Řídké hlavní bilance

K. Hron[1] C. Mert[2] P. Filzmoser[2]

[1]Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta, Univerzita Palackého, Olomouc

[2]Department of Statistics and Probability Theory
Vienna University of Technology, Austria

Robust 2014, 19. - 24. ledna 2014, Jetřichovice

Obsah

Definice

Data popisující koncentrace složek jsou **kompoziční data**:
D-složková kompozice $x = (x_1, \ldots, x_D)^t$ je prvkem simplexu jako výběrového prostoru (*reprezentací*) kompozičních dat,

$$S^D = \{(x_1, \ldots, x_D)^t \mid x_i > 0, \sum_{i=1}^{D} x_i = \kappa\},$$

kde κ je vhodně zvolená konstanta, např. 1 nebo 100.
Definice: **Kompoziční data** jsou reálné vektory $\mathbf{x} = (x_1, \ldots, x_D)^t$ s D kladnými složkami popisujícími kvantitativně relativní příspěvky částí na celku (Aitchison, 1986).
Kompoziční data se řídí Aitchisonovou geometrií na simplexu (a nikoli standardní euklidovskou geometrií).

Logratio souřadnice

(transformace) ze simplexu do euklidovského reálného prostoru:

- **alr (*aditivní logratio*) souřadnice:**
  nejsou ortonormální, dělíme j-tou složkou $j \in \{1, \ldots, D\}$:

$$\mathbf{x}^{(j)} = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j} \right)^t$$

- **clr (*centrované logratio*) souřadnice:**
  singulární varianční matice, izometrie s Aitchisonovou geom.:

$$\mathbf{y} = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)^t, \ \mathbf{y}^t \mathbf{1} = 0$$

- **ilr (*izometrické logratio*) souřadnice:**
  volbou ortonormální báze v clr-prostoru $\Longrightarrow$ **komplexní interpretace** (absence kanonické báze na simplexu)

Cíle

- Objekty: vysoce-dimenzionální kompoziční data
- Oblasti: chemometrie, proteomika, genomika, metabolomika
- Cíl: redukce dimenze
  - maximalizace vysvětlené variability
  - zjednodušení interpretace nových souřadnic (směrů)

Problémy

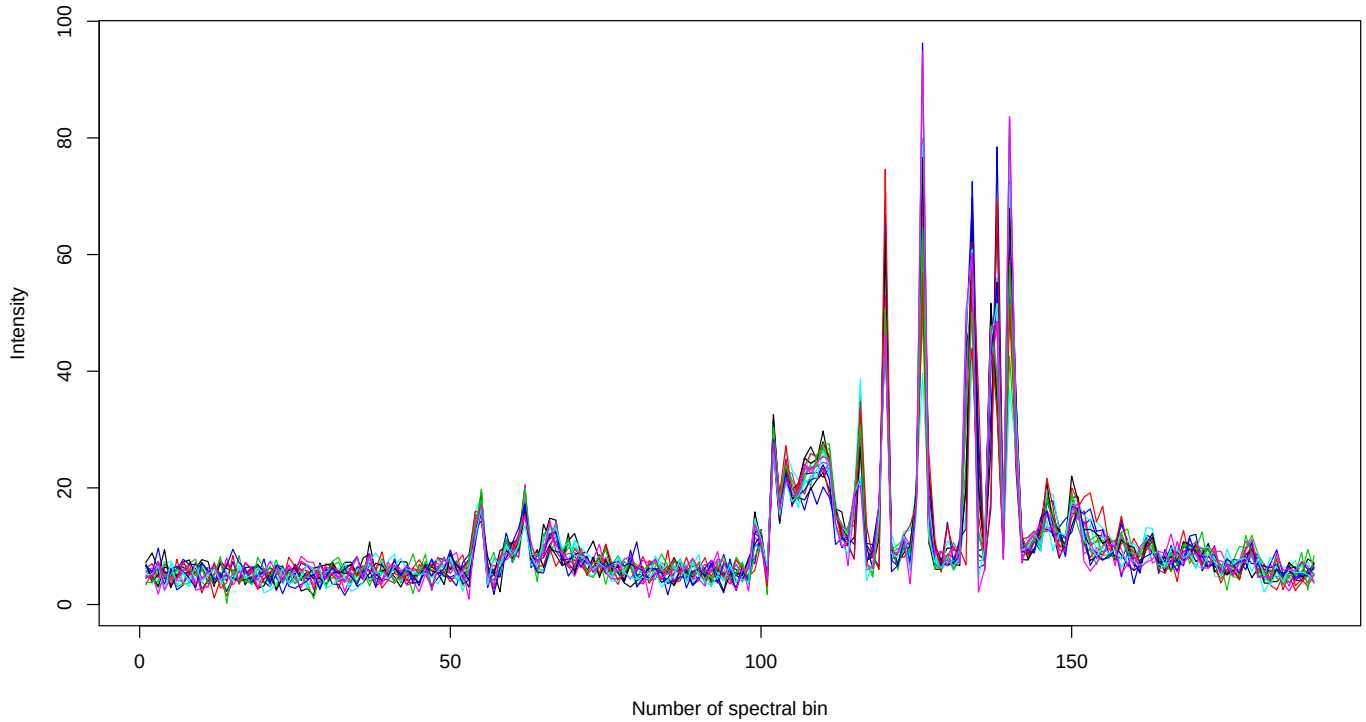
Současné metody často selhávají při řešení následujících problémů:

- nové směry jsou obtížně interpretovatelné
- velká ztráta informace (vysvětlené variability)
- metody nejsou použitelné pro vysoce-dimenzionální kompoziční data

NMR metabolomická spektra

- NMR metabolomická spektra vzorků moči od 18 myší
- každé spektrum má 189 spektrálních píků
- data jsou měřena v ppm (CoDa)
- detailní popis dat v Nyamundanda a kol. (2010)

NMR metabolomická spectra



Obrázek: Původní data (vzorky moči) se 189 spektrálními píky.

Methods

- Metoda hlavních komponent (PCA)
  - redukce dat při maximalizaci vysvětlené variability
  - nové směry jsou lineární kombinace všech proměnných: obtížná interpretovatelnost
- Bilance
  - charakterizují rovnováhu mezi disjunktními skupinami kompozičních složek
  - představují souřadnice vzhledem k ortonormální bázi na simplexu
  - bez ohledu na maximalizaci vysvětlené variability
  - bilance jsou konstruovány užitím postupného binárního dělení

Bilance

Užití **postupného binárního dělení** pro vytvoření disjunktních skupin kompozičních složek (Egozcue a Pawlowsky-Glahn, 2005).
Například pro $D = 5$

	x_1	x_2	x_3	x_4	x_5
1	$+1$	$+1$	-1	-1	-1
2	$+1$	-1	0	0	0
3	0	0	$+1$	-1	-1
4	0	0	0	$+1$	-1

Řádek 1: $G_1 = \{x_1, x_2\}$ a $G_2 = \{x_3, x_4, x_5\}$
Řádek 2: rozdělit G_1 na $\{x_1\}$ a $\{x_2\}$
atd.
Znaménka v $D - 1$ řádcích jsou použita ke kontrukci *ilr* báze V.

Obecně, **ortonormální báze na simplexu** může být definována vektory (sloupce $D \times (D-1)$ matice V)

$$v_i = \left(\underbrace{a_+, \ldots, a_+}_{r \text{ složek}}, \underbrace{a_-, \ldots, a_-}_{s \text{ složek}}, \underbrace{0, \ldots, 0}_{D-r-s \text{ složek}} \right)^t$$

pro $i = 1, \ldots, D-1$, kde

$$a_+ = \frac{\sqrt{s}}{\sqrt{r(r+s)}} \quad \text{and} \quad \mathrm{a}_- = \frac{-\sqrt{\mathrm{r}}}{\sqrt{\mathrm{s}(\mathrm{r}+\mathrm{s})}}$$

r je počet **kladných** a s počet **záporných** prvků v tabulce postupného binárního dělení (Egozcue a Pawlowsky-Glahn, 2005).

Hlavní bilance (PB)

- představují co nejlepší aproximaci hlavních komponent
- pokus splnit oba požadavky: maximalizace vysvětlené variability a jednoduchá interpretovatelnost
- obtížně použitelné pro vysoce-dimenzionální kompoziční data

Algoritmy pro konstrukci hlavních bilancí:

- úhlové přiblížení k hlavním komponentám (AV)
- hierarchické shlukování složek (HC)
- hierarchické bilance s maximální vysvětlenou variabilitou (MV)
- podrobný popis viz Pawlowsky-Glahn a kol. (2011)

Řídké hlavní bilance (SPB)

- řídké hlavní bilance představující kompromis mezi maximalizací vysvětlené variability a počtem zahrnutých kompozičních složek
- obsahují informaci pouze o několika málo kompozičních složkách s nulovým příspěvkem (většiny) ostatních složek
- obdoba cílů řídké PCA
- užijeme algoritmus z Witten a kol. (2012) založený na řídkém singulárním rozkladu (SVD)

Algoritmus pro konstrukci řídkých hlavních bilancí (SPB)

- aplikujeme řídkou PCA na clr-transformovanou datovou matici
- zvolíme k-komponent
- matice zátěží V má rozměry $D \times k$ s mnoha nulami
- $V = [v_{ij}]$ vyžaduje další modifikaci
  - dosažení nepřekrývajícího se efektu nenulových prvků matice
    - garance ortogonality hlavních směrů
    - zjednodušení intepretace

Algoritmus pro konstrukci řídkých hlavních bilancí (SPB)

- najdeme nejmenší j pro které $v_{ij} \neq 0$, a položíme všechny prvky v_{il}, $l > j$ rovny nule (v případě, že jsou nenulové)
- $v_l^* : d \leq D$ nenulových prvků v každém sloupci modifikované matice V
- projektujeme v_l^* na nadrovinou clr transformovaných dat
- užijeme modifikovanou matici ke konstrukci bilancí

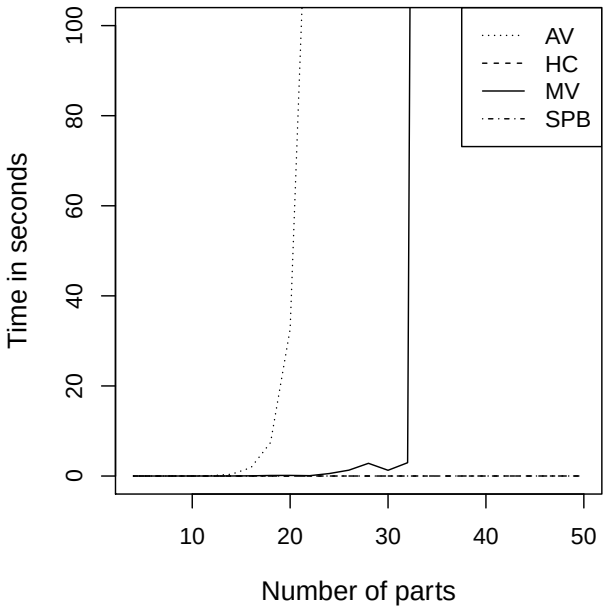
Algoritmus pro konstrukci řídkých hlavních bilancí (SPB)

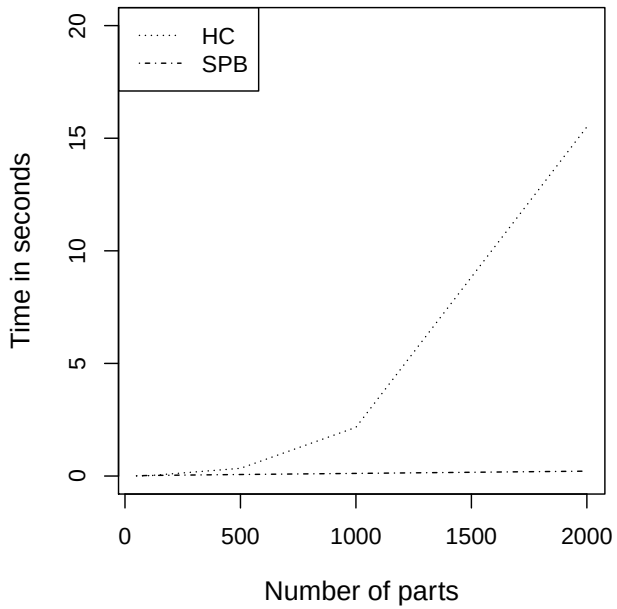
	1	2	3	4	5
1	-0.62	-0.11	0.00	0.25	0.12
2	0.00	0.42	0.72	0.00	-0.06
3	0.00	0.00	0.00	-0.36	-0.01
4	0.00	0.00	0.00	0.85	0.05
5	-0.26	0.00	0.00	0.00	0.18
6	0.00	0.81	0.00	-0.09	-0.03
7	0.00	0.00	-0.48	0.00	-0.73
8	0.22	0.00	-0.45	0.06	0.62
9	0.68	-0.13	0.00	0.20	-0.18
10	0.00	0.00	0.14	0.20	0.04

	1	2	3	4	5
1	-0.62	0.00	0.00	0.00	0.00
2	0.00	0.42	0.00	0.00	0.00
3	0.00	0.00	0.00	-0.36	0.00
4	0.00	0.00	0.00	0.85	0.00
5	-0.26	0.00	0.00	0.00	0.00
6	0.00	0.81	0.00	0.00	0.00
7	0.00	0.00	-0.48	0.00	0.00
8	0.22	0.00	0.00	0.00	0.00
9	0.68	0.00	0.00	0.00	0.00
10	0.00	0.00	0.14	0.00	0.00

	1	2	3	4	5
1	-0.62	0.00	0.00	0.00	0.00
2	0.00	-0.19	0.00	0.00	0.00
3	0.00	0.00	0.00	-0.60	0.00
4	0.00	0.00	0.00	0.60	0.00
5	-0.26	0.00	0.00	0.00	0.00
6	0.00	0.19	0.00	0.00	0.00
7	0.00	0.00	-0.31	0.00	0.00
8	0.21	0.00	0.00	0.00	0.00
9	0.68	0.00	0.00	0.00	0.00
10	0.00	0.00	0.31	0.00	0.00

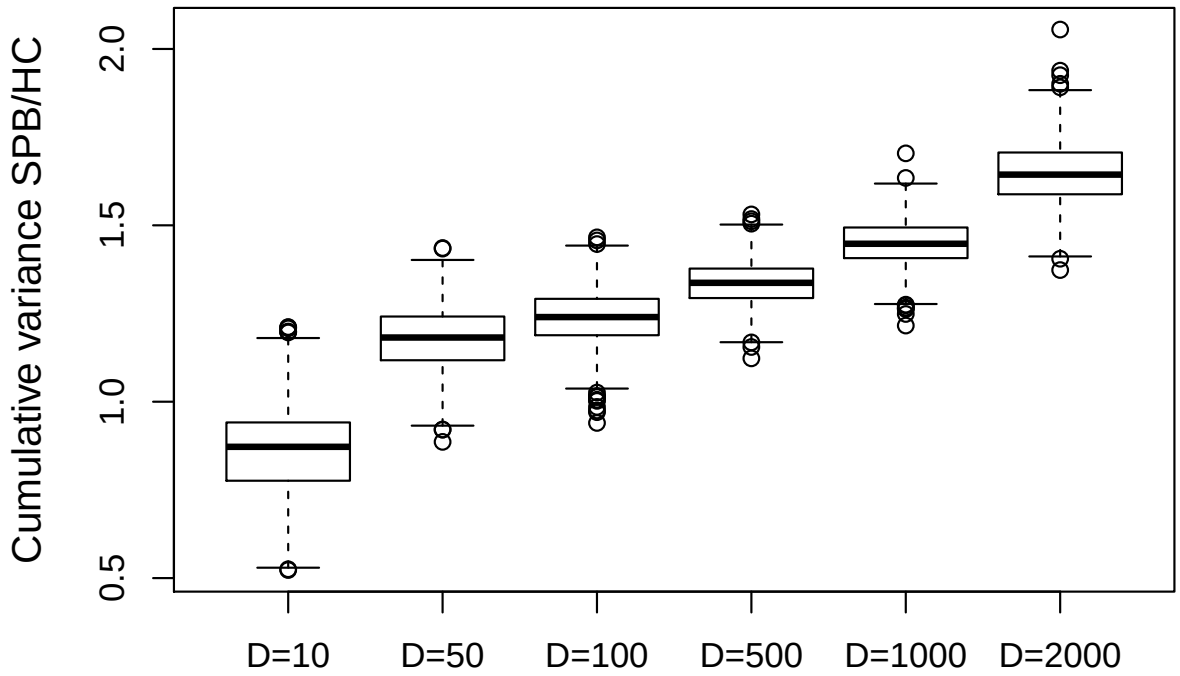
Porovnání časové náročnosti





Obrázek: Porovnání doby potřebné k výpočtu první bilance pomocí algoritmů AV (úhlové přiblížení k hlavním komponentám), HC (hierarchické shlukování složek), MV (hierarchické bilance s maximální vysvětlenou variabilitou) a SPB (řídké hlavní bilance).

Výsledky simulací



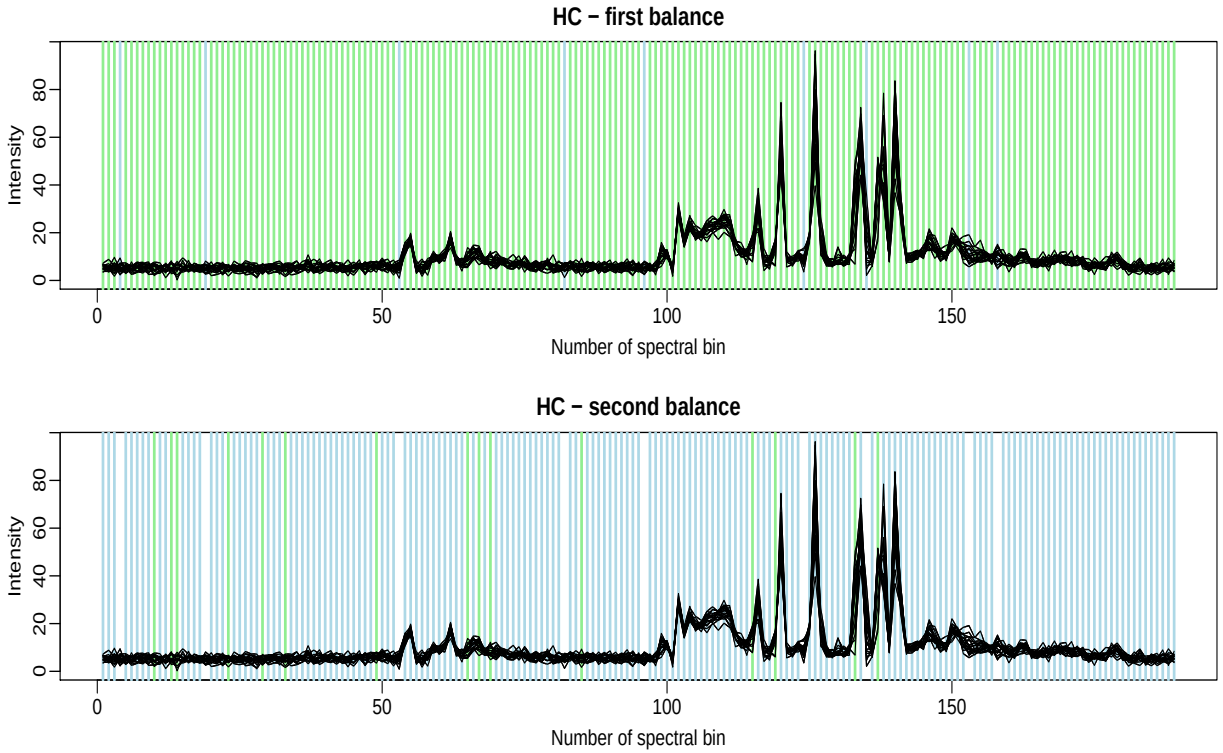
Obrázek: Kumulativní vysvětlená variabilita pro $k = 2$ komponent. Zobrazený je podíl mezi SPB a HC.

Výsledky pro reálný příklad

Tabulka: Kumulativní vysvětlená variabilita pro CoDa-PCA, hierarchické shlukování složek (HC) a řídké hlavní bilance (SPB) pro datových soubor močových vzorků.

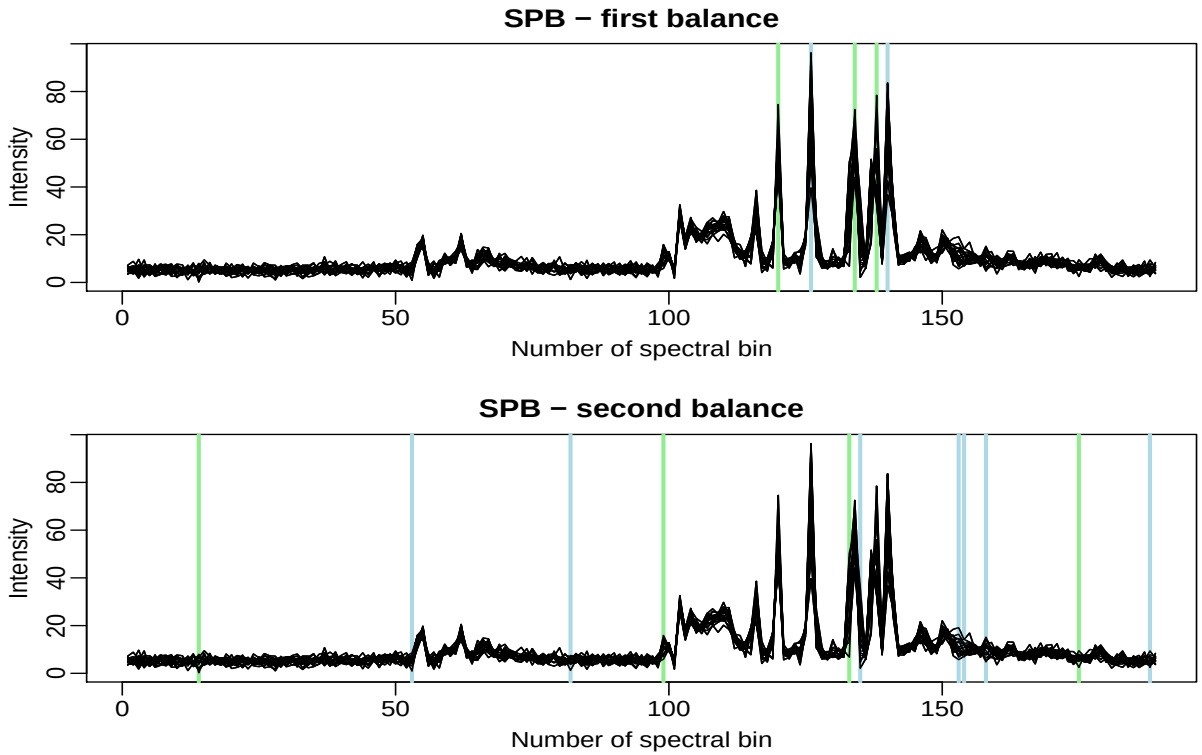
	Kumulativní vysvětlená variabilita [%]	
metoda	jedna komponenta	dvě komponenty
CoDa-PCA	28.1	38.5
HC	8.9	16.7
SPB	13.9	15.6

Výsledky pro reálný příklad



Obrázek: První dvě bilance z HC aplikované na reálná data (vzorky močí). Zobrazena jsou původní data (černá), a dále pozice kladných (zelené vertikály) a záporných (modré vertikály) znamének bilancí.

Výsledky pro reálný příklad



Obrázek: První dvě řídké hlavní bilance aplikované na reálná data (vzorky močí). Zobrazena jsou původní data (černá), a dále pozice kladných (zelené vertikály) a záporných (modré vertikály) znamének bilancí.

Závěr

- Řídké hlavní bilance jsou aplikovatelné pro vysoce-dimenzionální kompoziční data s možností rychlého výpočtu
- Umožňují dosáhnout vysoké úrovně vysvětlené variability (více než hlavní bilance)
- Výsledky jsou jednoduše interpretovatelné

Literatura

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London.

Egozcue, J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37, 795–820.

Mert, C., Filzmoser, P., Hron, K., 2014. Sparse principal balances. Statistical Modelling, přijato k tisku.

Nyamundanda, G., Brennan, L., Gormley, I., 2010. Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics 11, 1–11.

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2011. Principal balances, in: Egozcue, J., Tolosana-Delgado, R., Ortego, M. (Eds.), Proceedings of the 4th International Workshop on Compositional Data Analysis, Girona, Spain. pp. 1–10.

Witten, D., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.