Robustní metody pro kompoziční data

Karel Hron

Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta
Univerzita Palackého v Olomouci

Robust 2010
31. ledna – 5. února 2010, Králíky

Obsah

... aneb než se dostaneme k oněm slíbeným robustním metodám ...

Co jsou kompoziční data (kompozice)?

- **definice: části nějakého celku nesoucí pouze relativní informaci**

Co jsou kompoziční data (kompozice)?

- **definice: části nějakého celku nesoucí pouze relativní informaci**
- **obvyklé jednotky měření**: procenta, mg/kg (**konstantní součet složek**), moly na litr (**součet není konstantní**)

Co jsou kompoziční data (kompozice)?

- **definice: části nějakého celku nesoucí pouze relativní informaci**
- **obvyklé jednotky měření**: procenta, mg/kg (**konstantní součet složek**), moly na litr (**součet není konstantní**)
- **příklady**: geochemická data - proporcionální zastoupení minerálů v hornině; koncentrace fenolických kyselin ve víně (mg/l); zastoupení politických stran dle volebních výsledků; výdaje domácností na konečnou spotřebu (jídlo, ubytování, ošacení) a další

Co jsou kompoziční data (kompozice)?

- **definice: části nějakého celku nesoucí pouze relativní informaci**
- **obvyklé jednotky měření**: procenta, mg/kg (**konstantní součet složek**), moly na litr (**součet není konstantní**)
- **příklady**: geochemická data - proporcionální zastoupení minerálů v hornině; koncentrace fenolických kyselin ve víně (mg/l); zastoupení politických stran dle volebních výsledků; výdaje domácností na konečnou spotřebu (jídlo, ubytování, ošacení) a další
- problém zpracování dat s ***konstantním součtem*** byl v minulosti řešen pomocí **standardní statistiky**, tedy za předpokladu **euklidovské geometrie v reálném prostoru**

"Spurious correlation", Karl Pearson (1897)

Možné problémy tohoto přístupu si budeme ilustrovat na **příkladu**: Geologové A a B zkoumají proporce jednotlivých složek v půdních vzorcích. A obdrží kompozici ze čtyř složek (živočišná, rostlinná, neživá, voda), B po vysušení vzorků pouze první tři složky. Předpokládáme absenci chyb měření:

"Spurious correlation", Karl Pearson (1897)

Možné problémy tohoto přístupu si budeme ilustrovat na **příkladu**: Geologové A a B zkoumají proporce jednotlivých složek v půdních vzorcích. A obdrží kompozici ze čtyř složek (živočišná, rostlinná, neživá, voda), B po vysušení vzorků pouze první tři složky. Předpokládáme absenci chyb měření:

	geolog A				geolog B		
výběr	x_1	x_2	x_3	x_4	x_1'	x_2'	x_3'
1	0.1	0.2	0.1	0.6	0.25	0.50	0.25
2	0.2	0.1	0.2	0.5	0.40	0.20	0.40
3	0.3	0.3	0.1	0.3	0.43	0.43	0.14

"Spurious correlation", Karl Pearson (1897)

Možné problémy tohoto přístupu si budeme ilustrovat na **příkladu**: Geologové A a B zkoumají proporce jednotlivých složek v půdních vzorcích. A obdrží kompozici ze čtyř složek (živočišná, rostlinná, neživá, voda), B po vysušení vzorků pouze první tři složky. Předpokládáme absenci chyb měření:

	geolog A				geolog B		
výběr	x_1	x_2	x_3	x_4	x_1'	x_2'	x_3'
1	0.1	0.2	0.1	0.6	0.25	0.50	0.25
2	0.2	0.1	0.2	0.5	0.40	0.20	0.40
3	0.3	0.3	0.1	0.3	0.43	0.43	0.14

Cor A	x_1	x_2	x_3	x_4
x_1	1.00	0.50	0.00	-0.98
x_2		1.00	-0.87	-0.65
x_3			1.00	0.19
x_4				1.00

Cor B	x_1'	x_2'	x_3'
x_1'	1.00	-0.57	-0.05
x_2'		1.00	-0.79
x_3'			1.00

Logratio analýza kompozic

Příčina selhání v předchozím příkladu: **ignorování "přirozenosti" kompozičních dat** (resp. důsledků plynoucích z jejich definice). Obecně, každá statistická metoda by měla v takovém případě mít následující vlastnosti:

Logratio analýza kompozic

Příčina selhání v předchozím příkladu: **ignorování "přirozenosti" kompozičních dat** (resp. důsledků plynoucích z jejich definice). Obecně, každá statistická metoda by měla v takovém případě mít následující vlastnosti:

- invariantnost na změnu škály (scale invariance)
- invariantnost na permutaci (permutation invariance)
- subkompoziční soudržnost (subcompositional coherence)
  - subkompoziční dominance (subcompositional dominance)
  - zachování podílů (ratio preserving)

Logratio analýza kompozic

Příčina selhání v předchozím příkladu: ignorování "přirozenosti" kompozičních dat (resp. důsledků plynoucích z jejich definice). Obecně, každá statistická metoda by měla v takovém případě mít následující vlastnosti:

- invariantnost na změnu škály (scale invariance)
- invariantnost na permutaci (permutation invariance)
- subkompoziční soudržnost (subcompositional coherence)
  - subkompoziční dominance (subcompositional dominance)
  - zachování podílů (ratio preserving)

Řešení: John Aitchison (1986): "The statistical analysis of compositional data", zobrazení kompozic z D-složkového simplexu, výběrového prostoru kompozic, do reálného prostoru $\mathbb{R}^{D-1}$, resp. $\mathbb{R}^{D}$ pomocí ***log-ratio*** transformací.

Simplex jako výběrový prostor kompozic a Aitchisonova geometrie

- **výběrový prostor:** tradičně D-složkový simplex

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D), \, x_i > 0 \, \sum_{i=1}^{D} x_i = \kappa \right\};$$

κ je zvolená konstanta (1, 100) $\Rightarrow$ simplex je výběrovým prostorem reprezentací kompozic při daném součtu složek (bez ztráty informace)

Simplex jako výběrový prostor kompozic a Aitchisonova geometrie

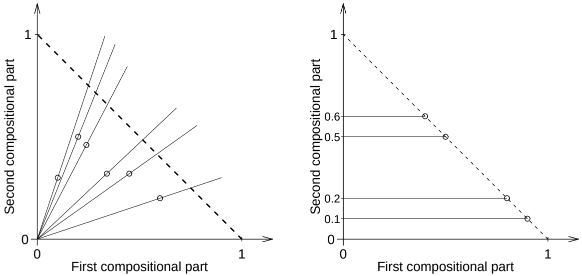
- **výběrový prostor:** tradičně D-složkový simplex

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D),\, x_i > 0 \sum_{i=1}^{D} x_i = \kappa \right\};$$

  κ je zvolená konstanta (1, 100) $\Rightarrow$ simplex je výběrovým prostorem reprezentací kompozic při daném součtu složek (bez ztráty informace)

- **"přirozená" geometrie kompozic:** **Aitchisonova geometrie**, má vlastnosti euklidovské geometrie, založena na operacích **perturbace**, **mocninná transformace** a na **Aitchisonově skalárním součinu**

Kompoziční data jsou třídy ekvivalence



$\Rightarrow \kappa$ není důležité, výběr reprezentanta = operace uzávěru:
pro $\mathbf{x} \in \mathbb{R}_+^D$

$$\mathcal{C}[x_1, \dots, x_D] = \left[\frac{\kappa x_1}{\sum_{i=1}^D x_i}, \dots, \frac{\kappa x_D}{\sum_{i=1}^D x_i}\right]$$

Aitchisonova geometrie na $\mathcal{S}^D$ a práce v souřadnicích

pro $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$ a pro $\mathcal{C}$ jako operaci uzávěru:

Aitchisonova geometrie na $\mathcal{S}^D$ a práce v souřadnicích

pro $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$ a pro $\mathcal{C}$ jako operaci uzávěru:

- **perturbace**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]$
- **mocninná transformace**: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$
- **Aitchisonův skalární součin**: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$

Aitchisonova geometrie na $\mathcal{S}^D$ a práce v souřadnicích

pro $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$ a pro $\mathcal{C}$ jako operaci uzávěru:

- **perturbace**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]$
- **mocninná transformace**: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$
- **Aitchisonův skalární součin**: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$

definují $(D-1)$**-dimenzionální Hilbertův prostor na** $\mathcal{S}^D$
$\Rightarrow$ **můžeme vytvořit ortonormální bázi na simplexu a vyjádřit kompozice v této bázi** $\Rightarrow$ $\mathbf{z} = (z_1, \ldots, z_{D-1}) \in \mathbb{R}^{D-1}$ (isometric logratio transformace)

Aitchisonova geometrie na $\mathcal{S}^D$ a práce v souřadnicích

pro **x**, **y** $\in \mathcal{S}^D$, $\alpha \in \mathbb{R}$ a pro $\mathcal{C}$ jako operaci uzávěru:

- **perturbace**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1y_1, \ldots, x_Dy_D]$
- **mocninná transformace**: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$
- **Aitchisonův skalární součin**: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$

definují $(D-1)$**-dimenzionální Hilbertův prostor na** $\mathcal{S}^D$
$\Rightarrow$ **můžeme vytvořit ortonormální bázi na simplexu a vyjádřit kompozice v této bázi** $\Rightarrow$ $\mathbf{z} = (z_1, \ldots, z_{D-1}) \in \mathbb{R}^{D-1}$ (isometric logratio transformace)

- **ortonormální souřadnice se řídí standardními pravidly euklidovské geometrie**
- **můžeme přímo aplikovat běžné statistické metody**
- **interpretace výsledků v souřadnicích nebo zpět na simplexu**
- **interpretace souřadnic: sequential binary partition**
- výjimka - kompoziční biplot: vyjádření v generujícím systému na simplexu (centred logratio transformace)

Robustní metody pro kompoziční data

Společně s P. Filzmoserem a M. Templem (TU Wien). Principielně se využívá práce v souřadnicích (resp. v generujícím systému) a vlastností ekvivariantních odhadů (např. MCD - Minimum Covariance Determinant), se specifikami co se postupu i interpretace výsledků týče. Takto lze robustifikovat např.

Robustní metody pro kompoziční data

Společně s P. Filzmoserem a M. Templem (TU Wien). Principielně se využívá práce v souřadnicích (resp. v generujícím systému) a vlastností ekvivariantních odhadů (např. MCD - Minimum Covariance Determinant), se specifikami co se postupu i interpretace výsledků týče. Takto lze robustifikovat např.

- **identifikaci odlehlých hodnot, imputaci chybějících hodnot v datech**
- **metodu hlavních komponent + biplot**
- **korelační analýzu** (**interpretace v souřadnicích!**)
- **faktorovou analýzu**
- **diskriminační analýzu**

Základní literatura a další informace

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

Software: R-knihovny `compositions`, **robCompositions**

Informační rozcestník: `http://compositions.sweb.cz/` (CoDaWork 2011)

Kontakt: `http://hronk.sweb.cz/` (mj. linky na články o robustních metodách pro kompoziční data), `hronk@seznam.cz`