VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

STATISTICKÉ MODELOVÁNÍ ZNEČIŠTĚNÍ OVZDUŠÍ PRAŠNÝM AEROSOLEM

STATISTICAL MODELLING OF AIR POLLUTION BY DUST AEROSOL

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. MARTINA MIKUŠKOVÁ
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. JAROSLAV MICHÁLEK, CSc.
SUPERVISOR

BRNO 2014

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2013/14

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Bc. Martina Mikušková

který/která studuje v **magisterském studijním programu**

obor: **Matematické inženýrství (3901T021)**

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Statistické modelování znečištění ovzduší prašným aerosolem

v anglickém jazyce:

Statistical Modelling of Air Pollution by Dust Aerosol

Stručná charakteristika problematiky úkolu:

Kvalita ovzduší je jedním ze základních ukazatelů kvality životního prostředí jako celku. Proto je zjišťování kvality ovzduší v popředí zájmů odborníků z celé řady environmentálních vědních oborů. V poslední době se hodně uplatňuje statistický přístup k vyhodnocování zjištěných dat o kvalitě ovzduší. Jedním z hlavních cílů těchto analýz je provést předpovědi denních koncentrací prachových částic v ovzduši (viz [1]) nebo identifikovat hlavní zdroje znečištění (viz [2]) na základě naměřených datových vzorků. K tomu se používají mnohorozměrné metody matematické statistiky, zejména regresní analýza, faktorová analýza a různé klasifikační techniky (klasifikační a regresní stromy, diskriminační analýza)(viz [3]). Pomocí těchto metod lze provést predikci znečištění s ohledem a sledované doprovodné faktory (např. klimatické) nebo stanovit zdroje zjištěných pevných částic vznášejících se v ovzduší.

Úkolem diplomanta bude pomocí mnohorozměrných statistických metod analyzovat složení prašného aerosolu, případně identifikovat zdroje znečištění a popsat statistickou vazbu mezi koncentrací prašného aerosolu a doprovodnými proměnými.

Cíle diplomové práce:

Popsat vybrané mnohorozměrné statistické metody (regresní techniky, faktorovou analýzu včetně robustního přístupu (viz [4])) vhodné pro analýzu dat získaných měřením prašného aerosolu PM1 v brněnské aglomeraci v letním a zimním období. Uvedené metody použít k analýze těchto dat a zhodnotit výsledky provedených analýz.

Seznam odborné literatury:

[1] Stadlober E.,Hübnerová Z., Michálek, J. and Kolář M.: Forecasting of Daily PM10 Concentrations in Brno and Graz by Different Regression Approaches. Austrian Journal of Statistics. Vol. 41 (2012) Number 4. p. 287-310

[2] Křůmal K.,Mikuška P., Vojtěšek M. a Večeřa Z. Seasonal variations of monosaccharide anhydrides in PM1
and PM2.5 aerosol in urban areas. Atmospheric Environment 44 (2010), p. 5148 - 5155

[3] Pison G., Rousseeuw P.J., Filzmoser P. and Croux C. Robust factor analysis. Journal of multivariate analysis.84 (2003) p. 145-172

[4]Johnson, R.A. and Wichern D.W. Applied multivariate statistical analysis. New Jersey Prentice Hall. 1992.

Vedoucí diplomové práce:doc. RNDr. Jaroslav Michálek, CSc.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2013/14.

V Brně, dne 22.11.2013

L.S.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA STROJNÍHO INŽENÝRSTVÍ DĚKANÁT

prof. RNDr. Josef Šlapal, CSc.
Ředitel ústavu

doc. Ing. Jaroslav Katolický, Ph.D.
Děkan

ABSTRAKT

Diplomová práce se zabývá mnohorozměrnými statistickými metodami a jejich environmentálními aplikacemi. Teoretická část práce je věnována vybraným metodám lineární regresní analýzy, metodě hlavních komponent a také je popsán model klasické a robustní faktorové analýzy. V praktické části práce jsou pomocí klasické faktorové analýzy stanoveny hlavní emisní zdroje PM1 aerosolů v letním a zimním období v Brně a ve Šlapanicích. Hlavní emisní zdroje aerosolů v létě a v zimě ve Šlapanicích jsou dále také identifikovány pomocí robustní faktorové analýzy. Dále je pomocí lineárního regresního modelu provedena predikce koncentrací PM1 aerosolů v letním a zimním období v Brně a ve Šlapanicích.

KLÍČOVÁ SLOVA

PM1 aerosoly, regresní analýza, metoda hlavních komponent, faktorová analýza, robustní faktorová analýza

ABSTRACT

The diploma thesis deals with multivariate statistical methods and their environmental applications. The theoretical part is devoted to selected methods of linear regression analysis, method of principal components and the model of classical and robust factor analysis is also described. In the practical part of thesis, the main emission sources of PM1 aerosols in summer and winter period in Brno and Šlapanice are determined by using the classical factor analysis. The main aerosol emission sources in summer and winter in Šlapanice are also identified by using the robust factor analysis. Furthermore, the prediction of concentrations of PM1 aerosols in summer and winter period in Brno and Šlapanice is performed by using the linear regression model.

KEYWORDS

PM1 aerosols, linear regression, principal component method, factor analysis, robust factor analysis

MIKUŠKOVÁ, Martina *Statistické modelování znečištění ovzduší prašným aerosolem*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, Ústav matematiky, 2014. 94 s. Vedoucí práce byl doc. RNDr. Jaroslav Michálek, CSc.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Statistické modelování znečištění ovzduší prašným aerosolem“ jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Ráda bych poděkovala doc. RNDr. Jaroslavu Michálkovi, CSc. za rady, připomínky a věnovaný čas při odborném vedení této diplomové práce.

Brno ....

(podpis autora)

OBSAH

ÚVOD

S rostoucí urbanizací se neustále zhoršuje kvalita životního prostředí. Znečištění ovzduší v důsledku vysokých koncentrací atmosferických aerosolů v okolním prostředí je v současnosti velmi diskutovaným tématem. Důležitým krokem ke zvýšení kvality životního prostředí je určení hlavních emisních zdrojů atmosferických aerosolů a predikování koncentrací těchto částic.

Hlavním cílem této diplomové práce je představit mnohorozměrné statistické metody, pomocí kterých je možné analyzovat data získaná měřením koncentrací a chemického složení PM1 aerosolů v letním a zimním období v Brně a ve Šlapanicích. Práce je rozčleněna do několika kapitol, jejichž obsah bude nyní stručně popsán.

Po úvodní kapitole následuje kapitola s názvem Základní pojmy, která je zpracovaná podle [1] a obsahuje pojmy z teorie pravděpodobnosti a matematické analýzy, které jsou potřebné v dalších kapitolách.

Ve třetí kapitole zpracované podle [1] a [9] je uvedena teorie, potřebná k sestavení lineárního regresního modelu.

Čtvrtá kapitola zpracovaná podle [9] popisuje princip metody hlavních komponent.

Pátá kapitola, která je zpracovaná podle [9], je věnována faktorové analýze. Je zde popsán ortogonální faktorový model, metody vhodné k odhadu parametrů tohoto modelu, kritérium pro faktorovou rotaci a statistické testy, které je vhodné před analýzou dat provést. Je také popsán test, pomocí kterého je možné ověřit, že extrahovaný počet faktorů je dostatečný.

V šesté kapitole zpracované podle [14] a [15] je popsán princip robustní faktorové analýzy.

V sedmé kapitole je stručný popis softwaru, který byl použit k praktickým výpočtům v následujících dvou kapitolách.

Osmá kapitola je věnována identifikaci emisních zdrojů PM1 aerosolů v Brně a ve Šlapanicích užitím vhodných metod z předchozích kapitol.

V deváté kapitole jsou s užitím teorie uvedené ve třetí kapitole sestaveny regresní modely umožňující predikovat denní koncentrace aerosolových částic.

Poslední, desátá kapitola s názvem Závěr obsahuje shrnutí diplomové práce a dosažených výsledků.

1 ZÁKLADNÍ POJMY A OZNAČENÍ

1.1 Označení a pomocné věty

V tomto odstavci bude uveden přehled základního značení, které je v matematice obvyklé, a jehož přehled je uveden na konci práce. Dále budou uvedeny pomocné věty, jejichž důkazy neuvádíme, lze je nalézt v monografii [1] (pokud nebude uvedeno jinak).

Množinu přirozených čísel budeme značit symbolem $\mathbb{N}$, množinu reálných čísel symbolem $\mathbb{R}$. Matice reálných čísel budeme značit velkými písmeny ze začátku abecedy, tedy $\mathbf{A}_{m\times n} = (a_{ij})_{i=1,\dots m,j=1\dots n}$ bude značit matici s m řádky a n sloupci. Symbol $'$ bude značit maticovou transpozici, $\det(\mathbf{A}) = |\mathbf{A}|$ determinant matice $\mathbf{A}$, $h(\mathbf{A})$ hodnost matice $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ stopu matice $\mathbf{A}$ a $\mathbf{A}^{-1}$ matici inverzní k matici $\mathbf{A}$. $\mathbf{A}^{-}$ bude značit pseudoinverzní matici k matici $\mathbf{A}$, tedy matici splňující $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$. Vektor jedniček dimenze n budeme značit $\mathbf{1}_n$, jednotkovou matici $\mathbf{I}$ a nulovou matici $\mathbf{0}$. $\Gamma(t)$ bude značit gama funkci v proměnné t, $\forall$ bude značit obecný kvantifikátor a symbolem $\blacksquare$ budeme značit konec důkazu.

Věta 1. Pro libovolnou matici $\mathbf{A}$ platí $h(\mathbf{A}) = h(\mathbf{A}\mathbf{A}')$.

Věta 2. Je-li $\mathbf{A}_{n\times n}$ idempotentní matice hodnosti r, potom $h(\mathbf{I}_n - \mathbf{A}) = n - r$.

Věta 3. Nechť $\mathbf{A}_{k\times k}$ je symetrická matice a $\mathbf{x}_{k\times 1}$ je vektor. Potom platí

1. $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathrm{Tr}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathrm{Tr}(\mathbf{A}\mathbf{x}\mathbf{x}')$.
2. $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{k} \lambda_i$, kde λ_i jsou vlastní čísla matice $\mathbf{A}$.

Dk: [9]

Věta 4. Nechť $\mathbf{A}_{k\times k}$, $\mathbf{B}_{k\times k}$ jsou matice a $c \in \mathbb{R}$. Potom platí

1. $\mathrm{Tr}(c\mathbf{A}) = c\mathrm{Tr}(\mathbf{A})$.
2. $\mathrm{Tr}(\mathbf{A} \pm \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) \pm \mathrm{Tr}(\mathbf{B})$.

Dk: Zřejmý

1.2 Výsledky z pravděpodobnosti a matematické statistiky

Tento odstavec, zpracovaný podle [1] a částečně podle [9], obsahuje základní pojmy z pravděpodobnosti a matematické statistiky, které jsou potřebné v dalších kapitolách. Uvedené citované věty jsou bez důkazu, lze ho rovněž nalézt v [1].

V práci budeme pracovat se spojitými náhodnými veličinami, které jsou definovány na stejném pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. Budeme je značit velkými písmeny z konce abecedy a jejich pravděpodobnostní chování v regulárních případech bude popsáno hustotou.

$\mathrm{E}(X)$ bude značit střední hodnotu náhodné veličiny X, $\mathrm{Var}(X)$ rozptyl náhodné veličiny X a $f(x)$ hustotu náhodné veličiny X. Dále, pro náhodné veličiny X, Y bude $\mathrm{Cov}(X, Y)$ značit kovarianci těchto náhodných veličin, ρ_{XY} korelační koeficient a r_{XY} výběrový korelační koeficient.

Nyní uvedeme přehled základních spojitých rozdělení pravděpodobnosti, která budou používaná v dalších kapitolách.

Normální rozdělení Nechť $\mu \in \mathbb{R}$ a $\sigma > 0$ jsou dané konstanty. Normální rozdělení je určeno hustotou

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty,$$

a značí se symbolem $N(\mu, \sigma^2)$. Pokud X má normální rozdělení $N(\mu, \sigma^2)$, potom $\mathrm{E}(X) = \mu$ a $\mathrm{Var}(X) = \sigma^2$.

Poznámka 5. V případě, že $\sigma^2 = 0$, říkáme, že náhodná veličina X má singulární normální rozdělení $N(\mu, 0)$, když $P(X = \mu) = 1$.

Pearsonovo χ^2 rozdělení Nechť $n \in \mathbb{N}$. Pearsonovo χ^2 rozdělení o n stupních volnosti je určeno hustotou

$$f(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{n/2-1} \mathrm{e}^{-x/2} \quad \text{pro } x > 0,$$

$$f(x) = 0 \quad \text{pro } x \leq 0,$$

a značí se symbolem χ^2_n. α kvantily Pearsonova rozdělení budeme značit $\chi^2_n(\alpha)$.

Studentovo rozdělení Nechť $n \in \mathbb{N}$. Studentovo rozdělení o n stupních volnosti je určeno hustotou

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < \infty,$$

a značí se symbolem t_n. α kvantily Studentova rozdělení budeme značit $t_n(\alpha)$.

Fisher-Snedecorovo rozdělení Nechť $m, n \in \mathbb{N}$. Fisher-Snedecorovo rozdělení o m a n stupních volnosti je určeno hustotou

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{m/2} x^{m/2-1} \left(1 + \frac{m}{n}x\right)^{-(m+n)/2} \quad \text{pro } x > 0,$$

$$f(x) = 0 \quad \text{pro } x \leq 0,$$

a značí se symbolem $F_{m,n}$. α kvantily Fisher-Snedecorova rozdělení budeme značit $F_{m,n}(\alpha)$.

Logaritmicko-normální rozdělení Nechť $\mu \geq 0$ a $\sigma > 0$ jsou dané konstanty. Logaritmicko-normální rozdělení je určeno hustotou

$$f(x) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left\{-\frac{[\ln(x) - \mu]^2}{2\sigma^2}\right\} \quad \text{pro } x > 0,$$

$$f(x) = 0 \quad \text{pro } x \leq 0,$$

a značí se symbolem $LN(\mu, \sigma^2)$. Jestliže X má normální rozdělení $LN(\mu, \sigma^2)$, potom $\ln(X)$ má logaritmicko-normální rozdělení $N(\mu, \sigma^2)$.

Zápisem $X \sim N(\mu, \sigma^2)$ budeme značit, že náhodná veličina X má normální rozdělení $N(\mu, \sigma^2)$, analogicky i pro další rozdělení pravděpodobnosti.

Věta 6. Nechť X, Z jsou nezávislé náhodné veličiny, přičemž $X \sim N(0, 1)$ a $Z \sim \chi^2(k)$. Potom náhodná veličina

$$T = \frac{X}{\sqrt{\frac{Z}{k}}} \sim t_k.$$

Dále budeme pracovat s náhodným vektorem. Jsou-li $X_1, \ldots, X_n$ náhodné veličiny, pak náhodným vektorem rozumíme n-tici $\mathbf{X} = (X_1, \ldots, X_n)'$. Při práci s náhodnými vektory budeme používat standartní označení (např. dle [1]), tedy střední hodnota náhodného vektoru $\mathbf{X}$ je dána vztahem

$$\mathrm{E}(\mathbf{X}) = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_n))',$$

(kde $\mathrm{E}(X_i)$ je střední hodnota náhodné veličiny X_i pro $i = 1, \ldots, n$). Poznamenejme, že v práci budeme pracovat pouze s náhodnými veličinami, jejichž střední hodnoty existují. Varianční matice náhodného vektoru $\mathbf{X}$ je dána vztahem

$$\mathrm{Var}(\mathbf{X}) = \mathrm{E}(\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{X} - \mathrm{E}\mathbf{X})',$$

a v následující větě budou uvedeny její dále potřebné vlastnosti.

Věta 7. Nechť $\mathbf{X}$ je náhodný vektor. Potom platí:

1. $$\mathrm{Var}(\mathbf{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \cdots & \cdots & \ddots & \cdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{pmatrix}.$$
2. Nechť $\mathbf{a} \in \mathbb{R}^m$ je vektor a $\mathbf{B}_{m\times n}$ je matice reálných čísel. Potom platí
   (a) $\mathrm{E}(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{a} + \mathbf{B}\mathrm{E}(\mathbf{X})$,
   (b) $\mathrm{Var}(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{B}\mathrm{Var}(\mathbf{X})\mathbf{B}'$.
3. $\mathrm{Var}(\mathbf{X})$ je pozitivně semidefinitní, tuto skutečnost budeme značit $\mathrm{Var}(\mathbf{X}) \geq 0$.
4. $\mathrm{Var}(\mathbf{X})$ je symetrická, tedy $\mathrm{Var}(\mathbf{X}) = (\mathrm{Var}(\mathbf{X}))'$.

Označíme-li $\mathbf{X} = (X_1, \ldots, X_n)'$ a $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ náhodné vektory splňující $\mathrm{E}(X_i^2) < \infty, i = 1, \ldots, n$ a $\mathrm{E}(Y_j^2) < \infty, j = 1, \ldots, m$, potom matici

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathrm{E}(\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{Y} - \mathrm{E}\mathbf{Y})'$$

nazýváme kovarianční maticí náhodných vektorů $\mathbf{X}, \mathbf{Y}$, a pokud $m = n$, matici

$$\mathrm{Cor}(\mathbf{X}, \mathbf{Y}) = [\mathrm{diag}(\mathrm{Var}(\mathbf{X}))]^{-1/2}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})[\mathrm{diag}(\mathrm{Var}(\mathbf{Y}))]^{-1/2}$$

nazýváme korelační maticí náhodných vektorů $\mathbf{X}, \mathbf{Y}$. Speciálně, $\mathrm{Cor}(\mathbf{X}, \mathbf{X})$ představuje korelační matici vektoru $\mathbf{X}$ a značíme ji $\mathrm{Cor}(\mathbf{X})$.

V následující větě jsou uvedeny dále potřebné vlastnosti kovarianční matice.

Věta 8. Nechť $\mathbf{X}, \mathbf{Y}$ jsou náhodné vektory. Potom platí:

1. $$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \mathrm{Cov}(X_1, Y_1) & \mathrm{Cov}(X_1, Y_2) & \cdots & \mathrm{Cov}(X_1, Y_m) \\ \mathrm{Cov}(X_2, Y_1) & \mathrm{Cov}(X_2, Y_2) & \cdots & \mathrm{Cov}(X_2, Y_m) \\ \cdots & \cdots & \ddots & \cdots \\ \mathrm{Cov}(X_n, Y_1) & \mathrm{Cov}(X_n, Y_2) & \cdots & \mathrm{Cov}(X_n, Y_m) \end{pmatrix}.$$
2. $\mathrm{Cov}(\mathbf{Y}, \mathbf{X}) = [\mathrm{Cov}(\mathbf{X}, \mathbf{Y})]'$.
3. Nechť $\mathbf{a} \in \mathbb{R}^p, \mathbf{c} \in \mathbb{R}^q$ jsou vektory a $\mathbf{B}_{p\times n}, \mathbf{D}_{q\times m}$ jsou matice reálných čísel. Potom platí $\mathrm{Cov}(\mathbf{a} + \mathbf{B}\mathbf{X}, \mathbf{c} + \mathbf{D}\mathbf{Y}) = \mathbf{B}[\mathrm{Cov}(\mathbf{X}, \mathbf{Y})]\mathbf{D}'$.
4. $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathrm{Var}(\mathbf{X})$.

Dále bude zavedeno obecné p-rozměrné normální rozdělení.

Definice 1. Mějme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)'$. Nechť $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ je daný vektor a $\boldsymbol{\Sigma}_{p\times p} = (\sigma_{jk})_{i,j=1,\ldots,p}$ symetrická, pozitivně semidefinitní matice. Řekneme, že $\mathbf{X}$ má p-rozměrné normální rozdělení s parametry $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, jestliže pro libovolný vektor $\mathbf{c} \in \mathbb{R}^p$ platí

$$\mathbf{c}'\mathbf{X} \sim \mathrm{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}).$$

Je-li $\boldsymbol{\Sigma}$ regulární, říkáme, že $\mathbf{X}$ má regulární p-rozměrné normální rozdělení. Pokud je $\boldsymbol{\Sigma}$ singulární, říkáme, že $\mathbf{X}$ má singulární p-rozměrné normální rozdělení a v obou případech značíme $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Při práci s náhodnými vektory, které mají p-rozměrné normální rozdělení, budeme potřebovat vlastnosti, které jsou uvedeny v následujících větách.

Věta 9. Jestliže $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, potom $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}, \mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$.

Věta 10. Nechť $\mathbf{X}$ má rozdělení $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Nechť $\mathbf{a} \in \mathbb{R}^m$ je vektor a $\mathbf{B}_{m\times p}$ je matice reálných čísel. Potom $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim \mathrm{N}_m(\mathbf{a} + \boldsymbol{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$.

Poznámka 11. Speciálně pokud $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a rozdělení je regulární, potom náhodný vektor $\mathbf{U} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I})$, kde $\mathbf{I}$ značí jednotkovou matici, nazýváme standardizovaný normální náhodný vektor.

Věta 12. Má-li náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)'$ regulární normální rozdělení $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, existuje jeho hustota a je dána vztahem

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

Věta 13. Nechť $\mathbf{X} = (X_1, \ldots, X_p)' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a $k \in \mathbb{N}$ splňuje $1 \leq k < n$. Položme

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

kde $\boldsymbol{\mu}_1$ má k složek a matice $\boldsymbol{\Sigma}_{11}$ je typu $k \times k$. Potom marginální vektor $\mathbf{X}_1 = (X_1, \ldots, X_k)' \sim N_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Věta 14. Nechť $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)' \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, $|\boldsymbol{\Sigma}_{22}| > 0$. Potom podmíněné rozdělení k-rozměrného vektoru $\mathbf{X}_1$ za podmínky $\mathbf{X}_2 = \mathbf{x}_2$ je k-rozměrné normální $N_k(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$.

Dále předpokládejme, že $\mathbf{X}_1, \ldots, \mathbf{X}_n$ jsou nezávislé náhodné vektory se stejným rozdělením pravděpodobnosti, jako má náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)'$, a

$$\mathbf{X}_D = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \ddots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix},$$

je datová matice. Potom lze střední hodnotu $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$ náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_p)'$ odhadnout výběrovým průměrem

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i = \frac{1}{n}\mathbf{X}_D'\mathbf{1}_n,$$

který je nestranným odhadem střední hodnoty $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, varianční matici $\mathrm{Var}(\mathbf{X})$ je možné odhadnout výběrovou varianční maticí

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_D' \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{X}_D,$$

která je nestranným odhadem varianční matice $\mathrm{Var}(\mathbf{X})$, a korelační matici $\mathrm{Cor}(\mathbf{X})$ lze odhadnout výběrovou korelační maticí

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2},$$

kde $\mathbf{D} = \mathrm{diag}\{\mathbf{S}\}$.

Věta 15. Nechť náhodné vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n$ jsou náhodným výběrem z nějakého spojitého p-rozměrného rozdělení. Jeli $n \geq p+1$, pak matice $\mathbf{S}_n = [(n-1)/n]\mathbf{S}$ je pozitivně definitní (a tedy regulární) s pravděpodobností 1.

Věta 16. Nechť $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ je náhodný výběr z vícerozměrného normálního rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Potom

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}, \quad \hat{\boldsymbol{\Sigma}} = \frac{(n-1)}{n}\mathbf{S}$$

jsou maximálně věrohodné odhady střední hodnoty $\boldsymbol{\mu}$ a varianční matice $\boldsymbol{\Sigma}$.

Důkaz. [9]

Věta 17. Nechť $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, přičemž $h(\boldsymbol{\Sigma}) = r \geq 1$. Potom náhodná veličina $\mathbf{Y} = (\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^-(\mathbf{X}-\boldsymbol{\mu})$ má χ^2_r rozdělení při libovolné volbě pseudoinverzní matice $\boldsymbol{\Sigma}^-$.

Věta 18. Nechť $\mathbf{X}$ je p-rozměrný náhodný vektor s konečnými druhými momenty a nechť $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$. Potom pro libovolnou matici $\mathbf{A}_{n\times n}$ platí $\mathrm{E}(\mathbf{X}'\mathbf{A}\mathbf{X}) = \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

Věta 19. Nechť $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a matice $\mathbf{A}_{n\times n} \geq 0$. Nechť $\mathbf{B}_{m\times n}$ je taková matice, že $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$. Potom pro libovolný p-rozměrný vektor $\mathbf{c}$ platí, že vektor $\mathbf{B}\mathbf{X}$ a náhodná veličina $(\mathbf{X}-\mathbf{c})'\mathbf{A}(\mathbf{X}-\mathbf{c})$ jsou nezávislé.

2 LINEÁRNÍ REGRESNÍ MODEL

Zavedení lineárního regresního modelu (LRM) budeme motivovat příkladem. Předpokládejme, že je třeba predikovat denní koncentrace PM1 aerosolů[1] v Brně. Aerosolové koncentrace jsou ovlivněny především teplotou, relativní vlhkostí, rychlostí a směrem větru, a také koncentracemi aerosolů předchozího dne. Cílem prováděné analýzy je sestavit model, který by umožňoval predikovat denní koncentrace aerosolů v závislosti na koncentracích aerosolů předchozího dne a hodnotách meteorologických veličin. Takovou predikci lze provést pomocí regresního modelu, o kterém pojednáno v této kapitole. Kapitola je zpracovaná převážně podle [1] a částečně podle [9]. Použité značení rovněž vychází z monografie [1].

2.1 Lineární regresní model

V tomto odstavci zavedeme obecný LRM. Označme Y závisle proměnnou (tzv. odezvu) a $x_1, \ldots, x_k$ reálná čísla (tzv. regresory). LRM můžeme zapsat:

$$Y = \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon,$$

kde $\beta_1 \ldots \beta_k$ jsou neznámé parametry a ε je náhodná chyba, která souvisí s pozorováním Y, a splňuje $\mathrm{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$, kde σ^2 značí neznámý parametr. Po provedení n nezávislých pozorování (za stejných podmínek) odezvy Y a regresorů $x_1, \ldots, x_k$ dostáváme

$$\begin{aligned}
Y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1, \\
Y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2, \\
&\vdots \\
Y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n,
\end{aligned}$$

kde x_{ij} značí i-té pozorování regresoru x_j, Y_i značí i-té pozorování odezvy Y, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, a $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ je vektor náhodných chyb. Předpokládáme, že chyby vyhovují následujícím předpokladům:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$,
2. $\mathrm{Var}\,(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

Předpoklad $\mathrm{Var}\,(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ říká, že chyby jsou nekorelované a měření jednotlivých složek vektoru $\mathbf{Y}$ jsou prováděna se stejnou přesností.

[1] Tento pojem je vysvětlen v kapitole 7

Maticový zápis LRM je:

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

kde $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ je náhodný vektor náhodných pozorování odezvy Y, $\mathbf{X} = (x_{ij})_{i=1,\ldots,n,j=1,\ldots,k}$ je tzv. matice plánu a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ je vektor neznámých parametrů.

Při sestavování LRM umožňujícího predikovat denní koncentrace aerosolů je náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ tvořen pozorovanými hodnotami denních koncentrací PM1 aerosolů a regresory $x_1, \ldots, x_k$ představují meteorologické faktory: např. průměrná denní teplota je označena proměnnou x_1, průměrná relativní vlhkost je značená proměnnou x_2 atd.

V následujícím odstavci bude popsáno, jak získat odhad neznámých parametrů $\beta_1, \beta_2, \ldots, \beta_k$ modelu (2.1).

2.2 Odhad a vlastnosti odhadovaných parametrů modelu

Předpokládejme, že je dán LRM (2.1). Užitím vlastností střední hodnoty a rozptylu můžeme odvodit střední hodnotu a rozptyl náhodného vektoru $\mathbf{Y}$. Platí

$$\mathrm{E}(\mathbf{Y}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{2.2}$$

$$\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}. \tag{2.3}$$

Dále budeme předpokládat, že $h(\mathbf{X}) = k < n$, a budeme říkat, že LRM je plné hodnosti. Podle věty 1 platí $h(\mathbf{X}) = h(\mathbf{X}\mathbf{X}') = h(\mathbf{X}'\mathbf{X}) = k$ a z teorie maticové algebry plyne, že matice $\mathbf{X}'\mathbf{X}$ je regulární.

Odhad $\mathbf{b} = (b_1, \ldots, b_k)'$ parametru $\boldsymbol{\beta}$ získáme metodou nejmenších čtverců, minimalizací výrazu $S(\boldsymbol{\beta}) = \sum_{j=1}^{n}(Y_j - \beta_1 x_{j1} - \ldots - \beta_k x_{jk})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. V následující větě je uveden výsledný tvar odhadu neznámých parametrů.

Věta 20. Odhad $\mathbf{b}$ parametru $\boldsymbol{\beta}$ metodou nejmenších čtverců je

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2.4}$$

Důkaz. Pro vektor $\mathbf{b}$ daný vztahem (2.4) platí

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Y} = \mathbf{0},$$

$$(\mathbf{Y} - \mathbf{X}\mathbf{b})'\mathbf{X} = (\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'\mathbf{X} = \mathbf{Y}'\mathbf{X} - \mathbf{Y}'\mathbf{X} = \mathbf{0}.$$

Dále odvodíme

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}) =$$

$$= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) + (\mathbf{b} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}) +$$

$$+(\mathbf{b} - \boldsymbol{\beta})'[\mathbf{X}'(\mathbf{Y} - \mathbf{Xb})] + [(\mathbf{Y} - \mathbf{Xb})'\mathbf{X}](\mathbf{b} - \boldsymbol{\beta}).$$

Protože $\mathbf{X}'(\mathbf{Y} - \mathbf{Xb}) = (\mathbf{Y} - \mathbf{bX})'\mathbf{X} = \mathbf{0}$ výraz se zjednodušší na

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) + (\mathbf{b} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}) \geq$$

$$\geq (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}).$$

Z předpokladu $h(\mathbf{X}) = k$ plyne, že matice $\mathbf{X}'\mathbf{X}$ je pozitivně definitní a proto rovnost nastává právě pro $\boldsymbol{\beta} = \boldsymbol{b}$.

■

Věta 21. Odhad $\mathbf{b}$ je nestranný a tedy $\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^k$, a $\mathrm{Var}\,(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Důkaz. Z vlastnosti střední hodnoty odvodíme

$$\mathrm{E}(\mathbf{b}) = \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^k,$$

a užitím vlastností varianční matice odvodíme

$$\mathrm{Var}\,(\mathbf{b}) = \mathrm{Var}\,[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}\,(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} =$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

■

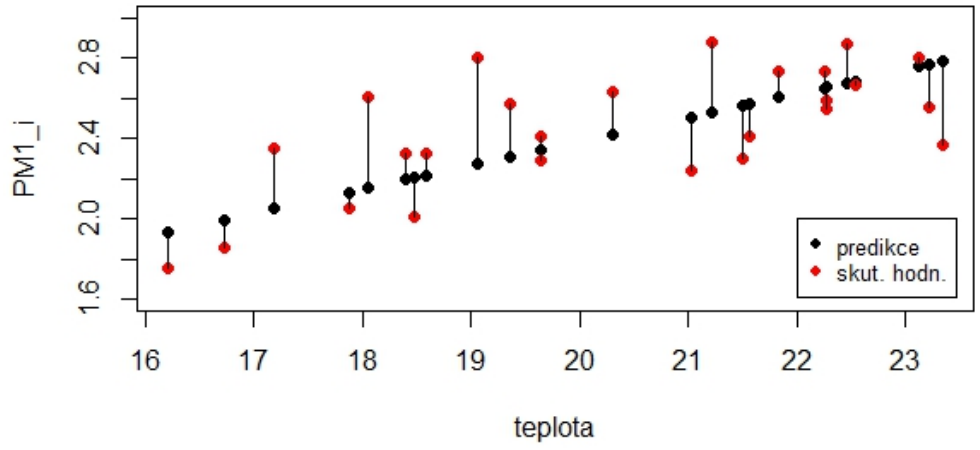
Soustava rovnic $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$, ze které je možné odhad $\mathbf{b}$ získat, se nazývá normální. Vektor $\hat{\mathbf{Y}} = \mathbf{Xb}$ obsahuje pro dané hodnoty regresorů $x_1, \ldots, x_k$ hodnoty predikované lineárním regresním modelem a je nejlepší aproximací vektoru $\mathbf{Y}$, kterou můžeme získat lineární kombinací sloupců matice $\mathbf{X}$.

Dále budeme pracovat s vektorem reziduí, který je definován $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, tedy j-té reziduum $e_j = Y_j - \hat{Y}_j = Y_j - b_1x_{j1} - \ldots - b_kx_{jk}$ pro $j = 1, \ldots, n$. Reziduální vektor $\mathbf{e}$ je odhadem vektoru chyb $\boldsymbol{\varepsilon}$, a výraz

$$S_e = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) \tag{2.5}$$

se nazývá reziduální součet čtverců.

LRM budeme ilustrovat na motivačním příkladu z úvodu této kapitoly. Označme $PM1_i$ i-té měření koncentrace PM1 aerosolů a T_i i-té měření teploty. Uvažujme model $PM1_i = \beta T_i + \varepsilon_i$ umožňující predikovat denní koncentrace aerosolů v závislosti na teplotě. Označme b odhad parametru β. Rezidua $e_i = PM1_i - bT_i$ znázorněná v



Obr. 2.1: Rezidua

grafu na obrázku 2.1 představují vzdálenosti mezi naměřenými hodnotami $PM1_i$ a hodnotami predikovanými podle modelu $PM1_i = \beta T_i + \varepsilon_i$.

V následujících dvou větách jsou uvedeny vlastnosti vektoru reziduí a reziduálního součtu čtverců.

Věta 22. Vektor reziduí $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$ splňuje

$$\mathbf{X}'\mathbf{e} = \mathbf{0}, \qquad \hat{\mathbf{Y}}'\mathbf{e} = 0.$$

Důkaz. Nejprve ukážeme, že

$$\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}.$$

Matice $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$ splňuje

1. $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$
2. $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] =$
   $= \mathbf{I} - 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$, což znamená, že je symetrická a idempotentní.

Dále platí $\mathbf{X}'\mathbf{e} = \mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Y} = \mathbf{0}$
a $\hat{\mathbf{Y}}'\mathbf{e} = (\mathbf{X}\mathbf{b})'\mathbf{e} = \mathbf{b}'\mathbf{X}'\mathbf{e} = \mathbf{b}'\mathbf{0} = \mathbf{0}$.

■

Věta 23. Pro reziduální součet čtverců S_e platí:

$$S_e = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}, \quad S_e = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}.$$

Důkaz. Podle věty 22 platí $\mathbf{Y}-\hat{\mathbf{Y}} = [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$. Matice $\mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ je idempotentní a symetrická, z čehož plyne $\mathbf{H}^2 = \mathbf{H}$. Dosazením do vztahu (2.5) dostaneme

$$S_e = (\mathbf{HY})'\mathbf{HY} = \mathbf{Y}'\mathbf{H}'\mathbf{HY} = \mathbf{Y}'\mathbf{HY} = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y},$$

což je první vztah. Roznásobením dostáváme druhý vztah:

$$S_e = [\mathbf{Y}' - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = [\mathbf{Y}' - \mathbf{b}'\mathbf{X}']\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}.$$

■

Kromě odhadů $b_1, \ldots, b_k$ neznámých parametrů $\beta_1, \ldots, \beta_k$ jsou pro statistickou interferenci potřebné také odhady rozptylů $\mathrm{Var}(b_j)$, které označíme $s^2(b_j)$. Nejdříve uvedeme odhad rozptylu σ^2.

Věta 24. Náhodná veličina $s^2 = \dfrac{S_e}{n-k}$ se nazývá reziduální rozptyl a je nestranným odhadem parametru σ^2.

Důkaz. Podle věty 23 je $S_e = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$. V důkazu věty 22 bylo ukázáno, že matice $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$ je idempotentní a z teorie matic je známo, že její stopa je rovna její hodnosti. Protože $\mathbf{X}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} = \mathbf{X}'\mathbf{X}$ a $h(\mathbf{X}'\mathbf{X}) = k$, je $h[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \geq k$. Protože $h(\mathbf{X}) = k$, musí současně být $h[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \leq k$, z čehož plyne $h[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = k$ a podle věty 2 je $h[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - k$. Protože $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}$, podle věty 18 platí

$$\mathrm{E}(S_e) = \mathrm{E}\left(\mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}\right) =$$

$$= \mathrm{Tr}\left([\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}\right) + (\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X}\boldsymbol{\beta} =$$

$$= \sigma^2\mathrm{Tr}\left([\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\right) + \mathbf{0} = \sigma^2 h[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma^2(n-k),$$

z čehož plyne, že $\mathrm{E}\left(\dfrac{S_e}{n-k}\right) = \sigma^2$ a náhodná veličina $s^2 = \dfrac{S_e}{n-k}$ je nestranným odhadem parametru σ^2.

■

Nahradíme-li v matici $\mathrm{Var}(\mathbf{b})$ parametr σ^2 reziduálním rozptylem s^2 a označíme-li v_{ij} prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$ na pozici (i,j), může být odhad rozptylů odhadů parametrů b_i stanoven podle vztahu:

$$s^2(b_j) = s^2 v_{jj}, \qquad j = 1, \ldots, k.$$

Poznamenejme, že $s(b_j)$ se nazývá směrodatná chyba, tedy

$$s(b_j) = \sqrt{s^2 v_{jj}}, \qquad j = 1, \ldots, k. \tag{2.6}$$

2.3 Testování hypotéz o parametrech regresního modelu

Při sestavování LRM je nutné ověřit, které z regresorů $x_1, \ldots, x_k$ mají na $\mathbf{Y}$ statisticky významný vliv. Například chceme ověřit, zda je pro predikci denních koncentrací aerosolů statisticky významná teplota, relativní vlhkost i rychlost a směr větru. V tomto odstavci je proto popsáno, jak testovat hypotézu, že vliv regresoru x_i na odezvu $\mathbf{Y}$ je statisticky nevýznamný.

Budeme předpokládat, že vektor chyb má n-rozměrné normální rozdělení, tedy

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}), \tag{2.7}$$

z čehož můžeme odvodit, že $\mathbf{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Nejdříve uvedeme větu o vlastnostech odhadů $\mathbf{b}$, S_e a s^2.

Věta 25. Za předpokladu (2.7) platí

1. $\mathbf{b} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$,
2. $\dfrac{S_e}{\sigma^2} \sim \chi^2_{n-k}$,
3. $\mathbf{b}$ a s^2 jsou nezávislé.

Důkaz.

1. Protože $\mathbf{Y}$ má normální rozdělení a $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, podle věty 10 má $\mathbf{b}$ také normální rozdělení. Podle věty 21 platí $\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$, $\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, z čehož je zřejmé, že $\mathbf{b} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
2. Nejprve ukážeme, že

$$\mathrm{E}(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathrm{E}(\mathbf{Y}) - \mathrm{E}(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta} - \mathrm{E}(\mathbf{X}\mathbf{b}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

a

$$\mathrm{Var}(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathrm{Var}\left([\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}\right) =$$
$$= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathrm{Var}(\mathbf{Y})[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] =$$
$$= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'].$$

Dále můžeme odvodit, že

$$\mathrm{Var}(\mathbf{Y} - \hat{\mathbf{Y}})(\sigma^{-2}\mathbf{I})\mathrm{Var}(\mathbf{Y} - \hat{\mathbf{Y}}) =$$
$$= \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'](\sigma^{-2}\mathbf{I})\sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] =$$
$$= \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathrm{Var}(\mathbf{Y} - \hat{\mathbf{Y}}),$$

z čehož je zřejmé, že matice $\sigma^{-2}\mathbf{I}$ je pseudoinverzní maticí k matici $\mathrm{Var}(\mathbf{Y}-\hat{\mathbf{Y}})$. Vektor $\mathbf{Y}-\hat{\mathbf{Y}}$ má normální rozdělení a proto podle věty 17 náhodná veličina

$$(\mathbf{Y}-\hat{\mathbf{Y}})'\sigma^{-2}\mathbf{I}(\mathbf{Y}-\hat{\mathbf{Y}}) = \frac{S_e}{\sigma^2} \sim \chi^2_m,$$

kde $m = h(\mathrm{Var}(\mathbf{Y}-\hat{\mathbf{Y}}))$. Dříve již bylo ukázáno, že $h[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n-k$, z čehož plyne $h(\mathrm{Var}(\mathbf{Y}-\hat{\mathbf{Y}})) = h\left(\sigma^2[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\right) = n-k$ a tedy $m = n-k$.

3. Protože $\mathbf{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\mathbf{I})$, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, $S_e = \mathbf{Y}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$ a $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'](\sigma^2\mathbf{I})[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{0}$, platí podle věty 19, že vektor $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{b}$ a náhodná veličina
$(\mathbf{Y}-\mathbf{0})'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'](\mathbf{Y}-\mathbf{0}) = S_e$ jsou nezávislé, z čehož plyne i nezávislost veličin $\mathbf{b}$ a $s^2 = S_e/(n-k)$.

■

Nyní bude uvedena věta, jejímž užitím je možné zjistit, které z regresorů $x_1, \ldots, x_k$ jsou pro predikci $\mathbf{Y}$ statisticky nevýznamné.

Věta 26. Nechť v_{ij} je prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$ na pozici (i,j). Pak pro každé $i = 1, \ldots, k$ náhodná veličina

$$T_i = \frac{b_i - \beta_i}{\sqrt{s^2 v_{ii}}} \sim t_{n-k}. \tag{2.8}$$

Důkaz. Podle vět 25 a 13 platí $b_i \sim N(\beta_i, \sigma^2 v_{ii})$ a je zřejmé, že náhodná veličina $(b_i-\beta_i)/\sqrt{\sigma^2 v_{ii}} \sim N(0,1)$. Podle věty 25 náhodná veličina $S_e/\sigma^2 \sim \chi^2_{n-k}$ a veličiny $\mathbf{b}$ a s^2 jsou nezávislé, z čehož užitím věty 6 můžeme odvodit, že

$$T_i = \frac{(b_i-\beta_i)/\sqrt{\sigma^2 v_{ii}}}{\frac{\sqrt{S_e/\sigma^2}}{\sqrt{n-k}}} = \frac{b_i-\beta_i}{\sqrt{s^2 v_{ii}}} \sim t_{n-k}.$$

■

Podle předcházející věty nulovou hypotézu $H_0 : \beta_i = 0$, že vliv regresoru x_i na $\mathbf{Y}$ je statisticky nevýznamný, zamítneme na hladině významnosti α, pokud

$$|T_i| = \frac{|b_i|}{\sqrt{s^2 v_{ii}}} \geq t_{n-k}(1-\alpha/2),$$

kde $t_{n-k}(1-\alpha/2)$ je $(1-\alpha/2)$ kvantil Studentova t rozdělení.

2.4 Regresní diagnostika

Máme-li sestavený regresní model a odhadnuté neznámé parametry, je třeba ohodnotit kvalitu modelu a posoudit, zda je dobře sestaven.

Předpokládáme-li v LRM, že regresory $x_1, \ldots, x_k$ jsou náhodné veličiny, potom lineární závislost mezi nimi a odezvou Y vyjádříme koeficientem mnohonásobné korelace $\rho_{Y,(x_1,\ldots,x_k)}$. Označíme-li $\mathrm{Cor}(Y, \mathbf{X})$ řádkový vektor obsahující korelační koeficienty $\rho_{Yx_1}, \ldots, \rho_{Yx_k}$, podobně $\mathrm{Cor}(\mathbf{X}, Y)$ sloupcový vektor obsahující korelační koeficienty $\rho_{Yx_1}, \ldots, \rho_{Yx_k}$ a předpokládáme-li, že korelační matice $\mathrm{Cor}(\mathbf{X})$ vektoru $(x_1, \ldots, x_k)'$ je regulární, potom koeficient mnohonásobné korelace počítáme podle vztahu

$$\rho^2_{Y,(x_1,\ldots,x_k)} = \mathrm{Cor}(Y, \mathbf{X})\mathrm{Cor}(\mathbf{X})^{-1}\mathrm{Cor}(\mathbf{X}, Y).$$

Užitím následující věty můžeme testovat statistickou významnost $\rho_{Y,(x_1,\ldots,x_k)}$.

Věta 27. Nechť náhodné vektory $(Y_i, x_{i1}, \ldots, x_{ik})'$ pro $i = 1, \ldots, n$, jsou výběrem z normálního rozdělení s regulární varianční maticí. Platí-li $\rho_{Y,(x_1,\ldots,x_k)} = 0$, pak při $n > p + 1$ veličina

$$Z = \frac{n-p-1}{p} \frac{\rho^2_{Y,(x_1,\ldots,x_k)}}{1-\rho^2_{Y,(x_1,\ldots,x_k)}} \sim F_{p,n-p-1}. \tag{2.9}$$

Důkaz: [1]

Pro testování nulové hypotézy, že koeficient mnohonásobné korelace $\rho_{Y,(x_1,\ldots,x_k)} = 0$, odhadneme koeficient mnohonásobné korelace $\rho_{Y,(x_1,\ldots,x_k)}$ výběrovým koeficientem mnohonásobné korelace

$$r^2_{Y,(x_1,\ldots,x_k)} = \mathbf{R}_{Y,\mathbf{X}}\mathbf{R}^{-1}_{\mathbf{X}}\mathbf{R}_{\mathbf{X},Y}, \tag{2.10}$$

kde $\mathbf{R}_{Y,\mathbf{X}}$ je řádkový vektor výběrových korelačních koeficientů $r_{Yx_1}, \ldots, r_{Y,x_k}$, podobně $\mathbf{R}_{\mathbf{x},Y}$ je sloupcový vektor výběrových korelačních koeficientů $r_{Yx_1}, \ldots, r_{Y,x_k}$ a $\mathbf{R}_{\mathbf{X}}$ značí výběrovou korelační matici vektoru $(x_1, \ldots, x_k)$. Dále stanovíme testovací statistiku (2.9) a hypotézu o lineární nezávislosti Y na regresorech $x_1, \ldots, x_k$ zamítneme, pokud je její realizace větší než $(1-\alpha)$ kvantil Fisher-Snedecorova rozdělení $F_{p,n-p-1}$.

Při testování hypotézy o lineární nezávislosti Y na jediném regresoru x_1 vycházíme z následující věty.

Věta 28. Nechť $(X_i, Y_i)'$ kde $i = 1, \ldots, n$ je výběr z dvojrozměrného normálního rozdělení, které má kladné rozptyly a korelační koeficient $\rho_{Y,X} = 0$. Nechť $n \geq 3$. Potom náhodná veličina

$$T = \frac{r_{YX}}{\sqrt{1-r^2_{YX}}}\sqrt{n-2} \sim t_{n-2}. \tag{2.11}$$

Důkaz: [1]

Podle předcházející věty nulovou hypotézu, že korelační koeficient $\rho_{Yx_1} = 0$ zamítáme, pokud je realizace testovací statistiky (2.11) $|T| > t_{n-2}(1-\alpha/2)$.

Kvalita regresního modelu s nenáhodnými regresory se v literatuře velmi často hodnotí koeficientem determinace, který nyní odvodíme. Vzhledem k definici vektoru reziduí platí

$$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \mathbf{Y} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}} + \mathbf{Y} - \hat{\mathbf{Y}}) = (\hat{\mathbf{Y}} + \mathbf{e})'(\hat{\mathbf{Y}} + \mathbf{e}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{e}'\mathbf{e}. \qquad (2.12)$$

Předpokládáme-li, že první sloupec matice plánu $\mathbf{X}$ je tvořen n-rozměrným vektorem jedniček, potom podle věty 22 platí $0 = \mathbf{1}'\mathbf{e} = \sum_{j=1}^{n} e_j = \sum_{j=1}^{n}(Y_j - \hat{Y}_j) = \sum_{j=1}^{n} Y_j - \sum_{j=1}^{n} \hat{Y}_j$, z čehož je zřejmé, že $(1/n)\sum_{j=1}^{n} Y_j = \bar{Y} = \bar{\hat{Y}} = (1/n)\sum_{j=1}^{n} \hat{Y}_j$. Odečtením výrazu $n(\bar{Y})^2 = n(\bar{\hat{Y}})^2$ od vztahu (2.12) dostáváme

$$\mathbf{Y}'\mathbf{Y} - n(\bar{Y})^2 = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n(\bar{\hat{\mathbf{Y}}})^2 + \mathbf{e}'\mathbf{e}.$$

S využitím $\bar{Y} = \bar{\hat{Y}}$, kde $\bar{}$ značí průměr hodnot daného vektoru, můžeme předchozí vztah zapsat ve tvaru

$$\frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\hat{Y}_j - \bar{Y})^2 + \frac{1}{n-1}\sum_{j=1}^{n} e_j^2$$

a koeficient determinace zavedeme jako R_D^2

$$R_D^2 = 1 - \frac{\sum_{j=1}^{n} e_j^2}{\sum_{j=1}^{n}(Y_j - \bar{Y})^2} = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}. \qquad (2.13)$$

Protože člen $\frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$ představuje celkovou variabilitu složek vektoru $\mathbf{Y}$ a člen $\frac{1}{n-1}\sum_{j=1}^{n} e_j^2$ značí variabilitu reziduí v modelu s k regresory $x_1, \ldots, x_k$, z předchozího odvození je zřejmé, že R_D^2 určuje, jaká část celkové variability v datech je vysvětlena modelem. Jestliže regresní model popisuje data přesně, potom je reziduální součet čtverců $S_e = \sum_{j=1}^{n} e_j^2$ nulový a koeficient determinace $R_D = 1$. Pokud je však první sloupec matice plánu tvořen vektorem jedniček a odhad $\mathbf{b} = (\bar{Y}, 0, \ldots, 0)$, potom je reziduální součet čtverců roven celkovému rozptylu $\mathbf{Y}$ a $R_D = 0$, což znamená, že regresory na předpověď proměnné $\mathbf{Y}$ nemají žádný vliv.

Nyní budeme předpokládat, že první sloupec matice plánu $\mathbf{X}$ je tvořen vektorem jedniček a ukážeme, že $R_D^2 = r^2_{Y,(x_1,\ldots,x_k)}$. Pomocí věty 22 můžeme odvodit, že $0 = \hat{\mathbf{Y}}'\mathbf{e} = \sum_{j=1}^{n} \hat{Y}_j e_j = \sum_{j=1}^{n} \hat{Y}_j(Y_j - \hat{Y}_j) = \sum_{j=1}^{n} \hat{Y}_j Y_j - \sum_{j=1}^{n} \hat{Y}_j^2$, z čehož je zřejmé, že $\sum_{j=1}^{n} \hat{Y}_j Y_j = \sum_{j=1}^{n} \hat{Y}_j^2$.

Teoretický koeficient mnohonásobné korelace $\rho_{Y,(x_1,\ldots,x_k)}$ je korelační koeficient mezi náhodnou veličinou Y a její nejlepší lineární aproximací založené na $x_1, \ldots, x_k$. Proto, označíme-li výběrový korelační koeficient mezi složkami vektorů $\mathbf{Y}$ a $\hat{\mathbf{Y}}$ symbolem $r_{Y,\hat{Y}}$, můžeme podle [1] psát

$$r^2_{Y,(x_1,\ldots,x_k)} = r^2_{Y,\hat{Y}} = \frac{(\sum_{i=1}^{n} Y_i\hat{Y}_i - n\bar{Y}\bar{\hat{Y}})^2}{(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2)(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{\hat{Y}}^2)}.$$

Protože $\bar{Y} = \bar{\hat{Y}}$ a $\sum_{j=1}^{n} \hat{Y}_j Y_j = \sum_{j=1}^{n} \hat{Y}_j^2$, můžeme dále odvodit

$$r^2_{Y,(x_1,\ldots,x_k)} = \frac{(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2)^2}{(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2)(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2)} = \frac{(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2)}{(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2)} =$$

$$= \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = R_D^2.$$

Za předpokladu, že první sloupec matice $\mathbf{X}$ je tvořen vektorem jedniček, $100R_D^2$ určuje, kolik procent variability dat je vysvětleno regresním modelem. Tedy $100r^2_{Y,(x_1,\ldots,x_k)}$ značí procenta variability dat vysvětlené regresním modelem i pokud předpoklad, že první sloupec matice $\mathbf{X}$ obsahuje samé jedničky, splněn není.

K ověření, že regresní model (2.1) je dobře sestaven, je možné užít grafické metody. Pokud platí lineární model (2.1), potom j-té reziduum e_j je odhadem j-té chyby ε_j. Protože předpokládáme, že vektor chyb $\boldsymbol{\varepsilon}$ má normální rozdělení s nulovou střední hodnotou a konstatním rozptylem, požadujeme, aby i vektor reziduí $\mathbf{e}$ měl nulovou střední hodnotu a konstantní rozptyl. Protože platí

$$\mathrm{E}(\mathbf{e}) = \mathrm{E}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathrm{E}(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0},$$

$$\mathrm{Var}(\mathbf{e}) = \mathrm{Var}\left([\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}\right) = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathrm{Var}(\mathbf{Y})[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'$$

$$= \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \sigma^2[\mathbf{I} - \mathbf{H}]$$

je zřejmé, že vektor reziduí má nulovou střední hodnotu, avšak jeho varianční matice $\sigma^2[\mathbf{I} - \mathbf{H}]$, kde $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, není diagonální, jednotlivá rezidua nemají stejný rozptyl a navíc mohou mít nenulové korelace. Pokud jsou diagonální prvky matice $\sigma^2[\mathbf{I} - \mathbf{H}]$ přibližně shodné, potom jsou rozptyly jednotlivých reziduí přibližně stejně velké a model je v pořádku. V opačném případě má matice plánu $\mathbf{X}$ na vlastnosti reziduí vliv, který je třeba v další analýze eliminovat.

Aby bylo možné užitím grafických metod model posoudit, budou zavedena standardizovaná rezidua

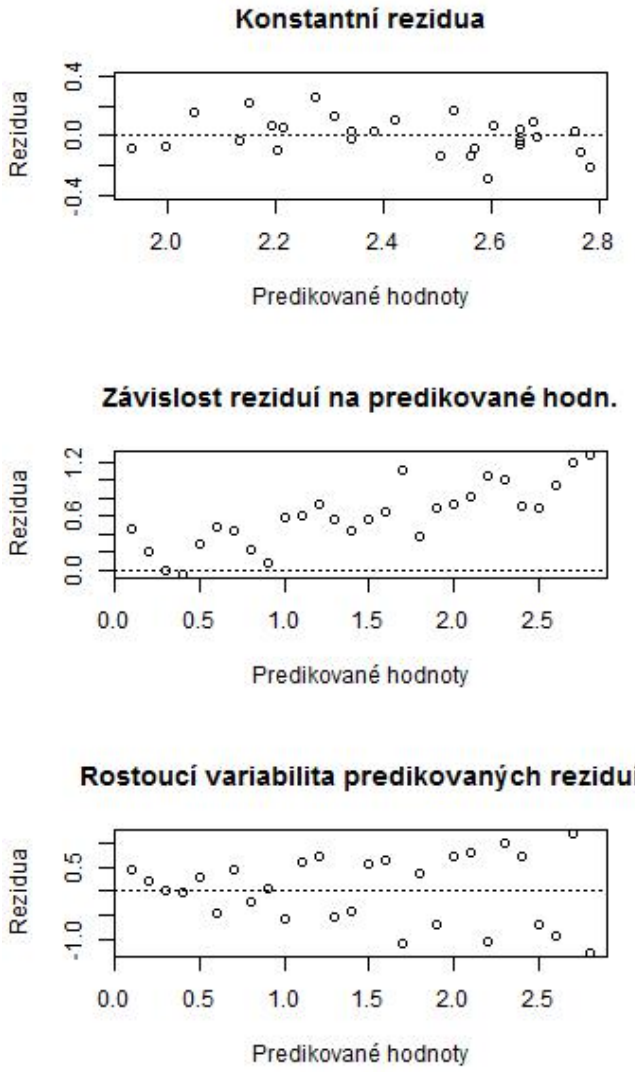
$$e_j^* = \frac{e_j}{\sqrt{s^2(1 - h_{jj})}}, \qquad j = 1, \ldots, n,$$

kde h_{jj} je prvek matice $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ na pozici (j, j). V ideálním případě by standardizovaná rezidua měla vypadat jako náhodný výběr z normovaného normálního rozdělení, což je obvykle stěží dosažitelná podmínka. Abychom mohli normalitu reziduí posoudit, vykreslíme následující grafy:

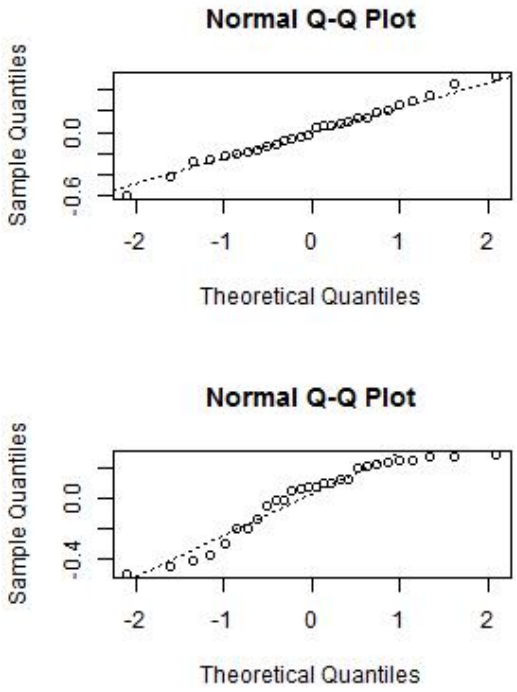
- Rezidua e_j vykreslíme na y-ové ose proti predikovaným hodnotám, vyneseným na x-ové ose. Podle toho, jak bude graf vypadat, můžeme rozlišit tři případy zobrazené na obrázku 2.2. Ideálně by všechna rezidua měla ležet v pásu ohraničeném dvěma rovnoběžnými přímkami $y = \pm\delta$, kde $\delta > 0$ je vhodná konstanta.

Pokud model není v pořádku, může se objevit závislost reziduí na predikované hodnotě, např. s rostoucí predikovanou hodnotou rostou i samotná rezidua. Další problém nastane, pokud rozptyly jednotlivých reziduí nejsou konstantní a s rostoucí predikovanou hodnotu roste i variabilita jednotlivých reziduí. Takový model můžeme opravit vhodnou transformací nebo užitím vážené metody nejmenších čtverců.

- Dalším grafem je tzv. Q-Q graf standardizovaných reziduí, podle kterého je možné usoudit, zda mají rezidua normální rozdělení. V Q-Q grafu jsou proti sobě vykresleny pozorované a teoretické kvantily. Na y-ové ose jsou vyneseny pozorované kvantily (vypočtené ze standardizovaných reziduí), na x-ové ose jsou vyneseny teoretické kvantily vypočtené z normálního normovaného rozdělení. Jestliže standardizovaná rezidua mají normální rozdělení, potom body v Q-Q grafu leží na přímce tak, jak je znázorněno v prvním grafu obrázku 2.3, který je stejně jako většina grafů uvedených v této práci výstupem softwaru R verze 3.0.1. popsaného v kapitole 6. Druhý graf obrázku 2.3 znázorňuje případ, kdy body netvoří přímku, ale tvar připomínající písmeno S. Znamená to, že normalita reziduí je narušena a je třeba model sestavit jiným způsobem.



Obr. 2.2: Grafy reziduí proti predikovaným hodnotám



Obr. 2.3: Q-Q grafy

3 METODA HLAVNÍCH KOMPONENT

Opět vyjdeme z motivačního příkladu. Předpokládejme, že máme náhodné veličiny $X_1, \ldots, X_p$ představující koncentrace kovů a iontů v PM1 aerosolech: např. X_1 představuje koncentraci chloridu (Cl^-) v aerosolech, X_2 představuje koncentraci fluoridu (F^-) v aerosolech atd. Cílem metody hlavních komponent je vysvětlit původních p veličin pomocí menšího počtu jejich lineárních kombinací, které nazýváme hlavní komponenty. Můžeme očekávat, že jednotlivé hlavní komponenty představují emisní zdroje PM1 aerosolů a původní veličiny $X_1, \ldots, X_p$ můžeme těmito komponentami nahradit, čímž dojde k redukci dat.

Metoda hlavních komponent patří mezi často využívané statistické metody, v praktické části této práce (v kapitole 7) je využita k nalezení skrytých faktorů, které ovlivňují koncentrace kovů a iontů v aerosolech a které mají stejnou interpretaci jako hlavní komponenty. Model zahrnující skryté faktory je podrobně popsán v následující kapitole.

V této kapitole, která je zpracovaná podle [9], je popsán princip metody hlavních komponent. Důkazy vět neuvádíme, lze je nalézt v [9].

Předpokládejme, že $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ je náhodný vektor pozorovaných veličin s varianční maticí $\mathbf{\Sigma} = (\sigma_{ij})_{i,j=1,\ldots,p}$ a nechť $\mathbf{a}_i \in \mathbb{R}^p$. Protože pro lineární kombinace

$$Y_i = \mathbf{a}_i'\mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p, \quad i = 1, \ldots, p$$

náhodného vektoru $\mathbf{X}$ podle vět 7 a 8 platí

$$\mathrm{Var}(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i, \quad \mathrm{Cov}(Y_i, Y_k) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k, \quad i, k = 1, 2, \ldots, p,$$

jsou hlavní komponenty hledány jako nekorelované lineární kombinace $Y_1, Y_2, \ldots, Y_p$ s největším možným rozptylem. Znamená to, že je třeba nalézt p-rozměrné vektory $\mathbf{a}_i$, které maximalizují $\mathrm{Var}(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i$. Protože pro vektor $\mathbf{a}_i$ splňující $\mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i > 0$ lze hodnotu výrazu $\mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i$ libovolně zvýšit vynásobením vektoru $\mathbf{a}_i$ libovolnou konstantou $c > 1$, je požadováno splnění podmínky $\mathbf{a}_i'\mathbf{a}_i = 1$. i-tá hlavní komponenta je tedy lineární kombinace $\mathbf{a}_i'\mathbf{X}$, která maximalizuje $\mathrm{Var}(\mathbf{a}_i'\mathbf{X})$ za podmínky $\mathbf{a}_i'\mathbf{a}_i = 1$ a $\mathrm{Cov}(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0$ pro $k < i$. V následující větě je popsána konstrukce hlavních komponent pomocí vlastních vektorů varianční matice $\mathbf{\Sigma}$.

Věta 29. Nechť $\mathbf{\Sigma}$ je varianční matice náhodného vektoru $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, která má dvojice vlastních čísel a vlastních vektorů $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, kde $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. Potom i-tá hlavní komponenta má tvar

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p, \quad i = 1, 2, \ldots, p$$

a platí

$$\mathrm{Var}(Y_i) = \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i, \quad \mathrm{Cov}(Y_i, Y_k) = \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_k = 0, \quad i = 1, 2, \ldots, p, i \neq k.$$

Jestliže $\lambda_i = \lambda_k$ pro $i \neq k$, potom volba odpovídajících $\mathbf{e}_i$ a Y_i není jednoznačná.

Věta 30. Nechť náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ má varianční matici $\boldsymbol{\Sigma}$ s dvojicemi vlastních čísel a vlastních vektorů $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, kde $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. Nechť $Y_i = \mathbf{e}_i'\mathbf{X}$, $i = 1, \ldots, p$ jsou hlavní komponenty. Potom platí

$$\mathrm{Tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i).$$

V následující větě je uvedeno, jak velký vliv má veličina X_k na i-tou hlavní komponentu.

Věta 31. Jestliže $Y_i = \mathbf{e}_i'\mathbf{X}$, $i = 1, \ldots, p$ jsou hlavní komponenty získané z varianční matice $\boldsymbol{\Sigma}$, potom korelační koeficient mezi komponentou Y_i a náhodnou veličinou X_k

$$\rho_{Y_i X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1, 2, \ldots, p,$$

kde $\sigma_{kk} \neq 0$.

Korelace mezi hlavními komponentami a pozorovanými náhodnými veličinami pomáhají interpretovat hlavní komponenty, ale neukazují vliv k-té veličiny X_k na i-tou komponentu Y_i v přítomnosti ostatních náhodných veličin $X_1, X_2, \ldots, X_{k-1}$, $X_{k+1}, \ldots, X_p$. Proto se doporučuje při interpretaci hlavních komponent uvažovat také koeficienty e_{ik}, které ukazují významnost veličiny X_k vzhledem ke komponentě Y_i.

Geometricky lze hlavní komponenty interpretovat jako výběr nového souřadnicového systému, který získáme rotací původního systému se souřadnicovými osami $X_1, X_2, \ldots, X_p$. Nové souřadnicové osy $\mathbf{e}_1'\mathbf{X}, \mathbf{e}_2'\mathbf{X}, \ldots, \mathbf{e}_p'\mathbf{X}$ reprezentují směry s maximální variabilitou, přičemž největší variabilita je ve směru souřadnicové osy $\mathbf{e}_1'\mathbf{X}$.

Hlavní komponenty jsou často získávány ze standardizovaných náhodných veličin $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, kde matice $\mathbf{V} = \mathrm{diag}(\boldsymbol{\Sigma})$ je regulární a $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$. Poznamenejme, že varianční matice standardizovaných veličin je rovna korelační matici původních veličin, tj. $\mathrm{Var}(\mathbf{Z}) = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2} = \mathrm{Cor}(\mathbf{X})$. V následující větě bude uveden vztah pro výpočet hlavních komponent standardizovaných veličin.

Věta 32. i-tá hlavní komponenta standartizovaných náhodných veličin $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)' = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ s varianční maticí $\mathrm{Var}(\mathbf{Z}) = \mathrm{Cor}(\mathbf{X})$ je dána

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'\mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, p$$

a platí

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p, \quad \rho_{Y_i Z_k} = e_{ik}\sqrt{\lambda_i}, \quad i,k = 1,2,\ldots,p,$$

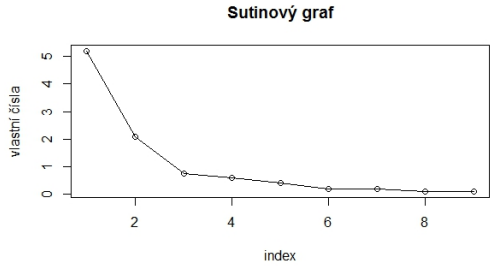
kde $\mathbf{V} = \mathrm{diag}(\mathbf{\Sigma})$, $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, a $(\lambda_1, \mathbf{e_1}), (\lambda_2, \mathbf{e_2}), \ldots, (\lambda_p, \mathbf{e_p})$ jsou dvojice vlastních čísel a vlastních vektorů matice $\mathrm{Cor}(\mathbf{X})$, splňujících $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$.

Máme-li n nezávislých pozorování náhodných veličin $X_1, \ldots, X_p$, potom lze střední hodnotu $\boldsymbol{\mu}$ odhadnout výběrovým průměrem $\bar{\mathbf{X}}$, varianční matici $\mathbf{\Sigma}$ lze odhadnout výběrovou varianční maticí $\mathbf{S}$, korelační matici $\mathrm{Cor}(\mathbf{X})$ lze odhadnout výběrovou korelační maticí $\mathbf{R}$ a při výpočtu hlavních komponent se vychází z matice $\mathbf{\Sigma}$ popř. matice $\mathbf{R}$, na kterou jsou aplikovány vztahy uvedené ve větě 29 popř. ve větě 32.

Předpokládejme, že jsou již "zkonstruované"hlavní komponenty, kterými je možné vysvětlit varianční strukturu náhodných veličin $X_1, \ldots, X_p$. Je třeba zjistit, jaký nejmenší počet $m < p$ komponent vysvětluje variabilitu dat dostatečně a těchto m komponent interpretovat. Při nesprávné volbě m by došlo ke ztrátě dat, nebo by některé z hlavních komponent vysvětlovaly pouze zanedbatelnou část variability a navíc by zřejmě byly těžce interpretovatelné. Podle vět 30 a 32 je poměr variability vysvětlené k-tou hlavní komponentou a celkové variability

$$\begin{aligned} &\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \text{ pro stanovení vlastních čísel z matice } \mathbf{S}, \\ &\frac{\lambda_k}{p} \text{ pro stanovení vlastních čísel z matice } \mathbf{R}. \end{aligned} \tag{3.1}$$

Jestliže prvních m hlavních komponent, $m < p$, vysvětluje podstatnou část variability (80 - 90%, [9]), potom tyto komponenty můžeme intepretovat a nahradit jimi původních p veličin.



Obr. 3.1: Sutinový graf

Další možností, jak určit optimální počet interpretovaných komponent, je využít tzv. "sutinový graf"vykreslený na obrázku 3.1. Je to graf vlastních čísel, které jsou

zobrazeny jako body o souřadnicích $[i, \lambda_i]$, přičemž $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. Body v grafu spojíme lomenou čarou a najdeme místo, kde dochází k největšímu poklesu mezi dvěma vlastními čísly. Komponenty odpovídající vlastním číslům ležícím nalevo od tohoto místa budou zahrnuty do modelu, zatímco komponenty odpovídající vlastním číslům ležícím napravo od tohoto místa jsou zanedbány. Z obrázku 3.1 je patrné, že dochází k největšímu poklesu mezi prvním a druhým a potom mezi druhým a třetím vlastním číslem. Na základě uvedeného grafu tedy může být učiněn závěr, že optimální je extrahovat jednu nebo dvě hlavní komponenty.

4 FAKTOROVÁ ANALÝZA

Zavedení faktorového modelu budeme opět motivovat příkladem. Stejně jako v předchozí kapitole předpokládejme, že náhodné veličiny $X_1, \ldots, X_p$ představují koncentrace kovů a iontů v PM1 aerosolech. Po stanovení výběrové korelační matice bylo zjištěno, že jednotlivé veličiny mají mezi sebou statisticky významné korelace a mohou být rozděleny do skupin tak, že veličiny patřící do stejné skupiny mají mezi sebou statisticky významné korelace, ale korelace s veličinami z ostatních skupin jsou statisticky nevýznamné. Je pravděpodobné, že tyto korelace jsou způsobeny společnými emisními zdroji a každá skupina náhodných veličin představuje skrytý faktor reprezentující jeden emisní zdroj. Cílem faktorové analýzy je tyto společné faktory nalézt a interpretovat.

V aplikační části této práce uvedené v kapitole 7 je faktorová analýza využita k určení hlavních emisních zdrojů aerosolů v Brně a ve Šlapanicích.

V této kapitole, která je zpracovaná podle [9], je popsán faktorový model a teorie, potřebná k praktickému výpočtu.

4.1 Ortogonální faktorový model

Nechť náhodný vektor pozorovaných veličin $\mathbf{X}_{p\times 1} = (X_1, X_2, \ldots, X_p)'$ má střední hodnotu $\boldsymbol{\mu}$ a varianční matici $\boldsymbol{\Sigma}$. Ortogonální faktorový model předpokládá, že $\mathbf{X}$ lineárně závisí na náhodném vektoru $\mathbf{F}_{m\times 1} = (F_1, F_2, \ldots, F_m)'$ nepozorovatelných náhodných veličin, které budeme nazývat faktory a vektoru p chyb, neboli specifických faktorů $\boldsymbol{\varepsilon}_{p\times 1} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)'$. Označme $\mathbf{L}$ matici reálných čísel, která bude nazývána matice faktorových zátěží a jejíž prvky l_{ij} značí zátěž faktoru F_j v proměnné X_i. Faktorový model může být zapsán

$$\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1,\\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2,\\
&\vdots\\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p.
\end{aligned}$$

Maticový zápis ortogonálního faktorového modelu je

$$\mathbf{X}_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = \mathbf{L}_{p\times m}\mathbf{F}_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1}. \tag{4.1}$$

Je zřejmé, že odchylky $X_i - \mu_i$ závisí na nepozorovatelných náhodných veličinách F_j, ε_i, kde $j = 1, \ldots, m$, $i = 1, \ldots, p$ a faktorový model (4.1) tudíž obsahuje velké množství nepozorovatelných proměnných (F_j, ε_i, kde $j = 1, \ldots, m$, $i = 1, \ldots, p$).

Dále, aby bylo možné odhadnout proměnné F_j, budeme pro zjednodušení úlohy předpokládat:

$$\mathrm{E}(\mathbf{F}) = \mathbf{0}_{m\times 1}, \quad \mathrm{Var}(\mathbf{F}) = \mathbf{I}_m, \tag{4.2}$$

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_{p\times 1}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}_{p\times p} = \mathrm{diag}(\Psi_1, \Psi_2, \ldots, \Psi_p), \tag{4.3}$$

$$\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{F}) = \mathbf{0}_{p\times m}. \tag{4.4}$$

Z modelu (4.1) odvodíme

$$(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = (\boldsymbol{LF} + \boldsymbol{\varepsilon})(\boldsymbol{LF} + \boldsymbol{\varepsilon})' = (\boldsymbol{LF} + \boldsymbol{\varepsilon})((\mathbf{LF})' + \boldsymbol{\varepsilon}') =$$

$$= \mathbf{LF}(\mathbf{LF})' + \boldsymbol{\varepsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}',$$

a dále užitím vlastností střední hodnoty odvodíme tvar varianční matice

$$\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X}) = \mathrm{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = \mathrm{E}(\boldsymbol{LF} + \boldsymbol{\varepsilon})(\boldsymbol{LF} + \boldsymbol{\varepsilon})' =$$

$$= \mathrm{E}\left[\mathbf{LF}(\mathbf{LF})' + \boldsymbol{\varepsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = \mathbf{L}\mathrm{E}(\mathbf{F}\mathbf{F}')\mathbf{L}' + \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{F}')\mathbf{L}' + \mathbf{L}\mathrm{E}(\boldsymbol{F}\boldsymbol{\varepsilon}') + \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') =$$

$$= \mathbf{L}\mathrm{Var}(\mathbf{F})\mathbf{L}' + \mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F})\mathbf{L}' + \mathbf{L}\mathrm{Cov}(\mathbf{F}, \boldsymbol{\varepsilon}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi},$$

a tvar kovarianční matice

$$\mathrm{Cov}(\mathbf{X}, \mathbf{F}) = \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})\mathbf{F}'] = \mathrm{E}[(\boldsymbol{LF} + \boldsymbol{\varepsilon})\mathbf{F}'] = \mathbf{L}\mathrm{E}(\mathbf{F}\mathbf{F}') + \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{F}') =$$

$$= \mathbf{L}\mathrm{Var}(\mathbf{F}) + \mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = \mathbf{L}.$$

Z předchozího odvození plyne následující varianční struktura modelu:

$$\mathrm{Var}(X_i) = \sigma_{ii} = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 + \Psi_i = h_i^2 + \Psi_i,$$

$$\mathrm{Cov}(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \ldots + l_{im}l_{km},$$

$$\mathrm{Cov}(X_i, F_j) = l_{ij}.$$

Výraz $h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2$ se nazývá i-tá komunalita a představuje část variability proměnné X_i, která je vysvětlena společnými faktory, zatímco výraz Ψ_i nazývaný specificita (jedinečnost) představuje část variability X_i, která náleží chybě ε_i.

Nyní ukážeme, že matice faktorových zátěží $\mathbf{L}$ a vektor společných faktorů $\mathbf{F}$ nejsou určeny jednoznačně. Nechť $\mathbf{T}_{m\times m}$ je ortogonální matice splňující $\mathbf{T}\mathbf{T}' = \mathbf{I}$. Potom

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon}.$$

Násobením ortogonální maticí $\mathbf{T}$ získáme nový vektor společných faktorů $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ a novou matici faktorových zátěží $\mathbf{L}^* = \mathbf{L}\mathbf{T}$, pro které platí $\mathrm{E}(\mathbf{F}^*) = \mathbf{T}'\mathrm{E}(\mathbf{F}) = \mathbf{T}'\mathbf{0} = \mathbf{0}$, $\mathrm{Var}(\mathbf{F}^*) = \mathbf{T}'\mathrm{Var}(\mathbf{F})\mathbf{T} = \mathbf{I}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}^*) = \mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F})\mathbf{T} = \mathbf{0}\mathbf{T} = \mathbf{0}$ a tudíž

nové faktory $\mathbf{F}^*$ také splňují předpoklady (4.2) a (4.4). Vektory $\mathbf{F}, \mathbf{F}^*$ a matice $\mathbf{L}, \mathbf{L}^*$ mají stejné statistické vlastnosti a generují stejnou varianční matici $\boldsymbol{\Sigma}$. Můžeme tedy psát

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)' + \boldsymbol{\Psi}. \tag{4.5}$$

Této vlastnosti se využívá při tzv. rotaci faktorů. Společné faktory reprezentují souřadnicové osy v kartézském souřadném systému, kde polohy proměnných X_i jsou určeny souřadnicemi $[l_{i1}, l_{i2}, \ldots, l_{im}]$ a násobení ortogonální maticí odpovídá rotaci souřadnicového systému. Podrobně bude faktorová rotace rozebrána později.

Máme-li datovou matici $\mathbf{X}_D$ obsahující n nezávislých pozorování $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ náhodného vektoru $\mathbf{X}$, varianční matice $\boldsymbol{\Sigma}$ je odhadnuta výběrovou varianční maticí $\mathbf{S}$ a můžeme přejít k odhadu parametrů faktorového modelu.

Poznamenejme, že náhodný výběr obvykle standardizujeme a následně pracujeme s náhodným výběrem standardizované náhodné veličiny $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, kde $\mathbf{V} = \mathrm{diag}(\boldsymbol{\Sigma})$. Varianční matice standardizovaného vektoru $\mathbf{Z}$ je rovna korelační matici vektoru $\mathbf{X}$ a protože $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$, ortogonální faktorový model můžeme zapsat ve tvaru

$$\mathbf{Z}_{p\times 1} = \mathbf{L}_{p\times m}\mathbf{F}_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1},$$

$$\mathrm{Var}(\mathbf{Z}) = \mathrm{Cor}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}.$$

Je tedy zřejmé, že při odhadu parametrů faktorového modelu můžeme vycházet z korelační matice $\mathrm{Cor}(\mathbf{X})$, kterou odhadneme výběrovou korelační maticí $\mathbf{R}$.

Než přejdeme k odhadu parametrů, je třeba posoudit, zda jsou data vhodná k užití faktorové analýzy. Jestliže jsou mimodiagonální prvky výběrové varianční matice $\mathbf{S}$ relativně malé nebo prvky výběrové korelační matice $\mathbf{R}$ ležící mimo diagonálu téměř nulové, potom nemá užití faktorové analýzy smysl, protože proměnné $X_1, X_2, \ldots, X_p$ jsou navzájem prakticky nekorelované a varianční strukturu faktorového modelu vysvětlují především specificity (specifické faktory), nikoliv společné faktory. Korelační strukturu pozorovaných proměných je možné posoudit užitím Bartlettova testu sféricity, který bude popsán v odstavci 4.5. V případě, že pozorované proměnné mají dobrou korelační strukturu, je třeba odhadnout parametry faktorového modelu a následně interpretovat společné faktory. K odhadu parametrů se nejčastěji používá metoda hlavních komponent a metoda maximální věrohodnosti. Konstrukce hlavních komponent již byla popsána v kapitole 3, v následujícím odstavci bude popsáno, jak je možné hlavní komponenty užít k nalezení a interpretaci společných faktorů. Také budou doplněny metody vhodné k určení počtu extrahovaných komponent (faktorů). Metoda maximální věrohodnosti pro odhad parametrů faktorového modelu bude popsána v odstavci 4.3.

4.2 Odhad parametrů faktorového modelu metodou hlavních komponent

Jak bylo popsáno v kapitole 3, princip metody hlavních komponent spočívá v hledání dvojic vlastních čísel a vlastních vektorů $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ varianční matice $\mathbf{\Sigma}$ uspořádaných tak, že $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Potom platí

$$\mathbf{\Sigma}_{p\times p} = [\sqrt{\lambda_1}\mathbf{e}_1 \quad \sqrt{\lambda_2}\mathbf{e}_2 \quad \ldots \quad \sqrt{\lambda_p}\mathbf{e}_p \quad] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \sqrt{\lambda_2}\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_p}\mathbf{e}_p' \end{bmatrix} = \mathbf{L}_{p\times p}\mathbf{L}'_{p\times p} + \mathbf{0}_{p\times} = \mathbf{L}_{p\times p}\mathbf{L}'_{p\times p}. \tag{4.6}$$

Ze vztahu (4.6) je zřejmé, že $\mathbf{\Sigma}$ má požadovaný tvar (4.5) vyplývající z kovarianční struktury faktorového modelu. Prvky matice faktorových zátěží l_{ij} jsou určeny $l_{ij} = \sqrt{\lambda_j}\mathbf{e}_{ji}$, specificity Ψ_i jsou nulové a komponenty odpovídající vlastním číslům $\mathbf{e}_i$ reprezentují společné faktory. Varianční matice $\mathbf{\Sigma}$ je takto odhadnuta naprosto přesně, ale počet společných faktorů je stejný jako počet původních proměnných $X_1, X_2, \ldots, X_p$. Protože je požadováno, aby variabilita pozorovaných proměnných byla vysvětlena pomocí menšího počtu $m < p$ společných faktorů, zanedbáme v (4.6) příspěvek do varianční matice $\mathbf{\Sigma}$ od $\lambda_{m+1}\mathbf{e}_{m+1}\mathbf{e}'_{m+1} + \ldots + \lambda_p\mathbf{e}_p\mathbf{e}'_p$, kde vlastní čísla $\lambda_{m+1} + \ldots + \lambda_p$ mají relativně malou hodnotu. Do aproximace matice $\mathbf{\Sigma}$ zahrneme i specifické faktory a dostaneme

$$\mathbf{\Sigma} \doteq [\sqrt{\lambda_1}\mathbf{e}_1 \quad \sqrt{\lambda_2}\mathbf{e}_2 \quad \ldots \quad \sqrt{\lambda_m}\mathbf{e}_m \quad] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \sqrt{\lambda_2}\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}_m' \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{bmatrix} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}. \tag{4.7}$$

Z aproximace 4.7 je zřejmé, že specificity jsou určeny

$$\Psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2, \quad i = 1, 2, \ldots, p$$

a matice faktorových zátěží má tvar

$$\mathbf{L}_{p\times m} = (l_{ij})_{i=1,\ldots,p;j=1,\ldots,m} = \sqrt{\lambda_j}e_{ji}.$$

Vztah (4.7) aplikujeme na výběrovou varianční matici $\mathbf{S}$, nebo výběrovou korelační matici $\mathbf{R}$ náhodného vektoru $\mathbf{X}$. Nejprve předpokládejme, že vycházíme z výběrové varianční matice $\mathbf{S}$. Označme $(\hat{\lambda}_1\hat{\mathbf{e}}_1), (\hat{\lambda}_2\hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_m\hat{\mathbf{e}}_m)(\hat{\lambda}_{m+1}\hat{\mathbf{e}}_{m+1}) \ldots (\hat{\lambda}_p\hat{\mathbf{e}}_p)$ dvojice vlastních čísel a vlastních vektorů matice $\mathbf{S}$ takové, že

$\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_m$ jsou relativně větší než $\hat{\lambda}_{m+1} > \ldots > \hat{\lambda}_p$. Podle (4.7) platí $\hat{\mathbf{L}} = (\hat{l}_{ij}) = [\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 \;\; \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 \;\; \ldots \;\; \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m]$, kde $\hat{\mathbf{L}} = (\hat{l_{ij}})$ je odhad matice faktorových zátěží a $m < p$ značí počet společných faktorů. Specificity jsou odhadnuty jako diagonální prvky matice $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$, tedy $\hat{\Psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}^2$, $\hat{\mathbf{\Psi}} = \mathrm{diag}(\hat{\Psi}_1, \hat{\Psi}_2, \ldots, \hat{\Psi}_p)$ a komunality odhadneme jako součet čtverců odhadů faktorových zátěží, tedy $\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \ldots + \hat{l}_{im}^2$. Při výpočtu parametrů z výběrové korelační matice $\mathbf{R}$ je postup analogický.

Poznamenejme, že při výpočtu metodou hlavních komponent se odhad faktorových zátěží nemění, jestliže počet faktorů postupně zvyšujeme.

V praxi je třeba zjistit, kolik společných faktorů je nutné extrahovat a interpretovat, aby byla variabilita dat dostatečně vysvětlena. Proto bude nyní uvedeno několik metod, pomocí kterých lze nalézt optimální počet extrahovaných faktorů.

Předpokládejme, že parametry faktorového modelu extrahujeme z výběrové varianční matice $\mathbf{S}$. Zavedeme reziduální matici $\mathbf{\Xi} = (\xi_{ij}) = \mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})$, jejíž diagonální prvky jsou v důsledku korelační struktury faktorového modelu nulové. Mimodiagonální prvky jsou nenulové, avšak pokud jsou malé, můžeme předpokládat, že extrahovaný počet společných faktorů je dostatečný. Protože platí

$$\sum_{i=1}^{p}\sum_{j=1}^{p} \xi_{ij}^2 \leq \hat{\lambda}_{m+1}^2 + \ldots + \hat{\lambda}_p^2,$$

můžeme počet m interpretovaných faktorů vhodně zvolit na základě odhadu vlastních čísel matice $\mathbf{S}$ (viz vztah (3.1)). Podíl variability vysvětlené pomocí společného faktoru F_j a celkové variability je dán

$$\frac{\hat{\lambda}_j}{s_{11} + s_{22} + \ldots + s_{pp}} \quad \text{pro stanovení vlastních čísel z matice } \mathbf{S},$$

$$\frac{\hat{\lambda}_j}{p} \quad \text{pro stanovení vlastních čísel z matice } \mathbf{R}.$$

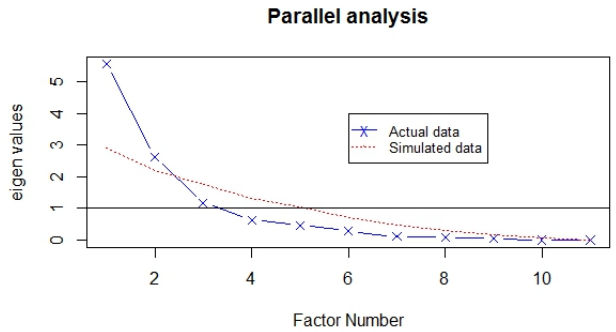
Počet m společných faktorů zvolíme tak, aby bylo vysvětleno alespoň 70% (ideálně 80% - 90%) celkové variability.

Počet společných faktorů můžeme také určit na základě Kaiserova kritéria, podle kterého by měly být interpretovány faktory odpovídají vlastním číslům matice $\mathbf{R}$ větším než 1 a faktory odpovídající vlastním číslům s hodnotou menší než 1 by měly být zanedbány. Přestože je tento přístup poměrně často využívaný, není zcela ideální. Pokud máme vlastní čísla s hodnotami 1,05 a 0,95, podle Kaiserova kritéria by měl být extrahován pouze faktor odpovídající prvnímu vlastnímu číslu, přestože faktor odpovídající vlastnímu číslu s hodnotou 0,95 vysvětluje téměř stejné procento variability.

Další možností, jak určit optimální počet interpretovaných společných faktorů, je využít tzv. sutinový graf, který byl popsán v kapitole 3. Protože kritérium vycházející ze sutinového grafu je z velké části intuitivní, uvedeme ještě metodu, která se nazývá paralelní analýza [11].

Paralelní analýza je založena na simulacích Monte Carlo, pomocí kterých je generován datový soubor stejného rozsahu a se stejným počtem proměnných jako je originální datový soubor, z kterého jsou extrahovány společné faktory. Generovaný datový soubor však obsahuje náhodné veličiny z normálního rozdělení. Z varianční matice souboru generovaných náhodných veličin jsou spočtena vlastní čísla a celý postup se několikrát opakuje (obvykle 50-100 replikací). Vlastní čísla získaná z varianční matice originálního datového souboru jsou potom porovnána s vlastními čísly, která odpovídají generovaným datovým souborům. Dříve se používal přístup, kdy byly extrahovány faktory odpovídající vlastním číslům s hodnotou větší než průměry vlastních čísel získaných z variančních matic simulovaných souborů. Tento přístup je využit v implementované funkci softwaru R, který byl použit k praktickému výpočtu a proto je také v aplikacích v odstavci 7.3. V současné době se však využívá přístup, podle kterého jsou extrahovány faktory odpovídající vlastním číslům s hodnotou větší než 95% kvantily vlastních čísel odpovídajících simulovaným souborům.

Příklad výsledku paralelní analýzy je zobrazen v grafu na obrázku 4.1. Vlastní čísla získaná z varianční matice originálního datového souboru jsou spojena modrou čarou, zatímco červená čára spojuje průměry vlastních čísel odpovídajících simulovaným datům. Je zřejmé, že by měly být extrahovány dva společné faktory, protože přesně dvě vlastní čísla odpovídající originálním datům jsou větší než průměry vlastních čísel vypočtených z variančních matic simulovaných dat.



Obr. 4.1: Výsledek paralelní analýzy

4.3 Odhad parametrů faktorového modelu metodou maximální věrohodnosti

Metoda maximální věrohodnosti je založena na maximalizaci věrohodnostní funkce, která je v tomto odstavci odvozena. Předpokladem pro použití metody maximální věrohodnosti je znalost rozdělení náhodného vektoru $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$. Zde budeme předpokládat, že náhodný vektor $\mathbf{X}$ má p-rozměrné normální rozdělení. Značí-li každý z p-rozměrných vektorů $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ nezávislá pozorování náhodného vektoru $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, potom za předpokladu normality vektory $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ reprezentují náhodný výběr z $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Pokud sdružené rozdělení $\mathbf{F}_j$ a $\boldsymbol{\varepsilon}_j$ (j-té pozorování) je mnohorozměrné normální, potom má mnohorozměrné normální rozdělení také $\mathbf{X}_j - \boldsymbol{\mu} = \mathbf{L}\mathbf{F}_j + \boldsymbol{\varepsilon}_j$, $j = 1, 2, \ldots, n$.

Nyní přejdeme k odvození věrohodnostní funkce $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Protože pozorování $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ p-rozměrného náhodného vektoru jsou nezávislá a $\mathbf{X}_j \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), j = 1, 2, \ldots, n$, jejich společná hustota $f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ je rovna součinu marginálních hustot. Hustotu $f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ zapíšeme v proměnných $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ a podle věty 12 dostaneme

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) =$$
$$= \prod_{j=1}^{n} \left[\frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -(\mathbf{X}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu})/2 \right\} \right] =$$
$$= \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp\left\{ -\sum_{j=1}^{n} (\mathbf{X}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu})/2 \right\}.$$

Podle prvního vztahu věty 3 platí

$$(\mathbf{X}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu}) = \mathrm{Tr}\left[(\mathbf{X}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu}) \right] = \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})' \right].$$

Dále, s využitím druhého vztahu věty 4 odvodíme, že

$$\sum_{j=1}^{n} (\mathbf{X}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu}) = \sum_{j=1}^{n} \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})' \right] =$$
$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})' \right) \right].$$

Připoměňme, že $\bar{\mathbf{X}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{X}_j$ značí výběrový průměr a předchozí vztah dále upravujeme:

$$\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})' \right) \right] =$$

$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}} + \bar{\mathbf{X}} - \boldsymbol{\mu})(\mathbf{X}_j - \bar{\mathbf{X}} + \bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right] =$$

$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' + \sum_{j=1}^{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right] +$$

$$+\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})(\mathbf{X}_j - \bar{\mathbf{X}})' + \sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right].$$

Protože $\sum_{j=1}^{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})(\mathbf{X}_j - \bar{\mathbf{X}})' = \sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}} - \boldsymbol{\mu})' = \mathbf{0}$, platí

$$\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})'\right)\right] =$$

$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' + \sum_{j=1}^{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})'(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right] =$$

$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' + n(\bar{\mathbf{X}} - \boldsymbol{\mu})'(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right].$$

Poslední vztah dosadíme do $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a dostaneme

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\times$$

$$\times\exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' + n(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right]\right\}.$$

Protože podle vět 3 a 4 platí:

$$\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' + n(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right)\right] =$$

$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'\right)\right] + n\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'\right] =$$

$$= \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'\right)\right] + n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}),$$

můžeme $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ napsat ve tvaru

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\times$$

$$\times\exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'\right)\right] - \frac{1}{2}n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})\right\} =$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' \right) \right] \right\} \times \\ \times \exp\left\{ -\frac{1}{2} n (\bar{\mathbf{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \right\}. \tag{4.8}$$

Z předpokladů faktorového modelu bylo odvozeno $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$, z čehož je zřejmé, že $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ závisí na $\mathbf{L}$ a $\boldsymbol{\Psi}$. V odstavci 4.1 byla zmíněna ortogonální transformace a bylo ukázáno, že matice $\mathbf{L}$ není jednoznačně určena. Aby úloha měla jednoznačné řešení, klademe doplňující podmínku

$$\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L} = \boldsymbol{\Delta}, \tag{4.9}$$

kde $\boldsymbol{\Delta}$ je diagonální matice. Podmínka (4.9) zajistí jednoznačné určení matice faktorových zátěží $\mathbf{L}$. Odhady matic $\hat{\mathbf{L}}$ a $\hat{\boldsymbol{\Psi}}$ metodou maximální věrohodnosti potom získáme numerickou maximalizací (4.8), kde varianční matici $\boldsymbol{\Sigma}$ nahradíme maticí $[(n-1)/n]\mathbf{S}$, a komunality odhadneme jako součet čtverců odhadů faktorových zátěží, tedy $\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \ldots + \hat{l}_{im}^2$. K numerické maximalizaci užijeme vhodný software, např. Statistica Cz 10 nebo R verze 2.15.2.

Pracujeme-li se standardizovanou náhodnou veličinou $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, platí

$$\mathrm{Var}(\mathbf{Z}) = \mathrm{Cor}(\mathbf{X}) = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2} = (\mathbf{V}^{-1/2}\mathbf{L})(\mathbf{V}^{-1/2}\mathbf{L})' + \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2}.$$

Je tedy zřejmé, že pokud vycházíme z korelační matice $\mathrm{Cor}(\mathbf{X})$, dostaneme matici faktorových zátěží $\mathbf{L}_z = \mathbf{V}^{-1/2}\mathbf{L}$ a matici specificit $\boldsymbol{\Psi}_z = \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2}$. Matice $\mathbf{V}$ je odhadnuta maticí $\mathbf{D} = \mathrm{diag}\{\mathbf{S}\}$, matice $\mathrm{Cor}(\mathbf{X})$ je odhadnuta výběrovou korelační maticí $\mathbf{R}$ a platí $\mathbf{R} = (\mathbf{D}^{-1/2}\hat{\mathbf{L}})(\mathbf{D}^{-1/2}\hat{\mathbf{L}})' + \mathbf{D}^{-1/2}\hat{\boldsymbol{\Psi}}\mathbf{D}^{-1/2} = \hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\boldsymbol{\Psi}}_z$. Odhady matic $\hat{\mathbf{L}}_z$ a $\hat{\boldsymbol{\Psi}}_z$ získáme numerickou maximalizací (4.8), kde varianční matici $\boldsymbol{\Sigma}$ nahradíme maticí $\mathbf{R}$.

Jsou-li $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$ odhady získané metodou maximální věrohodnosti vycházející z matice $\mathbf{S}$ a $\hat{\mathbf{L}}_z = \mathbf{V}^{-1/2}\mathbf{L}$, $\hat{\boldsymbol{\Psi}}_z = \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2}$ odhady získané metodou maximální věrohodnosti vycházející z matice $\mathbf{R}$, potom poměr variability vysvětlené pomocí faktoru F_j a celkové variability je dán vztahy

$$\frac{\hat{l}_{1j}^2 + \hat{l}_{2j}^2 + \ldots + \hat{l}_{pj}^2}{s_{11} + s_{22} + \ldots + s_{pp}}, \quad \text{pro odhad parametrů vycházející z matice } \mathbf{S},$$

$$\frac{\hat{l}_{z_{1j}}^2 + \hat{l}_{z_{1j}}^2 + \ldots + \hat{l}_{z_{1j}}^2}{p}, \quad \text{pro odhad parametrů vycházející z matice } \mathbf{R}.$$

Nyní bude uvedena věta, kterou užijeme v odstavci 4.6 k testování hypotézy, že počet extrahovaných faktorů je dostatečný.

Věta 33. Nechť $\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_n$ je náhodný výběr z p-rozměrného normálního rozdělení. Odhady $\hat{\mathbf{L}}$ a $\hat{\mathbf{\Psi}}$ (maximálně věrohodné) metodou maximální věrohodnosti získané maximalizací (4.8) vzhledem k podmínce (4.9) splňují

$$(\hat{\mathbf{\Psi}}^{-1/2}\mathbf{S}_n\hat{\mathbf{\Psi}}^{-1/2})(\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{L}}) = (\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{L}})(\mathbf{I} + \hat{\mathbf{\Delta}}).$$

Tedy j-tý sloupec $\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{L}}$ je vlastní vektor matice $\hat{\mathbf{\Psi}}^{-1/2}\mathbf{S}_n\hat{\mathbf{\Psi}}^{-1/2}$ odpovídající vlastnímu číslu $1 + \hat{\Delta}_i$. Dále $\mathbf{S}_n = \frac{n-1}{n}\mathbf{S}$ a $\hat{\Delta}_1 \geq \hat{\Delta}_2 \geq \ldots \geq \hat{\Delta}_m$.Také platí

$$\hat{\psi}_i = i\text{-tý diagonální prvek } \mathbf{S}_n - \hat{\mathbf{L}}\hat{\mathbf{L}}', \quad \mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}_n = p).$$

Důkaz: [9]

4.4 Rotace faktorů

V odstavci 4.1 bylo uvedeno, že po provedení ortogonální transformace získáme novou matici faktorových zátěží $\mathbf{L}^* = \mathbf{LT}$, která generuje stejnou varianční matici $\mathbf{\Sigma}$ (popř. korelační matici $\mathrm{Cor}(\mathbf{X})$) jako původní matice $\mathbf{L}$. Ortogonální transformace faktorových zátěží a společných faktorů se nazývá faktorová rotace.

Odhady faktorových zátěží $\hat{l}_{ij}$ jsou obvykle špatně interpretovatelné a proto jsou užitím faktorové rotace získány odhady lépe interpretovatelných rotovaných zátěží $\hat{l}^*_{ij}$. Ideální by bylo získat takový odhad matice $\mathbf{L}$, aby každá proměnná X_i, $i = 1, \ldots, p$ měla velkou zátěž pouze od jediného faktoru. V takovém případě říkáme, že byla získána jednoduchá struktura.

Je-li $\hat{h}_i$, $i = 1, \ldots, p$ odhad komunalit a $\hat{\mathbf{L}}$ odhad matice faktorových zátěží získaný užitím jedné ze dvou dříve popsaných metod, potom $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$ je odhad matice rotovaných faktorových zátěží. Navíc platí $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*\prime} + \hat{\mathbf{\Psi}}$. Nejčastěji užívanou ortogonální transformací je varimax kritérium, které hledá ortogonální transformaci $\mathbf{T}$ maximalizující výraz

$$V = \frac{1}{p}\sum_{j=1}^{m}\left\{\sum_{i=1}^{p}\left(\frac{\hat{l}^*_{ij}}{\hat{h}_i}\right)^4 - \frac{1}{p}\left[\sum_{i=1}^{p}\left(\frac{\hat{l}^*_{ij}}{\hat{h}_i}\right)^2\right]^2\right\}.$$

Maximalizace výrazu V odpovídá maximalizaci variability čtverců $(\hat{l}^*_{ij}/\hat{h}_i)$ na každém faktoru. Tato variabilita bude maximální, pokud bude získána jednoduchá struktura a každý sloupec matice $\hat{\mathbf{L}}^*$ bude obsahovat skupiny relativně velkých a skupiny zanedbatelných koeficientů. Znamená to, že varimax kritérium minimalizuje počet proměnných, které jsou vysvětlené jedním faktorem. Dělení koeficientů $\hat{l}^*_{ij}$ odhadem komunalit $\hat{h}_i$ má za následek, že veličiny, jejichž komunality jsou relativně malé, mají v určení jednoduché struktury větší váhu než veličiny s relativně velkými komunalitami.

Ortogonální transformace budeme dále používat v aplikační části práce k identifikaci emisních zdrojů atmosferických aerosolů, protože faktory reprezentující emisní zdroje jsou nezávislé. V některých situacích je však učiněn předpoklad závislosti společných faktorů. V takovém případě je vhodné užít šikmé faktorové rotace [19], ale v této práci se jimi nebudeme podrobně zabývat.

4.5 Bartlettův test sféricity

V odstavci 4.1 bylo uvedeno, že k odhadu parametrů faktorového modelu je možné přejít až po ověření, že data jsou adekvátní k užití faktorové analýzy.

Jsou-li mimodiagonálmí prvky výběrové korelační matice $\mathbf{R}$ analyzovaného datového souboru relativně malé (téměř nulové), potom jednotlivé proměnné nejsou vzájemně korelované a ve faktorovém modelu vystupují pouze specificity. Je proto třeba testovat hypozétu, zda jsou jednotlivé proměnné vstupující do modelu faktorové analýzy mezi sebou korelované, což je ekvivalentní s testováním hypotézy, že korelační matice $\mathrm{Cor}(\mathbf{X})$ datového souboru je jednotkovou maticí. Tuto hypotézu lze ověřit Bartlettovým testem sféricity [18], který dále uvedeme.

Je testována nulová hypotéza

$$\mathrm{H}_0 : \mathrm{Cor}(\mathbf{X}) = \mathbf{I},$$

oproti alternativní hypotéze $\mathrm{H}_1 : \mathrm{Cor}(\mathbf{X}) \neq \mathbf{I}$. Testovací statistika

$$B = -\left[(n-1) - \frac{2p+5}{6}\right] \ln|\mathbf{R}| \tag{4.10}$$

má asymptoticky $\chi^2(\nu)$ rozdělení, kde $\nu = p(p-1)/2$ jsou stupně volnosti (viz [18]). Nulová hypotéza je zamítnuta, jestliže realizace testovací statistiky je větší než příslušný kvantil.

4.6 Test o počtu společných faktorů

Máme-li odhad parametrů faktorového modelu, je vhodné otestovat hypotézu, že extrahovaný počet společných faktorů je dostatečný. Předpokládejme, že náhodný vektor pozorovaných veličin $\mathbf{X} = (X_1, \ldots, X_p)'$ má p-rozměrné normální rozdělení a platí ortogonální faktorový model (4.1). Test hypotézy, že počet společných faktorů m ve faktorovém modelu (4.1) je "dostatečný"je ekvivalentní s testováním hypotézy

$$H_0 : \mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$$

oproti alternativní hypotéze $H_1 :$ $\mathbf{\Sigma} =$ jiná libovolná pozitivně definitní matice.

Ze vztahu (4.8) a věty 16 je zřejmé, že $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ a $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_n = [(n-1)/n]\mathbf{S}$ jsou maximálně věrohodné odhady střední hodnoty $\boldsymbol{\mu}$ a varianční matice $\boldsymbol{\Sigma}$, které maximalizují věrohodnostní funkci $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a platí

$$L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \frac{1}{(2\pi)^{\frac{np}{2}}} e^{\frac{-np}{2}} \frac{1}{|\hat{\boldsymbol{\Sigma}}|^{\frac{n}{2}}} = \frac{1}{(2\pi)^{\frac{np}{2}}} e^{\frac{-np}{2}} \frac{1}{|\hat{\mathbf{S}}_n|^{\frac{n}{2}}}, \tag{4.11}$$

což znamená, že $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ může být zapsáno ve tvaru

$$L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = c \times |\mathbf{S}_n|^{-\frac{n}{2}} e^{-\frac{np}{2}}, \tag{4.12}$$

kde c je vhodná konstanta. Pokud platí nulová hypotéza H_0, je $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$ a pro odhady $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$, $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$ ($\hat{\mathbf{L}}, \hat{\boldsymbol{\Psi}}$ jsou odhady matic $\mathbf{L}, \boldsymbol{\Psi}$ získané metodou maximální věrohodnosti) ze vztahu (4.8) odvodíme

$$\begin{aligned} L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= c \times |\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}^{-1} \left(\sum_{j=1}^{n} (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' \right) \right] \right\} \\ &= c \times |\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} n \mathrm{Tr}\left[\left(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}} \right)^{-1} \mathbf{S}_n \right] \right\}. \end{aligned} \tag{4.13}$$

Pro testování nulové hypotézy H_0 bude použita testovací statistika $-2\ln\boldsymbol{\Lambda}$, která vychází z teorie maximální věrohodnosti, kde

$$\boldsymbol{\Lambda} = \left(\frac{\max L(\boldsymbol{\Sigma}) \text{ za předpokladu } H_0}{\max L(\boldsymbol{\Sigma})} \right) \tag{4.14}$$

se nazývá věrohodnostní poměr. Pro velký rozsah výběru n má statistika $-2\ln\boldsymbol{\Lambda}$ asymptoticky χ^2 rozdělení.

Užitím vztahů (4.12), (4.13) a (4.14) odvodíme, že statistika

$$-2\ln\boldsymbol{\Lambda} = -2\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\mathbf{S}_n|} \right)^{-\frac{n}{2}} + n\left[\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}_n) - p \right]$$

má asymptoticky χ^2_ν rozdělení, kde počet stupňů volnosti $\nu = \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)] = \frac{1}{2}[(p-m)^2 - p - m]$. Podle věty 33 pro odhad $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$ získaný metodou maximální věrohodnosti platí $\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}_n) - p = 0$, z čehož odvodíme

$$-2\ln\boldsymbol{\Lambda} = n\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\mathbf{S}_n|} \right). \tag{4.15}$$

Bartlett ukázal, že χ^2 aproximaci rozdělení $-2\ln\boldsymbol{\Lambda}$ lze zpřesnit nahrazením konstanty n v (4.15) hodnotou $n - 1 - (2p + 4m + 5)/6$. S užitím tohoto poznatku hypotézu H_0 zamítáme na hladině významnosti α, pokud

$$(n - 1 - (2p + 4m + 5)/6)\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\mathbf{S}_n|} \right) \geq \chi^2_{[(p-m)^2-p-m]/2}(1-\alpha).$$

Pokud při odhadu parametrů faktorového modelu vycházíme z výběrové korelační matice $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, kde $\mathbf{D} = \mathrm{diag}(\mathbf{S})$, můžeme podle [9] ukázat, že platí

$$\frac{|\hat{\mathbf{\Sigma}}|}{|\mathbf{S}|} = \frac{|\mathbf{D}^{-1/2}||\hat{\mathbf{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}||\mathbf{D}^{-1/2}|}{|\mathbf{D}^{-1/2}||\mathbf{S}||\mathbf{D}^{-1/2}|} =$$

$$= \frac{|(\mathbf{D}^{-1/2}\hat{\mathbf{L}})(\mathbf{D}^{-1/2}\hat{\mathbf{L}})' + \mathbf{D}^{-1/2}\hat{\mathbf{\Psi}}\mathbf{D}^{-1/2}|}{|\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}|} = \frac{|\hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\mathbf{\Psi}}_z|}{|\mathbf{R}|},$$

a nulovou hypotézu H_0 zamítneme na hladině významnosti α, pokud testovací statistika

$$(n - 1 - (2p + 4m + 5)/6)\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\mathbf{R}|}\right) \tag{4.16}$$

má větší hodnotu než kvantil $\chi^2_{[(p-m)^2-p-m]/2}(1-\alpha)$.

4.7 Odhad společných faktorů regresní metodou

Opět uvažujme motivační příklad z úvodu této kapitoly. Kromě odhadů faktorových zátěží a komunalit jsou požadovány pro jednotlivá pozorování $\mathbf{X}_j$, $j = 1, \ldots, n$, také odhady společných faktorů (představujících společné emisní zdroje), které nazýváme faktorové skóre. Pro odhad hodnot faktorového skóre se užívá vážená metoda nejmenších čtverců, nebo regresní metoda, která je dále užita v této práci při analýze reálných dat a nyní bude popsána. Odhad j-tého faktorového skóre $\mathbf{f}_j$, dosaženého vektorem společných faktorů $\mathbf{F}_j$ pro j-té pozorování $\mathbf{X}_j$, $j = 1, \ldots, n$, budeme značit $\hat{\mathbf{f}}_j$.

Uvažujme ortogonální faktorový model (4.1) a předpokládejme, že známe matici faktorových zátěží $\mathbf{L}$ i matici specificit $\mathbf{\Psi}$. Jestliže náhodné vektory $\mathbf{F} = (F_1, F_2, \ldots, F_m)$ a $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)$ mají mnohorozměrné normální rozdělení a jejich střední hodnoty a varianční matice splňují předpoklady (4.2), (4.2), potom $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$ má p-rozměrné normální rozdělení $\mathrm{N}_p(\mathbf{0}, \mathbf{L}\mathbf{L}' + \mathbf{\Psi})$. Předpokládáme, že sdružené rozdělení marginálních vektorů $\mathbf{X} - \boldsymbol{\mu}$ a $\mathbf{F}$ je $\mathrm{N}_{m+p}(\mathbf{0}, \mathbf{\Sigma}^*)$, kde

$$\mathbf{\Sigma^*}_{(m+p)\times(m+p)} = \begin{bmatrix} \mathbf{\Sigma}_{p\times p} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi} & \mathbf{L}_{p\times m} \\ \mathbf{L}'_{m\times p} & \mathbf{I}_{m\times m}. \end{bmatrix}$$

Užitím věty (14) lze odvodit, že podmíněné rozdělení vektoru $\mathbf{F}$ za podmínky $\mathbf{X} = \mathbf{x}$ je mnohorozměrné normální s následující podmíněnou střední hodnotou a podmíněným rozptylem:

- $\mathrm{E}(\mathbf{F}|\mathbf{x}) = \mathbf{L}'\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{L}'(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})^{-1}(\mathbf{X} - \boldsymbol{\mu})$
- $\mathrm{Var}(\mathbf{F}|\mathbf{x}) = \mathbf{I} - \mathbf{L}'\mathbf{\Sigma}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{L}'(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})^{-1}\mathbf{L}$

Hodnoty $\mathbf{L}'(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})^{-1}$ představují koeficienty pro lineární regresi faktorů v závislosti na regresorech $X_1 - \mu_1, \ldots, X_p - \mu_p$. Odhady těchto koeficientů použijeme k sestrojení m-rozměrného vektoru faktorového skóre. Máme-li daný vektor některého pozorování $\mathbf{X}_j$, odhady matic $\hat{\mathbf{L}}$ a $\hat{\mathbf{\Psi}}$ získané metodou maximální věrohodnosti, potom j-tý vektor faktorového skóre je dán:

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{X}_j - \bar{\mathbf{X}}) = \hat{\mathbf{L}}'(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})^{-1}(\mathbf{X}_j - \bar{\mathbf{X}}), \quad j = 1, \ldots, n.$$

Abychom zamezili chybám vlivem nesprávné volby počtu společných faktorů, použijeme k odhadu faktorového skóre místo matic $\hat{\mathbf{L}}$ a $\hat{\mathbf{\Psi}}$ výběrovou varianční matici $\mathbf{S}$, čímž dostaneme:

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'\mathbf{S}^{-1}(\mathbf{X}_j - \bar{\mathbf{X}}), \quad j = 1, \ldots, n.$$

Jestliže při odhadu faktorového skóre vycházíme z výběrové korelační matice $\mathbf{R}$, platí:

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}_z'\mathbf{R}^{-1}\mathbf{Z}_j, \quad j = 1, \ldots, n,$$

kde $\mathbf{Z}_j = \mathbf{D}^{-1/2}(\mathbf{X}_j - \bar{\mathbf{X}})$. Pokud odhad vychází z rotované matice faktorových zátěží $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$, potom je získáno rotované faktorové skóre $\hat{\mathbf{f}}_j^{\,*}$, splňující:

$$\hat{\mathbf{f}}_j^{\,*} = \mathbf{T}'\hat{\mathbf{f}}_j, \quad j = 1, \ldots, n.$$

4.8 Boxův test

Opět budeme uvažovat model faktorové analýzy, kde náhodné veličiny $X_1, X_2, \ldots, X_p$, představují koncentrace kovů a iontů v PM1 aerosolech v zimním a letním období. Pokud by bylo prokázáno, že náhodné veličiny $X_1, X_2, \ldots, X_p$ nejsou ročním obdobím ovlivněny, data z letního a zimního období by bylo možné analyzovat společně. Základním předpokladem pro testování hypotézy, že koncentrace analyzovaných kovů a iontů v aerosolových částicích nejsou ovlivněny ročním obdobím, je homogenita variančních matic vektorů náhodných veličin získaných z letního a zimního měření. Homogenitu variančních matic lze testovat užitím Boxova testu popsaného v tomto odstavci.

Obecně předpokládejme, že máme m datových souborů obsahujících nezávislá pozorování p-rozměrného náhodného vektoru a označme $\mathbf{\Sigma}_l$ varianční matici l-tého souboru, $l = 1, 2, \ldots, g$, a $\mathbf{\Sigma}$ společnou varianční matici za předpokladu, že se jednotlivé varianční matice mezi sebou neliší. Je testována nulová hypotéza

$$\mathrm{H}_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g = \mathbf{\Sigma},$$

oproti alternativní hypotéze H_1: alespoň dvě varianční matice nejsou shodné. Za předpokladu mnohorozměrného normálního rozdělení jednotlivých datových souborů lze test nulové hypotézy H_0 založit na věrohodnostním poměru

$$\Lambda = \prod_{l=1}^{g} \left(\frac{|\mathbf{S}_l|}{|\mathbf{S}_{pool}|} \right)^{(n_l-1)/2}, \tag{4.17}$$

kde n_l značí rozsah l-tého souboru, $\mathbf{S}_l$ značí výběrovou varianční matici pro l-tý soubor a $\mathbf{S}_{pool}$ představuje sdruženou výběrovou varianční matici danou váženým průměrem jednotlivých variančních matic. Tedy

$$\mathbf{S}_{pool} = \frac{1}{\sum_{l=1}^{g}(n_l-1)}[(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2 + \ldots + (n_g-1)\mathbf{S_g}].$$

Opět z teorie maximální věrohodnosti plyne, že statistika $-2\ln\Lambda$ má asymptoticky χ^2 rozdělení. Položíme-li $-2\ln\Lambda = M$, kde M je tzv. Boxova statistika, můžeme odvodit, že

$$M = -2\ln \prod_{l=1}^{g} \left(\frac{|\mathbf{S}_l|}{|\mathbf{S}_{pool}|} \right)^{(n_l-1)/2} = \left[\sum_{l=1}^{g}(n_l-1) \right] \ln|\mathbf{S}_{pool}| - \sum_{l=1}^{g} \left[(n_l-1)\ln|\mathbf{S}_l| \right].$$

Za platnosti nulové hypotézy se výběrové varianční matice $\mathbf{S}_1, \ldots, \mathbf{S}_g$ příliš neliší od sdružené varianční matice $\mathbf{S}_{pool}$, poměr determinantů v (4.17) je blízký 1 a Boxova statistika M má relativně malou hodnotu. V opačném případě je hodnota Boxovy statistiky M relativně velká.

Položme

$$u = \left[\sum_{l=1}^{g} \frac{1}{(n_l-1)} - \frac{1}{\sum_{l=1}^{g}(n_l-1)} \right] \left[\frac{2p^2+3p-1}{6(p+1)(g-1)} \right].$$

Zavedeme testovací statistiku

$$C = (1-u)M = (1-u) \left\{ \left[\sum_{l=1}^{g}(n_l-1) \right] \ln|\mathbf{S}_{pool}| - \sum_{l=1}^{g}[(n_l-1)\ln|\mathbf{S}_l|] \right\}, \tag{4.18}$$

která má χ^2_ν rozdělení, kde $\nu = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1)(g-1)$. Hypotéza H_0 je zamítnuta na hladině významnosti α, jestliže $C > \chi^2_\nu(1-\alpha)$.

Test je asymptotický a doporučuje se jej užívat v případě, že $n_l > 20$ pro $l = 1, 2, \ldots, g$ a $p < 5, g < 5$. Pokud tyto podmínky splněny nejsou, lze test nulové hypotézy založit na testovací statistice s F rozdělením [20], kterou nyní zavedeme. Položme

$$u_2 = \frac{(p-1)(p+2)}{6(g-1)} \left(\sum_{l=1}^{g} \frac{1}{(n_l-1)^2} - \frac{1}{(\sum_{l=1}^{g} n_l - g)^2} \right),$$

a dále zaveďme

$$\nu_2 = \frac{\nu+2}{|u_2-u^2|}, \quad a^+ = \frac{\nu}{1-u-\frac{\nu}{\nu_2}}, \quad a^- = \frac{\nu_2}{1-u+\frac{2}{\nu_2}}.$$

Jestliže $u_2 > u^2$, položíme testovací statistiku

$$F = \frac{M}{a^+}, \tag{4.19}$$

v opačném případě položíme

$$F = \frac{\nu_2 M}{\nu(a^- - M)}. \tag{4.20}$$

Protože testovací statistika F má asymptoticky F_{ν,ν_2} rozdělení, nulovou hypotézu H_0 zamítáme, pokud je $F > F_{\nu,\nu_2}(1-\alpha)$.

5 ROBUSTNÍ FAKTOROVÁ ANALÝZA

Opět uvažujme motivační příklad, kde $X_1, X_2, \ldots, X_p$ jsou náhodné veličiny představující koncentrace kovů a iontů v PM1 aerosolech. Označme $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ nezávislá náhodná pozorování p-rozměrného náhodného vektoru $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$. Užitím modelu faktorové analýzy je třeba určit hlavní emisní zdroje aerosolů. Protože klasická faktorová analýza může být ovlivněna odlehlými pozorováními, bude v této kapitole pojednáno o robustní faktorové analýze.

V prvním kroku klasické faktorové analýzy, uvedené v předchozí kapitole, je počítán výběrový průměr $\bar{\mathbf{X}}$ a výběrová varianční matice $\mathbf{S}$, popř. výběrová korelační matice $\mathbf{R}$, kterými je odhadnuta střední hodnota a varianční matice, popř. korelační matice pozorovaného náhodného vektoru $\mathbf{X}$. Dále jsou z matice $\mathbf{S}$ popř. z matice $\mathbf{R}$ extrahovány faktorové zátěže, které však mohou být ovlivněny odlehlými pozorováními, neboť již samotné odhady střední hodnoty a varianční matice jsou k odlehlým pozorováním velmi citlivé. V této kapitole je proto popsán princip robustní faktorové analýzy [14], která odlehlými pozorováními ovlivněna není.

Ortogonální faktorový model pro robustní faktorovou analýzu i předpoklady o vektorech $\mathbf{F}$ a ε jsou stejné jako pro klasickou faktorovou analýzu, postup výpočtu se liší odhadem střední hodnoty $\boldsymbol{\mu}$ a varianční matice $\boldsymbol{\Sigma}$ náhodného vektoru $\mathbf{X}$. Místo klasických odhadů $\bar{\mathbf{X}}$ a $\mathbf{S}$ jsou užity robustní odhady střední hodnoty a varianční matice vycházející z „minimum covariance determinant estimator“ (MCD odhad), který je nalezen užitím FAST-MCD algoritmu [15]. MCD algoritmus hledá podmnožinu h pozorování náhodného vektoru $\mathbf{X}$ takových, že jejich varianční matice má nejmenší determinant. Střední hodnota a varianční matice je potom odhadnuta výběrovým průměrem a výběrovou varianční maticí těchto h pozorování. Obvyklá volba počtu pozorování MCD podmnožiny je $h = 3n/4$.

Základní krok FAST-MCD algoritmu je založen na tom, že pokud máme libovolnou podmnožinu pozorování (aproximaci MCD), můžeme najít novou podmnožinu pozorování takovou, že její varianční matice bude mít stejný nebo nižší determinant než varianční matice původní podmnožiny. Zvolme $H_{old} \subset \{1, \ldots n\}$, kde $\mathrm{card}(H_{old}) = h$ a stanovme

$$\bar{\mathbf{X}}_{old} = \frac{1}{h} \sum_{i \in H_{old}} \mathbf{X}_i, \quad \mathbf{S}_{old} = \frac{1}{h} \sum_{i \in H_{old}} (\mathbf{X}_i - \bar{\mathbf{X}}_{old})(\mathbf{X}_i - \bar{\mathbf{X}}_{old})'.$$

Dále jsou stanoveny relativní vzdálenosti vektorů $\bar{\mathbf{X}}_{old}$ a $\mathbf{X}_j$ $(j = 1, 2, \ldots, n)$ v metrice dané maticí $\mathbf{S}_{old}$ podle vztahu

$$d_{old}(i) = \sqrt{(\mathbf{X}_i - \bar{\mathbf{X}}_{old})' \mathbf{S}_{old}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}_{old})} \quad \text{pro } j = 1, \ldots, n. \tag{5.1}$$

Pozorování jsou přečíslována tak, že platí $d_{old}(1) \leq d_{old}(2) \leq \ldots \leq d_{old}(n)$ a je položeno $H_{new} = \{1, 2, \ldots h\}$. Opakováním tohoto kroku je iterována posloupnost

$\det(\mathbf{S}_1) \geq \det(\mathbf{S}_2) \geq \det(\mathbf{S}_3) \geq \ldots$. V okamžiku, kdy je nalezeno m, pro které platí

$$\det(\mathbf{S}_m) = 0, \text{ nebo } \det(\mathbf{S}_m) = \det(\mathbf{S}_{m-1}),$$

jsou iterace zastaveny, protože dalším opakováním základního kroku už by nebyla nalezena podmnožina, jejíž varianční matice by měla menší determinant. Poznamenejme, že posloupnost $\det(\mathbf{S}_1) \geq \det(\mathbf{S}_2) \geq \det(\mathbf{S}_3) \geq \ldots$ je nezáporná a tedy konvergentní. Konvergence této posloupnosti plyne také z toho, že může být vybrán pouze konečný počet H podmnožin a tedy existuje takové m, pro které platí $\det(\mathbf{S}_m) = 0$, nebo $\det(\mathbf{S}_m) = \det(\mathbf{S}_{m-1})$. Dále poznamenejme, že ani jedna z podmínek $\det(\mathbf{S}_m) = 0$, $\det(\mathbf{S}_m) = \det(\mathbf{S}_{m-1})$ není postačující pro nalezení podmnožiny h pozorování s nejmenším determinantem varianční matice, jedná se pouze o nutné podmínky.

Základní myšlenka algoritmu tedy je zvolit několik počátečních H_{old} podmnožin a na každé z nich opakovat základní krok tak dlouho, dokud nebude nalezeno m splňující některou z podmínek pro ukončení iterace. Ze všech získaných podmnožin bude vybrána ta, jejíž varianční matice má nejmenší determinant.

Protože datový soubor, na který je v této práci FAST-MCD algoritmus aplikován není rozsáhlý, nebudeme dále zmiňovat další kroky algoritmu, které vedou ke zkrácení času výpočtu. Poznamenejme však, že obvykle stačí 2 - 3 iterace základního kroku k tomu, abychom poznali, zda je řešení dobré. Tento poznatek je velmi užitečný při aplikaci algoritmu na rozsáhlý datový soubor, kdy je každá počáteční podmnožina H_{old} iterována pouze $2\times$ a je ponecháno řekněme 10 nejlepších řešení. Tato řešení jsou dále iterována až do okamžiku, kdy jsou splněny podmínky pro ukončení iterace.

Robustní výběrový průměr $\bar{\mathbf{X}}^r$, robustní výběrová varianční matice $\mathbf{S}^r$, popř. robustní výběrová korelační matice $\mathbf{R}^r = (\boldsymbol{\Delta}^r)^{-1/2}\mathbf{S}^r(\boldsymbol{\Delta}^r)^{-1/2}$, kde $\boldsymbol{\Delta}^r = \operatorname{diag} \mathbf{S}^r$, získané z MCD podmnožiny jsou použity k odhadu parametrů faktorového modelu metodou hlavních komponent nebo metodou maximální věrohodnosti.

Protože robustní odhady střední hodnoty a varianční matice nejsou ovlivněny odlehlými pozorováními, rovněž ani na robustní odhady matice faktorových zátěží $\hat{\mathbf{L}}^r$, matice specificit $\hat{\boldsymbol{\Psi}}^r$ a společné faktory odlehlá pozorování nemají vliv.

6 SOFTWARE

Všechny dále uvedené výpočty byly provedeny užitím softwaru R verze 3.0.1. R je volně dostupný programovací jazyk specializovaný na statistické výpočty a analýzy. Vychází z jazyka S, který byl vyvinut Johnem Chambersem a jeho kolegy. R poskytuje všechny prostředky potřebné pro vizualizaci a analýzu dat, v přídavných balíčcích je implementováno velké množství pokročilých statistických metod a funkcí. Jazyk umožňuje psaní vlastních funkcí a skriptů a poskytuje prostředky pro tvorbu nejrůznějších diagramů, 2D i 3D grafů. R se stal velmi populárním jazykem, který má kromě statistiky využití ve např. ve finančnictví, průmyslu nebo ve vědecké sféře.

7 URČENÍ EMISNÍCH ZDROJŮ PM1 AEROSOLŮ

Atmosferický aerosol je směs pevných a kapalných částic suspendovaných ve vzduchu s velikostí částic pohybující se v rozmezí 1 nm – 100 µm. Aerosolové částice mají nepříznivý vliv na zemské klima, snižují dohlednost, podílejí se na produkci smogu a negativně ovlivňují lidské zdraví ([16]). Několik epidemiologických studií ([4], [5]) ukázalo statistickou vazbu mezi vysokou koncentrací aerosolů ve vzduchu a negativními vlivy na lidské zdraví.

V této práci je uvedena analýza chemického složení PM1 aerosolů, což jsou částice s aerodynamickým průměrem [1] menším než 1 µm. Částice PM1 aerosolů pronikají až do alveolární oblasti plic a způsobují dýchací a kardiovaskulární onemocnění.

Pro určení hlavních emisních zdrojů PM1 aerosolů v zimním a v letním období v Brně a ve Šlapanicích je užita klasická faktorová analýza. Protože klasická analýza může být ovlivněna odlehlými pozorováními, jsou emisní zdroje ve Šlapanicích určeny také užitím robustní faktorové analýzy.

K určení emisních zdrojů je nutné znát chemické složení aerosolových částic, které je velmi různorodé, protože částice obsahují organické i anorganické sloučeniny. Faktorová analýza uvedená v této diplomové práci vychází z koncentrací kovů a iontů v PM1 částicích.

Faktorová analýza chemického složení aerosolů uvedená v této kapitole vychází z modelů, vztahů, poznatků a statistik uvedených v kapitolách 4 a 5.

7.1 Data

Všechna data, která jsou v této diplomové práci analyzována, byla změřena Ústavem analytické chemie AV ČR, v.v.i, Veveří 97 Brno, a poskytnuta pro účely této práce. PM1 aerosoly byly vzorkovány na odběrových lokalitách umístěných v Brně a ve Šlapanicích. Brno s 370 000 obyvateli je druhým největším městem České republiky, zatímco Šlapanice představují malé město s 6000 obyvateli. Odběrová lokalita ve Šlapanicích byla umístěna na zahradě rodinného domu, v Brně se aerosoly vzorkovaly na balkoně v prvním poschodí Ústavu analytické chemie na ulici Veveří. Vzorkování v obou lokalitách neprobíhalo ve stejných termínech.

Mezi hlavní znečišťovatele ovzduší ve Šlapanicích patří cihelna, místní drobné provozovny, doprava a spalování v domácnostech. Méně než 6 km od města se nachází cementárna v Mokré, dálnice, městská spalovna v Brně-Židenicích a letiště v

[1] aerodynamický průměr je průměr ekvivalentní koule o jednotkové hustotě (1g/cm^3) a stejné rychlosti sedimentace jako má studovaná částice

Brně-Tuřanech. Dalším významným znečišťovatelem ovzduší je regionální a dálkový transport. V Brně patří mezi nejvýznamnější znečišťovatele ovzduší v místě odběrové lokality doprava na ulici Veveří a lokální topeniště. Regionální znečišťovatelé zahrnují dopravu, spalování v domácnostech a továrny v dalších částech města a okolních vesnicích. Stejně jako ve Šlapanicích je i v Brně významným znečišťovatelem ovzduší dálkový a regionální transport.

Na obou odběrových lokalitách byly PM1 aerosoly vzorkovány denně po dobu 24 hodin během jednoho týdne v létě a v zimě roku 2009 a 2010. PM1 částice byly analyzovány na přítomnost kovů a iontů. Pro analýzu kovů byly aerosoly vzorkovány na nitrátcelulózové filtry (pórovitost 3 µm, průměr filtru 150 mm, Sartorius) užitím velkoobjemového vzorkovače (DHA-80, Digitel, 30 m^3/h), a pro analýzu iontů byly částice vzorkovány nízkoobjemovým vzorkovačem (NILU držák filtrů, 1 m^3/h) na Teflonové filtry (Zefluor,pórovitost 1 µm, průměr filtru 47 mm, PALL). Z nitrátcelulózových filtrů bylo analyzováno 14 kovů (Al, K, Ca, Fe, Mn, Zn, Cu, Cd, Ba, As, Pb, V, Ni, Sb) metodou indukčně vázané plazmy. Iontovou chromatografiií bylo z teflonových filtrů analyzováno dohromady 10 iontů (dusičnan, dusitan, síran, šťavelan, fluorid, chlorid, Na^+, K^+, NH_4^+, Ca^{2+}). Dohromady bylo ze všech kampaní získáno 52 filtrů (24 filtrů pro Brno a 28 filtrů pro Šlapanice) a z každého filtru bylo získáno 24 veličin.

V grafu na obrázku 7.1 jsou zobrazeny průměrné koncentrace PM1 aerosolů během jednotlivých kampaní. Je zřejmé, že kromě brněnské kampaně v zimě roku 2009 byly zimní koncentrace aerosolů větší než letní koncentrace. Je to způsobeno odlišnými meteorologickými podmínkami a také odlišným složením emisí v létě a zimě (v zimě dochází k většímu spalování dřeva a uhlí v rámci domácích topenišť).

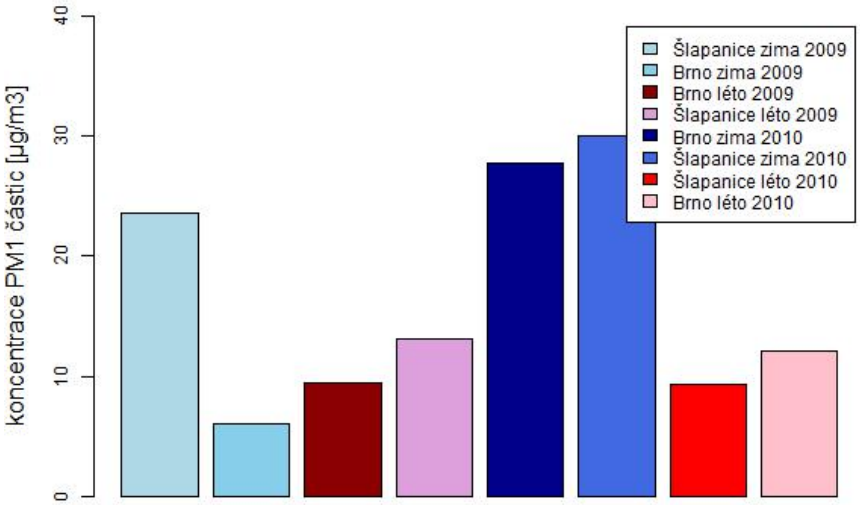
Na obrázcích 7.3 a 7.2 jsou zobrazeny průměrné koncentrace kovů v PM1 aerosolech. Je patrné, že v zimě měly největší zastoupení v PM1 aerosolech dva kovy, konkrétně Pb a K. V létě byly kromě vysokých koncentrací těchto dvou kovů nalezeny také vysoké koncentrace Ca a Zn.

Obrázky 7.4 a 7.5 zobrazují průměrné koncentrace iontů v PM1 aerosolech. Je zřetelné, že v zimním období byly zjištěny největší koncentrace dusičnanu, síranu a amonia, zatímco v létě byly zjištěny největší koncentrace pouze u síranu a amonia.

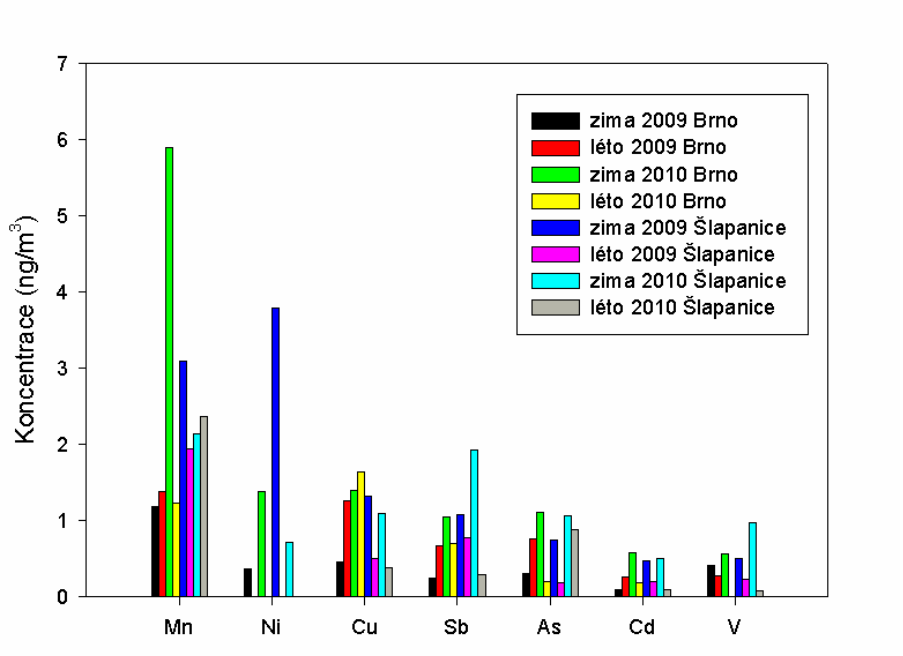
7.2 Klasická faktorová analýza

Ověření homogenity variančních matic

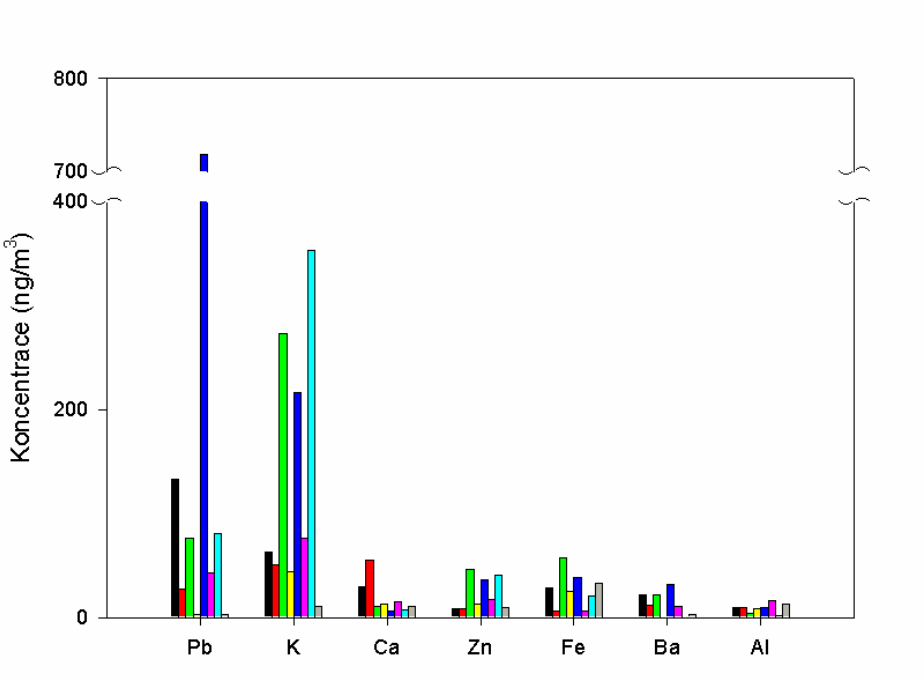
Pro ověření, zda je možné letní a zimní data získaná měřením ve Šlapanicích analyzovat společně, bylo třeba provést Boxův test popsaný v odstavci 4.8. Na hladině významnosti $\alpha = 0,05$ měla být testována hypotéza, že varianční matice $\mathbf{\Sigma}_{SL}$ a



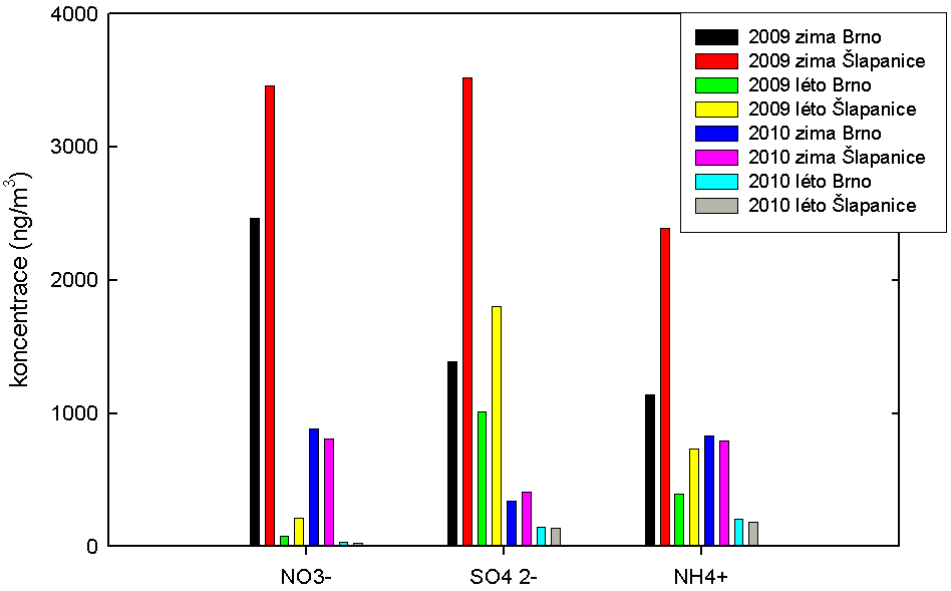
Obr. 7.1: Průměrné koncentrace PM1 aerosolů



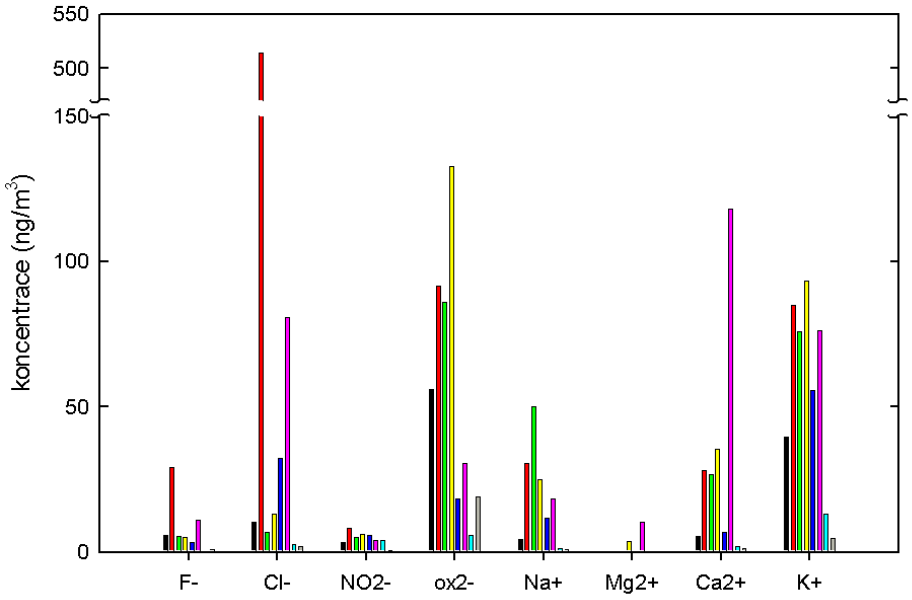
Obr. 7.2: Průměrné koncentrace minoritních kovů v PM1 aerosolech



Obr. 7.3: Průměrné koncentrace majoritních kovů v PM1 aerosolech



Obr. 7.4: Průměrné koncentrace majoritních iontů v PM1 aerosolech



Obr. 7.5: Průměrné koncentrace minoritních iontů v PM1 aerosolech

$\mathbf{\Sigma}_{SZ}$ datových souborů Šlapanice léto a Šlapanice zima jsou homogenní. Protože oba datové soubory obsahovaly 14 měření na 24 veličinách (představujících koncentrace 10 iontů a 14 kovů v PM1 aerosolech), determinanty výběrových variančních matic těchto souborů byly nulové a nebylo možné spočítat testovací statistiku (4.18) nutnou pro provedení Boxova testu. Z tohoto důvodu byla homogenita variančních matic testována na podsouborech obsahujících 13 proměnných, které byly postupně obměňovány.

První vytvořené podsoubory obsahovaly koncentrace následující podmnožiny kovů a iontů: F^-, Cl^-, SO_4^{2-}, ox^{2-}, NH_4^+, K^+, V, Sb, Mn, Fe, Zn, K, Pb. Pro testování nulové hypotézy $H_0 : \mathbf{\Sigma}_{SL1} = \mathbf{\Sigma}_{SZ1}$ o homogenitě variančních matic prvních dvou podsouborů byla spočítána testovací statistika (4.18) $C = 295,34.$ a kvantil $\chi^2_{91}(0,95) = 114.27$. Protože hodnota testovací statistiky C byla větší než $0,95$ kvantil χ^2 rozdělení, nulová hypotéza o homogenitě variančních matic $\mathbf{\Sigma}_{SL1}$ a $\mathbf{\Sigma}_{SZ1}$ byla zamítnuta. Testované podsoubory obsahovaly 13 proměnných a proto byla spočítána také testovací statistika (4.19) $F = 2,96$ a kvantil $F_{91,2119}(0,95) = 1,26$. Hypotéza o homogenitě variančních matic $\mathbf{\Sigma}_{SL1}$ a $\mathbf{\Sigma}_{SZ1}$ byla opět zamítnuta, protože hodnota testovací statistiky F byla větší než $0,95$ kvantil F rozdělení.

Druhé datové podsoubory obsahovaly koncentrace následujících 13 iontů a kovů: F^-, NO_2^-, NO_3^-, Na^+, NH_4^+, Cd, As, Sb, Cu, Ba, Zn, Ca, Pb. Pro testování nulové hypotézy $H_0 : \mathbf{\Sigma}_{SL2} = \mathbf{\Sigma}_{SZ2}$ o homogenitě variančních matic druhých podsouborů byly opět spočítány testovací statistiky (4.18) $C = 228,30 > \chi^2_{91}(0,95) = 114.27$ a

(4.19) $F = 2,29 > F_{91,2119}(0.95) = 1,26$ s hodnotami většími než příslušné kvantily, z čehož je zřejmé, že nulová hypotéza byla zamítnuta.

Třetí podsoubory byly vytvořeny následující kombinací iontů a kovů: Cl^-, NO_2^-, SO_4^{2-}, ox^{2-}, Ca^{2+}, K^+, Cd, As, Ni, Mn, Ba, Fe, Zn. Po spočítání testovacích statistik (4.18) $C = 431,52 > \chi^2_{91}(0,95) = 114.27$ a (4.19) $F = 4,32 > F_{91,2119}(0.95) = 1,26$ byla nulová hypotéza $H_0 : \mathbf{\Sigma}_{SL3} = \mathbf{\Sigma}_{SZ3}$ o homogenitě variančních matic třetích podsouborů zamítnuta.

Protože byly zamítnuty hypotézy o homogenitě variančních matic jednotlivých podsouborů Šlapanice léto a Šlapanice zima, je velice pravděpodobné, že varianční matice originálních datových souborů Šlapanice léto a Šlapanice zima nejsou homogenní.

Podobně byla testována a zamítnuta nulová hypotéza $H_0 : \mathbf{\Sigma}_{BLi} = \mathbf{\Sigma}_{BZi}$, $i = 1, 2, 3$ o homogenitě variančních matic i-tých podsouborů Brno léto a Brno zima. Je tedy pravděpodobné, že varianční matice originálních souborů Brno léto a Brno zima jsou nehomogenní.

Protože homogenita variančních matic je základním předpokladem pro vytvoření sdruženého náhodného výběru z letních a zimních hodnot, bude nutné oba soubory letních a zimních hodnot zpracovávat odděleně.

Je zřejmé, že pozorované koncentrace kovů a iontů v PM1 aerosolech záleží na ročním období a datové soubory z letního a zimního období je třeba analyzovat individuálně.

7.2.1 Analýza pro zimní období ve Šlapanicích

Nyní bude následovat podrobná analýza datového souboru Šlapanice zima, jehož příslušná výběrová korelační matice bude dále značena $\mathbf{R}$.

Ověření korelační struktury dat

Pro ověření, že data získaná analyzováním koncentrací kovů a iontů v PM1 aerosolech jsou korelovaná a vhodná k aplikaci faktorové analýzy, byla na hladině významnosti $\alpha = 0,05$ užitím Bartlettova testu (popsaného v odstavci 4.5) testována nulová hypotéza $H_0 : \mathbf{R} = \mathbf{I}$ o shodnosti výběrové korelační matice s jednotkovou maticí. Protože datový soubor Šlapanice zima obsahoval při daném počtu 14 pozorování příliš velké množství proměnných, determinant matice $\mathbf{R}$ byl nulový a proto stejně jako při testování homogenity variančních matic byl Bartlettův test proveden na několika podsouborech obsahujících podmnožiny proměnných originálního datového souboru.

První podsoubor byl vytvořen následující kombinací 13 iontů a kovů: F^-, Cl^-, SO_4^{2-}, ox^{2-}, NH_4^+, K^+, V, Sb, Mn, Fe, Zn, K, Pb. Pro testování nulové hypotézy,

že výběrová korelační matice prvního podsouboru $\mathbf{R}_1 = \mathbf{I}$, byla spočítána testovací statistika (4.10) $B = 225,76$ a kvantil $\chi^2_{78}(0,95) = 99,62$. Protože hodnota testovací statistiky B byla větší než $0,95$ kvantil χ^2 rozdělení, nulová hypotéza byla zamítnuta.

Vybráním následující podmnožiny veličin: F^-, NO_2^-, NO_3^-, Na^+, NH_4^+, Cd, As, Sb, Cu, Ba, Zn, Ca, Pb byl vytvořen druhý podsoubor. Po spočítání testovací statistiky (4.10) $B = 190,23 > \chi^2_{78}(0,95) = 99,62$ byla hypotéza, že výběrová korelační matice druhého podsouboru $\mathbf{R}_2 = \mathbf{I}$, zamítnuta.

Třetí podsoubor byl vytvořen kombinací následujících iontů a kovů: Cl^-, NO_2^-, SO_4^{2-}, ox^{2-}, Ca^{2+}, K^+, Cd, As, Ni, Mn, Ba, Fe, Zn. Protože testovací statistika (4.10) $B = 221,34 > \chi^2_{78}(0,95) = 99,62$, nulová hypotéza $H_0 : \mathbf{R}_3 = \mathbf{I}$ o shodnosti výběrové korelační matice třetího podsouboru s jednotkovou maticí byla zamítnuta.

Protože byly zamítnuty hypotézy, že výběrové korelační matice i-tých podsouborů $\mathbf{R}_i = \mathbf{I}$ pro $i = 1, 2, 3$, je očekávané, že výběrová korelační matice originálního datového souboru $\mathbf{R} \neq \mathbf{I}$. Znamená to, že náhodné veličiny představující koncentrace kovů a iontů v aerosolech jsou navzájem korelované a je možné přejít k modelu faktorové analýzy.

Určení počtu extrahovaných faktorů

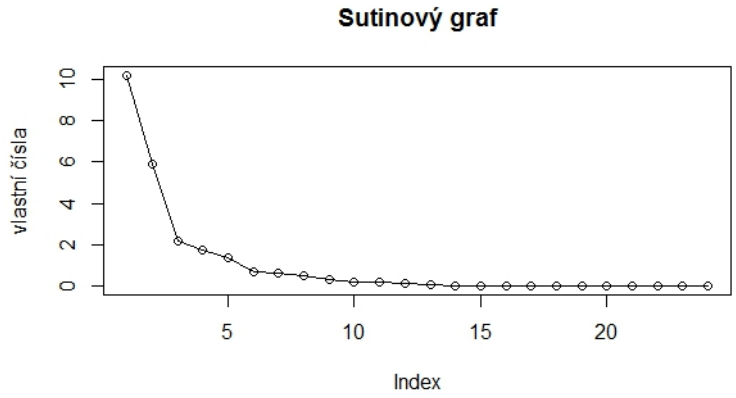
K určení správného počtu extrahovaných faktorů byl použit sutinový graf a Kaiserovo kritérium, jejichž princip byl popsán v kapitole 3 a v odstavci 4.2.

Z výběrové korelační matice $\mathbf{R}$ byla vypočtena vlastní čísla s následujícími hodnotami: 10,19; 5,87; 2,18; 1,74; 1,45; 0,71; 0,60; 0,51; 0,29; 0,19; 0,17; 0,12; 0,04, 0,..., 0. Protože přesně 5 vlastních čísel má větší hodnotu než 1, na základě Kaiserova kritéria by mělo být extrahováno 5 faktorů.

Z hodnot vlastních čísel byl vykreslen sutinový graf zobrazený na obrázku 7.6. Z grafu je patrné, že pokles nastává mezi prvním až třetím a mezi pátým a šestým vlastním číslem. Protože cílem faktorové analýzy je identifikace emisních zdrojů a je požadována interpretace více než dvou faktorů, bylo rozhodnuto, že bude extrahováno 5 společných faktorů, což je závěr odpovídající Kaiserovu kritériu.

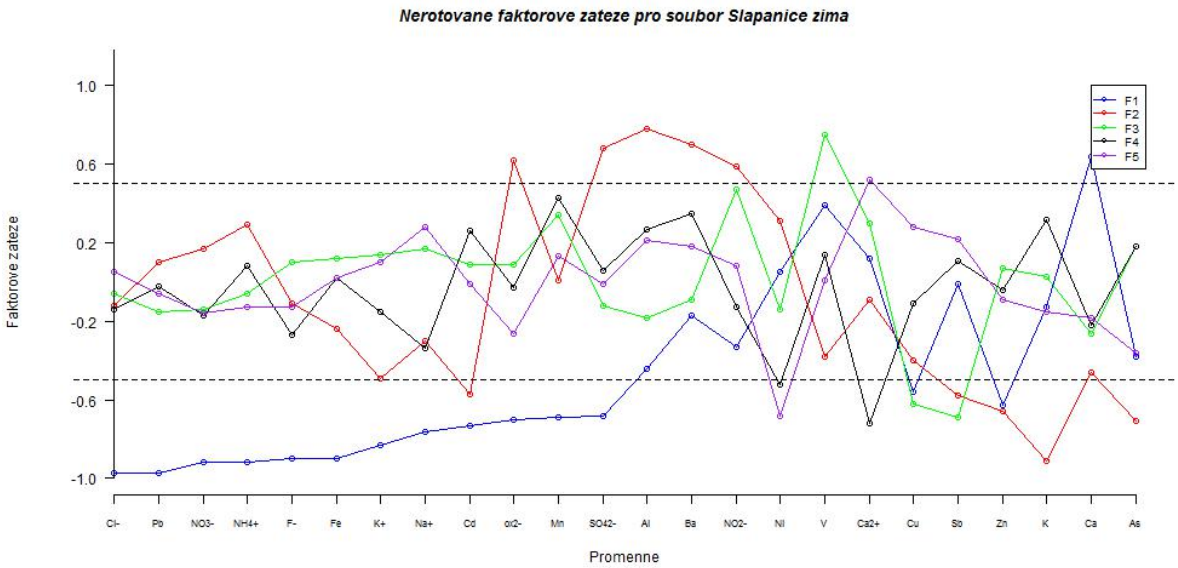
Odhad nerotovaných faktorových zátěží a komunalit

Pět společných faktorů bylo extrahováno užitím metody hlavních komponent, jejíž princip byl popsán v odstavci 4.2. Výsledné odhady nerotovaných faktorových zátěží a komunalit jsou zobrazeny v tabulce 7.1. První sloupec obsahuje studované ionty a kovy, následující sloupce s hlavičkami F1, F2 ... obsahují odhady faktorových zátěží a poslední sloupec s hlavičkou h_i^2 obsahuje odhady komunalit. Řádek s názvem "Prop. Var."značí poměr variability dat vysvětlené jednotlivými faktory, zatímco řádek s názvem "Cum. Var."obsahuje poměr variability vysvětlené více faktory najednou.



Obr. 7.6: Sutinový graf pro datový soubor Šlapanice zima

Pro větší přehled jsou odhady nerotovaných faktorových zátěží zobrazeny v grafu na obrázku 7.7. Z tabulky 7.1 a grafu 7.7 je patrné, že 8 z 24 veličin má $|\text{zátěž}| > 0,5$ od dvou společných faktorů, zatímco zbývajících 16 veličin má $|\text{zátěž}| > 0,5$ pouze od jednoho společného faktoru. Protože extrahované faktory nejsou jednoznačně určeny, bylo přistoupeno k faktorové rotaci, jejíž cílem je získat jednodušší strukturu a lépe interpretovatelné faktory.



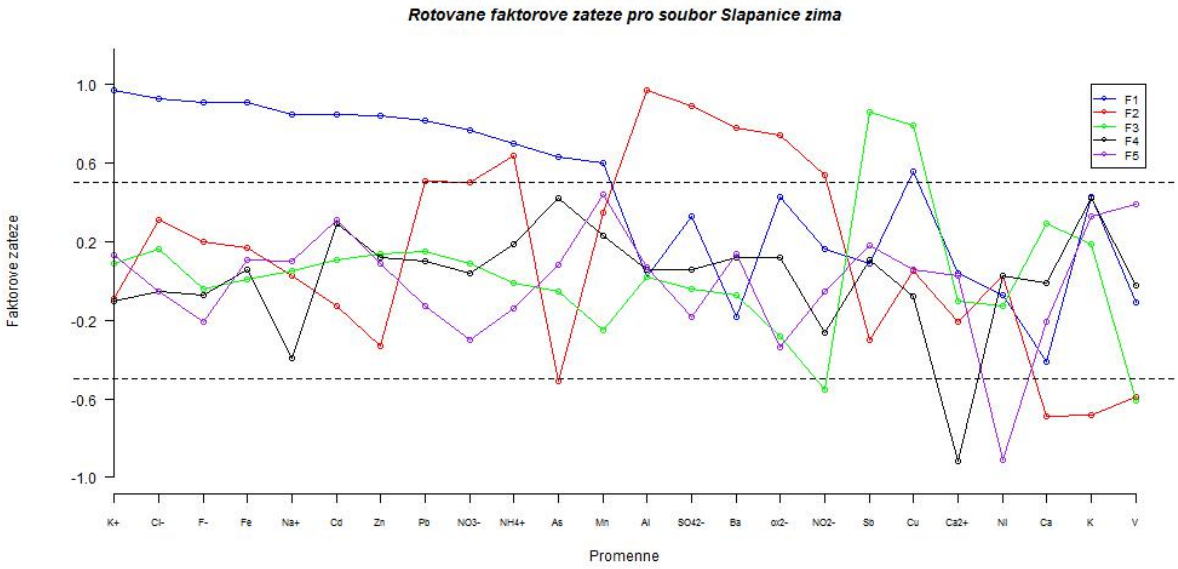
Obr. 7.7: Nerotované faktorové zátěže pro datový soubor Šlapanice zima

Tab. 7.1: Výsledky klasické faktorové analýzy bez faktorové rotace pro datový soubor Šlapanice zima

Proměnné	**F1**	**F2**	**F3**	**F4**	**F5**	$\mathbf{h}_i^2$
F^-	**-0,90**	-0,11	0,10	-0,27	-0,13	0,92
Cl^-	**-0,97**	-0,12	-0,06	-0,14	0,05	0,98
NO_2^-	-0,33	**0,59**	0,47	-0,13	0,08	0,70
NO_3^-	**-0,92**	0,17	-0,14	-0,17	-0,16	0,94
SO_4^{2-}	**-0,68**	**0,68**	-0,12	0,06	-0,01	0,94
ox^{2-}	**-0,70**	**0,62**	0,09	-0,03	-0,26	0,95
Na^+	**-0,76**	-0,30	0,17	-0,34	0,28	0,89
NH_4^+	**-0,92**	0,29	-0,06	0,08	-0,13	0,96
Ca^{2+}	0,12	-0,09	0,30	**-0,72**	**0,52**	0,90
K^+	**-0,83**	-0,49	0,14	-0,15	0,10	0,98
V	0,39	-0,38	**0,75**	0,14	0,01	0,88
Cd	**-0,73**	**-0,57**	0,09	0,26	-0,01	0,93
As	-0,38	**-0,71**	0,18	0,18	-0,36	0,85
Sb	-0,01	**-0,58**	**-0,69**	0,11	0,22	0,88
Cu	**-0,56**	-0,40	**-0,62**	-0,11	0,28	0,95
Ni	0,05	0,31	-0,14	**-0,52**	**-0,68**	0,84
Mn	**-0,69**	0,01	0,34	0,43	0,13	0,78
Al	-0,44	**0,78**	-0,18	0,27	0,21	0,95
Ba	-0,17	**0,70**	-0,09	0,35	0,18	0,68
Fe	**-0,90**	-0,24	0,12	0,02	0,02	0,88
Zn	**-0,63**	**-0,66**	0,07	-0,04	-0,09	0,85
Ca	**0,64**	-0,46	-0,26	-0,22	-0,18	0,77
K	-0,13	**-0,91**	0,03	0,32	-0,15	0,96
Pb	**-0,97**	0,10	-0,15	-0,02	-0,06	0,98
Prop. Var.	0,42	0,24	0,09	0,07	0,06	
Cum. Var.	0,42	0,67	0,76	0,83	0,89	

Odhad rotovaných faktorových zátěží a komunalit

Bylo vyzkoušeno několik ortogonálních rotací implementovaných v softwaru R, nejjednodušší strukturu se však podařilo získat užitím Varimax kritéria popsaného v odstavci 4.4. V tabulce 7.2 jsou zobrazeny odhady rotovaných faktorových zátěží a komunalit, poměry vysvětlené variability a emisní zdroje reprezentované jednotlivými faktory. Odhady rotovaných faktorových zátěží jsou vykresleny v grafu na obrázku 7.8, z kterého je patrné, že i po užití faktorové rotace má několik proměnných $|\text{zátěž}| > 0,5$ od dvou společných faktorů. Rotované faktory jsou však lépe interpretovatelné a v následujícím odstavci jsou identifikovány jednotlivé emisní zdroje.



Obr. 7.8: Rotované faktorové zátěže pro datový soubor Šlapanice zima

Interpretace společných faktorů

Podle údajů uvedených v tabulce 7.2 je zřejmé, že pět extrahovaných faktorů vysvětluje 89% variability dat.

První faktor, který vysvětluje 42% variability dat, má vysoké nebo střední zátěže F^-, Cl^-, NO_3^-, Na^+, NH_4^+, K^+, Cd, Fe, Zn, Pb a Cu a představuje směs tří emisních zdrojů. Prvním z nich jsou emise z městské spalovny (F^-, Cl^-, NO_3^-, Na^+, K^+, Zn, Pb a Cu), dále faktor reprezentuje emise ze spalování uhlí (NO_3^-, Na^+, K^+, Zn, Pb, Cd a Fe) a ze spalování biomasy (NO_3^-, NH_4^+ a K^+). Emise ze spalování biomasy a uhlí pravděpodobně pochází z vytápění domácností.

Druhý faktor s vysokými nebo středními zátěžemi SO_4^{2-}, ox^{2-}, NH_4^+, NO_3^-, Al a Ba reprezentuje sekundární aerosoly (SO_4^{2-}, ox^{2-}, NH_4^+ a NO_3^-), které vznikají

Tab. 7.2: Výsledky klasické faktorové analýzy po použití Varimax rotace pro datový soubor Šlapanice zima

Proměnné	**F1**	**F2**	**F3**	**F4**	**F5**	$\mathbf{h}_i^2$
F^-	**0,91**	0,20	-0,04	-0,07	-0,21	0,92
Cl^-	**0,93**	0,31	0,16	-0,05	-0,05	0,98
NO_2^-	0,16	**0,54**	**-0,55**	-0,26	-0,05	0,70
NO_3^-	**0,77**	**0,50**	0,09	0,04	-0,30	0,94
SO_4^{2-}	0,33	**0,89**	-0,04	0,06	-0,18	0,94
ox^{2-}	0,43	**0,74**	-0,28	0,12	-0,34	0,95
Na^+	**0,85**	0,03	0,05	-0,39	0,10	0,89
NH_4^+	**0,70**	**0,64**	-0,01	0,19	-0,14	0,96
Ca^{2+}	0,04	-0,21	-0,10	**-0,92**	0,03	0,90
K^+	**0,97**	-0,09	0,09	-0,10	0,13	0,98
V	-0,11	**-0,59**	**-0,61**	-0,02	0,39	0,88
Cd	**0,85**	-0,13	0,11	0,29	0,31	0,93
As	**0,63**	**-0,51**	-0,05	0,42	0,08	0,85
Sb	0,09	-0,30	**0,86**	0,11	0,18	0,88
Cu	**0,56**	0,05	**0,79**	-0,08	0,06	0,95
Ni	-0,07	0,03	-0,13	0,03	**-0,91**	0,84
Mn	**0,60**	0,35	-0,25	0,23	0,44	0,78
Al	0,02	**0,97**	0,02	0,06	0,07	0,95
Ba	-0,18	**0,78**	-0,07	0,12	0,14	0,68
Fe	**0,91**	0,17	0,01	0,06	0,11	0,88
Zn	**0,84**	-0,33	0,14	0,12	0,09	0,85
Ca	-0,41	**-0,69**	0,29	-0,01	-0,21	0,77
K	0,43	**-0,68**	0,19	0,42	0,33	0,96
Pb	**0,82**	**0,51**	0,15	0,10	-0,13	0,98
Prop. Var.	0,42	0,24	0,09	0,07	0,06	
Cum. Var.	0,42	0,67	0,76	0,83	0,89	
Zdroje	spalovna, spalování biomasy a uhlí	sekundární aerosoly, spalování uhlí, doprava	doprava	produkce cementu	průmysl	

sekundární reakcí primárních polutantů, jako jsou např. oxid siřičitý SO_2, oxidy dusíku NO_x a amoniak NH_3. Dále faktor představuje emise ze spalování uhlí (SO_4^{2-} a Al) a z dopravy (Ba).

Třetí faktor indikuje emise z dopravy spojené s vysokými zátěžemi Sb a Cu a čtvrtý faktor má vysokou zátěž Ca^{2+}, což pravděpodobně indikuje emise z produkce cementu. Poslední, pátý faktor má vysokou zátěž Ni a zřejmě představuje průmyslové emise.

Odhad faktorového skóre

Pro získání přehledu o odhadu společných faktorů bylo regresní metodou, jejíž princip je popsán v odstavci 4.7, vypočteno faktorové skóre zobrazené v tabulce 7.3.

Tab. 7.3: Odhady faktorového skóre pro datový soubor Šlapanice zima

Pozorování	**F1**	**F2**	**F3**	**F4**	**F5**
pozorování 1	1,28	0,38	0,37	0,46	-0,19
pozorování 2	2,61	0,54	0,59	-0,09	0,07
pozorování 3	0,63	1,17	0,16	0,36	0,06
pozorování 4	-0,98	0,78	-0,20	-0,39	0,07
pozorování 5	-1,00	2,01	-0,26	0,24	1,30
pozorování 6	-0,77	1,08	-0,34	-0,58	-0,52
pozorování 7	-0,38	0,04	-0,35	0,12	-3,16
pozorování 8	0,06	-1,18	-0,86	1,21	0,20
pozorování 9	-0,55	-0,66	0,08	0,55	0,16
pozorování 10	-0,21	-0,74	-1,20	0,54	0,65
pozorování 11	0,41	-0,92	-1,48	0,29	0,55
pozorování 12	0,20	-0,87	-0,29	-3,14	0,35
pozorování 13	-0,64	-0,76	2,44	0,20	0,29
pozorování 14	-0,68	-0,89	1,33	0,23	0,17

7.2.2 Analýza pro letní období ve Šlapanicích

Datové soubory Šlapanice léto, Brno léto a Brno zima byly analyzovány stejným způsobem jako soubor Šlapanice zima. Pro stručnost jsou v této práci uvedeny pouze výsledné odhady faktorových zátěží a komunalit získané užitím metody hlavních komponent a Varimax kritéria.

Odhady parametrů faktorového modelu pro datový soubor Šlapanice léto jsou zobrazeny v příloze v tabulce A.1. Dohromady bylo extrahováno 6 faktorů, které společně vysvětlují 92% celkové variability dat.

První faktor vysvětlující 33% variability má vysoké nebo střední zátěže ox^{2-}, Ca^{2+}, SO_4^{2-}, K^+ a NH_4^+, což značí emise ze spalování biomasy. Protože tento faktor má také vysoké zátěže Sb, Ba, F^- a Cl^-, představuje rovněž emise z dopravy (Sb, Ba) a z městské spalovny (F^- a Cl^-).

Druhý faktor má vysoké zátěže SO_4^{2-} a NH_4^+ a střední zátěž ox^{2-}, což indikuje sekundární aerosoly. Tento faktor představuje také emise z městské spalovny, jak indikují vysoké nebo střední zátěže F^-, Zn, K, a Pb. Vysoká zátěž Cd, představuje kromě již zmíněných zdrojů také průmyslové emise.

Třetí faktor s vysokými nebo středními zátěžemi Cu, Fe, Mn a Al reprezentuje emise z průmyslu a čtvrtý faktor s vysokými zatěžemi NO_3^- a Na^+ pravděpodobně představuje emise pocházející z dopravy (NO_3^-) a ze spalování biomasy (Na^+).

Pátý faktor má vysokou zátěž As, což indikuje emise ze spalování uhlí. Poslední, šestý faktor je špatně interpretovatelný, ale protože má vysokou zátěž V a střední zátěž Mn, zřejmě představuje průmyslové emise.

7.2.3 Analýza pro zimní období v Brně

Odhady parametrů faktorového modelu pro datový soubor Brno zima jsou zobrazeny v příloze v tabulce A.2. Dohromady bylo interpretováno 6 faktorů, které společně vysvětlují 88% variability dat.

První faktor vysvětlující 27% variability má vysoké nebo střední zátěže Cl^-, K^+, Mn, Sb, Cu, Zn, Fe a K a představuje emise pocházející z městské spalovny.

Druhý faktor s vysokými zátěžemi NO_3^-, ox^{2-}, NH_4^+ a Pb představuje směs dvou emisních zdrojů: emise ze spalování biomasy a sekundární aerosoly.

Třetí faktor má vysoké zátěže V a Ni, což jsou kovy, které můžeme společně nalézt v emisích ze spalování olejů.

Čtvrtý faktor s vysokými zátěžemi F^- a Ca^+ a pátý faktor s vysokými nebo středními zátěžemi NO_2^-, Na^+ a Cd indikují emise ze spalování uhlí. Poslední, šestý faktor představuje emise z dopravy spojené s vysokými zátěžemi Cu a Ba.

7.2.4 Analýza pro letní období v Brně

Odhady parametrů faktorového modelu pro datový soubor Brno léto jsou zobrazeny v příloze v tabulce A.3, z které je zřejmé, že 5 extrahovaných faktorů společně vysvětluje 91% variability dat.

První faktor vysvětlující 33% variability má vysoké zátěže SO_4^{2-}, ox^{2-}, Na^+, NH_4^+, Ca^{2+}, K^+, Ba, Ca a Pb a reprezentuje směs tří emisních zdrojů. Kombinace iontů SO_4^{2-}, ox^{2-}, Na^+, NH_4^+, Ca^{2+} a K^+ indikuje emise pocházející ze spalování biomasy, zatímco kovy Ba a Pb můžeme nalézt v emisích z dopravy. Tento faktor reprezentuje také sekundární aerosoly, což indikují ionty SO_4^{2-} a NH_4^+.

Druhý faktor s vysokými zátěžemi Cd, Cu, Sb, Mn, Al, Zn a K reprezentuje emise z dopravy.

Třetí faktor má vysoké zátěže F^- a Cl^-, což jsou ionty vyskytující se v emisích z městské spalovny. Čtvrtý faktor pravděpodobně reprezentuje průmyslové emise spojené s vysokou zátěží NO_3^- a poslední, pátý faktor s vysokou zátěží As představuje emise pocházející ze spalování uhlí.

7.2.5 Zhodnocení výsledků

Z dat získaných ve Šlapanicích v zimním období bylo užitím klasické faktorové analýzy extrahováno 5 společných faktorů, zatímco v letním období bylo interpretováno o 1 společný faktor více. V letním období je společnými faktory vysvětleno 89% variability dat, zimní faktory vysvětlují o 3% varability dat více.

V letním i zimním období byly nalezeny faktory reprezentující sekundární aerosoly, emise z městské spalovny, spalování uhlí, emise z dopravy, ze spalování biomasy a z průmyslu. Letní i zimní faktor vysvětlující největší část variability dat představuje emise pocházející ze spalování biomasy a městské spalovny. V zimě tento faktor reprezentuje také emise ze spalování uhlí, zatímco v létě faktor vysvětlující největší část variability představuje kromě již zmíněných zdrojů také emise z dopravy. V reprezentaci zimních faktorů byly dále nalezeny emise z produkce cementu, což je zdroj, který v letním období nalezen nebyl. Emise z produkce cementu zřejmě pochází z cementárny umístěné cca 6 km od Šlapanic.

Z datového souboru Brno zima bylo extrahováno 6 společných faktorů, o 1 společný faktor více než z letního brněnského datového souboru. Zimní faktory dohromady vysvětlují 88% variability dat, zatímco v letním období je společnými faktory vysvětleno 91% celkové variability.

V obou obdobích byly nalezeny faktory představující sekundární aerosoly, emise ze spalování biomasy, městské spalovny, spalování uhlí a z dopravy. Faktor, který v zimě vysvětluje největší část variability dat představuje emise pocházející z městské spalovny, zatímco letní faktor vysvětlující největší část variability indikuje sekundární aerosoly a emise ze spalování biomasy a z dopravy. V letním období byl také interpretován faktor představující průmyslové emise, zatímco v zimě byl nalezen faktor představující spalování olejů.

7.3 Robustní faktorová analýza

Aby mohly být užitím robustní faktorové analýzy určeny hlavní emisní zdroje PM1 aerosolů v letním a zimním období ve Šlapanicích, bylo nutné vytvořit datové soubory, které by obsahovaly méně proměnných než originální datové soubory. Maximální počet pozorování, která mohou být z datového souboru odstraněna užitím FAST-MCD algoritmu, je roven číslu $n-1/2(n+p+1)$, kde n představuje počet pozorování a p značí počet proměnných. Protože originální datové soubory obsahovaly 14 měření na 10 iontech a 14 kovech, nebylo možné odstranit žádné pozorování. Aby mohlo být z datových souborů Šlapanice léto a Šlapanice zima odstraněno alespoň jedno odlehlé pozorování, byly vytvořeny "robustní datové soubory" s 11 veličinami, z nichž šest bylo vytvořeno sečtením koncentrací kovů a iontů, které se v emisních zdrojích vyskytují společně. Konkrétně byly sečteny koncentrace iontů F^- a Cl^-, NO_3^- a SO_4^{2-} a dále byly sečteny koncentrace kovů V a Ni, Sb a Ba, Cd a As, Al a Fe. Zbývajících pět veličin představovalo koncentrace následujících iontů a kovů: NH_4^+, Cu, Zn, K a Pb. Veličiny tvořící "robustní datový soubor" jsou uvedeny v prvním sloupci tabulky 7.4.

7.3.1 Analýza pro zimní období ve Šlapanicích

Nyní bude uvedena podrobná analýza pro "robustní datový soubor" Šlapanice zima, jehož výběrová korelační matice bude dále značena $\mathbf{R}$.

Ověření korelační struktury dat

Pro ověření, že veličiny tvořící robustní datový soubor jsou vzájemně korelované a na data může být aplikována faktorová analýza, byla užitím Bartlettova testu sféricity popsaného v odstavci 4.5 na hladině významnosti $\alpha = 0,05$ testována nulová hypotéza, že výběrová korelační matice $\mathbf{R} = \mathbf{I}$. Byla spočítána testovací statistika (4.10) $B = 172,78$ a kvantil $\chi^2_{55}(0,95) = 73,31$. Protože testovací statistika měla větší hodnotu než 0,95 kvantil χ^2 rozdělení, nulová hypotéza byla zamítnuta a bylo přistoupeno k hledání MCD podmnožiny.

Nalezení MCD podmnožiny

Pomocí funkce implementované v přídavném balíčku softwaru R byla nalezena podmnožina s minimálním determinantem varianční matice tvořená 13 pozorováními. Graf na obrázku 7.9 zobrazuje relativní vzdálenosti (5.1) všech 14 pozorování spočítaných na základě MCD podmnožiny. Z grafu vidíme, že největší vzdálenost má

druhé pozorování, které v MCD podmnožině obsaženo není. Výběrová korelační matice MCD podmnožiny, která je robustním odhadem korelační matice, bude dále značena $\mathbf{R}_r$.

Obr. 7.9: Graf relativních vzdáleností pro "robustní datový soubor" Šlapanice zima

Určení počtu extrahovaných faktorů

Z výběrové korelační matice MCD podmnožiny $\mathbf{R}_r$ byla spočítána následující vlastní čísla: 4,57; 3,28; 1,06; 0,96; 0,54; 0,30; 0,18; 0,09; 0,02; 0,01; 0,00. Protože tři vlastní čísla mají hodnotu větší než 1, podle Kaiserova kritéria by měly být extrahovány 3 společné faktory. Hodnota čtvrtého vlastního čísla je však velmi blízká 1 a je zřejmé, že faktory příslušné třetímu a čtvrtému vlastnímu číslu vysvětlují téměř stejné procento variability.

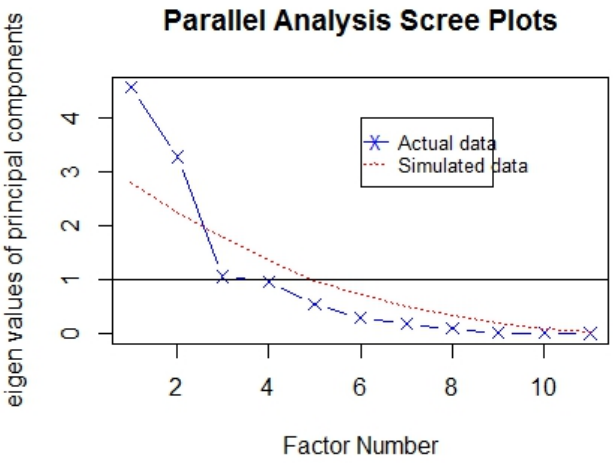
Dále byla provedena paralelní analýza popsaná v odstavci 4.2, jejíž výsledek je zobrazen v grafu na obrázku 7.3.1. Modrá lomená čára tvoří sutinový graf vlastních čísel matice $\mathbf{R}_r$, zatímco červená lomená čára spojuje průměry vlastních čísel získaných z náhodně generovaných dat. Je zřejmé, že pouze dvě vlastní čísla získaná z matice $\mathbf{R}_r$ jsou větší než průměry vlastních čísel simulovaných dat a podle kritéria paralelní analýzy by měly být extrahovány dva společné faktory.

V sutinovém grafu vlastních čísel matice $\mathbf{R}_r$ nastává pokles mezi druhým a třetím a čtvrtým a pátým vlastním číslem, z čehož je zřejmé, že by měly být extrahovány 2 nebo 4 společné faktory.

Protože pro identifikaci emisních zdrojů je požadována interpretace většího počtu společných faktorů, byl učiněn závěr, že budou extrahovány 4 společné faktory.

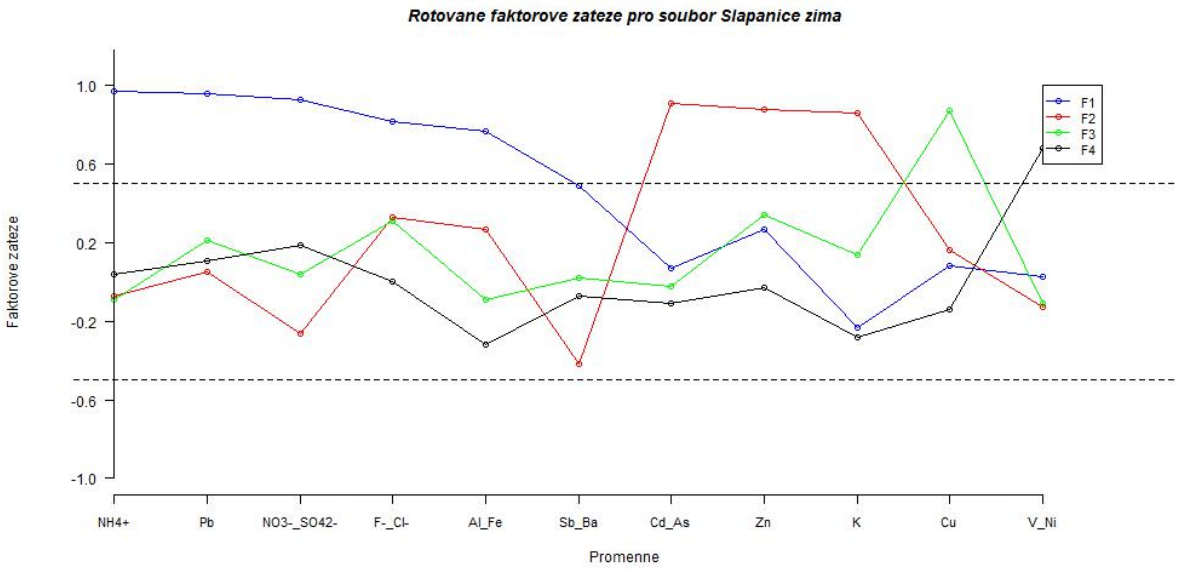
Odhad rotovaných faktorových zátěží a komunalit

Společné faktory byly extrahovány užitím metody hlavních komponent popsané v odstavci 4.2, a rotovány užitím Varimax kritéria (viz. kapitola 4.4). V tabulce 7.4



Obr. 7.10: Výsledky paralelná analýzy pro "robustní datový soubor" Šlapanice zima

jsou zobrazeny odhady rotovaných faktorových zátěží, odhady komunalit a poměry variability vysvětlené jednotlivými faktory. Odhady faktorových zátěží jsou také znázorněny v grafu na obrázku 7.11, z kterého je patrné, že se podařilo získat jednoduchou strukturu a jednotlivé veličiny mají zátěž $> 0,5$ právě od jednoho ze čtyř společných faktorů.



Obr. 7.11: Rotované faktorové zátěže pro "robustní soubor" Šlapanice zima

Tab. 7.4: Výsledky robustní faktorové analýzy pro datový soubor Šlapanice zima

Proměnné	**F1**	**F2**	**F3**	**F4**	$\mathbf{h}_i^2$
$F^-\_Cl^-$	**0,82**	0,33	0,31	0,00	0,88
NH_4^+	**0,97**	-0,07	-0,09	0,04	0,96
Cu	0,08	0,16	**0,87**	-0,14	0,81
Zn	0,27	**0,88**	0,34	-0,03	0,96
K	-0,23	**0,86**	0,14	-0,28	0,89
Pb	**0,96**	0,05	0,21	0,11	0,98
V_Ni	0,03	-0,13	-0,11	**0,68**	0,49
Sb_Ba	0,49	-0,42	0,02	-0,07	0,42
Cd_As	0,07	**0,91**	-0,02	-0,11	0,85
Al_Fe	**0,77**	0,27	-0,09	-0,32	0,78
$NO_3^-\_SO_4^{2-}$	**0,93**	-0,26	0,04	0,19	0,97
Prop. Var.	0,40	0,26	0,10	0,07	
Cum. Var.	0,40	0,66	0,76	0,83	
Zdroje	spalovna, sekundární aerosoly	spalování uhlí a dřeva	doprava	spalování olejů	

Interpretace společných faktorů

Z tabulky 7.4 je zřejmé, že 4 extrahované faktory dohromady vysvětlují 83% variability dat.

První faktor vysvětlující 40% variability dat má vysoké zátěže veličin $F^-\_Cl^-$, NH_4^+, Pb, Al_Fe a $NO_3^-\_SO_4^{2-}$ a představuje směs dvou emisních zdrojů. Ionty NH_4^+, NO_3^- a SO_4^{2-} indikují sekundární aerosoly, zatímco kombinace F^-, Cl^-, Pb, Al a Fe se vyskytuje v emisích z městské spalovny.

Druhý faktor s vysokými zátežemi Zn, K, Cd a As reprezentuje emise pocházející ze spalování dřeva a uhlí. Třetí faktor má vysokou zátěž Cu, což pravděpodobně indikuje emise z dopravy, a čtvrtý faktor s vysokou zátěží V a Ni reprezentuje emise ze spalování olejů.

Odhad faktorového skóre

Regresní metodou popsanou v odstavci 4.7, bylo spočítáno faktorové skóre, zobrazené v tabulce 7.5.

Tab. 7.5: Odhady faktorového skóre pro "robustní datový soubor" Šlapanice zima

Pozorování/Faktory	**F1**	**F2**	**F3**	**F4**
Pozorování 1	1,86	1,37	1,52	0,55
Pozorování 2	2,13	-0,10	-0,17	-1,30
Pozorování 3	0,00	-1,44	-0,30	-0,17
Pozorování 4	0,40	-1,48	-0,17	-0,93
Pozorování 5	0,37	-1,36	-0,55	0,28
Pozorování 6	0,05	-0,40	-0,24	3,24
Pozorování 7	-0,40	1,70	-1,42	0,73
Pozorování 8	-0,54	0,12	-0,99	-0,73
Pozorování 9	-0,49	0,50	-0,77	0,12
Pozorování 10	-0,31	1,10	-1,06	-1,20
Pozorování 11	-0,99	0,21	0,93	0,77
Pozorování 12	-1,12	-0,05	2,18	-0,76
Pozorování 13	-0,96	-0,18	1,05	-0,60

Testování hypotézy o dostatečném počtu extrahovaných faktorů

Na závěr byla užitím testu popsaného v odstavci 4.6 testována hypotéza, že 4 extrahované faktory vysvětlují variabilitu "robustního datového souboru" Šlapanice zima dostatečně. Byla spočítána testovací statistika (4.16) vycházející z robustní výběrové korelační matice $\mathbf{R}^r$:

$$(n-1-(2p+4m+5)/6)\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\mathbf{R}^r|}\right) = 34,06.$$

Protože hodnota 0,95 kvantilu χ^2 rozdělení $\chi^2_{17}(0,95) = 27,58 < 34,06$, měla by hypotéza o tom, že extrahovaný počet faktorů je dostatečný, být zamítnuta. Vzhledem k malému rozsahu analyzovaného datového souboru je však pravděpodobné, že test nefunguje dobře a vysoká hodnota testovací statistiky je způsobena nízkou hodnotou $\det(\mathbf{R}^r)$.

7.3.2 Analýza pro letní období ve Šlapanicích

"Robustní datový soubor" Šlapanice léto byl analyzován stejným způsobem jako datový soubor Šlapanice zima. Odhady faktorových zátěží a komunalit získané užitím robustní analýzy na data z letního období ve Šlapanicích, jsou zobrazeny v příloze v tabulce A.4.

Z uvedených výsledků je zřejmé, že 3 extrahované faktory společně vysvětlují 81% variability dat.

První faktor vysvětlující 49% variability dat má vysoké zátěže veličin F^-_Cl^-, NO_3^-_SO_4^{2-}, NH_4^+, Zn, Pb a K a představuje směs tří emisních zdrojů. Ionty NO_3^-, SO_4^{2-} a NH_4^+ se vyskytují v sekundárních aerosolech, kombinaci F^- a Cl^- s kovy Zn a Pb lze nalézt v emisích z městské spalovny a K indikuje emise pocházející ze spalování dřeva.

Druhý faktor s vysokými zátěžemi Cu, Zn, Fe a Al reprezentuje emise z dopravy a třetí faktor představuje emise ze spalování olejů, což značí vysoká zátěž V a Ni.

7.3.3 Zhodnocení výsledků

Z dat získaných v zimním období byly užitím robustní faktorové analýzy extrahovány 4 společné faktory, zatímco v letním období bylo extrahováno a interpretováno o 1 faktor méně. Zimní i letní faktory však vysvětlují téměř stejnou část variability dat. V obou obdobích byly interpretovány faktory reprezentující sekundární aerosoly, emise z městské spalovny, spalování dřeva, olejů a z dopravy. V zimním období byl dále nalezen faktor představující emise pocházející ze spalování uhlí, což je emisní zdroj, který v letním období nalezen nebyl. Letní i zimní faktor vysvětlující největší část variability dat představuje sekundární aerosoly a emise z městské spalovny, přičemž v létě tento faktor reprezentuje také emise pocházející ze spalování dřeva.

7.4 Zhodnocení klasické a robustní faktorové analýzy

Užitím klasické faktorové analýzy byly nalezeny hlavní emisní zdroje PM1 aerosolů v letním a zimním období v Brně a ve Šlapanicích, zatímco robustní faktorová analýza byla aplikována na data získaná v létě a v zimě ve Šlapanicích. Na data získaná z Brna robustní analýza aplikována nebyla, protože počet měření v této lokalitě byl příliš malý (problémy se vzorkováním v důsledku deštivého počasí) a nebylo možné vytvořit robustní datový soubor, aniž by došlo k degeneraci dat.

Aplikováním obou metod na zimní datový soubor ve Šlapanicích byly nalezeny faktory představující sekundární aerosoly, emise z městské spalovny, spalování uhlí a z dopravy. Dále byly klasickou metodou interpretovány faktory představující emise ze spalování biomasy, z produkce cementu a z průmyslu. Robustní metodou byly naopak nalezeny emise pocházející ze spalování dřeva a olejů, což jsou emisní zdroje, které klasickou analýzou interpretovány nebyly.

Z dat získaných v letním období ve Šlapanicích byly užitím klasické i robustní analýzy extrahovány faktory představující sekundární aerosoly, emise z městské spalovny a z dopravy. První klasický faktor představuje především emise ze spalování

biomasy, což je emisní zdroj, který robustní metodou nenalezen nebyl. Spalování biomasy je reprezentováno také čtvrtým klasickým faktorem. Další zdroj nalezený pouze klasickou metodou je spalování uhlí indikované pátým faktorem a průmyslové emise reprezentované druhým, třetím a šestým faktorem. Naopak robustní metodou byly v reprezentaci prvního faktoru nalezeny emise ze spalování dřeva a také byl nalezen faktor představující emise pocházející ze spalování olejů.

Výhodou robustní faktorové analýzy je, že vychází z robustního odhadu výběrové korelační matice, která není ovlivněna odlehlými pozorováními a tedy ani odhady faktorových zátěží a samotné faktory odlehlými pozorováními ovlivněny nejsou. Odstraněním odlehlých pozorování získáme lepší korelační strukturu dat. Z výsledků představených v této práci je zřejmé, že užitím klasické analýzy se podařilo vysvětlit větší část variability dat než užitím robustní analýzy. Je to patrně způsobeno tím, že klasická analýza byla aplikována na datový soubor obsahující jiné proměnné než datový soubor, z kterého vycházela robustní analýza.

Obecně nevýhodou robustní faktorové analýzy je delší výpočetní čas, protože FAST-MCD algoritmus je výpočetně náročný. Datové soubory analyzované v této práci však nebyly rozsáhlé a proto výpočetní čas algoritmu pro hledání MCD byl zanedbatelný. Další nevýhodou robustní analýzy je potřeba většího datového souboru, což může být problém, zvláště pokud jednotlivá měření jsou finančně nákladná.

Klasická ani robustní faktorová analýza nedokázala úplně oddělit emisní zdroje v reprezentaci jednotlivých faktorů. Některé robustní i klasické faktory reprezentují směs dvou až tří emisních zdrojů, což je pravděpodobně způsobeno tím, že emise z různých zdrojů jsou v ovzduší smíchané dohromady a také tím, že pro statistickou analýzu bylo k dispozici jen omezené množství vstupních dat.

8 PREDIKCE KONCENTRACÍ PM1 AEROSOLŮ

V této kapitole jsou představeny lineární regresní modely, které umožňují predikovat denní koncentrace PM1 aerosolů v zimním i letním období v Brně a ve Šlapanicích v roce 2009 a 2010. Koncentrace aerosolů jsou predikovány na základě naměřených koncentrací PM1 aerosolů z předchozího dne a hodnot meteorologických veličin. Analýza v této kapitole vychází z modelů, vztahů, poznatků a statistik uvedených v kapitole 2.

8.1 Data

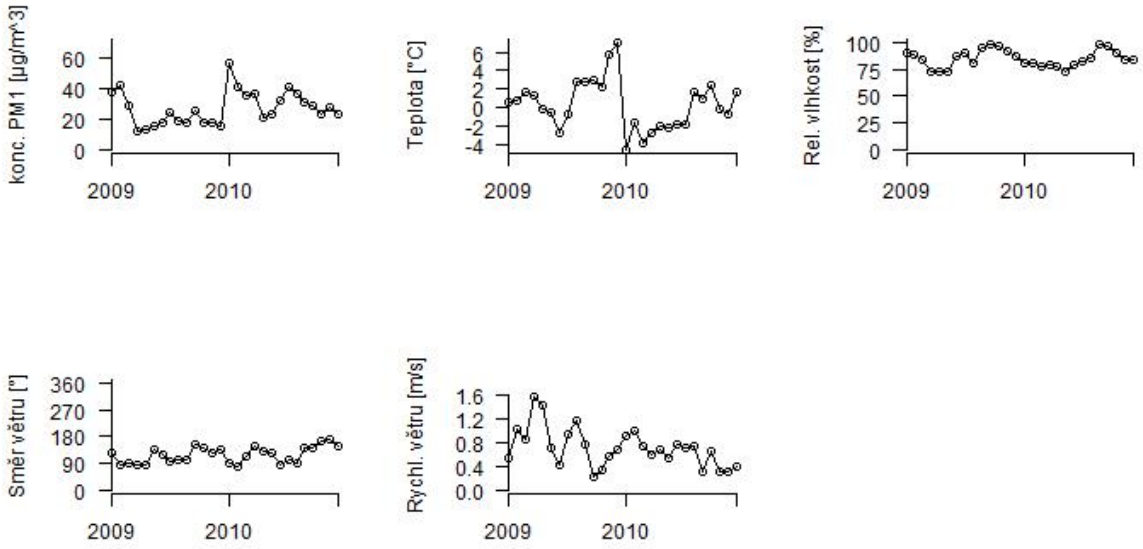
Regresní analýza je založena na datech, která byla získaná během experimentálních kampaní na stejných odběrových lokalitách jako data pro faktorovou analýzu (Brno, Šlapanice). Denní koncentrace PM1 aerosolů byly měřeny po dobu dvou týdnů během zimního a letního období roku 2009 a 2010. Současně s koncentrací aerosolů byla měřena rychlost větru V_t [ms^{-1}], směr větru D_t [°], teplota T_t [°C] a relativní vlhkost RH_t [%]. Meteorologická data použitá v analýze představují denní průměry 1-minutových měření a index t značí den. Obrázky 8.1 - 8.4 zobrazují pro danou lokalitu a dané období průběhy denních koncentrací PM1 aerosolů a průměrných denních hodnot teploty, relativní vlhkosti, rychlosti a směru větru během jednotlivých dnů všech kampaní.

8.2 Regresní model

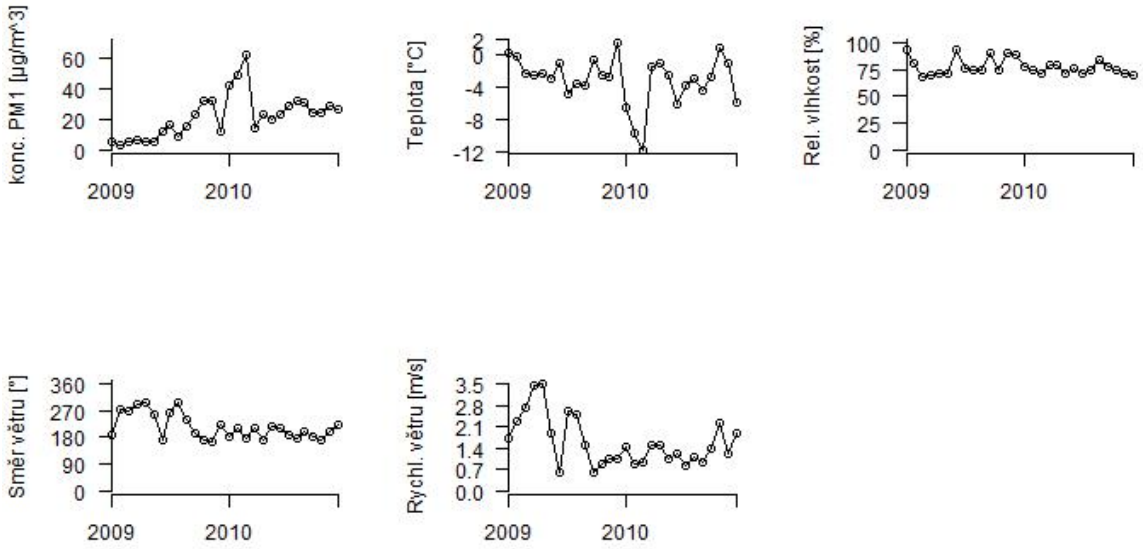
Pro predikci denních koncentrací aerosolů $PM1_t$ v dané lokalitě (Brno, Šlapanice) a daném období (zima, léto) je užit lineární regresní model zavedený v odstavci 2.1. Jako regresory jsou použity následující faktory:

- $PM1_{t-1}$: denní koncentrace PM1 aerosolů měřená dne $t-1$
- T_t: Průměr 24 1-hodinových měření teploty ze dne t.
- T_{t-1}: Průměr 24 1-hodinových měření teploty ze dne $t-1$.
- RH_t: Průměr 24 1-hodinových měření relativní vlhkosti ze dne t.
- RH_{t-1}: Průměr 24 1-hodinových měření relativní vlhkosti ze dne $t-1$.

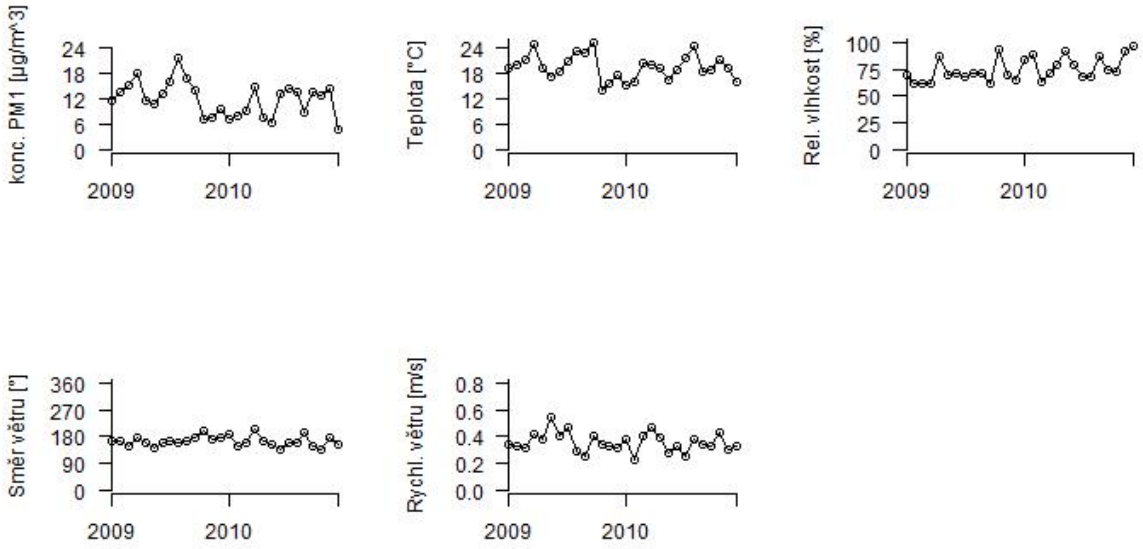
Analogicky jako v [8] byly kromě uvedených regresorů pomocí proměnných D_t - směr větru [°] (orientovaný úhel mezi vektorem směřujícím severně od odběrové lokality



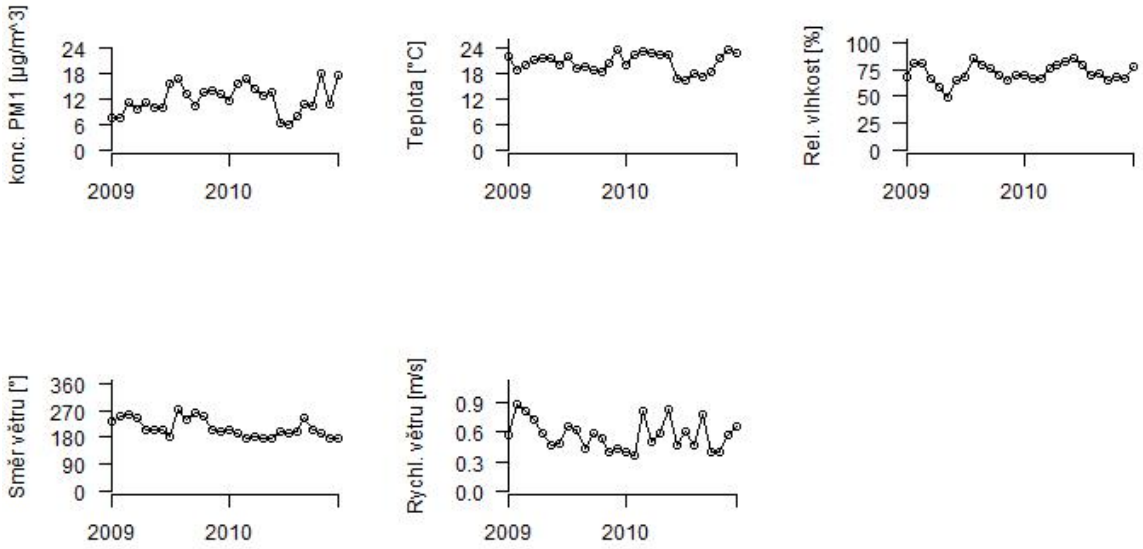
Obr. 8.1: Hodnoty PM1 koncentrací a meteorologických faktorů pro datový soubor Šlapanice zima



Obr. 8.2: Hodnoty PM1 koncentrací a meteorologických faktorů pro datový soubor Brno zima



Obr. 8.3: Hodnoty PM1 koncentrací a meteorologických faktorů pro datový soubor Šlapanice léto



Obr. 8.4: Hodnoty PM1 koncentrací a meteorologických faktorů pro datový soubor Brno léto

a vektorem směřujícím v pozorovaném směru větru) a V_t - rychlost větru [ms^{-1}] zavedeny další dva regresory:

- $V_t\sin(D_t)$: průměr 24 1-hodinových měření veličiny $V_t\sin(D_t)$.
- $V_t\cos(D_t)$: průměr 24 1-hodinových měření veličiny $V_t\cos(D_t)$.

Protože v několika výzkumech ([17], [8]) se ověřil předpoklad, že měření koncentrací PM1 aerosolů má logaritmicko-normální rozdělení, je uvažován následující model

$$\begin{aligned}\ln(PM1_t) = \beta_0 + \beta_1\ln(PM_{t-1}) + \beta_2 RH_t + \beta_3 RH_{t-1} + \beta_4 T_t + \\ +\beta_5 T_{t-1} + \beta_6 V_t\sin(D_t) + \beta_7 V_t\cos(D_t) + \varepsilon_t,\end{aligned} \tag{8.1}$$

kde $t = 2, \ldots, 14, 16, \ldots, n$, dále ε_t představuje náhodnou chybu a $\beta_0, \beta_1, \ldots, \beta_7$ jsou neznámé parametry, které je třeba odhadnout. Po provedení n pozorování náhodné veličiny $\ln(PM1_t)$, regresorů RH_t, T_t, $V_t\sin(D_t)$, $V_t\cos(D_t)$ a $n-2$ pozorování regresorů $\ln(PM_{t-1})$, RH_{t-1}, T_{t-1}, je zaveden náhodný vektor pozorovaných hodnot $\mathbf{Y} = (\ln(PM1_2), \ldots, \ln(PM1_{14}), \ln(PM1_{16}), \ldots, \ln(PM1_n))'$ a matice plánu $\mathbf{X}_{(n-2)\times 8}$, jejíž první sloupec je tvořen vektorem jedniček a ostatní sloupce jsou tvořeny naměřenými hodnotami jednotlivých regresorů pro $t = 2, \ldots, 14, 16, \ldots, n$. Model (8.1) potom může být zapsán ve tvaru uvedeném v odstavci 2.1, tedy

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\varepsilon} = (\varepsilon_2, \ldots, \varepsilon_{14}, \varepsilon_{16}, \ldots, \varepsilon_n)'$ je náhodný vektor chyb s předpoklady uvedenými v odstavci 2.1.

8.3 Odhad parametrů a ohodnocení modelu

Pro konkrétní lokalitu a období je třeba odhadnout neznámé parametry modelu (8.1) a postupným odstraňováním statisticky nevýznamných regresorů vytvořit jednodušší podmodel. Nyní bude následovat podrobné sestavování podmodelu na datech z datového souboru Brno léto.

Užitím softwaru R byly metodou nejmenších čtverců, která je popsána v odstavci 2.2, odhadnuty neznámé parametry modelu (8.1). Pro každý parametr byl spočítán odhad směrodatné chyby (2.6) a pomocí testu popsaného v odstavci 2.3 byla na hladině významnosti $\alpha = 0,05$ testována nulová hypotéza $\mathrm{H}_0 : \beta_i = 0$, že vliv i-tého regresoru na $\ln(PM1_t)$ je statisticky nevýznamný.

Odhady parametrů a směrodatných chyb modelu (8.1) jsou zobrazeny v tabulce 8.1, v které je pro každý parametr zobrazena také testovací statistika (2.8) T_i potřebná pro testování nulové hypotézy $\mathrm{H}_0 : \beta_i = 0$ a tzv. p-hodnota, která je součástí výstupu většiny statistických softwarů. Výhodou p-hodnoty je to, že dává podrobnější informaci o výsledku testu než kombinace hodnoty testovací statistiky a

příslušného kvantilu. Je to pravděpodobnost, s jakou testovací statistika nabývá extrémnějších hodnot než je pozorovaná hodnota statistiky. Ve výstupech statistických testů se p-hodnota interpretuje jako nejnižší možná hladina významnosti, na které můžeme nulovou hypotézu zamítnout. V tabulce 8.1 jsou pro větší přehled u některých p-hodnot zobrazeny hvězdičky, jejichž počet se liší v závislosti na hladině významnnosti, na které můžeme zamítnout nulovou hypotézu. Interpretace je následující: * - zamítnutí H_0 na $\alpha = 0,05$, ** - zamítnutí H_0 na $\alpha = 0,001$, *** - zamítnutí H_0 na $\alpha = 0,0001$.

Tab. 8.1: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat.** T_i	**p-hodnota**
konstanta β_0	0,950	0,870	1,093	0,289
$\ln(PM1_{t-1})$	0,324	0,247	1,310	0,207
RH_t	0,016	0,008	2,024	0,058
RH_{t-1}	-0,016	0,008	-2,017	0,059
$V_t\sin(D_t)$	-0,075	0,179	-0,420	0,680
$V_t\cos(D_t)$	0,023	0,222	0,102	0,920
T_t	0,114	0,034	3,385	0,003 **
T_{t-1}	-0,079	0,038	-2,093	0,051

V odstavci 2.3 bylo popsáno, že nulová hypotéza $H_0 : \beta_i = 0$ je zamítnuta na hladině významnosti α, pokud je realizace testovací statistiky $|T_i| > t_{n-k}(0,975)$, kde $t_{n-k}(0,975)$ je kvantil Studentova t rozdělení. Z tabulky 8.1 je zřejmé, že hodnotu kvantilu $t_{18}(0,975) = 2,101$ překračuje pouze realizace testovací statistiky $|T_i|$ regresoru T_t. Znamená to, že na $\ln(PM1_t)$ má statisticky významný vliv pouze T_t, vliv ostatních regresorů na denní koncentrace aerosolů nebyl prokázán jako statisticky významný. Tento závěr je potvrzen také tím, že p-hodnota odpovídající T_t splňuje nerovnost p-hodnota $= 0,003 < 0,05 = \alpha$, zatímco pro p-hodnoty ostatních regresorů platí p-hodnota $> 0,05 = \alpha$.

Statisticky nevýznamné regresory budou postupně z modelu (8.1) odstraňovány. Po odstranění regresoru $V_t\cos(D_t)$, který má nejnižší hodnotu testovací statistiky $|T_i|$, je uvažován regresní model

$$\ln(PM1_t) = \beta_0+\beta_1\ln(PM1_{t-1})+\beta_2 RH_t+\beta_3 RH_{t-1}+\beta_4 T_t+\beta_5 T_{t-1}+\beta_6 V_t\sin(D_t)+\varepsilon_t, \tag{8.2}$$

kde $t = 2,\ldots,14,16,\ldots,n$. Odhady neznámých parametrů, směrodatných chyb, testovacích statistik a p-hodnot modelu (8.2) jsou zobrazeny v tabulce 8.2.

Protože kvantil Studentova t rozdělení $t_{19}(0,975) = 2,093$, z tabulky 8.2 je zřejmé, že po odstranění regresoru $V_t\cos(D_t)$ byl potvrzen statisticky významný

Tab. 8.2: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat.** T_i	**p-hodnota**
konstanta β_0	0,955	0,845	1,131	0,272
$\ln(PM1_{t-1})$	0,340	0,188	1,809	0,086
RH_t	0,016	0,008	2,077	0,052
RH_{t-1}	-0,016	0,007	-2,218	0,039 *
$V_t\sin(D_t)$	-0,090	0,101	-0,895	0,382
T_t	0,113	0,032	3,538	0,002 **
T_{t-1}	-0,080	0,034	-2,366	0,029 *

vliv T_t na $\ln(PM1_{t-1})$ a navíc bylo prokázáno, regresory že RH_{t-1} a T_{t-1} mají na denní koncentraci aerosolů rovněž statisticky významný vliv. Odstraněním statisticky nevýznamného regresoru $V_t\sin(D_t)$ je model (8.2) zjednodušen na model

$$\ln(PM1_t) = \beta_0 + \beta_1\ln(PM1_{t-1}) + \beta_2 RH_t + \beta_3 RH_{t-1} + \beta_4 T_t + \beta_5 T_{t-1} + \varepsilon_t, \quad (8.3)$$

pro $t = 2, \ldots, 14, 16, \ldots, n$. Výsledky odhadu parametrů modelu (8.3) jsou zobrazeny v tabulce 8.3.

Tab. 8.3: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat. T**	**p-hodnota**
konstanta β_0	1,227	0,784	1,564	0,134
$\ln(PM1_{t-1})$	0,337	0,187	1,803	0,086
RH_t	0,017	0,008	2,252	0,036 *
RH_{t-1}	-0,017	0,007	-2,284	0,034 *
T_t	0,101	0,029	3,492	0,002 **
T_{t-1}	-0,082	0,034	-2,447	0,024 *

Z odhadů uvedených v tabulce 8.3 je zřejmé, že hodnotu kvantilu Studentova t rozdělení $t_{19}(0,975) = 2,086$ překračují testovací statistiky $|T_i|$ regresorů RH_t, RH_{t-1}, T_t a T_{t-1}, které mají po odstranění regresoru $V_t\sin(D_t)$ na $\ln(PM1_t)$ statisticky významný vliv. Po odstranění statisticky nevýznamné konstanty β_0 je uvažován model

$$\ln(PM1_1) = \beta_1\ln(PM1_{t-1}) + \beta_2 RH_t + \beta_3 RH_{t-1} + \beta_4 T_t + \beta_5 T_{t-1} + \varepsilon_t, \quad (8.4)$$

kde $t = 2, \ldots, 14, 16, \ldots, n$.

Tab. 8.4: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat. T**	**p-hodnota**
$\ln(PM1_{t-1})$	0.307	0.192	1.598	0.125
RH_t	0.018	0.008	2.292	0.032 *
RH_{t-1}	-0.010	0.006	-1.636	0.117
T_t	0.123	0.026	4.638	0.000 ***
T_{t-1}	-0.066	0.033	-1.999	0.059

Z tabulky 8.4, kde jsou zobrazeny odhady parametrů, směrodatných chyb, hodnoty testovacích statistik a p-hodnot modelu (8.4) je zřejmé, že po odstranění konstanty bylo prokázáno, že na $\ln(PM1_t)$ má statisticky významný pouze RH_t a T_t. Zamítnutí nulových hypotéz $H_0 : \beta_2 = 0, \beta_4 = 0$ je potvrzeno tím, že hodnoty testovacích statistik pro $i = 2, 4$ splňují $|T_i| > t_{21}(0,975) = 2,080$. Odstraněním statisticky nevýznamného regresoru RH_{t-1} je model (8.4) zjednodušen na model

$$\ln(PM1_t) = \beta_1 \ln(PM1_{t-1}) + \beta_2 RH_t + \beta_3 T_t + \beta_4 T_{t-1} + \varepsilon_t, \tag{8.5}$$

kde $t = 2, \ldots, 14, 16, \ldots, n$.

Odhady parametrů modelu (8.5) jsou zobrazeny v tabulce 8.5, z které je zřejmé,

Tab. 8.5: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat. T**	**p-hodnota**
$\ln(PM1_{t-1})$	0,254	0,196	1,294	0,209
RH_t	0,007	0,005	1,605	0,123
T_t	0,107	0,025	4,185	0,000 ***
T_{t-1}	-0,042	0,031	-1,375	0,183

že hodnotu kvantilu $t_{22}(0,975) = 2,074$ překračuje pouze testovací statistika regresoru T_t. Znamená to, že po odstranění regresoru RH_{t-1} má na $\ln(PM1_t)$ statisticky významný vliv pouze T_t . Ostatní regresory jsou považovány za statisticky nevýznamné a model (8.5) je dále zjednodušen odstraněním regresoru $\ln(PM1_{t-1})$. Je proto uvažován model

$$\ln(PM1_t) = \beta_1 RH_t + \beta_2 T_t + \beta_3 T_{t-1} + \varepsilon_t, \tag{8.6}$$

kde $t = 2, \ldots, 14, 16, \ldots, n$.

Protože kvantil Studentova t rozdělení $t_{23}(0,975) = 2,069$, z tabulky 8.6, kde jsou zobrazeny odhady parametrů modelu (8.6) je zřejmé, že i po odstranění $\ln(PM1_{t-1})$

Tab. 8.6: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat. T**	**p-hodnota**
RH_t	0,008	0,005	1,868	0,074
T_t	0,121	0,023	5,233	2,62e-05 ***
T_{t-1}	-0,031	0,030	-1,031	0,314

má na koncentraci aerosolů statisticky významný vliv pouze T_t. Vliv ostatních regresorů nebyl prokázán jako statisticky významný, z modelu (8.6) je dále odstraněn regresor T_{t-1} a je uvažován model

$$\ln(PM1_t) = \beta_1 RH_t + \beta_2 T_t + \varepsilon_t, \tag{8.7}$$

kde $t = 2, \ldots, 14, 16, \ldots, n$. Odhady parametrů tohoto modelu jsou zobrazeny v tabulce 8.7.

Tab. 8.7: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat. T**	**p-hodnota**
RH_t	0,006	0,004	1,717	0,098
T_t	0,097	0,013	7,361	8,11e-08 ***

Kvantil $t_{24}(0,975) = 2,064$ a z tabulky 8.7 je zřejmé, že vliv regresoru RH_t je statisticky nevýznamný a je třeba jej z modelu (8.7) odstranit, čímž je získán konečný model

$$\ln(PM1_t) = \beta_1 T_t + \varepsilon_t, \tag{8.8}$$

kde $t = 2, \ldots, 14, 16, \ldots, n$. V tabulce 8.8 je zobrazen odhad parametru, směrodatné chyby, testovací statistiky a p-hodnoty výsledného modelu (8.8).

Tab. 8.8: Průběžné odhady koeficientů regresního modelu pro Brno léto

Koeficienty	**Odhad**	**Směr. chyba**	**Testovací stat. T**	**p-hodnota**	R_D^2
T_t	0,119	0,002	50,54	<2e-16	0,990

Podle odhadu z tabulky 8.8 lze původní model (8.1) zjednodušit a zapsat jeho podmodel ve tvaru:

$$\ln(PM1_t) = 0,119 T_t, \tag{8.9}$$

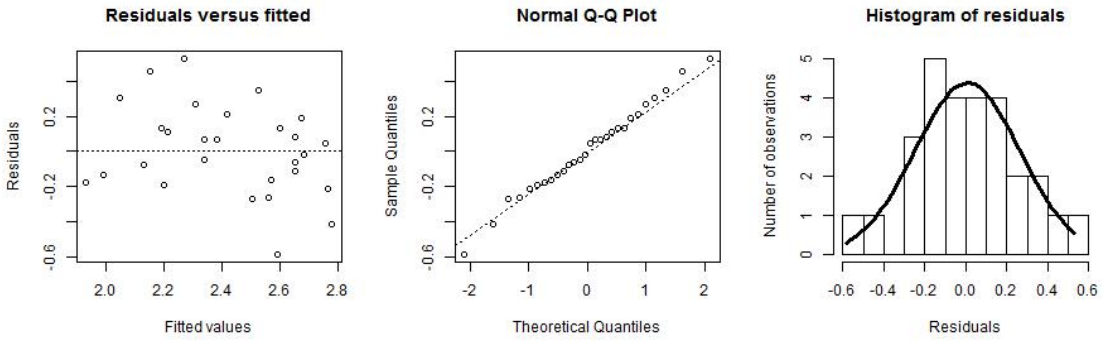
což znamená, že denní koncentrace PM1 aerosolů je ovlivněna pouze teplotou ve sledovaný den, jejíž růst má za následek rostoucí koncentraci aerosolů. Vliv ostatních meteorologických faktorů a koncentrace aerosolů předchozího dne nebyl prokázán

jako statisticky významný. Je to zřejmě důsledek toho, že směr větru D_t byl během měření prakticky konstantní a rychlost větru V_t se v průběhu období, kdy měření probíhalo, pohybovala v rozmezí 0,3-0,9 ms^{-1}, což jsou relativně malé hodnoty (viz. obr. 8.4). Můžeme se domnívat, že statistická nevýznamnost regresorů RH_t a RH_{t-1} je způsobena tím, že absolutní obsah vodní páry ve vzduchu je v letním období vyšší než v zimním období, a proto změna relativní vlhkosti vzduchu nemá v létě na změnu aerosolových koncentrací tak významný vliv jako v zimě.

Byl stanoven výběrový korelační koeficient $r_{\ln(PM1_t)T_t} = 0,652$ a pro posouzení lineární závislosti veličiny $\ln(PM1_t)$ na regresoru T_t byla testována nulová hypotéza, že korelační koeficient $\rho_{\ln(PM1_t)T_t} = 0$. Protože hodnota testovací statistiky (2.11) $T = 4,218 > t_{24}(0,975) = 2,064$, byla hypotéza o nulovosti korelačního koeficientu zamítnuta (p-hodnota=0,000), což potvrzuje lineární závislost $\ln(PM1_t)$ na T_t.

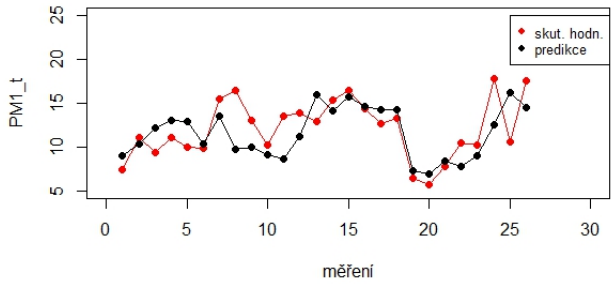
Pro ohodnocení modelu (8.9) byly použity grafické metody popsané v odstavci 2.4. Graf na obrázku 8.5 zobrazuje rezidua $e_t = \ln(PM1_t) - 0,119(T_t)$ vykreslená proti predikovaným hodnotám, Q-Q graf a histogram reziduí. Z prvního grafu je patrné, že rezidua jsou náhodně rozdělená kolem přímky $y = 0$, což znamená, že předpoklad homogenity rozptylu není narušen. Z Q-Q grafu a histogramu je možné usoudit, že rezidua nevykazují významné odchylky od normálního rozdělení a může tedy být učiněn závěr, že model (8.9) je dobře sestaven.

Na obrázku 8.6 jsou zobrazeny naměřené hodnoty koncentrací PM1 aerosolů a hodnoty predikované podle modelu (8.9).



Obr. 8.5: Závislost reziduí na predikovaných hodnotách, Q-Q graf a histogram reziduí (Brno léto)

Regresní modely umožňující predikovat koncentrace PM1 aerosolů v zimním období v Brně a ve Šlapanicích a v letním období ve Šlapanicích byly sestaveny stejným způsobem. V tabulce 8.9 jsou pro jednotlivé regresní modely uvedeny odhady parametrů a směrodatných chyb, výběrové korelační koeficienty r_{YX_i} popř. výběrové



Obr. 8.6: Predikované a skutečné hodnoty $PM1_t$ v letním období v Brně

koeficienty mnohonásobné korelace $r^2_{Y(X_1,\ldots,X_k)}$ a p-hodnoty stanovené pro rozhodnutí o zamítnutí hypotézy, že korelační koeficient popř. koeficient mnohonásobné korelace je nulový.

Tab. 8.9: Odhady parametrů včetně směrodatných chyb

	Šlap. léto		**Šlap. zima**		**Brno zima**	
	Odhad	**S. chyba**	**Odhad**	**S. chyba**	**Odhad**	**S. chyba**
konst.					2,093	0,337
$\ln(PM1_{t-1})$			0,463	0,107	0,256	0,110
RH_t			0,020	0,004		
$V_t\sin(D_t)$					0,409	0,077
T_t	0,124	0,002			-0,111	0,023
T_{t-1}			-0,088	0,022		
r_{YX_i}	0,783					
$r^2_{Y(X_1,\ldots,X_k)}$			0,710		0,870	
p-hodnota	0,000		0,000		0,000	

Pomocí odhadu v tabulce 8.9 lze podmodel pro letní období ve Šlapanicích napsat ve tvaru

$$\ln(PM1_t) = 0,124T_t.$$

Znamená to, že stejně jako v letním období v Brně jsou denní koncentrace aerosolů ovlivněny pouze teplotou ve sledovaný den, jejíž rostoucí hodnota má za následek růst koncentrace PM1 aerosolů. Z obrázku 8.3 je patrné, že stejně jako v letním období v Brně měla rychlost větru v průběhu měření relativně malé hodnoty (0,2-0,6 ms^{-1}) a směr větru byl konstantní, což zřejmě způsobilo statistickou nevýznamnost V_t a D_t na denní koncentraci aerosolů. Statisticky nevýznamný vliv regresorů RH_t

a RH_{t-1} je pravděpodobně opět způsoben tím, že absolutní obsah vodní páry ve vzduchu je v letním období vyšší než v zimním období.

Byl stanoven výběrový korelační koeficient $r_{YX_i} = 0,783$ a z p-hodnoty zobrazené v tabulce 8.9 je zřejmé zamítnutí hypotézy, že korelační koeficient $\rho_{YX_i} = 0$, což potvrzuje lineární $PM1_t$ na regresoru T_t.

Podmodel umožňující predikovat denní koncentrace aerosolů v zimním období ve Šlapanicích může být zapsán ve tvaru

$$\ln(PM1_t) = 0,463\ln(PM1_{t-1}) + 0,20RH_t - 0,088T_{t-1},$$

což znamená, že s rostoucí koncentrací aerosolů předchozího dne roste koncentrace aerosolů ve sledovaný den a podobně růst relativní vlhkosti ve sledovaný den má za následek růst koncentrace aerosolů. Rostoucí teplota předchozího dne naopak způsobuje pokles aerosolových koncentrací. Statisticky nevýznamný vliv regresorů $V_t\cos(D_t)$ a $V_t\sin(D_t)$ můžeme zřejmě opět přičítat konstantnímu směru větru a relativně nízkým hodnotám rychlosti větru (0,2-1,6 ms-1) během období, kdy měření probíhalo (viz. obr. 8.1).

Byl stanoven výběrový koeficient mnohonásobné korelace $r^2_{Y(X_1,\ldots,X_3)} = 0,710$ a z p-hodnoty zobrazené v tabulce 8.9 je zřejmé zamítnutí hypotézy o nulovosti koeficientu mnohonásobné korelace $\rho_{Y(X_1,\ldots,X_3)}$, což potvrzuje lineární závislost denních koncentrací aerosolů na regresorech $\ln(PM1_{t-1})$, RH_t a T_{t-1}, které vysvětlují 71% variability $\ln(PM1_t)$.

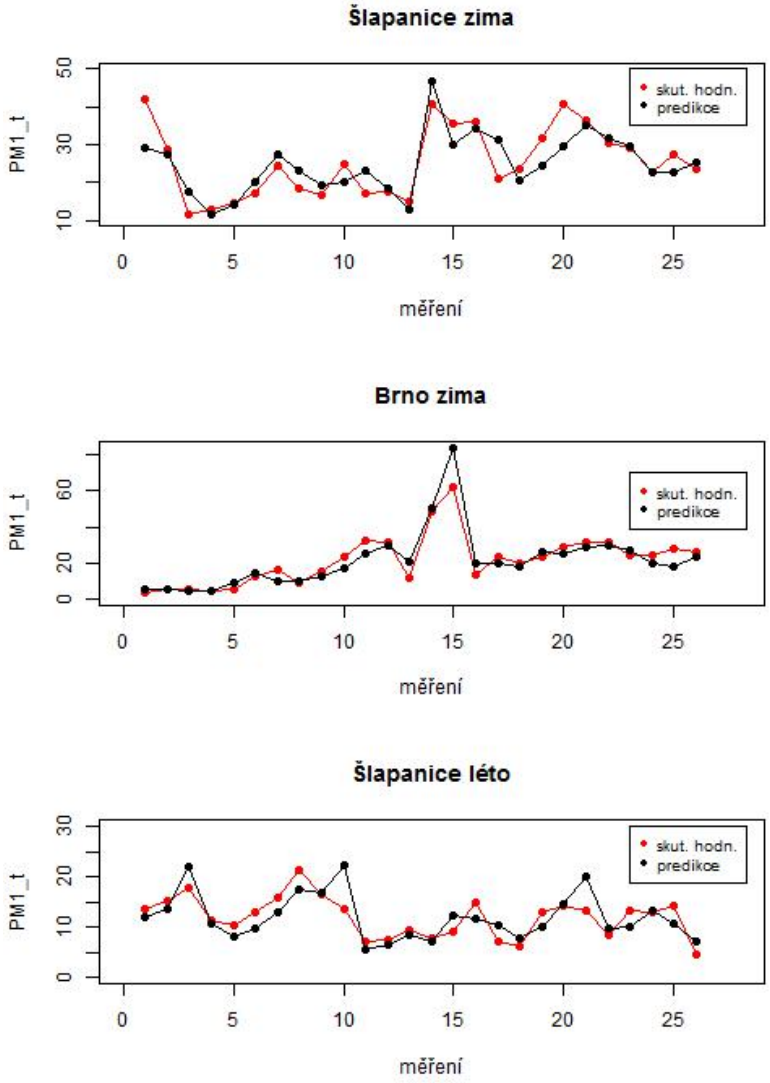
Podmodel, který umožňuje predikovat denní koncentrace aerosolů v zimním období v Brně lze pomocí odhadů v tabulce 8.9 napsat ve tvaru

$$\ln(PM1_t) = 2,093 + 0,256\ln(PM1_{t-1}) + 0,409V_t\sin(D_t) - 0,111T_t.$$

Znamená to, že růst aerosolových koncentrací předchozího dne má za následek růst koncentrací aerosolů ve sledovaný den, zatímco s rostoucí teplotou koncentrace sledovaných aerosolů klesá. Dále byla prokázána statistická významnost regresoru $V_t\sin(D_t)$, jehož růst má za následek růst aerosolových koncentrací.

V tabulce 8.9 je uvedena hodnota stanoveného výběrového koeficientu mnohonásobné korelace $r^2_{Y(X_1,\ldots,X_3)}$ a je zřejmé zamítnutí hypotézy, že koeficient mnohonásobné korelace $\rho_{Y(X_1,\ldots,X_3)} = 0$, čímž je potvrzena lineární závislost $PM1_t$ na regresorech $\ln(PM1_{t-1})$, $V_t\sin(D_t)$ a T_t vysvětlujících 87% variability $\ln(PM1_t)$.

Na obrázku 8.7 jsou zobrazené naměřené a predikované hodnoty koncentrací PM1 aerosolů v letním a zimním období ve Šlapanicích a v zimním období v Brně.



Obr. 8.7: Predikované a skutečné hodnoty $PM1_t$ v jednotlivých lokalitách a období

8.4 Shrnutí

V letním období v Brně a ve Šlapanicích jsou denní koncentrace PM1 aerosolů ovlivněny pouze teplotou ve sledovaný den, vliv koncentrace aerosolů předchozího dne ani vliv ostatních meteorologických faktorů není statisticky významný. Je to pravděpodobně způsobeno nízkými rychlostmi větru v uvažovaném krátkém obodobí, kdy měření probíhalo, a tím, že v letním období je ve vzduchu obsaženo vyšší množství vodní páry než v zimním období a proto nejsou aerosolové koncentrace ovlivněny změnou relativní vlhkosti tak jako v zimě.

Denní koncentrace aerosolů v zimním období ve Šlapanicích jsou ovlivněny teplotou a koncentrací aerosolů předchozího dne a relativní vlhkostí ve sledovaný den. Rychlost ani směr větru nemá na aerosolové koncentrace statisticky významný vliv, což je zřejmě opět způsobeno nízkými hodnotami rychlosti větru v uvažovaném zimním období.

Koncentrace PM1 aerosolů v zimním období v Brně jsou ovlivněny koncentrací aerosolů předchozího dne, teplotou ve sledovaný den a regresorem $V_t\sin(D_t)$, který můžeme interpretovat jako rychlost větru ve směru od západu na východ ve sledovaný den.

Zatímco v letním období v Brně i ve Šlapanicích má teplota na aerosolové koncentrace pozitivní vliv, v zimním období (v obou lokalitách) s rostoucí teplotou nastává pokles koncentrací aerosolů. Pravděpodobně je to způsobeno tím, že v zimním období s rostoucí teplotou klesá intenzita vytápění v domácnostech, což vede k poklesu emisí z jednoho z nejvýznamnějších zdrojů. Koncentrace aerosolů předchozí den, relativní vlhkost a regresor $V_t\sin(D_t)$ má na aerosolové koncentrace pozitivní vliv.

Při konstrukci modelů byla při výběru regresorů uplatňována metoda postupného odstraňování uvažovaných regresorů, aby byla ilustrována použitá teorie. Je ovšem možné používat i jiné metody pro vylučování regresorů z modelu. Jiný postup vylučování regresorů by mohl vést k jiným modelům, jejichž srovnání by mohlo být cílem dalších analýz.

9 ZÁVĚR

V první části diplomové práce je uvedena teorie popisující mnohorozměrné statistické metody potřebné k analýze dat získaných měřením koncentrací PM1 aerosolů. Je zde popsán model lineární regresní analýzy, metody hlavních komponent, klasické a robustní faktorové analýzy.

V praktické části diplomové práce jsou užitím klasické faktorové analýzy identifikovány hlavní emisní zdroje PM1 aerosolů v Brně a ve Šlapanicích a užitím robustní faktorové analýzy jsou nalezeny hlavní emisní zdroje ve Šlapanicích. Jsou srovnány výsledky získané z letních a zimních měření a dále jsou uvedeny hlavní přednosti a nevýhody robustní faktorové analýzy. Uvedená srovnání jsou poněkud poplatná malému rozsahu reálných dat. Dále jsou v praktické části sestaveny lineární regresní modely umožňující predikovat denní koncentrace aerosolů v letním a zimním období v Brně a ve Šlapanicích v závislosti na hodnotách meteorologických veličin a koncentrací aerosolů předchozího dne.

LITERATURA

[1] Anděl, J.: *Matematická statistika.* SNTL/Alfa Praha,1985

[2] Anděl, J.: *Základy matematické statistiky.* MATFYZPRESS Praha, 2011.

[3] Dobson, A.: *An Introduction to Generalized Linear Models.* Chapman and Hall, London, 1999.

[4] Dockery D.W., Pope C.A., XU X., Spengler J.D., Ware J.H., Fay M.E., Ferris Jr., B.G., Speizer F.E.:*An association between air pollution and mortality in six U.S. cities.* The New England Journal of Medicine (1993), 329, 1753-1759.

[5] Dockery, D.W., Stone, P.H.:*Cardiovascular risks from fine particulate air pollution.* New England Journal of Medicine (2007) 356, 511–513.

[6] Hebák, P., Hustopecký J., Malá I.: *Vícerozměrné statistické metody [2].* INFORMATORIUM, vydání první, Praha, 2005.

[7] Hebák, P., Hustopecký J., Pecáková I., Jarošová E.: *Vícerozměrné statistické metody [1].* INFORMATORIUM, vydání první, Praha, 2004.

[8] Hrdličková Z., Michálek J., Kolář M., Veselý V.: *Identification of factors affecting air pollution by dust aerosol PM_{10} in Brno City, Czech Republic.* Atmospheric Environment. Vol. 42 (2008) Number 37. p. 8661-8673

[9] Johnson, R.A., Wichern, D.W.: *Applied Multivariate Statistical Analysis.* New Jersey Prentice-Hall. 1992.

[10] Křůmal K., Mikuška P., Vojtěšek M. a Večeřa Z.: *Seasonal variations of monosaccharide anhydrides in PM1 and PM2.5 aerosol in urban areas.* Atmospheric Environment 44 (2010), p. 5148-5155

[11] Ledesma R. D., Valero-Mora P.:*Determining the Number of Factors to Retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis.* Practical Assesment, Research & Evaluation. Vol. 12 (2007) Number 2

[12] Mikušková, M., Michálek, J., Mikuška, P., Ivanov, T.: *Statistical analysis of chemical composition of PM1 aerosols.* AIP Conf. Proc. Vol. 1558 (2013), p. 1897-1900

[13] Mikušková, M., Mikuška, P., Michálek, J.: *Predikce koncentrací PM1 aerosolů ve Šlapanicích.* přijato do sborníku konference XXXII. mezinárodní kolokvium o řízení vzdělávacího procesu

[14] Pison G., Rousseeuw P.J., Filzmoser P. and Croux C.: *Robust factor analysis.* Journal of multivariate analysis 84 (2003) p. 145-172

[15] Rousseeuw P. J., Driessen K.V.: *A Fast Algorithm for the Minimum Covariance Determinant Estimator.* Technometrics 41 (1999) p. 212-223

[16] Seinfeld J.H., Pandis S.N.:*Atmospheric Chemistry and Physics. From Air Pollution to Climate Change.* Wiley & Sons, 1998, New York

[17] Stadlober E., Hübnerová Z., Michálek, J. and Kolář M.: *Forecasting of Daily PM10 Concentrations in Brno and Graz by Different Regression Approaches.* Austrian Journal of Statistics. Vol. 41 (2012) Number 4. p. 287-310

[18] Tobias, S., Carlson, J.E.: *Brief report: Bartlett´s test of sphericity and chance findindgs in factor analysis* Multivariate Behavioral Research, University of California at Davis, 1969, Volume 4, Issue 3, pages 375 - 374

[19] Überla, K.: *Faktorová analýza.* Alfa Bratislava,1976

[20] Zaiontz Ch.: *Real Statistics using Excel* [online]. poslední aktualizace 11. 11. 2013 [cit. 17. 2. 2014]. Dostupné z URL: <`www.real-statistics.com/multivariate-statistics/boxs-test-equality-covariance-matrices/boxs-test-basic-concepts/`>.

[21] Zvára, K.: *Regresní analýza.* Academia Praha, 1989

SEZNAM ZÁKLADNÍCH POUŽITÝCH SYMBOLŮ, VELIČIN A ZKRATEK

$\mathbb{N}$ množina přirozených čísel

$\mathbb{R}$ množina reálných čísel

$\Gamma(t)$ gama funkce v proměnné t

$\mathbf{A}_{m\times n} = (a_{ij})_{i=1,\dots m, j=1\dots n}$ matice s m řádky a n sloupci

$\mathbf{A}'$ matice transponovaná k matici $\mathbf{A}$

$\mathbf{A}^{-1}$ matice inverzní k matici $\mathbf{A}$

$\mathbf{A}^{-}$ pseudoinverzní matice k matici $\mathbf{A}$

$\det(\mathbf{A})$ determinant matice $\mathbf{A}$

$\mathbf{1}_n$ n-rozměrný vektor jedniček

$\mathbf{I}$ jednotková matice

$\mathbf{0}$ nulová matice

$h(\mathbf{A})$ hodnost matice $\mathbf{A}$

$\mathrm{Tr}(\mathbf{A})$ stopa matice $\mathbf{A}$

$\forall$ obecný kvantifikátor

$\blacksquare$ konec důkazu

$N(\mu, \sigma^2)$ normální rozdělení s parametry μ, σ^2

χ^2_n χ^2 rozdělení o n stupních volnosti

t_n Studentovo rozdělení o n stupních volnosti

F_{n_1,n_2} Fisher-Snedecorovo rozdělení o n_1, n_2 stupních volnosti

$\chi^2_n(\alpha)$ α kvantil χ^2 rozdělení o n stupních volnosti

$t_n(\alpha)$ α kvantil Studentova t rozdělení o n stupních volnosti

$F_{n_1,n_2}(\alpha)$ α kvantil Fisher-Snedecorova F rozdělení o n_1, n_2 stupních volnosti

$N_n(\boldsymbol{\mu}, \mathbf{V})$ normální n-rozměrné rozdělení s parametry $\boldsymbol{\mu}, \mathbf{V}$

X, Y, Z náhodné veličiny

$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ náhodné vektory

$\mathrm{E}(X)$ střední hodnota náhodné veličiny X

$\mathrm{Var}(X)$ rozptyl náhodné veličiny X

$f(x)$ hustota náhodné veličiny X

$\mathrm{Cov}(X, Y)$ kovariance náhodných veličin X, Y

ρ_{XY} korelační koeficient náhodných veličin X, Y

r_{XY} výběrový korelační koeficient náhodných veličin X, Y

$\mathrm{E}(\mathbf{X})$ střední hodnota náhodného vektoru $\mathbf{X}$

$\mathrm{Var}(\mathbf{X})$ rozptyl náhodného vektoru $\mathbf{X}$

$\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$ kovarianční matice náhodných vektorů $\mathbf{X}, \mathbf{Y}$

$\mathrm{Cor}(\mathbf{X}, \mathbf{Y})$ korelační matice náhodných vektorů $\mathbf{X}, \mathbf{Y}$

$\mathrm{Cor}(\mathbf{X})$ korelační matice náhodného vektoru $\mathbf{X}$

$f(\mathbf{x})$ hustota náhodného vektoru $\mathbf{X}$

$\bar{\mathbf{X}}$ výběrový průměr

$\mathbf{S}$ výběrová varianční matice

$\mathbf{R}$ výběrová korelační matice

$\rho_{Y,(x_1,\ldots,x_k)}$ koeficient mnohonásobné korelace náhodné veličiny Y a vektoru $(x_1, \ldots, x_k)$

$r_{Y,(x_1,\ldots,x_k)}$ výběrový koeficient mnohonásobné korelace náhodné veličiny Y a vektoru $(x_1, \ldots, x_k)$

R_D^2 koeficient determinace

$L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ věrohodnostní funkce s parametry $\boldsymbol{\mu}, \boldsymbol{\Sigma}$

A VÝSLEDKY FAKTOROVÉ ANALÝZY

Tab. A.1: Výsledky klasické faktorové analýzy pro datový soubor Šlapanice léto

Proměnné	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	$\mathbf{h}_i^2$
F^-	**0,73**	**0,56**	0,05	0,27	0,05	0,00	0,92
Cl^-	**0,86**	0,25	0,16	0,33	0,05	0,11	0,95
NO_2^-	**0,89**	0,30	0,11	0,22	0,09	0,07	0,95
NO_3^-	0,26	0,24	-0,23	**0,87**	0,02	-0,14	0,95
SO_4^{2-}	**0,61**	**0,71**	-0,08	0,21	0,13	-0,03	0,95
ox^{2-}	**0,77**	**0,54**	0,04	0,25	0,10	0,01	0,95
Na^+	0,48	0,11	0,04	**0,84**	0,17	0,10	0,99
NH_4^+	**0,58**	**0,71**	-0,05	0,25	0,03	0,08	0,91
Ca^{2+}	**0,80**	0,29	0,18	0,40	0,17	0,19	0,98
K^+	**0,61**	**0,59**	0,00	0,49	0,12	0,06	0,97
V	0,10	-0,06	0,04	-0,06	-0,12	**0,95**	0,93
Cd	0,19	**0,84**	0,01	0,15	-0,31	-0,17	0,89
As	-0,21	-0,03	0,15	-0,11	**-0,89**	0,17	0,91
Sb	**0,89**	0,21	0,15	0,05	0,12	-0,03	0,88
Cu	0,22	0,07	**0,80**	-0,08	-0,37	-0,29	0,93
Ni	**0,71**	0,27	-0,32	0,33	0,27	-0,10	0,87
Mn	0,11	-0,16	**0,73**	0,21	0,01	0,48	0,83
Al	0,14	-0,01	**0,87**	-0,14	0,22	0,18	0,87
Ba	**0,94**	-0,06	-0,05	0,05	0,04	0,05	0,91
Fe	-0,35	-0,11	**0,70**	-0,19	-0,41	-0,26	0,90
Zn	-0,13	**0,92**	0,03	-0,05	0,15	-0,02	0,89
Ca	**0,53**	0,17	0,37	0,10	0,49	0,02	0,71
K	**0,57**	**0,79**	-0,07	0,12	0,06	0,00	0,97
Pb	0,45	**0,85**	-0,12	0,16	0,15	-0,04	0,99
Prop. Var.	0,33	0,22	0,12	0,11	0,07	0,06	
Cum. Var.	0,33	0,56	0,68	0,78	0,86	0,92	
Zdroje	spalování biomasy, doprava, spalovna	sekundární aerosoly, spalovna, průmysl	průmysl	doprava, spalování biomasy	spalování uhlí	průmysl	

Tab. A.2: Výsledky klasické faktorové analýzy pro datový soubor Brno zima

Proměnné	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	$\mathbf{h}_i^2$
F^-	0,01	0,20	**-0,50**	**0,74**	-0,07	-0,09	0,85
Cl^-	**0,88**	-0,11	0,28	0,22	0,15	0,16	0,95
NO_2^-	0,40	-0,17	0,13	0,27	**0,74**	-0,03	0,82
NO_3^-	-0,04	**-0,93**	-0,17	-0,16	-0,19	0,14	0,97
SO_4^{2-}	**-0,71**	**-0,57**	-0,18	0,21	-0,04	-0,11	0,93
ox^{2-}	-0,41	**-0,79**	-0,34	0,13	-0,11	-0,11	0,95
Na^+	0,10	0,37	**0,68**	-0,15	0,48	0,22	0,91
NH_4^+	-0,10	**-0,90**	0,01	-0,13	0,06	0,03	0,84
Ca^{2+}	-0,06	0,16	0,10	**0,81**	0,03	-0,01	0,69
K^+	**0,81**	-0,19	-0,06	0,05	0,28	0,21	0,83
V	-0,07	0,22	**0,83**	-0,08	0,01	-0,06	0,76
Cd	0,25	0,47	0,12	-0,34	**0,72**	0,01	0,93
As	0,36	**0,59**	0,34	-0,49	0,01	-0,03	0,83
Sb	**0,59**	0,45	0,24	-0,13	0,44	0,40	0,98
Cu	**0,58**	0,22	0,21	0,11	0,26	**0,66**	0,94
Ni	0,13	0,15	**0,79**	0,00	0,22	-0,08	0,73
Mn	**0,78**	0,21	**0,52**	0,04	-0,06	0,03	0,92
Al	**-0,79**	-0,21	0,43	0,15	-0,03	0,17	0,91
Ba	0,12	-0,35	-0,14	-0,10	-0,09	**0,86**	0,92
Fe	**0,58**	0,24	**0,69**	0,01	-0,16	0,14	0,91
Zn	**0,70**	0,44	0,29	-0,09	0,26	0,36	0,97
Ca	**-0,85**	-0,16	0,13	0,27	-0,20	0,01	0,88
K	**0,69**	0,45	0,28	-0,18	0,32	0,25	0,95
Pb	0,02	**-0,82**	-0,26	-0,28	-0,01	0,18	0,85
Prop. Var.	0,27	0,21	0,16	0,09	0,08	0,07	
Cum. Var.	0,27	0,48	0,64	0,73	0,81	0,88	
Zdroje	spalovna	spalování biomasy, sekundární aerosoly	spalování olejů	spalování uhlí	spalování uhlí	doprava	

Tab. A.3: Výsledky klasické faktorové analýzy pro datový soubor Brno léto

Proměnné	**F1**	**F2**	**F3**	**F4**	**F5**	$\mathbf{h}_i^2$
F^-	0,18	0,09	**0,93**	0,15	0,03	0,93
Cl^-	-0,07	0,08	**0,94**	0,10	0,19	0,93
NO_2^-	-0,21	-0,14	-0,14	**-0,90**	-0,21	0,93
NO_3^-	0,15	0,00	0,00	**-0,93**	-0,18	0,91
SO_4^{2-}	**0,92**	-0,03	0,21	0,27	-0,13	0,97
ox^{2-}	**0,97**	0,05	0,11	0,18	-0,04	0,98
Na^+	**0,82**	0,15	-0,08	-0,22	0,38	0,89
NH_4^+	**0,90**	0,09	0,10	0,34	-0,10	0,95
Ca^{2+}	**0,88**	0,20	0,32	0,00	0,25	0,98
K^+	**0,87**	0,24	-0,09	-0,17	0,27	0,92
V	0,41	-0,13	0,19	0,24	**0,70**	0,77
Cd	0,28	**0,73**	0,35	0,42	-0,05	0,92
As	0,18	-0,14	0,16	0,26	**0,92**	0,99
Sb	-0,13	**0,75**	0,29	-0,02	**0,56**	0,97
Cu	0,07	**0,93**	-0,19	-0,09	0,23	0,97
Ni	-0,43	0,13	-0,49	0,36	-0,11	0,58
Mn	0,34	**0,89**	0,16	0,19	0,03	0,96
Al	0,09	**0,76**	0,17	-0,08	-0,49	0,87
Ba	**0,94**	0,15	0,08	-0,12	0,07	0,94
Fe	**-0,57**	**0,58**	-0,46	0,08	-0,13	0,89
Zn	-0,11	**0,89**	-0,20	0,25	-0,20	0,95
Ca	**0,82**	-0,32	-0,15	-0,08	0,20	0,84
K	0,15	**0,84**	0,11	-0,13	-0,17	0,79
Pb	**0,73**	0,36	**0,51**	-0,14	0,01	0,94
Prop. Var.	0,33	0,24	0,13	0,11	0,10	
Cum. Var.	0,33	0,57	0,69	0,80	0,91	
Zdroje	spalování biomasy, sekundární aerosoly, doprava	doprava	spalovna	průmysl	spalování uhlí	

Tab. A.4: Výsledky robustní faktorové analýzy pro datový soubor Šlapanice léto

Proměnné	**F1**	**F2**	**F3**	$\mathbf{h}_i^2$
F^-_Cl^-	**0,89**	0,09	0,03	0,80
NH_4^+	**0,96**	0,01	0,07	0,93
Cu	0,18	**0,92**	-0,19	0,91
Zn	0,35	**0,75**	0,06	0,69
K	**0,97**	0,05	0,03	0,94
Pb	**0,99**	-0,05	-0,06	0,99
V_Ni	0,10	0,01	**0,99**	0,99
Sb_Ba	**0,74**	-0,02	0,12	0,56
Cd_As	-0,16	0,46	0,20	0,28
Al_Fe	-0,25	**0,86**	-0,04	0,80
NO_3^-_SO_4^{2-}	**0,97**	-0,06	-0,07	0,95
Prop. Var.	0,49	0,22	0,10	
Cum. Var.	0,49	0,71	0,81	
Zdroje	spalovna, sekundární aerosoly, spalování dřeva	doprava	spalování olejů	