Metoda dílčích nejmenších čtverců pro kompoziční data s aplikací v metabolomice

Alžběta Kalivodová[a,b]

Karel Hron[a,b], Peter Filzmoser[c], Lukáš Najdekr[d]

[a] Katedra matematické analýzy a aplikací matematiky
[b] Katedra geoinformatiky
PŘF UPOL
[c] Department of Statistics and Probability Theory
Vienna University of Technology
[d] Laboratoř metabolomiky
Ústav molekulární a translační medicíny, UPOL a FNOL

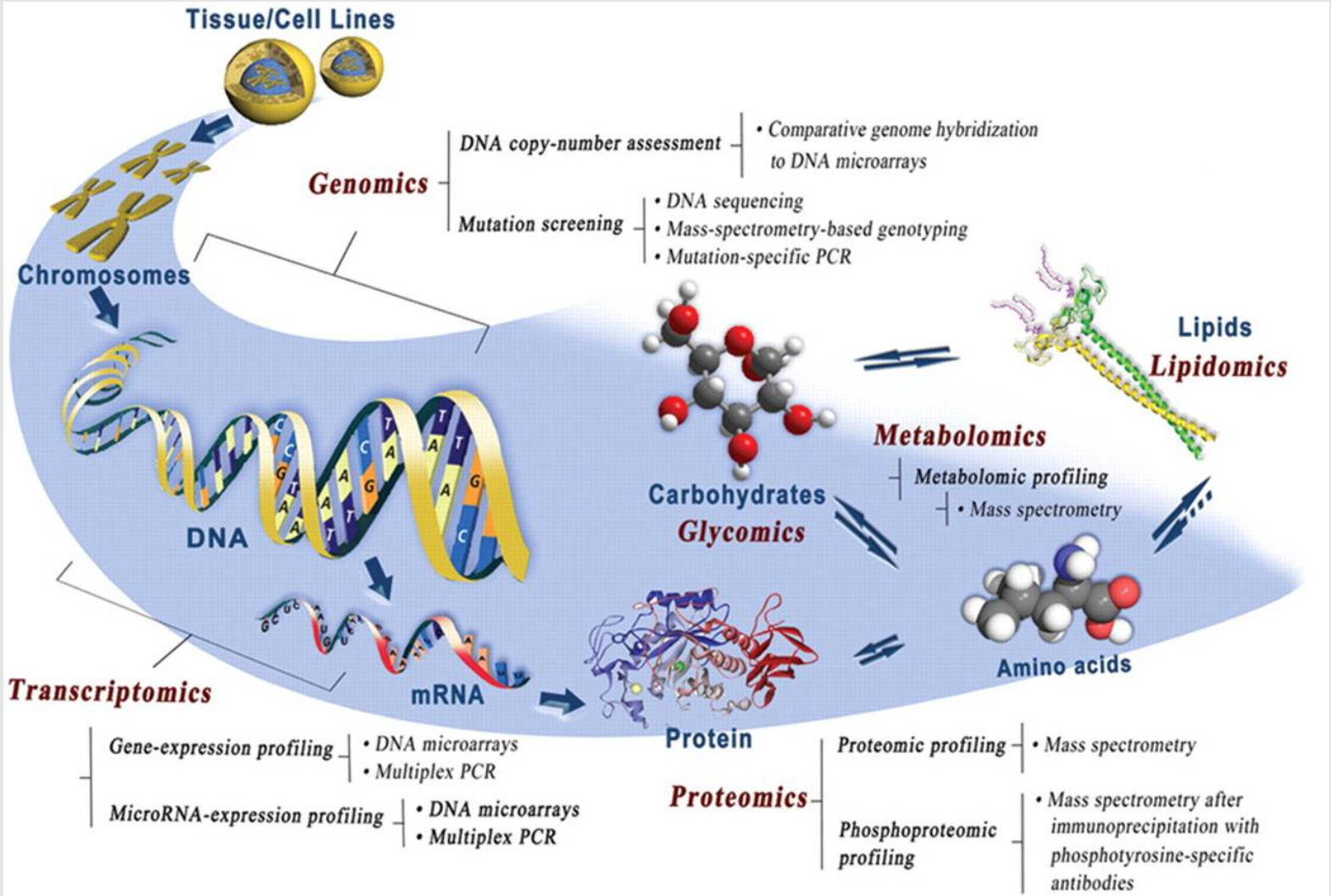
Robust 2014, 19. - 24.1. 2014

Obsah

Úvod

- Metabolomická data mají speciální strukturu.
- Na metabolomická data aplikujeme přístup tzv. kompozičních dat.
- Speciální struktura dat $\rightarrow$ užití metody dílčích nejmenších čtverců - diskriminační analýzy (PLS-DA).
- Teoretické aspekty jsou dále ukázány na konkrétních datových souborech.
- Porovnání standardního a kompozičního přístupu.

Metabolomika



(Wu et al., 2011)

Metabolomika

- **Metabolom** - soubor organických sloučenin, které jsou obsaženy v daném biologickém materiálu. Jejich velikost je na úrovni molekul.
- **Metabolity** - molekuly tvořící metabolom.
- **Metabolomika** - analýza metabolomu za daných podmínek.
- Metabolity jsou měřeny na biologických materiálech (buňky, krev, moč, ...) a nesou pouze relativní informaci.

Metabolomický datový soubor

- Pozorování - **pacienti vs. kontroly** . Proměnné - jednotlivé **metabolity**.
- Velký datový soubor.
- Více proměnných než pozorování.
- Malý počet pacientů (nejvíce desítky, spíše méně). $\times$ Velký počet metabolitů (stovky) $\rightarrow$ užití vhodných metod.

Kompoziční data

- Vícerozměrný statistický soubor, jehož prvky jsou kvantitativně vyjádřené části celku.

Definice

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

- Možnost reprezentovat kompozice jako data s konstantním součtem (k).
- Např. procenta ($k = 100$) nebo podíly na celku velikosti k (nejčastěji $k = 1$).
- **Příklad**: koncentrace chemických sloučenin v horninách.

Centrované logratio (clr) souřadnice

- Geometrie pro práci s kompozičními daty - tzv. Aitchisonova geometrie.
- Převod na známou euklidovskou geometrii přes tzv. logratio souřadnice.
- Clr souřadnice (centred logratio coordinations) - zachovává vzdálenosti mezi daty, vede k singulární varianční matici.
- Clr souřadnice vyjádřené po složkách:

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

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

Isometrické logratio (ilr) souřadnice

- Principem je vytvořit ortonormální bázi pro kompozici $\mathbf{x} = (x_1, \ldots, x_D)'$ na simplexu (výběrovém prostoru).

Ilr souřadnice

$(D-1)$-dimenzionální reálný vektor $ilr(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$, jehož složky jsou definovány jako

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1. \qquad (1)$$

- Proměnná z_1 nese všechnu relevantní informaci o části kompozice x_1, vysvětluje všechny podíly mezi x_1 a ostatními částmi $\mathbf{x}$.

Isometrická logratio (ilr) souřadnice

- Chceme vytvořit ortonormální bázi, ve které první ilr souřadnice vysvětluje veškerou důležitou informaci o zvolené složce.
- Konstruujeme D různých ilr souřadnic, kde je D-tice $(x_1, \ldots, x_D)$ v (1) nahrazena pro $l = 1, \ldots, D$ pomocí $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D) =: (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$ [3].

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \qquad (2)$$

Vztah mezi clr a ilr souřadnicemi

$$\mathbf{y} = \mathbf{V}\mathbf{z}. \tag{3}$$

Matice $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ má dimenzi $D \times (D-1)$ a její sloupce jsou tvořeny clr souřadnicovými vektory, které tvoří ortonormální bázi (vzhledem k Aitchisonově geometrii). Pro $i = 1, \ldots, D-1$

$$\mathbf{v}_{D-i} = \sqrt{\frac{D-i}{D-i+1}}\left(0, \ldots, 0, 1, -\frac{1}{D-i}, \ldots, -\frac{1}{D-i}\right)'. \tag{4}$$

PLS-DA

- Metoda dílčích nejmenších čtverců - lineární diskriminace (partial least squares regression - discriminant analysis).
- Třída metod, která slouží k modelování vztahů mezi pozorováními (vysvětlujícími proměnnými) a skupinou tzv. latentních proměnných.
- Kombinace metody hlavních komponent a mnohonásobné regrese.
- **Cíl** - predikovat množinu závislých proměnných na základě informace ze skupiny latentních proměnných (prediktorů). Predikce je prováděna extrakcí ortogonálních faktorů z prediktorů.
- **Cíl** - maximalizovat kovarianci mezi skupinami závislých a latentních proměnných.

PLS-DA

- Závislé a latentní proměnné - blokové matice, centrovány vzhledem k Aitchisonově geometrii.
- Regrese je prováděna na latentních proměnných.
- PLS1 - pouze jedna závisle proměnná, PLS2 - více závislých proměnných (užívá se častěji).

Datové matice $\mathbf{X}_{n\times D}$ a $\mathbf{Y}_{n\times q}$. Matice $\mathbf{X}$ je transformována na $\mathbf{Z}$ pomocí ilr souřadnic (2). Výsledkem je lineární vztah

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E}, \tag{5}$$

kde $\boldsymbol{\Gamma}$ je $(D-1)\times q$ dimenzionální matice regresních koeficientů a $\mathbf{E}$ je matice chyb. $\mathbf{Y}$ je tvořena binárními proměnnými reprezentujícími příslušnost ke skupinám.

PLS-DA

Místo řešení rovnice (5) jsou matice $\mathbf{Z}$ a $\mathbf{Y}$ vyjádřeny pomocí latentních proměnných:

$$\mathbf{Z} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_X,$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{E}_Y.$$

- $\mathbf{E}_X$, $\mathbf{E}_Y$ - matice reziduí
- $\mathbf{T}$, $\mathbf{U}$ - matice skórů
- $\mathbf{P}$, $\mathbf{Q}$ - matice zátěží
- Všechny matice mají a sloupců, $a \leq \min(D, q, n)$ je počet PLS komponent.

PLS-DA

- Kovariance mezi x a y je maximalizována.
- Kovariance mezi $\mathbf{t}$ a $\mathbf{u}$: $cov(\mathbf{t}, \mathbf{u}) = \mathbf{t}^T\mathbf{u}/(n-1)$, za podmínky $\|\mathbf{t}\| = \|\mathbf{u}\| = 1$.
- Vektory vah $\mathbf{w}$ a $\mathbf{c}$: $\mathbf{t} = \mathbf{Zw}$, $\mathbf{u} = \mathbf{Yc}$.

Maximalizace:

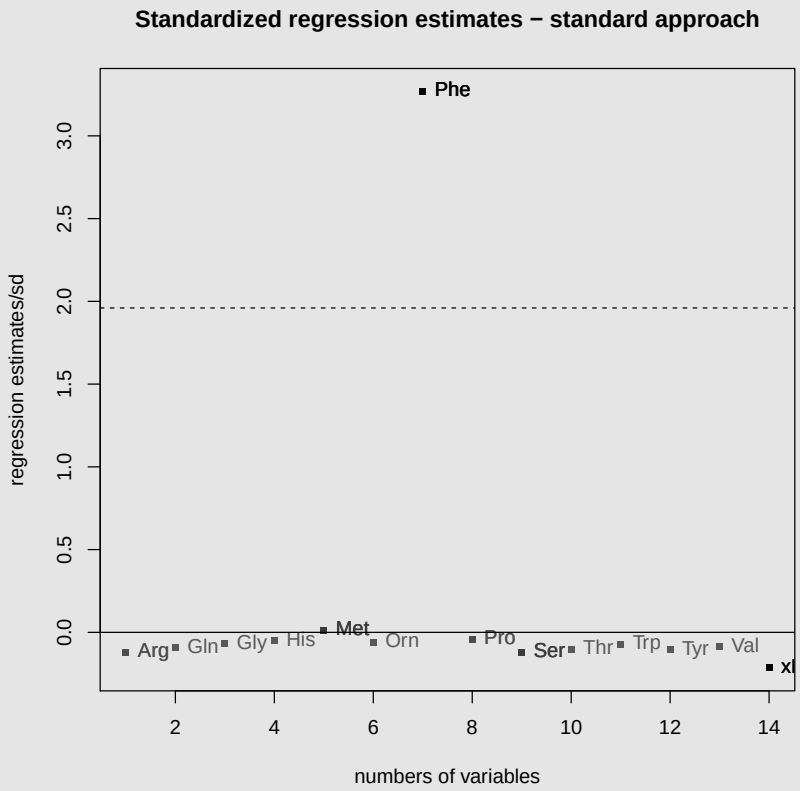
$$cov(\mathbf{t}, \mathbf{u}) = cov(\mathbf{Zw}, \mathbf{Yc}) \rightarrow \max_{\|\mathbf{t}\|=\|\mathbf{u}\|=1}.$$

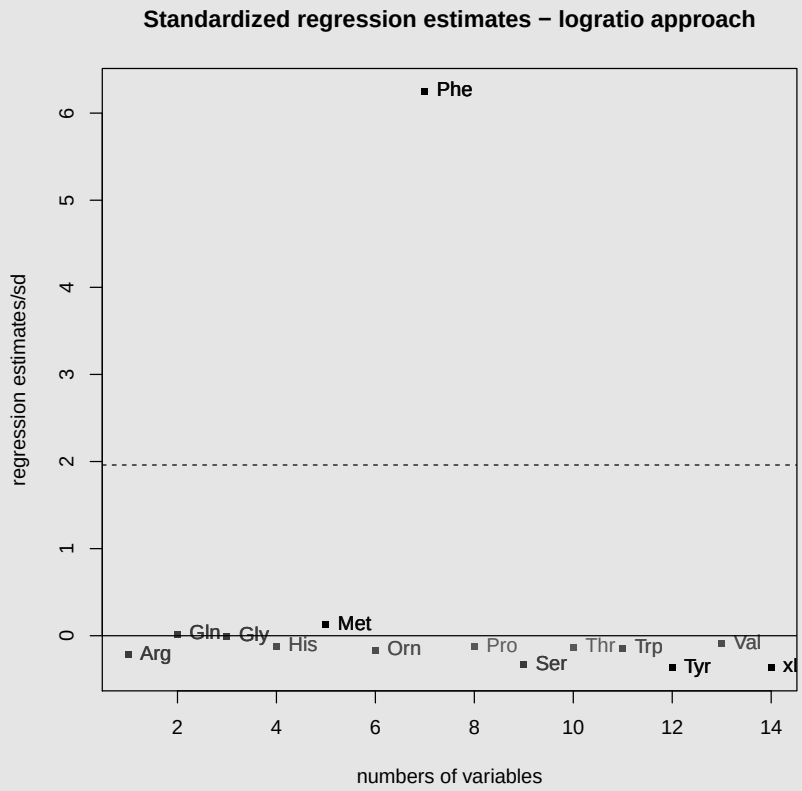
Řešení: první dva vektory skórů $\mathbf{t}_1$ and $\mathbf{u}_1$.

PLS-DA - další postup

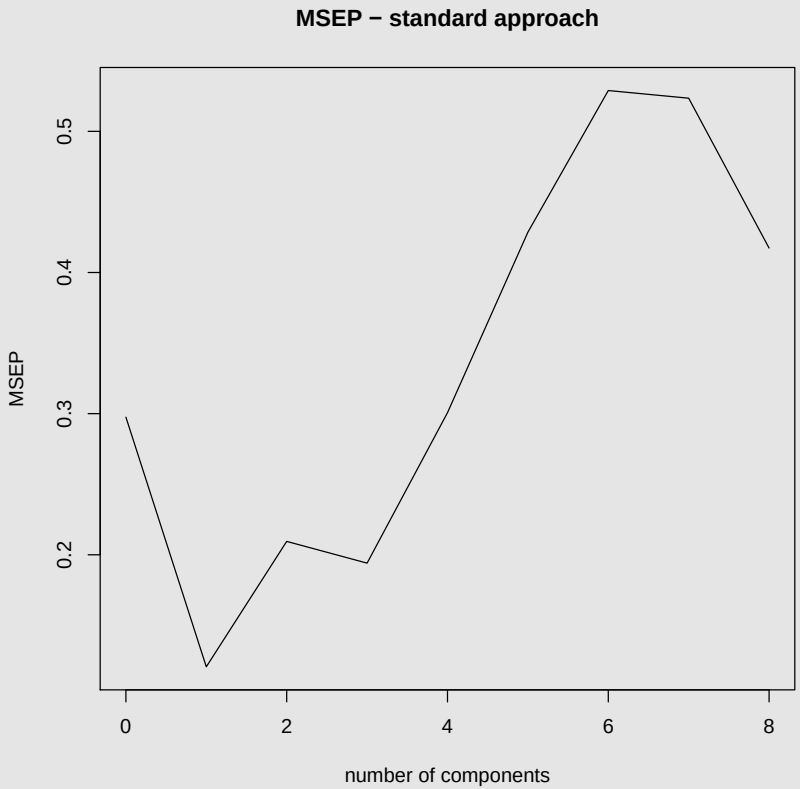
- Vyšetřování statistické významnosti jednotlivých regresních parametrů pomocí metody **bootstrap**.
- V každém kroku je proveden náhodný výběr s opakováním prvků z jednotlivých skupin.
- Velikost výběru = počet prvků ve skupinách.
- Na tomto výběru je následně proveden odhad regresních parametrů.
- Postup je opakován, následně jsou zjištěny směrodatné odchylky výsledných odhadů parametrů.

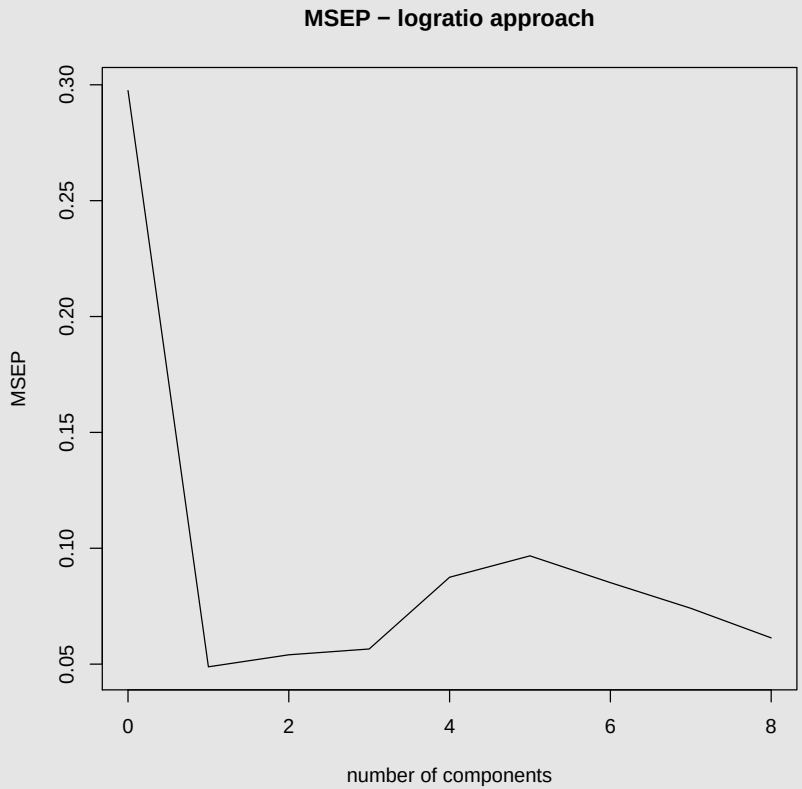
Praktický příklad I.



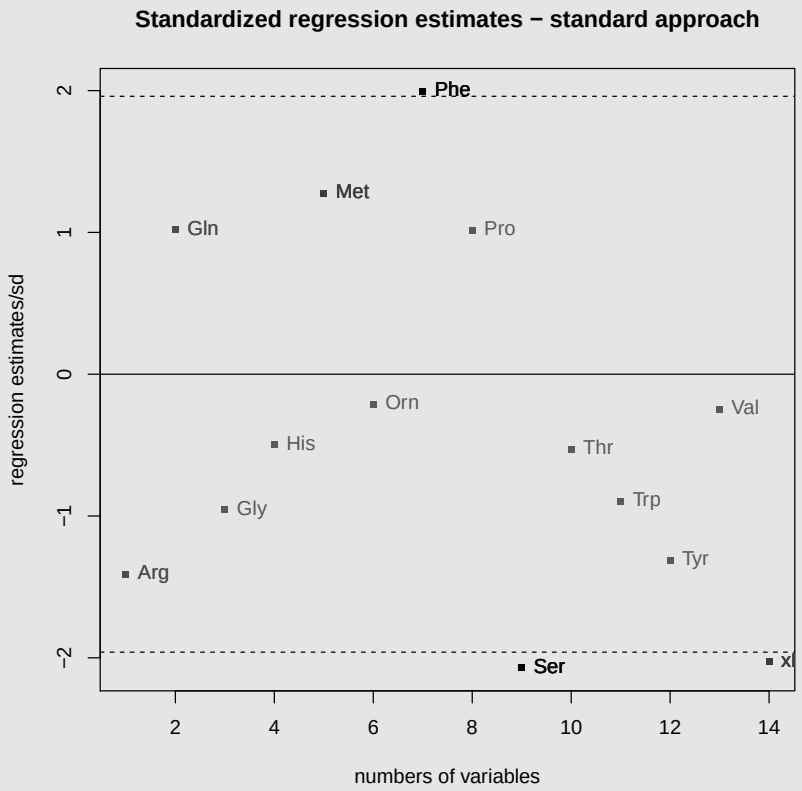


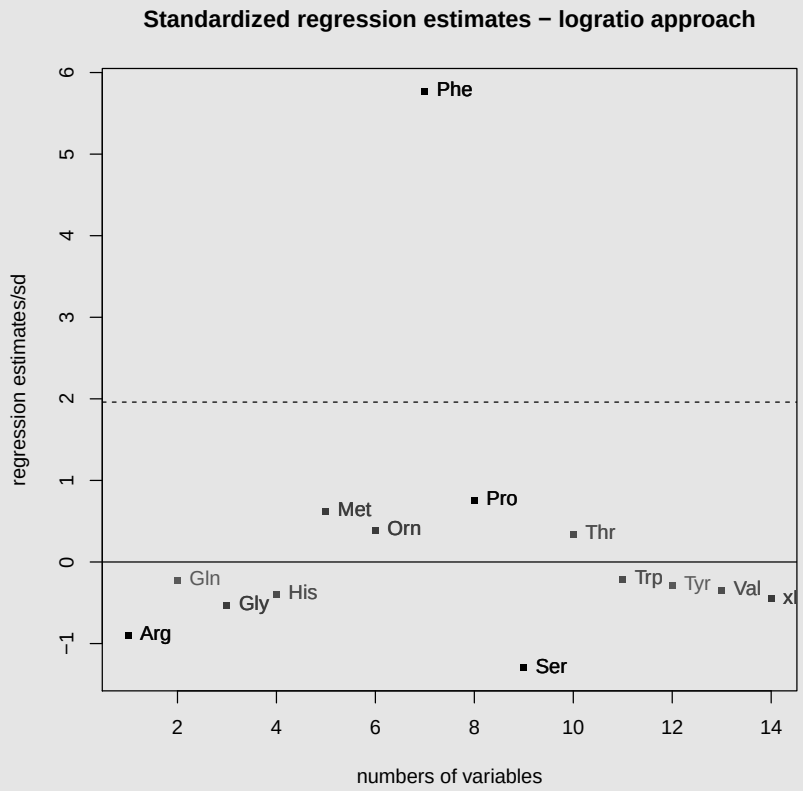
Praktický příklad I.



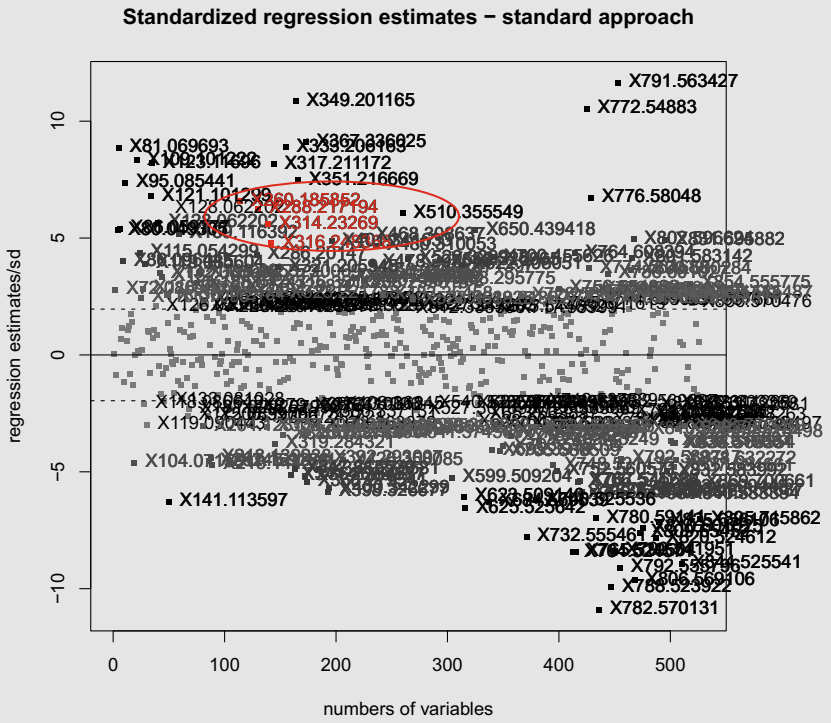


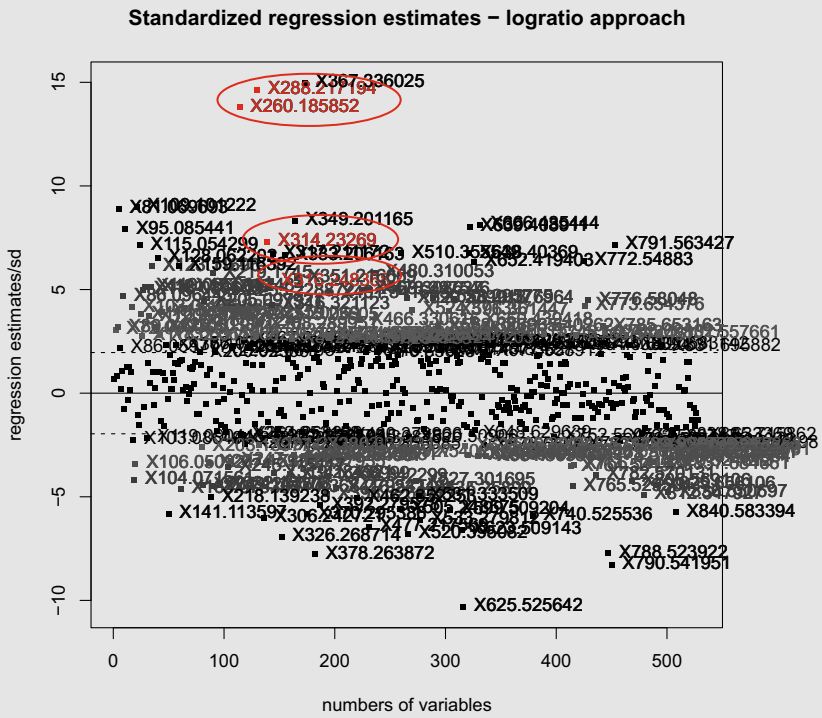
Praktický příklad I.





Praktický příklad II.





Závěr

- Při zpracování metabolomických dat je potřeba vzít v potaz jejich specifické vlastnosti.
- Pro praktické účely je třeba zajistit interpretovatelné ortonormální souřadnice.
- Úprava PLS-DA pro kompoziční data probíhá v logratio souřadnicích.
- Kompoziční přístup k PLS-DA dává velmi dobré výsledky - ukazuje signifikantní metabolity lépe než standardní metodika.

Literatura

- Roux, A., Lison, D., Junot, Ch. and Heilier, J. (2011) *Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review*. Clinical Biochemistry 44 , 119–135.
- Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.
- Hron, K., Filzmoser, P., Thompson, K. (2012) *Linear regression with compositional explanatory variables*. Journal of Applied Statistics 39 (5), 1115–1128.
- Egozcue, J. J., Pawlowsky-Glahn, V. (2005) *Groups of Parts and Their Balances in Compositional Data Analysis*. Mathematical Geology, 37 (7), 795–828.

Literatura

- Filzmoser, P., Hron, K., Reimann, C. (2012) *Interpretation of multivariate outliers for compositional data*. Computers & Geosciences 39, 77–85.
- Roux, A., Lison, D., Junot, Ch., Heilier, J. (2011) *Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review*. Clinical Biochemistry 44 , 119–135.