf{D}^{\mathrm{T}}\mathbf{D}$.

Provedeme-li rozklad na singulární hodnoty na transponované matici **D** (tj. $\mathbf{D}^{\mathrm{T}}$), dostaneme výsledné matice **V**, **U** a **Γ** příliš velké a obsahující velké množství nesmyslných čísel nebo nul. Proto se doporučuje původní matici orientovat, tak jak je uvedeno v rovnici (10.13).

Vícerozměrné statistické metody v biologii
RNDr. Danka Haruštiaková, PhD., RNDr. Jiří Jarkovský, PhD., Mgr. Simona Littnerová, Doc. RNDr. Ladislav Dušek, Dr.

Recenzenti: Doc. RNDr. Eva Bulánková, PhD. druhý recenzent …
Obálka: Radim Šustr, DiS
Vydalo: AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o. Brno,
Purkyňova 95a, 612 00 Brno
www.cerm.cz
Tisk: FINAL TISK s.r.o. Olomučany
Náklad: 200 ks
Vydání: první
Vyšlo v roce 2011

ISBN XXX-YY-XYXY-ZZZ-X