Užití kompozičního biplotu při analýze medicínských dat

ALŽBĚTA KALIVODOVÁ
kalivodovaa@gmail.com
Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta, UP Olomouc
Laboratoř dědičných metabolických poruch, Ústav molekulární a translační medicíny, UPOL

Biplot je v současnosti hojně užívaný grafický nástroj mnohorozměrné statistické analýzy. Zobrazuje objekty a proměnné do jednoho grafu. Často se aplikuje také při statistické analýze speciálních typů dat, tzv. kompozičních dat, nesoucích pouze relativní informaci (speciálně procenta, proporce). V tomto příspěvku bude uvedena aplikace standardního a kompozičního biplotu na data z oblasti lékařství.

STANDARDNÍ BIPLOT

Mějme dánu datovou matici $\mathbf{X}$ o n řádcích a D sloupcích, kde u každého z n objektů bylo provedeno D měření. Pro konstrukci biplotu je důležitý singulární rozklad matice $\mathbf{X}$ pomocí matic $\mathbf{U}$, $\mathbf{D}$ a $\mathbf{V}$:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{1}$$

kde $\mathbf{U}$ a $\mathbf{V}$ jsou ortogonální matice. Sloupce matice $\mathbf{U}$ se nazývají skóry (scores), odpovídající sloupce matice $\mathbf{V}$ se nazývají zátěže (loadings). Matice $\mathbf{D}$ tvoří singulární hodnoty nacházející se na hlavní diagonále.

Princip konstrukce biplotu je založen na nahrazení matice $\mathbf{X}$ pomocí její aproximace $\mathbf{X}_{(2)}$ s hodností rovnou dvěma. Tu volíme tak, aby byla optimální z hlediska minimalizace součtu čtverců odchylek jejích prvků od příslušných prvků matice $\mathbf{X}$. Ve vyjádření matice $\mathbf{X}_{(2)}$ přitom použijeme pouze první dva sloupce matice $\mathbf{U}$ a první dva sloupce matice $\mathbf{V}$ ze singulárního rozkladu. Maticově lze tuto skutečnost zapsat jako

$$\mathbf{X} \approx \mathbf{X}_{(2)} = \mathbf{U}\mathbf{D}\mathbf{V}^T = (\mathbf{u}_1, \mathbf{u}_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}\begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix}. \tag{2}$$

Je zřejmé, že $\mathbf{X}_{(2)}$ je opět rozměrů $n \times D$. Můžeme ji rozdělit podle metody hlavních komponent takto:

$$\mathbf{X}_{(2)} = \mathbf{G}\mathbf{H}^T, \tag{3}$$

kde

$$\mathbf{G} = (\mathbf{u}_1, \mathbf{u}_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c}, \ \mathbf{H} = (\mathbf{v}_1, \mathbf{v}_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{4}$$

pro $0 \leq c \leq 1$. Řádky matice $\mathbf{G}$ pak využijeme v biplotu k zobrazení objektů a řádky matice $\mathbf{H}$ k zobrazení proměnných.

Grafická interpretace

Ve standardním biplotu zobrazujeme pozorování pomocí bodů (skórů) a proměnné pomocí paprsků (zátěží) vycházejících z počátku. Délky paprsků potom aproximují směrodatnou odchylku příslušných proměnných, kosinus úhlu mezi paprsky zobrazuje korelaci těchto proměnných. Vzdálenost mezi body aproximuje Mahalanobisovu vzdálenost mezi pozorováními. Podrobnější interpretace bude uvedena u praktických příkladů.

KOMPOZIČNÍ DATA

Mějme dán vícerozměrný statistický soubor, jehož složky představují kvantitativně vyjádřené části celku. Tato data potom označujeme jako kompoziční (nebo zkráceně kompozice). Data jsou vyjádřena jako podíly na celku velikosti k (nejčastěji $k = 100$ nebo $k = 1$). Formální definice zní následovně:

Definice 1 *Sloupcový vektor* $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T$ *se nazývá D-složková kompozice, jestliže jsou všechny jeho prvky kladná reálná čísla nesoucí pouze relativní informaci.*

Informace obsažená v kompozici se změnou měřítka nezmění. Vycházíme z faktu, že kompoziční data nesou pouze relativní informaci. Tuto vlastnost požadujeme také od příslušných statistických metod, které užíváme pro analýzu kompozic. Dále požadujeme invarianci vůči permutaci, tedy aby se při změně pořadí složek vektoru nezměnily podíly mezi nimi. Poslední vlastností je tzv. podkompoziční soudružnost. Ta nám říká, že podíly mezi složkami v podkompozici jsou vždy stejné jako podíly v rámci celé kompozice. Tedy z celého datového souboru bychom měli obdržet stejné informace o prvcích nějaké podkompozice jako při analýze pouze této podkompozice. Ovšem standardní statistické metody, jako např. analýza hlavních komponent založená na kovariancích původních složek kompozičních dat, tuto vlastnost nemají. Škála kompozic je relativní.

Clr transformace

Práce s kompozičními daty probíhá na tzv. Aitchisonově geometrii, která se chová analogicky jako euklidovský prostor. Pro možnost práce s kompozičními daty přímo v euklidovském prostoru byly sestrojeny tzv. logratio transformace ze simplexu do reálného prostoru. V tomto příspěvku zmíníme pouze jednu z nich - centrovanou logratio (clr) transformaci. Ta zachovává vzdálenosti mezi daty, ale vede k singulární varianční matici. Při použití této transformace zůstává konstrukce biplotu analogická, změní se ovšem jeho interpretace. Bližší podrobnosti zmíníme u příkladu.

Clr transformace vyjádřená po složkách má následující tvar:

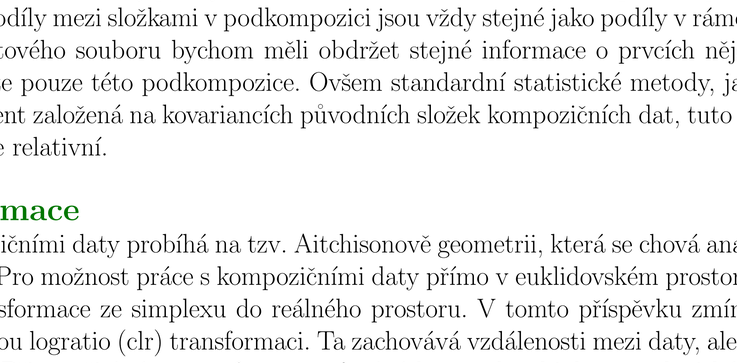
$$\mathrm{clr}(\mathbf{x}) = \left(\ln\frac{x_1}{g(\mathbf{x})}, \ln\frac{x_2}{g(\mathbf{x})}, \ldots, \ln\frac{x_D}{g(\mathbf{x})}\right)^T, \tag{5}$$

kde $g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$ značí geometrický průměr složek kompozice $\mathbf{x}$.

PRAKTICKÉ PŘÍKLADY

Úmrtí v důsledku nemoci

Datový soubor je převzatý z databáze Eurostatu a jsou v něm zkoumány podíly nemocí na úmrtnost v každém z vybraných 24 států Evropy. Úkolem je zjistit, jaký je vzájemný vztah mezi jednotlivými státy a také mezi jednotlivými nemocemi. Zjištěné skutečnosti budou porovnány na standardním a kompozičním biplotu (s užitím clr transformace). Pro výpočty byl použit software R.



Obrázek 1a (vlevo): standardní biplot , 1b (vpravo): kompoziční biplot.

Ve standardním biplotu se nám z pohledu paprsků vytvořily čtyři skupiny. Jednu z nich tvoří poruchy oběhového systému, ischemická choroba srdeční a cerebrovaskulární nemoci. Jejich paprsky navíc míří opačným směrem než ostatní šipky. Tyto choroby spolu tedy souvisejí a naopak se liší od ostatních. Zajímavá je zde nekorelovanost cévních chorob se srdečními daná ortogonoalitou příslušných paprsků. Velmi blízko k sobě má cukrovka a metabolické poruchy, nebo také nemoci nervového systému a mentální poruchy. Všechny paprsky jsou zhruba stejně dlouhé, takže mají jednotlivé nemoci na celkové uspořádáni v grafu stejný vliv. V grafu 1a si můžeme také všimnout negativní korelace mezi nemocemi srdce a cukrovkou. Naopak kompoziční biplot se může zdát z hlediska interpretace paprsků trochu nepřehledný, jejich interpretace je ovšem relevantní. Došlo zde k diferenciaci vlivu jednotlivých proměnných, jež je viditelná na různých délkách paprsků. Nejvýznamnějším markerem s nejdelší šipkou je proměnná mentální poruchy. Ty tedy značně ovlivňují celý soubor. Potom zde můžeme pozorovat rozdělení zbylých nemocí do dvou skupin s navzájem stabilními podíly v rámci datového souboru. Například tu menší z nich tvoří již dříve zmíněné poruchy oběhového systému, ischemická choroba srdeční a cerebrovaskulární nemoci (jejich šipky jsou opět stejně dlouhé).

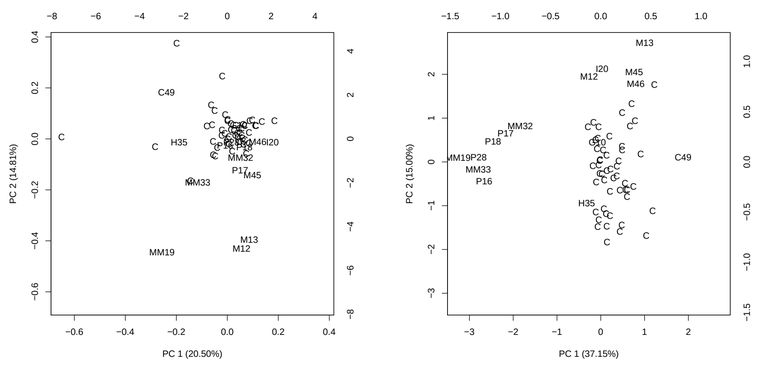
Z pohledu vzdáleností jednotlivých bodů můžeme u obou biplotů říci, že čím jsou jednotlivé státy v grafu blíže u sebe, tím jsou si podobnější. V obrázku 1a můžeme například vidět skupinku států Francie (FR), Dánsko (DK) a Španělsko (ES). Další výraznou charakteristikou tohoto biplotu je např. odlehlé pozorování Portugalsko (PT). To má nejvyšší výskyt cukrovky a druhý nejvyšší u metabolických poruch a zápalu plic. Na druhou stranu má nejnižší přítomnost ischemické choroby srdeční. Tyto extrémy zřejmě způsobily jeho oddělení se od ostatních. Toto odlehlé pozorování (PT) zůstalo zachováno i v kompozičním biplotu. Přibylo k němu ještě jedno významné - SK. Vychýlení Slovenska může být způsobeno velmi nízkou proporcí u mentálních poruch. Tato skutečnost se také projevuje tak, že SK je velmi daleko od vrcholu paprsku reprezentujícího danou chorobu.

Vertikální uspořádání bodů není ve standardním biplotu jasně dáno. Zřetelné je spíše uspořádání do úhlopříčky (z levého dolního rohu) ovlivněné výše zmínenými negativně korelovanými nemocemi srdce a cukrovkou. Vertikální rozložení v obrázku 1b je dáno výskytem mentálních poruch. Při průběhu zleva doprava hodnoty klesají. Horizontální uspořádání států v 1b ovlivňuje celá řada nemocí. Při postupování zdola nahoru postupně roste výskyt zhoubných novotvarů, metabolických chorob, cukrovky, poruch nervového systému, nemocí dýchací soustavy a zápalu plic. Při tomto postupu naopak klesá zastoupení poruch oběhového systému a ischemické choroby srdeční. Při přechodu shora dolů je postup opačný.

Při použití kompozičního biplotu se nám značně zvýšilo procento vysvětlené variability souboru prvními dvěma hlavními komponentami.

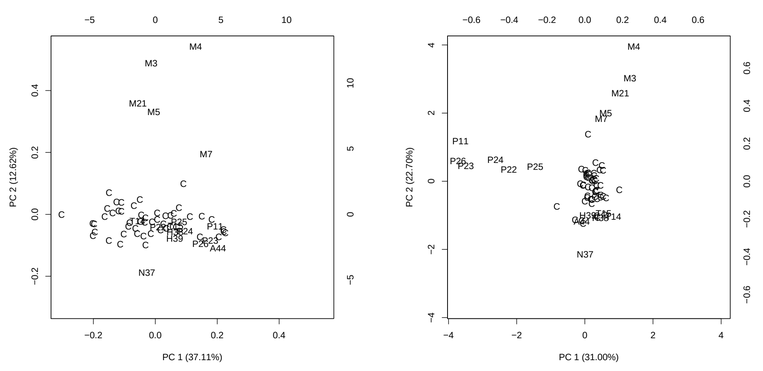
Metabolomická data

Druhý datový soubor je tvořen hodnotami naměřenými v rámci výzkumu v Laboratoři dědičných metabolických poruch Ústavu molekulární a translační medicíny ve Fakultní nemocnici Olomouc. Jedná se o hladiny acylovaných karnitinů a aminokyselin u pacientů, jež trpí poruchami souvisejícími s těmito látkami. Dále jsou zahrnuty kontrolní vzorky. Protože byla data srovnávána na základě velkého počtu proměnných, byly v grafickém výstupu vynechány paprsky reprezentující tyto proměnné a zůstanou zobrazena pouze pozorování (pacienti a kontroly) v podobě bodů. Cílem je vyšetřit datovou strukturu, zejména výskyt shluků.



Obrázek 2a (vlevo): skóry standardního biplotu pro acylované karnitiny , 2b (vpravo): skóry kompozičního biplotu.

Při porovnání grafů vidíme, že při použití kompozičních dat se nám soubor lépe rozdělil do skupin. Shluk kontrolních vzorků (C) se sice více rozptýlil. Naopak se ale lépe oddělily dvě zřetelné skupiny nemocí. První tvoří choroby s označeními MM a P. Druhá je tvořena pacienty s označením M a I. Vzorek C 49 si zachoval v obou grafech pozici samostatného pozorování, v 1b je ale zřetelnější.



Obrázek 3a (vlevo): skóry standardního biplotu pro aminokyseliny , 3b (vpravo): skóry kompozičního biplotu.

Ve standardním biplotu můžeme pozorovat oddělení skupiny nemocí s označením M a vytvoření samostatného pozorování N37. Tyto vlastnosti se zachovaly i v kompozičním biplotu. Navíc se vytvořil ještě shluk chorob s označením P (jiné než výše). Skupina kontrolních vzorků se naopak více „semkla“.

Literatura:

[1] J. Aitchison (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.
[2] J. Aitchison, M. Greenacre (2002) *Biplots of compositional data*. Journal of the Royal Statistical Society, 51 (4), 375–392.
[3] K. R. Gabriel (1971) *The biplot graphic display of matrices with application to principal component analysis*. Biometrika, 58 (3), 453–467.