Bayesovský přístup k t-testům v kompoziční analýze metabolomických dat

Julie Rendlová[1,2,3], Karel Hron[1], Ondřej Vencálek[1], David Friedecký[2,3]

ROBUST 2018

[1]KMAAM, PřF, Univerzita Palackého
[2]OKB, Fakultní nemocnice Olomouc
[3]Laboratoř metabolomiky, ÚMTM, Univerzita Palackého

Jak vznikají metabolomická data

Proč dělat bayesovský t-test

Proč uplatňovat kompoziční přístup

Co je na posteru

Metabolomika

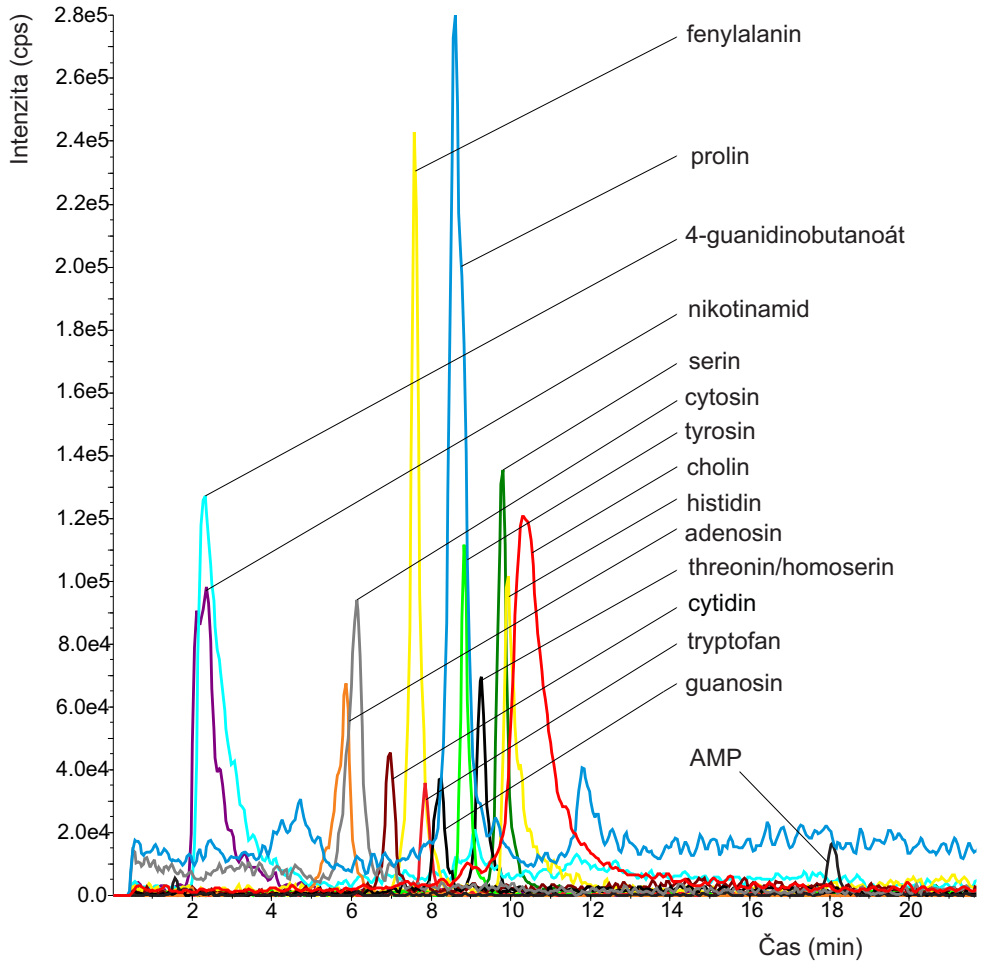
- zabývá se malými molekulami (aminokyseliny, sacharidy, . . .)
- soubor všech molekul v daném biologickém materiálu (krev, moč, buňky, . . .) je metabolom
- **Cílená metabolomická analýza**
  - seznam metabolitů před analýzou
  - méně naměřených látek
  - jednodušší interpretace
- **Necílená metabolomická analýza**
  - měří se vše, co se ve vzorku najde
  - vhodné, když se hledají nové markery nemocí
  - složitá interpretace (metabolity, fragmenty, adukty, šum)

Postup měření

- měření odebraných vzorků pacientů a kontrol v náhodném pořadí
- vzorky prochází
  1) **kapalinovým chromatografem**
     - průchod molekul určí RT metabolitů → „seřadí“ metabolity
  2) **hmotnostním spektrometrem** – štěpení na fragmenty dané hmoty
     - metabolity se stejným RT se rozbijí na fragmenty
     - přechod z dané hmoty na fragmenty určí původní metabolit
     - nejspecifičtější přechod se vykreslí (AUC odpovídá intenzitě původního metabolitu)

Postup měření – přístroje

Postup měření – chromatografické píky



Předzpracování dat

- píky se zintegrují → AUC
- náhodné vlivy ovlivňují měření → narovnání trendu pomocí LOESS regrese, kontrola CV
- předzpracováním se značný počet metabolitů vyřadí z dat
- vstupní data pro statistickou analýzu:
  - **pozorování** – pacienti a kontroly (řádky)
  - **proměnné** – plochy intenzit metabolitů (sloupce)
- stovky proměnných, jednotky až desítky pozorování → vysoce dimenzionální data

Tradiční přístup k jednorozměrným statistickým metodám

- parametrický t-test nebo neparametrický Wilcoxonův test na hladině významnosti $\alpha = 0.05$
- mnohonásobné testování $\rightarrow$ růst pravděpodobnosti získání falešně pozitivních výsledků
- při stovkách proměnných je Bonferroniho korekce příliš konzervativní
- při stovkách proměnných Bonferroniho korekce neumožní u Wilcoxna zamítat
- výsledky t-testů prudce ovlivněny přítomností odlehlých hodnot
- výsledky t-testů a Wilcoxonových testů jsou „chudé“ – pouze p-hodnoty $\rightarrow$ neposkytují možnost seřadit metabolity dle nejlepšího

Bayesovský t-test

- apriori předpokládáme data z t-rozdělení s těžkými chvosty → přirozeně robustní metoda (Kruschke, 2012)
  - **parametry pro apriorní rozdělení:** střední hodnoty (μ_1, μ_2) a směrodatné odchylky (σ_1, σ_2), parametr normality ($\nu \equiv$ počet stupňů volnosti)
- **Bayesova věta** – aposteriorní rozdělení je proporcionální věrohodnostní funkci (podmíněná pravděpodobnost dat za podmínky daných parametrů) vynásobené apriorním rozdělením (Kruschke, 2011) → pro aposteriorní rozdělení Bayesovského t-testu platí:

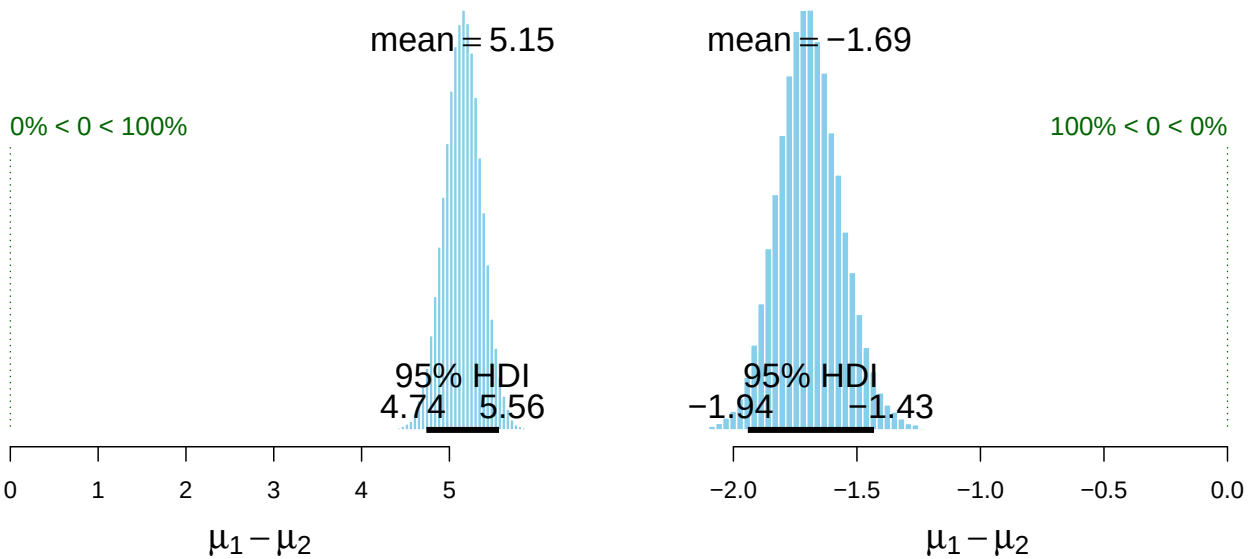
$$p(\mu_1, \sigma_1, \mu_2, \sigma_2, \nu | D) \propto p(D | \mu_1, \sigma_1, \mu_2, \sigma_2, \nu) \times p(\mu_1, \sigma_1, \mu_2, \sigma_2, \nu)$$

- k přibližnému výpočtu aposteriorního rozdělení použijeme MCMC – generuje kombinace parametrů

$$\left\langle \mu_1^j, \sigma_1^j, \mu_2^j, \sigma_2^j, \nu^j \right\rangle, \quad j = 1, \ldots, N$$

Bayesovský t-test

- zajímá nás především rozdíl μ_1 a $\mu_2 \rightarrow$ získáme aposteriorní histogramy MCMC rozdílů mezi pacienty a kontrolami
- konstrukce **HDI** intervalů – intervaly s 95 % nejčastějších aposteriorních hodnot
- pokud HDI neobsahuje nulu, „hypotéza“ o rovnosti parametrů se zamítá



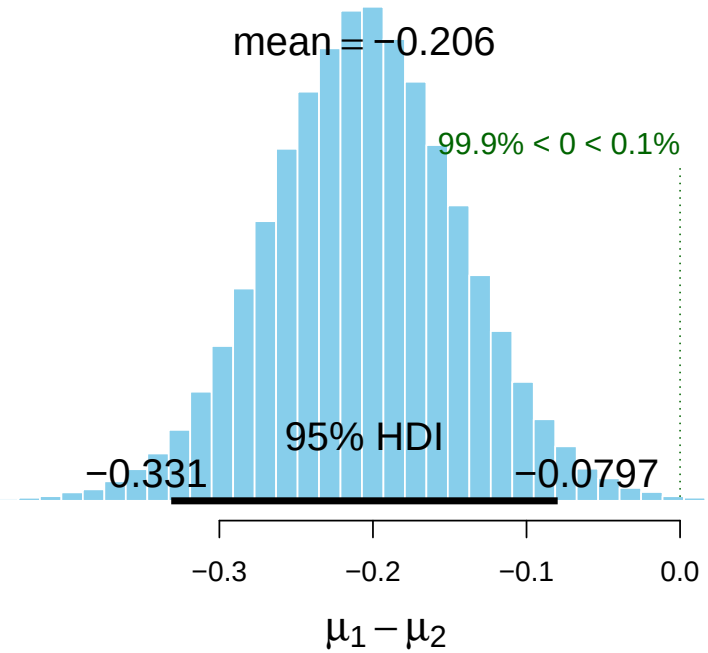
Kompoziční data

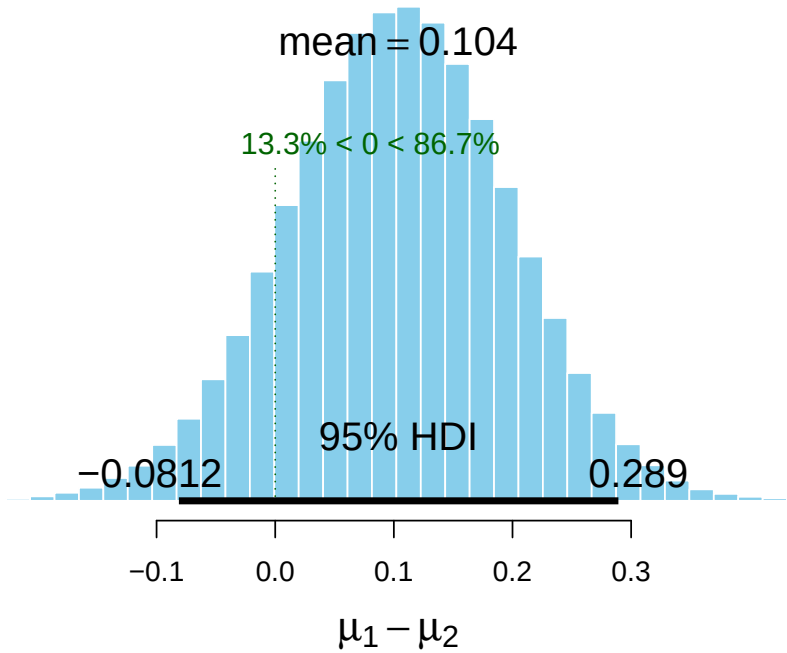
- metabolom v libovolném biologickém materiálu je složený z mnoha metabolitů → relativní struktura metabolomu
- relevantní informace pouze v poměrech mezi intenzitami metabolitů → kompoziční data na simplexu
- simplex se řídí Aitchisonovou geometrií namísto euklidovské (Aitchison, 1986; Pawlowsky-Glahn et al., 2015)
  → vyjádřit kompozice ze simplexu v reálných souřadnicích
- **centrované logpodílové koeficienty**

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_1^D x_i}}, \ln \frac{x_2}{\sqrt[D]{\prod_1^D x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_1^D x_i}} \right)$$

  - odpovídají centrování log transformovaných dat podle všech proměnných
  - jsou jednoduše interpretovatelné – dominance daného metabolitu vůči průměrnému chování celého metabolomu

Rozdíl mezi nekompozičním a kompozičním přístupem





Co je na posteru

- jak je konstruovaný **bayesovský t-test** – princip testu, apriorní rozdělení, MCMC
- jak to funguje u **vícenásobného testování** – hledání markerů, řazení markerů podle středních hodnot aposteriorních rozdělení a tzv. b-hodnot
- co to je **volcano graf** – bayesovská verze, pásma podle vzdálenosti HDI aposteriorních rozdělní od nuly
- **aplikace** na metabolomických datech z necílené analýzy – krevní skvrny pacientů s dědičným metabolomickým onemocněním (deficit acyl-CoA dehydrogenázy se středně dlouhým řetězcem; Najdekr et al., 2015) a zdravých kontrol
- **simulace** úbytku pozorování na polovinu v obou skupinách – porovnání bayesovského a tradičního přístupu
- **simulace** výskytu systematické chyby měření – porovnání kompozičního a nekompozičního přístupu

Literatura


Aitchison, J. (1986) *The statistical analysis of compositional data.* Chapman and Hall, London.


Kruschke, J.K. (2011) *Doing Bayesian Data Analysis.* Academic Press, New York.


Kruschke, J.K. (2012) Bayesian Estimation Supersedes the T Test. *Journal of Experimental Psychology: General 142*(2), pp.73–603.


Najdekr, L., Gardlo, A., Mádrová, L., Friedecký, D., Janečková, H., Correa, E. S., Goodacre, R., Adam, T. (2015) Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chainacyl-CoAdehydrogenase deficiency. *Talanta 139*, pp.2–66.


Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2015) *Modeling and analysis of compositional data*. Wiley, Chichester.