Aplikace $\mathcal{T}$-prostorů při modelování kompozičních časových řad

P. Kynčlová[1,3]
P. Filzmoser[1], K. Hron[2,3]

[1]Department of Statistics and Probability Theory
Vienna University of Technology

[2]Katedra matematické analýzy a aplikací matematiky
Univerzita Palackého v Olomouci

[3]Katedra geoinformatiky
Univerzita Palackého v Olomouci

ROBUST 2014, 23. ledna 2014

Obsah

Kompoziční časové řady

VAR model

$\mathcal{T}$-prostory

Praktický příklad

Kompoziční data

- data nesoucí výhradně **relativní** informaci, jejichž výběrovým prostorem je simplex $\mathcal{S}^D$

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

- **Aitchisonova geometrie na simplexu**

  $\Longrightarrow$ logratio transformace ze simplexu do reálného prostoru

- **izometrická logratio transformace**: $\mathrm{ilr}(\mathbf{x}) = (z_1, \ldots, z_{D-1})'$

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{l=j+1}^{D} x_l}}, \quad j = 1, \ldots, D-1 \quad (1)$$

$\Longrightarrow$ souřadnice z_1 obsahuje veškerou relativní informaci o x_1 vzhledem k ostatním složkám $x_2, \ldots, x_D$

Kompoziční časové řady

Definice

- mnohorozměrné časové řady pro kompoziční data

$$\{\mathbf{x}_t : t = 1, \ldots, n\}, \qquad \mathbf{x}_t = (x_{1t}, \ldots, x_{Dt})' \in \mathcal{S}^D$$

- časové řady s kladnými reálnými prvky

$$x_{1t}, \ldots, x_{Dt} > 0$$

- **problém:** konstatní součet v každém čase $t \Longleftrightarrow$ často roven 1 (proporcionální data)
- **řešení:** logratio transformace ze simplexu do prostoru souřadnic ($\mathcal{S}^D \to \mathbb{R}^{D-1}$)

Kompoziční časové řady

Princip modelování kompozičních časových řad

- transformace do prostoru souřadnic:

  izometrická logratio (ilr) transformace (1)

  $\Longrightarrow$ výhodné interpretační vlastnosti

- aplikace standardních metod pro analýzu mnohorozměrných časových řad v prostoru souřadnic
- návrat zpět na simplex $\Longrightarrow$ inverzní ilr transformace

Vektorový autoregresní proces VAR(p)

Definice

- **VAR(p) model** v redukovaném tvaru

$$\mathbf{z}_t = \mathbf{c} + \mathbf{A}_{\mathbf{V}}^{(1)}\mathbf{z}_{t-1} + \mathbf{A}_{\mathbf{V}}^{(2)}\mathbf{z}_{t-2} + \cdots + \mathbf{A}_{\mathbf{V}}^{(p)}\mathbf{z}_{t-p} + \boldsymbol{\epsilon}_t, \qquad (2)$$

  kde $\mathbf{B} = [\mathbf{c}_{\mathbf{V}}, \mathbf{A}_{\mathbf{V}}^{(1)}, \ldots, \mathbf{A}_{\mathbf{V}}^{(p)}]$ jsou parametry a $\boldsymbol{\epsilon}_t$ chybová složka, $\boldsymbol{\epsilon}_t \sim \mathcal{WN}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$
- p se nazývá řád modelu
- pozorování $\mathbf{z}_t$ je modelováno pomocí p předchozích pozorování $\mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-p}$.

- $\mathbf{z}_t = \mathrm{ilr}_{\mathbf{V}}(\mathbf{x}_t)$ jsou ilr souřadnice kompozice $\mathbf{x}_t \in \mathcal{S}^D$ dané maticí $\mathbf{V}$ a $\mathbf{z}_t^* = \mathrm{ilr}_{\mathbf{V}^*}(\mathbf{x}_t)$ ilr souřadnice dané maticí $\mathbf{V}^*$
- VAR(p) modely jsou kompozičně ekvivalentní při použití libovolné ilr transformace

$$\mathbf{z}_t = \mathbf{c}_{\mathbf{V}} + \mathbf{A}_{\mathbf{V}}^{(1)}\mathbf{z}_{t-1} + \mathbf{A}_{\mathbf{V}}^{(2)}\mathbf{z}_{t-2} + \ldots \mathbf{A}_{\mathbf{V}}^{(p)}\mathbf{z}_{t-p},$$

$$\mathbf{z}_t^* = \mathbf{c}_{\mathbf{V}^*} + \mathbf{A}_{\mathbf{V}^*}^{(1)}\mathbf{z}_{t-1}^* + \mathbf{A}_{\mathbf{V}^*}^{(2)}\mathbf{z}_{t-2}^* + \ldots \mathbf{A}_{\mathbf{V}^*}^{(p)}\mathbf{z}_{t-p}^*.$$

Vektorový autoregresní proces VAR(p)

Endomorfismus na simplexu $\mathcal{S}^D$

$$\mathbf{A} \boxdot \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^{D} x_j^{a_{1j}}, \ldots, \prod_{j=1}^{D} x_j^{a_{Dj}}\right)^\top \qquad \mathbf{A} \in \mathbb{R}_{D\times D}, \mathbf{x} \in \mathcal{S}^D$$

- představuje lineární transformaci vzhledem k Aitchisonově geometrii
  - za podmínky $\mathbf{A}\mathbf{1}_D = \mathbf{0}_D$ (jinak $\mathbf{A} \boxdot \mathbf{x} \neq \mathbf{A} \boxdot (k\mathbf{x})$)
- pokud navíc platí $\mathbf{A}^\top \mathbf{1}_D = \mathbf{0}_D$, pak funkce $\mathbf{x} \to \mathbf{A} \boxdot \mathbf{x}$ se nazývá **endomorfismus** na simplexu $\mathcal{S}^D$
- máme-li endomorfismus $\mathbf{y} = \mathbf{A} \boxdot \mathbf{x}$ na simplexu, pak ho můžeme též vyjádřit i v prostoru souřadnic
  - $\mathrm{clr}(\mathbf{y}) = \mathbf{A} \cdot \mathrm{clr}(\mathbf{x})$
  - $\mathrm{ilr}_\mathbf{V}(\mathbf{y}) = \mathbf{A}_\mathbf{V} \cdot \mathrm{ilr}_\mathbf{V}(\mathbf{x}) \Leftrightarrow \mathrm{ilr}(\mathbf{y}) = \mathbf{V}^\top \mathbf{A}\mathbf{V} \cdot \mathrm{ilr}(\mathbf{x})$

Vektorový autoregresní proces VAR(p)

VAR(p) na simplexu $\mathcal{S}^D$

- výsledný VAR model na simplexu **nezávisí** na volbě konkrétní transformace

$$\mathbf{x}_t = \mathbf{b} \oplus \left(\mathbf{A}^{(1)} \boxdot \mathbf{x}_{t-1}\right) \oplus \cdots \oplus \left(\mathbf{A}^{(p)} \boxdot \mathbf{x}_{t-p}\right) \oplus \mathbf{w}_t$$

- modelování kompozičních časových řad přímo na simplexu $\Longrightarrow$ Barcelo-Vidal et al.(2011)

Vektorový autoregresní proces VAR(p)

Odhad parametrů

- stacionární VAR(p) model lze rovněž zapsat v maticovém tvaru

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{E}$$

kde $\mathbf{Y} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$, $\mathbf{Z}_t = (1, \mathbf{z}'_{t-1}, \ldots, \mathbf{z}'_{t-p})'$ je t-tý řádek matice $\mathbf{Z} \in \mathbb{R}^{n \times [(D-1)p+1]}$

- parametry $\mathbf{B}$ lze odhadnout pomocí MNČ pro každou rovnici zvlášť na základě
  - "sample" $\rightarrow \mathbf{z}_1, \ldots, \mathbf{z}_n$
  - "presample" $\rightarrow z_{-p+1}, \ldots, \mathbf{z}_0$

Vektorový autoregresní proces VAR(p)

Testování hypotéz: Grangerova kauzalita

- testování kauzality založené na predikci
- x_1 „Granger causes“ x_2
  $\Longrightarrow$ minulé (zpožděné) hodnoty x_1 obsahují statisticky významnou informaci o budoucích hodnotách x_2
- test hypotézy $\mathrm{H}_0 : \mathbf{R}\hat{\beta} = \mathbf{r}$
- **Waldova statistika**

$$(\mathbf{R}\hat{\beta}-\mathbf{r})'\left\{\mathbf{R}\left[\widehat{avar}(\hat{\beta})\right]\mathbf{R}'\right\}^{-1}(\mathbf{R}\hat{\beta}-\mathbf{r}) \sim F\left(q, n-(D-1)p-1\right)$$

q . . . hodnost matice $\mathbf{R}$

Teorie $\mathcal{T}$-prostorů

- **kompoziční data** = data nesoucí výhradně **relativní** informaci $\rightarrow$ v podílech mezi složkami
- **kompoziční časové řady** $\rightarrow$ zajímá nás též **absolutní** informace o celkovém množství v čase t (např. při predikci budoucích hodnot)
- teorie $\mathcal{T}$-prostorů definuje rozšířený vektorový prostor $\mathcal{T} = \mathbb{R}_+^D \times \mathcal{S}^D \Longrightarrow$ umožňuje modelovat podíly mezi jednotlivými složkami s úhrnnými hodnotami těchto složek současně v jednom modelu

Teorie $\mathcal{T}$-prostorů

Aplikace teorie $\mathcal{T}$-prostorů na časové řady

- vektor $\tilde{\mathbf{x}} = [t(\mathbf{x}), \mathcal{C}(\mathbf{x})] = [t_x, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D]$ je prvkem prostoru $\mathcal{T} = \mathbb{R}_+^D \times \mathcal{S}^D$
  - $\mathcal{C}(\mathbf{x})$ ..... kompozice, $\mathbf{x} \in \mathcal{S}^D$
  - $t(\mathbf{x})$ ..... celkový součet hodnot složek kompozice v daném okamžiku t

$$t(\mathbf{x}) = \sum_{i=1}^{D} x_i$$

- kompozice $\mathbf{x}$ jsou modelovány pomocí logratio transformací
- úhrnné hodnoty $t(\mathbf{x})$ jsou uvažovány v jejich zlogaritmované podobě

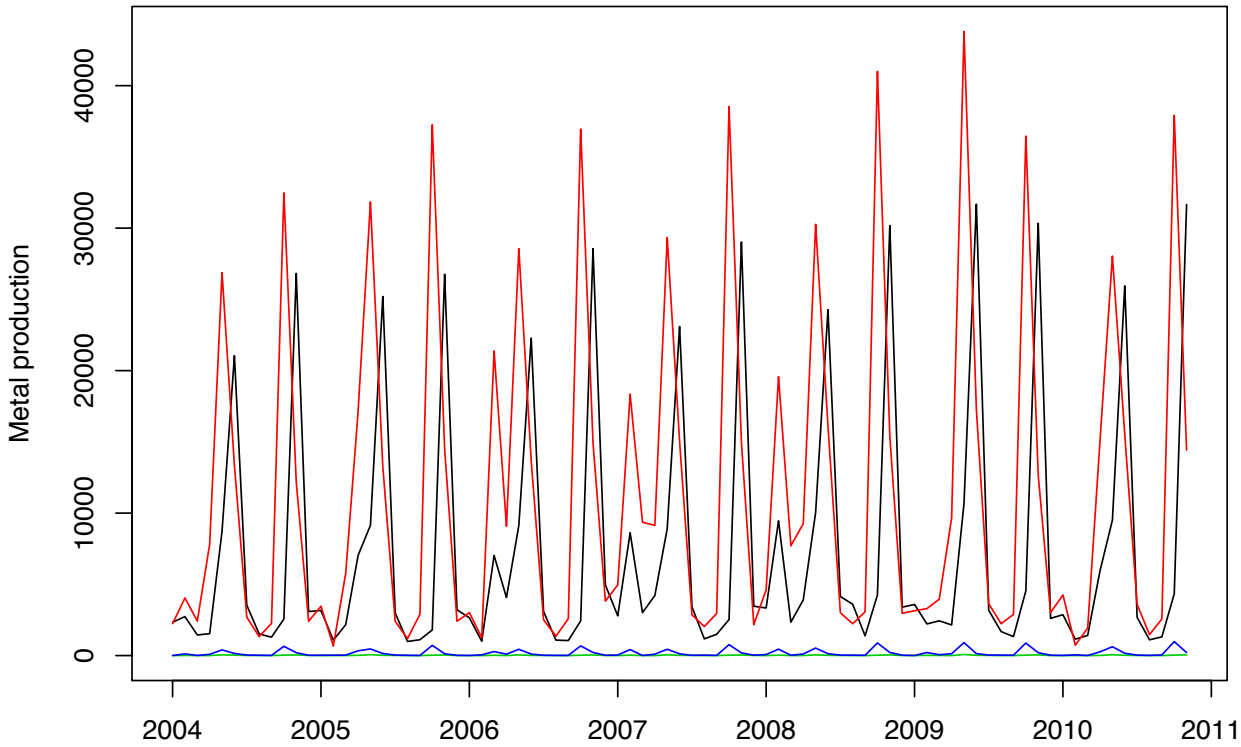
Praktický příklad

Hrubé bonusy k příjmům zaměstnanců

- hrubé bonusy k příjmům zaměstnanců v odvětví zpracovávání kovů v Rakousku (v tisících eurech) $\Longrightarrow$ leden 2004 - listopad 2011
  - x_1 ..... tzv. „bílé límečky“
  - x_2 ..... dělníci
  - x_3 ..... učni v komerční sféře
  - x_4 ..... učni v industriální sféře
  - X_t ..... celkový úhrn hodnot v čase $t \rightarrow x_1 + x_2 + x_3 + x_4$

	x1	x2	x3	x4	Xt
1	1494.70	610.50	2.00	19.40	2126.60
2	723.40	785.20	3.70	14.50	1526.80
3	2168.10	1171.90	3.00	9.90	3352.90
4	2293.70	3173.30	4.80	60.60	5532.40
5	12497.20	12782.20	21.10	97.80	25398.30
6	13851.50	23369.40	176.00	380.30	37777.20

Praktický příklad



Kompoziční časové řady - absolutní hodnoty, x_1 (—), x_2 (—), x_3 (—), x_4 (—)

Praktický příklad

Volba modelu

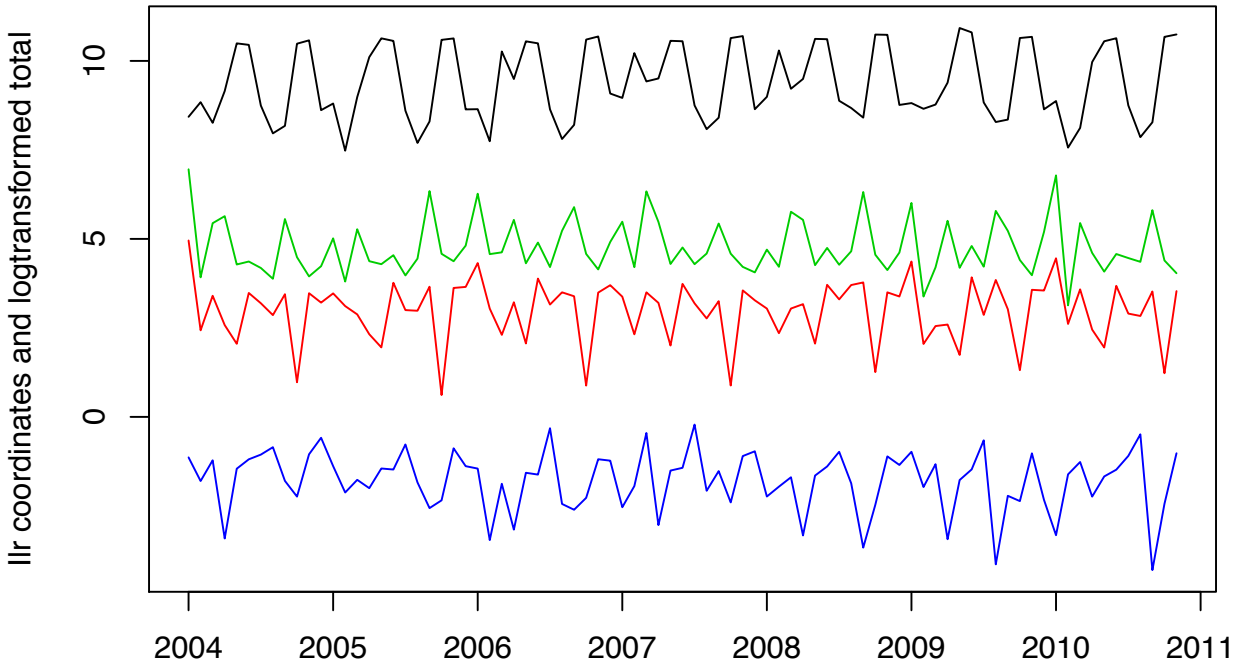
Dva přístupy:

- aplikace VAR modelu standardním způsobem na proměnné x_1, x_2, x_3, x_4
- s využitím teorie $\mathcal{T}$-prostorů byly proměnné transformovány pomocí ilr transformace a zlogaritmovaný celkový úhrn hodnot $\log(X_t)$ byl zahrnut do modelu jako další proměnná

Přístup	AIC(p)	HQ(p)	SC(p)	FPE(p)
standardní	10	1	1	2
kompoziční	10	1	1	1

$\Longrightarrow$ **VAR(1) model**

Praktický příklad



Ilr souřadnice pro kompoziční časové řady (z_1 (——), z_2 (——), z_3 (——))

a celkový úhrn hodnot složek X_t po zlogaritmování.

Praktický příklad

Testování Grangerovy kauzality

- na rozdíl od standardního přístupu nám použití ilr transformace a teorie $\mathcal{T}$-prostorů umožňuje testovat kauzalitu nejen mezi proměnnými, ale i mezi úhrnnými hodnotami složek
- vzhledem k výhodným interpretačním vlastnostem dané ilr transformace lze otestovat nulovou hypotézu

  „z_1 nemá signifikantní vliv na $z_2, z_3, \log(X_t)$“
- testování Grangerovy kauzality kompozičním přístupem je **nesrovnatelné** s aplikací VAR modelu klasickým způsobem
  - samotné testování probíhá v prostoru ortonormálních souřadnic
  - celkové úhrnné hodnoty jsou zahrnuty do modelu jako další proměnná

Praktický příklad

Testování Grangerovy kauzality

nulová hypotéza	**p-hodnota**
z_1 nemá vliv na $z_2, z_3, \log(X_t)$	
$z_1 =$ rel. informace o x_1 vzhledem k x_2, x_3, x_4	0.526
$z_1 =$ rel. informace o x_2 vzhledem k x_1, x_3, x_4	0.852
$z_1 =$ rel. informace o x_3 vzhledem k x_1, x_2, x_4	0.659
$z_1 =$ rel. informace o x_4 vzhledem k x_1, x_2, x_3	0.043
$\log(X_t)$ nemá vliv na z_1, z_2, z_3	0.279

$\Longrightarrow$ relativní informace obsažená v x_4 má dle testu významný vliv na předpověď budoucích hodnot složek x_1, x_2, x_3 i celkových úhrnných hodnot X_t

Závěr

- určení předpovědí **nezávisí** na typu zvolené logratio transformace
- výběr ilr transformace usnadňuje možnost **interpretace** (např. testování hypotéz)
- teorie $\mathcal{T}$-prostorů umožňuje modelovat nejen proporcionální strukturu dat, ale i celkové absolutní hodnoty **v jednom modelu**
- použití $\mathcal{T}$-prostorů poskytuje **nové možnosti testování** Grangerovy kauzality (testování kauzálních vztahů i mezi kompozicemi a úhrnnými hodnotami)
- kompoziční přístup ukázal vyšší přesnost předpovědí na základě RMSEP

Reference

Barceló-Vidal, C., Aguilar, L., Martín-Fernández, J. A. (2011)

Compositional VARIMA time series. In V. Pawlowsky-Glahn and A. Buccianti, eds.,

Compositional Data Analysis. Theory and Applications.

John Wiley & Sons, Chichester, pp. 87–103.

Fišerová, E., Hron, K. (2012).

On interpretation of orthonormal coordinates for compositional data.

Mathematical Geosciences 43(4), 455–468.

Pawlowsky-Glahn, V., Egozcue, J.J., Lovell, D. (2013).

The product space $\mathcal{T}$ (tools for compositional data with a total). In K. Hron, P. Filzmoser and M. Templ, editors,

Proceedings of CoDaWork'13, The 5th Compositional Data Analysis Workshop. Vorau, Austria.

Reference

Aitchison, J. (1986).
The Statistical Analysis of Compositional Data.
Chapman and Hall Ltd., London, UK.

Lütkepohl, L. (2005).
New Introduction to Multiple Time Series Analysis.
Springer, Berlin.

Pawlowsky-Glahn, V., Buccianti, A., eds. (2011).
Compositional Data Analysis: Theory and Applications.
Wiley, Chichester.

$$\hat{\beta} = \mathrm{vec}(\hat{\mathbf{B}}) = \begin{pmatrix} \hat{\boldsymbol{c}}^\top \\ (\hat{\mathbf{A}}^{(1)})^\top \\ \vdots \\ (\hat{\mathbf{A}}^{(p)})^\top \end{pmatrix} \qquad \widehat{avar}(\hat{\beta}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} \otimes (\mathbf{Z}\mathbf{Z}^\top)^{-1}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \frac{1}{n-(D-1)p-1} \sum_{t=1}^{n} (\mathbf{z}_t - \mathbf{Z}\hat{\mathbf{B}})(\mathbf{z}_t - \mathbf{Z}\hat{\mathbf{B}})^\top.$$

Testování Grangerovy kauzality ve standardním VAR modelu

nulová hypotéza	**p-hodnota**
x_1 nemá vliv na x_2, x_3, x_4	0.749
x_2 nemá vliv na x_1, x_3, x_4	0.190
x_3 nemá vliv na x_1, x_2, x_4	0.085
x_4 nemá vliv na x_1, x_2, x_3	$6.158 \cdot 10^{-6}$

$\Longrightarrow$ i standardně docházíme k výsledku, že x_4 má signifikantní vliv na x_1, x_2, x_3, a tedy minulé hodnoty x_4 mají významný efekt na odhad budoucích hodnot ostatních proměnných

Přesnost předpovědí pro použité modely

Root mean squared error of prediction (RMSEP)

$$\mathrm{RMSEP} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2},$$

kde m je počet predikovaných hodnot.

přístup	RMSEP
standardní	170.94
kompoziční	167.02