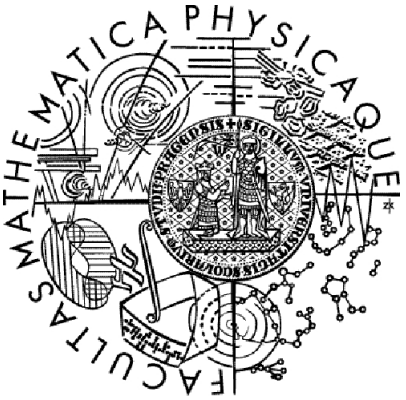
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Tomáš Coufal

Mnohorozměrné statistické metody s aplikací ve financích

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr.
Studijní program: matematika, finanční matematika

2008

Děkuji vedoucí bakalářské práce RNDr. Jitce Zichové, Dr. za vstřícný přístup, zapůjčení materiálů a podnětné připomínky.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 28. května 2008 Tomáš Coufal

Obsah

Název práce: Mnohorozměrné statistické metody s aplikací ve financích
Autor: Tomáš Coufal
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr
e-mail vedoucího: Jitka.Zichova@mff.cuni.cz

Abstrakt: Práce pojednává o některých metodách mnohorozměrné statistické analýzy, konkrétně o analýze hlavních komponent, testování nezávislosti kategoriálních dat metodou kontingenčních tabulek a o grafických modelech. Vícerozměrná statistika je vhodným nástrojem pro analýzu finančních dat, používají ji například banky při tzv. kreditscoringu, tj. posuzování bonity žadatelů o úvěr. Zahrnuje škálu metod umožňujících třídění databází, posuzování důležitosti sledovaných znaků i zkoumání jejich vzájemných souvislostí. Právě na datech z oblasti kreditscoringu jsou jednotlivé metody demostrovány a následně jsou komentovány jejich závěry.

Klíčová slova: analýza hlavních komponent, snížení dimenze, kontingenční tabulky, grafické modely, kreditscoring

Title: Multivariate statistical methods with an application in finance
Author: Tomáš Coufal
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Jitka Zichová, Dr
Supervisor's e-mail address: Jitka.Zichova@mff.cuni.cz

Abstract: In the present work we study some areas of multivariate statistical methods, in the concrete analysis of principal components, independence tests of categorical variables using contingency tables and graphical models. Multivariate statistics is a suitable instrument for analysis of financial data. It is used for example for so called Credit Scoring. Multivariate statistics comprises various methods enabling data sorting, significance assessment of monitored characteristics and enquiry into their mutual connection. Those methods are illustrated on Credit Scoring data. The comments on the outputs are provided.
Keywords: principal components, reduction of dimension, contingency tables, graphical models, Credit Scoring

Úvod

Metody mnohorozměrné statistické analýzy mají své uplatnění nejen v oblasti finančnictví. V této práci se zaměříme na popis některých těchto metod, konkrétně na popis analýzy hlavních komponent, testu nezávislosti znaků pomocí kontingenčních tabulek a grafických modelů. Tyto metody následně aplikujeme na finanční data, konkrétně na data z oblasti tzv. kreditscoringu, tedy hodnocení solventnosti žadatelů o úvěr.
V první kapitole se blíže seznámíme s metodou analýzy hlavních komponent a možnostmi jejího využití. Jednou z nich je například zmenšení databáze historických dat (v důsledku snížení dimenze prostoru pozorování) při ztrátě minimální možné informace o původních datech. Na základě analýzy hlavních komponent provedeme také úsudek o důležitosti jednotlivých složek vícerozměrného pozorování na variabilitu celého souboru.
Ve druhé kapitole si představíme některé pomocné statistické metody, které nám pomohou objasnit závěry analýzy hlavních komponent na finančních datech. Konkrétně se soustředíme na výběrovou korelační matici znaků, o kterých jsme na základě analýzy hlavních komponent usoudili, že mají největší vliv na variabilitu celého souboru pozorování, a na test jejich nezávislosti metodou kontingenčních tabulek kategoriálních dat.
Na závěr se seznámíme s tzv. grafickými modely, možnostmi zobrazení struktury závislostí statistických dat do matematického grafu a testováním hypotéz o vhodnosti různých grafických modelů. Také tuto metodu aplikujeme na naše finanční data.

Kapitola 1

Analýza hlavních komponent

1.1 Teoretické zázemí

Analýza hlavních komponent je vícerozměrnou statistickou metodou související s tzv. problémem snížení dimenze. Uvažujme skupinu n objektů, na kterých pozorujeme p znaků, kde n i p jsou velké. Z důvodů snadného zálohování dat, ale také z důvodu jejich větší přehlednosti, je pro nás často výhodné jejich velikost (ve smyslu snížení p na nějaké q, $q < p$) omezit. Při takovémto omezení však chceme ztratit jen minimum informace v původních datech obsažené.
Analýza hlavních komponent snížení počtu uchovánaných znaků u každého objektu z p na q optimalizuje ve smyslu uchování informace o rozptylu původních proměnných.

Předpokládejme, že náhodné veličiny $X^1, X^2, ..., X^p$ mají nějaké mnohorozměrné rozdělení s vektorem středních hodnot μ a kovarianční maticí Σ, jejíž hodnost je p. Nechť $\boldsymbol{X} = X^1, X^2, ..., X^p$.
Hlavními komponentami nazveme množinu znaků $Y^1, Y^2, ...Y^p$, které tvoříme jako vhodnou lineární kombinaci výchozích $X^1, X^2, ...X^p$ nebo jejich normovaných zástupců $Z^1, Z^2, ...Z^p$, kde

$$Z^i = (X^i - \mu_i)/\sigma_i,\ \mu_i = E(X^i),\ \sigma_i = \sqrt{\Sigma_{i,i}}. \tag{1.1}$$

Toto normování způsobí, že procedura hledání hlavních komponent proběhne nezávisle na použitých měrných jednotkách.
Jelikož jsou prvky kovarianční matice Σ invariantní vůči záměně znaků X^j za $\tilde{X}^j = X^j - c(j)$, $j = 1, ..., p$, kde $c(j)$ je vektor libovolných konstant,

můžeme dále předpokládat, že vektor středních hodnot μ je nulový, respektive můžeme v každém případě provést alespoň transformaci $Z^j = X^j - \mu_j$ aniž by tím byla dotčena snadnost zpětné interpretace významu jednotlivých hlavních komponent, neboť na jejich určení má vliv právě kovarianční matice.

Ve výsledku chceme $Y^1, Y^2, ..., Y^p$ vzájemně nekorelované seřazené podle variability kolísání.

1.1.1 Volba hlavních komponent

První hlavní komponentu nalezeme jako lineární kombinaci

$$Y^1 = v_1^1 X^1 + v_1^2 X^2 + ... + v_1^p X^p$$

(tedy vektorově $Y^1 = \boldsymbol{v}_1^T \boldsymbol{X}$), která má mezi všemi lineárními kombinacemi splňujícími $\mathbf{v}_1^T \cdot \mathbf{v}_1 = 1$ největší rozptyl $Var(Y^1) = \boldsymbol{v}_1^T \Sigma \boldsymbol{v}_1$.

Druhou hlavní komponentu volíme jako takovou lineární kombinaci

$$Y^2 = v_2^1 X^1 + v_2^2 X^2 + ... + v_2^p X^p,$$

která má mezi všemi lineárními kombinacemi nekorelovanými s Y^1 splňujícími $\boldsymbol{v}_2^T \cdot \boldsymbol{v}_2 = 1$ největší rozptyl.

Obecně *i*-tou hlavní komponentu najdeme jako lineární kombinaci

$$Y^i = v_i^1 X^1 + v_i^2 X^2 + ... + v_i^p X^p,$$

která má mezi všemi lineárními kombinacemi nekorelovanými se všemi Y^j, $j < i$ splňujícími $\boldsymbol{v}_i^T \cdot \boldsymbol{v}_i = 1$ největší rozptyl.

Z výše popsaného postupu plyne, že hlavní komponenty jsou seřazeny sestupně podle velikosti jejich rozptylu, $Var(Y^1) \geq Var(Y^2) \geq ... \geq Var(Y^p)$.

Z rovnice pro rozptyl *i*-té hlavní komponenty:

$$Var(Y^i) = E(\boldsymbol{v}_i^T \boldsymbol{X})^2 = E(\boldsymbol{v}_i^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{v}_i) = \boldsymbol{v}_i^T \Sigma \boldsymbol{v}_i \tag{1.2}$$

a z vlastností vlastních čísel a vlastních vektorů (viz níže) plyne, že za *i*-tý transformační vektor $\boldsymbol{v}_i$ volíme vlastní vektor kovarianční matice Σ příslušící k *i*-tému (dle velikosti) vlastnímu číslu λ_i.

Jak víme z lineární algebry [3, str.104], vlastní čísla matice hledáme jako kořeny charakteristického polynomu, v našem případě

$$|\Sigma - \lambda \cdot \boldsymbol{I}| = 0. \tag{1.3}$$

Je-li matice regulární (což o matici Σ předpokládáme), má právě tolik vlastních čísel, kolik je její hodnost (tedy p). Je-li λ_i jednonásobný kořen charakteristického polynomu, pak k němu připadá právě jeden normovaný vlastní vektor $\boldsymbol{v}_i$ ($\boldsymbol{v}_i^T \cdot \boldsymbol{v}_i = 1$), který se určí jako řešení soustavy

$$(\Sigma - \lambda_i \cdot \boldsymbol{I}) \cdot \boldsymbol{v}_i = 0, \tag{1.4}$$

kde I je jednotková diagonální matice typu $p \times p$.

O normovaných vlastních vektorech dále víme, že jsou každé dva na sebe kolmé. Tedy, že

$$\boldsymbol{v}_i^T \cdot \boldsymbol{v}_j = 0, \ i \neq j.$$

Z rovnic 1.2 a 1.4 plyne, že v souladu s požadavkem na maximalizaci rozptylu komponent s nízkými indexovými čísly volíme za vektor $\boldsymbol{v}_1$ vlastní vektor příslušející k největšímu vlastnímu číslu matice Σ, $\boldsymbol{v}_2$ bude vlastní vektor příslušející k druhému největšímu vlastnímu číslu matice Σ atd.

Označme

$$\boldsymbol{V} = \begin{pmatrix} v_1^1 & v_1^2 & \cdots & v_1^p \\ v_2^1 & v_2^2 & \cdots & v_2^p \\ \vdots & \vdots & \ddots & \vdots \\ v_p^1 & v_p^2 & \cdots & v_p^p \end{pmatrix} \tag{1.5}$$

matici transformace, potom celý proces přechodu k hlavním komponentám můžeme popsat maticovou operací

$$\boldsymbol{Y} = \boldsymbol{V} \cdot \boldsymbol{X}.$$

V tomto maticovém zápisu můžeme díky ortogonalitě matice $\boldsymbol{V}$ z hlavních komponent $Y^1, Y^2, ..., Y^p$ snadno vyjádřit původní $X^1, X^2, ..., X^p$, neboť

$$\boldsymbol{X} = \boldsymbol{V}^T \cdot \boldsymbol{Y}.$$

1.1.2 Snížení dimenze

Jak plyne z rovnic 1.2, 1.3 a 1.4, pro rozptyl i-té hlavní komponenty platí

$$Var(Y^i) = \lambda_i,$$

kde λ_i je i-té vlastní číslo kovarianční matice Σ. S touto znalostí můžeme snadno spočítat koeficient r_i,

$$r_i = \frac{Var(Y^i)}{|\Sigma|} = \frac{\lambda_i}{|\Sigma|} = \frac{\lambda_i}{tr(\Sigma)} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

(kde $tr(\Sigma)$ je stopa matice Σ)udávající poměr, jakým se i-tá hlavní komponenta podílí na celkovém rozptylu souboru.
Chceme-li pak, aby nám prvních q hlavních komponent reprezentovalo například alespoň 90%, 95% či jakoukoliv jinou část l ($l \in \langle 0, 1 \rangle$) rozptylu celého souboru , zvolíme q tak, aby platilo

$$q = min \left\{ q' : \frac{\sum_{i=1}^{q'} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq l \right\},$$

neboli

$$q = min \left[j : \sum_{i=1}^{j} r_i \geq l \right].$$

Takto určené q nyní představuje nejnižší možnou dimenzi prostoru generovaného hlavními komponentami při podmínce zachování dané části informace o rozptylu původního souboru. Ortonormální bází tohoto prostoru je prvních q hlavních komponent.

Poznámka: Jsou-li hlavní komponenty získávány z korelační matice (odpovídá situaci, kdy původní $X^1, X^2, ..., X^p$ transformujeme na $Z^1, Z^2, ..., Z^p$ ve smyslu popsaném výše), pro součet vlastních čísel korelační matice platí

$$\sum_{i=1}^{p} \lambda_i = p.$$

Podíl i-té hlavní komponenty na celkovém rozptylu je tedy dán poměrem λ_i/p.

1.1.3 Náhodný výběr, výběrové hlavní komponenty

Mějme nyní náhodný výběr o n p-rozměrných pozorováních. Ta můžeme zapsat do matice $\underline{\boldsymbol{X}}$ typu n × p . Uvažujme situaci, že přesné p-rozměrné rozdělení $X^1, X^2, ..., X^p$, z něhož byl výběr n prvků proveden, neznáme. Předpokládejme však existenci takového rozdělení, vektoru středních hodnot μ a kovarianční matice Σ (hodnosti p). Matici Σ odhadneme výběrovou kovarianční maticí $\hat{\Sigma}$.

$$\hat{\Sigma_{i,j}} = \frac{1}{n} \cdot \sum_{k=1}^{n} (X_k^i - \bar{X^i}) \cdot (X_k^j - \bar{X^j}),\ \bar{X^i} = \frac{1}{n} \cdot \sum_{k=1}^{n} X_k^i,\ i,j = 1, ..., p. \quad (1.6)$$

Rozhodnutí zda budeme nadále v analýze hlavních komponent pracovat s naměřenými X_i^j či s normovanými Z_i^j, kde

$$Z_i^j = \frac{X_i^j - \bar{X^j}}{\sqrt{\hat{\Sigma}_{jj}}},\ i = 1, ..., n,\ j = 1, ..., p \quad (1.7)$$

je věcí zvážení. Oba přístupy mají svá pro i proti. Práce s nenormovanými X_i^j má zajisté výhodu lepší zpětné interpretace hlavních komponent. Na druhou stranu normováním předejdeme problémům spojeným s použitím různých měrných jednotek, v nichž byla pozorování realizována.

Hlavní komponenty získané pomocí výběrové kovarianční matice nazveme ***výběrové hlavní komponenty***.

1.1.4 Vlastnosti hlavních komponent

Vraťme se nyní k případu, kdy známe rozdělení náhodného vektoru $X^1, X^2, ..., X^p$ a zejména pak jeho kovarianční matici Σ. Následující text lze snadno interpretovat pro případ výběrových hlavních komponent jako vyjádření o nejlepších odhadech zmiňovaných pojmů. Nyní, když již máme určené hlavní komponenty, je vhodné zamyslet se nad jejich vztahem k původním proměnným. Mějme transformační matici $\boldsymbol{V}$ z vyjádření 1.5. Jak plyne ze způsobu výběru hlavních komponent (pomocí vlastních vektorů kovarianční matice Σ), kovarianční matici hlavních komponent $Cov(\boldsymbol{Y})$ získáme jako

$$Cov(\boldsymbol{Y}) = \boldsymbol{V} \Sigma \boldsymbol{V}^T. \quad (1.8)$$

Ta je, jak plyne z výkladu výše, diagonální maticí, kde prvky diagonály tvoří vlastní čísla matice Σ seřazená sestupně podle velikosti. Pro kovarianci

původních proměnných a hlavních komponet platí

$$Cov(\boldsymbol{X}, \boldsymbol{Y}) = \Sigma \boldsymbol{V}^T. \tag{1.9}$$

Vzhledem ke vztahu 1.4, je i-tý sloupec matice $Cov(\boldsymbol{X}, \boldsymbol{Y})$ roven násobku i-tého vlastního čísla kovarianční matice Σ a příslušného vlastního vektoru, tedy

$$Cov_i(\boldsymbol{X}, \boldsymbol{Y}) = \Sigma \boldsymbol{v}_i = \lambda_i v_i.$$

Pro kovarianci i-té složky náhodného vektoru $X^1, X^2, ..., X^p$ s j-tou hlavní komponentou pak dostáváme

$$cov(X^i, Y^j) = \lambda_j v_j^i$$

a pro jejich korelaci

$$cor(X^i, Y^j) = \frac{\sqrt{\lambda_j} v_j^i}{\sqrt{\Sigma_{i,i}}}.$$

S vektory $\sqrt{\lambda_j}\boldsymbol{v}_j$ budeme pracovat i nadále. Jelikož kovarianční matici Σ lze zapsat jako

$$\Sigma = \boldsymbol{V}^T \Lambda \boldsymbol{V},$$

kde Λ je diagonální matice, která má na i-tém místě na diagonále hodnotu i-tého vlastního čísla λ_i, dostáváme

$$\Sigma = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$$

[4, str.377].
Tento rozklad nám umožňuje kovarianční matici Σ zpětně konstruovat a přes postupné součty například sledovat míru reprodukce kovarinční matice Σ prvními q hlavními komponentami.

1.1.5 Geometrický význam výběrových hlavních komponent

Náhodný výběr o n pozorováních z p-rozměrného rozdělení si lze představit jako "shluk" n bodů v p-rozměrném (euklidovském) prostoru, jehož osy odpovídají jednotlivým proměnným $X^1, X^2, ..., X^p$. Bez újmy na obecnosti můžeme za střed souřadnic položit bod odpovídající výběrovým průměrům

jednotlivých X^i. Nalezení hlavních komponet $Y^1, Y^2, ..., Y^p$ jako lineárních kombinací původních $X^1, X^2, ..., X^p$ metodou popsanou výše si můžeme představit jako rotaci souřadnic do směru největšího rozptylu shluku bodů. Tak můžeme jen na několika málo prvních souřadnicích zachytit většinu informace o našem souboru p-rozměrných pozorování.
Konkrétně, zvolíme-li l takové (l má stejný význam jako v kapitole 1.1.2), aby nám na reprezentaci této části celkového rozptylu stačilo q hlavních komponent, kde $q \leq 3$, můžeme schéma našich pozorování snadno zobrazit (na přímku, do roviny či do prostoru), aniž bychom tímto zobrazením ztráceli velké množství informace obsažené v původně p-rozměrném pozorování.

1.1.6 Vztah teoretických a výběrových hlavních komponent

Na závěr si řekněme něco o vztahu teoretických hlavních komponent (tedy těch určených na základě známé kovarianční matice Σ) a jejich výběrových protějšků. Jak již bylo napsáno výše, výběrové hlavní komponenty jsou nejlepším odhadem teoretických hlavních komponent získaným na základě provedených n pozorování. Jsou totiž odhadem konzistentním a asymptoticky vydatným.

Poznámka: Vydatnost (eficience) odhadu je určena poměrem nejmenšího možného rozptylu odhadu ke skutečnému rozptylu odhadu parametru. V případě, že skutečný rozptyl je stejný jako nejmenší možný rozptyl odhadu parametru, je vydatnost rovna 1. Odhad s nejmenším možným rozptylem se nazývá vydatným odhadem. Asymptoticky vydatný je odhad, který se za vydatný dá považovat při velkém počtu pozorování n.

Víme tedy, že když budeme přidávat další pozorování (zvyšovat n), budou se odhady vlastních čísel $\hat{\lambda}_i$ a vlastních vektorů $\hat{\boldsymbol{v}}_i$ měnit a při $n \to \infty$ konvergovat k teoretickým λ_i a $\boldsymbol{v}_i$. Podívejme se na tuto konvergenci trochu blíže a uveďme některé související testy hypotéz.
Předpokládejme, že $X^1, X^2, ..., X^p$ mají vícerozměrné normální rozdělení s nulovou střední hodnotou (nulovosti střední hodnoty dosáhneme transformací původních $X^1, X^2, ..., X^p$ na centrované veličiny $\tilde{X}^1, \tilde{X}^2, ..., \tilde{X}^p$ odečtením středních hodnot od jednotlivých složek: $\tilde{X}^i = X^i - \mu_i,\ \mu_i = E(X^i)$), tedy vektor $\boldsymbol{X} = (X^1, X^2, ..., X^p) \in N(0, \Sigma)$. Předpokládejme dále, že provedená p-rozměrná pozorování $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ vektoru $\boldsymbol{X}$ jsou vzájemně nezávislá.

Nechť jsou vlastní čísla $\lambda_1, \lambda_2, ..., \lambda_p$ kovarianční matice Σ různá, pak platí:

- vlastní čísla $\hat{\lambda_1}, \hat{\lambda_2}, ..., \hat{\lambda_p}$ a jim příslušné vlastní vektory $\hat{v_1}, \hat{v_2}, ..., \hat{v_p}$ výběrové kovarianční matice $\hat{\Sigma}$ jsou maximálně věrohodnými odhady odpovídajících teoretických vlastních čísel $\lambda_1, \lambda_2, ..., \lambda_p$ repektive jim příslušejících vlastních vektorů $v_1, v_2, ..., v_p$. Mají vlastnosti konzistence a asymptotické vydatnosti, které po takovýchto odhadech požadujeme. Výběrové hlavní komponenty jsou pak odhady teoretických hlavních komponent.

- Veličiny
$$\sqrt{n-1} \cdot (\lambda_i - \hat{\lambda_i}),\ i = 1, 2, ..., p$$
mají pro $n \to \infty$ normální rozdělení $N(0, 2\lambda_i^2)$ a jsou vzájemně nezávislé.

- Vektory
$$\sqrt{n-1} \cdot (\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i),\ i = 1, 2, ..., p$$
mají pro $n \to \infty$ vícerozměrné normální rozdělení s nulovým vektorem středních hodnot a s kovarianční maticí
$$\lambda_i \cdot \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2} \boldsymbol{v}_j \boldsymbol{v}_j^T$$

- Kovariance mezi k-tou složkou výběrového vlastního vektoru $\hat{\boldsymbol{v}}_i$ a l-tou složkou výběrového vlastního vektoru $\hat{\boldsymbol{v}}_j$ je rovna
$$-\frac{\lambda_i \lambda_j \boldsymbol{v}_i^k \boldsymbol{v}_j^l}{(n-1)(\lambda_i - \lambda_j)^2}$$

[1, str.159-160].

Na úplný závěr zmíníme některé testy hypotéz týkající se hlavních komponent. Jako první uveďme test hypotézy, že nejmenších $p - q$ vlastních čísel teoretické kovarianční matice Σ (kterou neznáme) si je rovno. V tomto případě je pouze prvních q vlastních vektorů určeno jednoznačně. Máme tedy následující hypotézu:

$$H : \lambda_{q+1} = \lambda_{q+2} = ... = \lambda_p$$

alternativou je v tomto případě existence alespoň jedné dvojice $\lambda_i \neq \lambda_j,\ i,j = q+1, q+2, ..., p$. Za platnosti hypotézy má veličina

$$A = -(n-1) \cdot \sum_{i=q+1}^{p} ln(\hat{\lambda}_i) + (n-1)(p-q) \cdot ln \left(\frac{\sum_{i=q+1}^{p} \hat{\lambda}_i}{p-q} \right)$$

rozdělení $\chi^2_{\frac{1}{2}(p-q)(p-q+1)-1}$.
Hypotézu tedy zamítáme, pokud A bude větší, než $(1-\alpha)$-tý kvantil výše zmíněného rozdělení, α je hladina testu. [4, str.382]

Druhým testem, který uvedeme, bude test hypotézy o diagonalitě teoretické kovarianční matice Σ. Jinými slovy za hypotézu vezmeme:

$$H : cov(X^i, X^j) = 0,\ i \neq j,\ i,j = 1,2,...,p.$$

Mějme $\left|\hat{R}\right|$ determinant výběrové korelační matice. Při $n \to \infty$ má veličina

$$B = -\left(n - \frac{2p+11}{6}\right) \cdot ln \left|\hat{R}\right|$$

za platnosti hypotézy rozdělení $\chi^2_{\frac{p(p-1)}{2}}$.

Hypotézu tedy zamítáme, pokud $B > \chi^2_{\frac{p(p-1)}{2}}(1-\alpha)$ ($(1-\alpha)$-tý kvantil výše zmíněného rozdělení), α je hladina testu. [1, str.161]

1.2 Zpracování dat

Jako podklad pro zpracování konkrétních dat nám poslouží databáze německé banky [7] sestavená pro účel takzvaného kreditscoringu (tedy hodnocení kredibility, solventnosti) žadatelů o úvěr. Jedná se o databázi sledující 21 znaků na 1000 osobách (tedy v návaznosti na předchozí text $p = 21, n = 1000$). Význam jednotlivých znaků vysvětluje následující tabulka:

Proměnná	**Popis**	**Hodnoty**	**Skóre**
Kredit	vyhodnocení klienta	špatný	0
		dobrý	1
Stav konta	zůstatek na účtě žadatele	nemá běžný účet	1
		BÚ $\leq$ 0 DM	2
		0 $<$ BÚ $<$ 200 DM	3
		BÚ $\geq$ 200 DM	4
Splatnost	doba splatnosti úvěru	$\leq$ 6 měsíců	10
		$6 < ... \leq 12$	9
		$12 < ... \leq 18$	8
		$18 < ... \leq 24$	7
		$24 < ... \leq 30$	6
		$30 < ... \leq 36$	5
		$36 < ... \leq 42$	4
		$42 < ... \leq 48$	3
		$48 < ... \leq 54$	2
		$>$ 54 měsíců	1
Moralka	splácení dřívějších půjček	váhavé	0
		další půjčky jinde	1
		žádné dřívější půjčky	2
		žádné problémy	3
		dřívější půj. splaceny	4
Ucel	použití úvěru na:	nové auto	1
		použité auto	2
		nábytek	3
		rádio/televize	4
		jiné spotřebiče	5
		opravy	6
		vzdělání	7
		dovolená	8
		rekvalifikace	9
		obchod	10
		ostatní	0

Proměnná	**Popis**	**Hodnoty**	**Skóre**
Vyse pujcky	v DM	≤ 500	10
		$500 < ... \leq 1000$	9
		$1000 < ... \leq 1500$	8
		$1500 < ... \leq 2000$	7
		$2000 < ... \leq 2500$	6
		$2500 < ... \leq 5000$	5
		$5000 < ... \leq 7500$	4
		$7500 < ... \leq 10000$	3
		$10000 < ... \leq 15000$	2
		$15000 < ... \leq 20000$	1
		> 20000	0
Vklady	hodnota úspor	žádné	1
	a cenných papírů	< 100	2
	(v DM)	$100 \leq ... < 500$	3
		$500 \leq ... < 1000$	4
		≥ 1000	5
Delka zam	počet let	nezaměstnaný	1
	v současném	≤ 1	2
	zaměstnání	$1 \leq ... < 4$	3
		$4 \leq ... < 7$	4
		≥ 7	5
Pomer k prijmu	splátka v %	≥ 35	1
	z disponibilního	$25 \leq ... < 35$	2
	příjmu	$20 \leq ... < 25$	3
		< 20	4
Rodinny stav	pohlaví:rodinný stav	M:rozvedený/	
		svobodný	1
		Ž:rozvedená/	
		vdaná/vdova	2
		M:ženatý/vdovec	3
		Ž:svobodná	4
Garance	existence garantů	nejsou	1
	nebo ručitelů	ručitel	2
		garant	3
Doba byd	počet let	< 1	1
	v současném	$1 \leq ... < 4$	2
	bydlišti	$4 \leq ... < 7$	3
		≥ 7	4

Proměnná	**Popis**	**Hodnoty**	**Skóre**
Nejc aktiva	nejcennější aktiva	žádná	1
		auto / ostatní	2
		spořící kontrakt	
		se stavební spol./	
		životní pojištění	3
		dům / pozemek	4
Vek	věk žadatele	$0 \leq ... \leq 25$	1
		$26 \leq ... \leq 39$	2
		$40 \leq ... \leq 59$	3
		$60 \leq ... \leq 64$	4
		≥ 65	5
Dalsi uvery	další úvěry	v jiných bankách	1
		v obchodech nebo	
		zásilkových službách	2
		žádné	3
Typ bydleni	typ bydlení	bezplatné	1
		pronajatý byt	2
		vlastní byt	3
Uver zde	počet dřívějších	1	1
	půjček v této	2 nebo 3	2
	bance	4 nebo 5	3
	(včetně současné)	6 a více	4
Zamestnani	druh zaměstnání	nezam./nevyučen	
		bez trv. místa	1
		nevyučen	
		s trv. místem	2
		kvalifikovaný	
		pracovník, nižší	
		civilní zaměstnanec	3
		nadřízený,samostatně	
		zaměstnaný, vyšší	
		civilní zaměstnanec	4
Vel domácnosti	společně	$2 \leq$	1
	posuzované osoby	≥ 3	2
Telefon	telefon	ne	1
		ano	2
Cizinec	cizinec	ano	1
		ne	2

Pro zpracování dat bylo použito statistického softwaru NCSS, který má implementovány procedury pro mnohorozměrnou statistickou analýzu. Výstupem dané procedury je tabulka se jmény původních proměnných, jejich středními hodnotami a směrodatnými odchylkami, dále výběrová korelační matice, matice vlastních vektorů (tedy transformační matice $\boldsymbol{V}$) a tabulka ukazující podíl jednotlivých hlavních komponent na celkovém rozptylu.

1.2.1 Výsledky programu NCSS

Prvním výstupem programu je tabulka výběrových průměrů jednotlivých znaků (složek vektoru $X^1, X^2, ..., X^p$) a jejich výběrových směrodatných odchylek.

Proměnná	***n***	**výb. průměr**	**výb. směr. odchylka**
Kredit	1000	0.7	0.4584869
Stav konta	1000	2.577	1.257638
Splatnost	1000	7.397	1.962458
Moralka	1000	2.545	1.08312
Ucel	1000	2.828	2.744439
Vyse pujcky	1000	6.658	1.552884
Vklady	1000	2.105	1.580023
Delka zam	1000	3.384	1.208306
Pomer k prijmu	1000	2.973	1.118715
Rodinny stav	1000	2.682	0.7080801
Garance	1000	1.145	0.4777062
Doba byd	1000	2.845	1.103718
Nejc aktiva	1000	2.358	1.050209
Vek	1000	2.188	0.864525
Dalsi uvery	1000	2.675	0.7056011
Typ bydleni	1000	1.928	0.5301859
Uver zde	1000	1.407	0.5776545
Zamestnani	1000	2.904	0.653614
Vel domacnosti	1000	1.845	0.3620858
Telefon	1000	1.404	0.490943
Cizinec	1000	1.963	0.1888562

Jak ukazuje předchozí tabulka, průměry i výběrové směrodatné odchylky

byly pro jednotlivé složky p-rozměrných pozorování různé, proto bylo přistoupeno ke standardizaci dat a hlavní komponenty byly počítány z výběrové korelační matice. To odpovídá situaci, kdy byla před samotným hledáním hlavních komponent provedena transformace dat 1.7.

Poznámka: Analýza hlavních komponent vypočtených z výběrové kovarianční matice, která byla také provedena, dosahovala podobných výsledků.

Prvky výběrové korelační matice představují odhady korelací dvojic složek vektoru $X^1, X^2, ..., X^{21}$ získané na základě provedených $n = 1000$ pozorování. Čím je absolutní hodnota korelace blíže 1, tím větší je lineární závislost mezi těmito složkami. V případě blízkosti 1 se jedná o přímou úměrnost (nárůst i-té složky ovlivní j-tou k růstu a naopak), v případě blízkosti -1 o nepřímou úměrnost. Jelikož je vypočtená výběrová korelační matice typu 21×21 , je nutné ji z prostorových důvodů rozdělit. Je tedy uvedena ve formě několika tabulek:

Proměnná	**Kredit**	**St konta**	**Splatnost**	**Moralka**	**Ucel**
Kredit	1.000000	0.350847	0.208152	0.228785	-0.017979
Stav konta	0.350847	1.000000	0.070138	0.192191	0.028783
Splatnost	0.208152	0.070138	1.000000	0.075177	-0.150492
Moralka	0.228785	0.192191	0.075177	1.000000	-0.090336
Ucel	-0.017979	0.028783	-0.150492	-0.090336	1.000000
Vyse pujcky	0.100385	0.015035	0.632230	0.037131	-0.049753
Vklady	0.178943	0.222867	-0.049937	0.039058	-0.018684
Delka zam	0.116002	0.106339	-0.060555	0.138225	0.016013
Pomer k prijmu	-0.072404	-0.005280	-0.068064	0.044375	0.048369
Rodinny stav	0.088184	0.043261	-0.012069	0.042171	0.000157
Garance	0.025137	-0.127737	0.015413	-0.040676	-0.017607
Doba byd	-0.002967	-0.042234	-0.039035	0.063198	-0.038221
Nejc aktiva	-0.142612	-0.032260	-0.300217	-0.053777	0.010966
Vek	0.096975	0.083342	0.024405	0.137410	0.015752
Dalsi uvery	0.109844	0.068274	0.061464	0.159957	-0.100230
Typ bydleni	0.018119	0.023335	-0.151445	0.061428	0.013495
Uver zde	0.045732	0.076005	0.006553	0.437066	0.054935
Zamestnani	-0.032735	0.040663	-0.212179	0.010350	0.008085
Vel domacnosti	-0.003015	0.014145	-0.026012	-0.011550	0.032577
Telefon	0.036466	0.066296	-0.163521	0.052370	0.078371
Cizinec	-0.082079	0.035187	-0.125080	-0.028554	0.113244

Proměnná	V. puj.	Vklady	Del. zam	P. k prij.	Rod.stav
Kredit	0.100385	0.178943	0.116002	-0.072404	0.088184
Stav konta	0.015035	0.222867	0.106339	-0.005280	0.043261
Splatnost	0.632230	-0.049937	-0.060555	-0.068064	-0.012069
Moralka	0.037131	0.039058	0.138225	0.044375	0.042171
Ucel	-0.049753	-0.018684	0.016013	0.048369	0.000157
Vyse pujcky	1.000000	-0.059193	-0.004626	0.292577	0.023891
Vklady	-0.059193	1.000000	0.120950	0.021993	0.017349
Delka zam	-0.004626	0.120950	1.000000	0.126161	0.111278
Pomer k prijmu	0.292577	0.021993	0.126161	1.000000	0.119308
Rodinny stav	0.023891	0.017349	0.111278	0.119308	1.000000
Garance	0.018338	-0.105069	-0.008116	-0.011398	0.050634
Doba byd	-0.027455	0.091424	0.245081	0.049302	-0.027269
Nejc aktiva	-0.320131	0.018948	0.087187	0.053391	-0.006940
Vek	-0.032587	0.078602	0.242254	0.054933	0.024175
Dalsi uvery	0.046454	0.001908	-0.007279	0.007894	-0.026747
Typ bydleni	-0.113829	0.006644	0.115077	0.091229	0.098934
Uver zde	-0.016522	-0.021644	0.125791	0.021669	0.064672
Zamestnani	-0.299646	0.011709	0.101225	0.097755	-0.011956
Vel domacnosti	0.037368	-0.027514	-0.097192	0.071207	-0.122165
Telefon	-0.266318	0.087208	0.060518	0.014413	0.027275
Cizinec	-0.029538	-0.010450	0.022845	0.094762	-0.073103

Proměnná	Garance	Doba b.	Nej. Akt.	Vek	Dalsi uv.
Kredit	0.025137	-0.002967	-0.142612	0.096975	0.109844
Stav konta	-0.127737	-0.042234	-0.032260	0.083342	0.068274
Splatnost	0.015413	-0.039035	-0.300217	0.024405	0.061464
Moralka	-0.040676	0.063198	-0.053777	0.137410	0.159957
Ucel	-0.017607	-0.038221	0.010966	0.015752	-0.100230
Vyse pujcky	0.018338	-0.027455	-0.320131	-0.032587	0.046454
Vklady	-0.105069	0.091424	0.018948	0.078602	0.001908
Delka zam	-0.008116	0.245081	0.087187	0.242254	-0.007279
Pomer k prijmu	-0.011398	0.049302	0.053391	0.054933	0.007894
Rodinny stav	0.050634	-0.027269	-0.006940	0.024175	-0.026747
Garance	1.000000	-0.025678	-0.155450	-0.017597	-0.038235
Doba byd	-0.025678	1.000000	0.147231	0.203664	0.022654
Nejc aktiva	-0.155450	0.147231	1.000000	0.060303	-0.107593
Vek	-0.017597	0.203664	0.060303	1.000000	-0.040860
Dalsi uvery	-0.038235	0.022654	-0.107593	-0.040860	1.000000
Typ bydleni	-0.065449	0.009990	0.342969	0.282891	-0.097398
Uver zde	-0.025447	0.089625	-0.007765	0.145288	-0.055810
Zamestnani	-0.057963	0.012655	0.276149	0.030200	0.006077
Vel domacnosti	-0.020400	-0.042643	-0.011872	-0.117869	0.076891
Telefon	-0.075035	0.095359	0.196802	0.136903	-0.025140
Cizinec	-0.140190	0.039691	0.132462	-0.012532	-0.007700

Proměnná	Typ byd.	Uv. zde	Zamest.	V. dom.	Telefon
Kredit	0.018119	0.045732	-0.032735	-0.003015	0.036466
Stav konta	0.023335	0.076005	0.040663	0.014145	0.066296
Splatnost	-0.151445	0.006553	-0.212179	-0.026012	-0.163521
Moralka	0.061428	0.437066	0.010350	-0.011550	0.052370
Ucel	0.013495	0.054935	0.008085	0.032577	0.078371
Vyse pujcky	-0.113829	-0.016522	-0.299646	0.037368	-0.266318
Vklady	0.006644	-0.021644	0.011709	-0.027514	0.087208
Delka zam	0.115077	0.125791	0.101225	-0.097192	0.060518
Pomer k prijmu	0.091229	0.021669	0.097755	0.071207	0.014413
Rodinny stav	0.098934	0.064672	-0.011956	-0.122165	0.027275
Garance	-0.065449	-0.025447	-0.057963	-0.020400	-0.075035
Doba byd	0.009990	0.089625	0.012655	-0.042643	0.095359
Nejc aktiva	0.342969	-0.007765	0.276149	-0.011872	0.196802
Vek	0.282891	0.145288	0.030200	-0.117869	0.136903
Dalsi uvery	-0.097398	-0.055810	0.006077	0.076891	-0.025140
Typ bydleni	1.000000	0.050020	0.104243	-0.115549	0.100327
Uver zde	0.050020	1.000000	-0.026321	-0.109667	0.065553
Zamestnani	0.104243	-0.026321	1.000000	0.093559	0.383022
Vel domacnosti	-0.115549	-0.109667	0.093559	1.000000	0.014753
Telefon	0.100327	0.065553	0.383022	0.014753	1.000000
Cizinec	0.083336	0.018893	0.092835	0.077071	0.075012

Proměnná	Cizinec
Kredit	-0.082079
Stav konta	0.035187
Splatnost	-0.125080
Moralka	-0.028554
Ucel	0.113244
Vyse pujcky	-0.029538
Vklady	-0.010450
Delka zam	0.022845
Pomer k prijmu	0.094762
Rodinny stav	-0.073103
Garance	-0.140190
Doba byd	0.039691
Nejc aktiva	0.132462
Vek	-0.012532
Dalsi uvery	-0.007700
Typ bydleni	0.083336
Uver zde	0.018893
Zamestnani	0.092835
Vel domacnosti	0.077071
Telefon	0.075012
Cizinec	1.000000

Jak je vidět z hodnot výběrové korelační matice, znaky "Ucel", "Rodinny stav", "Vel domacnosti" a "Cizinec" jsou jen nízce korelované s většinou ostatních znaků. Následující tabulka vlastních čísel výběrové korelační matice ukazuje, jakou část celkového rozptylu popisují jednotlivé hlavní komponenty. K tomu, abychom popsali alespoň 90 % celkového rozptylu potřebujeme uvažovat 17 hlavních komponent, což není příliš uspokojivý výsledek. První tři hlavní komponenty reprezentují jen necelých 30 % celkového rozptylu, a tak jakákoliv (3 a méně rozměrná) grafická reprezentace našich pozorování ztrácí smysl.

Pořadí	**Vl. čísla**	**Individuální procento**	**Kumulativní procento**
1	2.538127	12.09	12.09
2	2.123567	10.11	22.20
3	1.485116	7.07	29.27
4	1.365792	6.50	35.77
5	1.231385	5.86	41.64
6	1.183251	5.63	47.27
7	1.142412	5.44	52.71
8	1.090361	5.19	57.90
9	1.004269	4.78	62.69
10	0.926294	4.41	67.10
11	0.874881	4.17	71.26
12	0.817028	3.89	75.15
13	0.789802	3.76	78.92
14	0.755604	3.60	82.51
15	0.709184	3.38	85.89
16	0.643904	3.07	88.96
17	0.579377	2.76	91.72
18	0.537647	2.56	94.28
19	0.489625	2.33	96.61
20	0.468257	2.23	98.84
21	0.244116	1.16	100.00

Podívejme se na na tabulku ukazující nám koeficienty lineárních kombinací pro určení prvních pěti hlavních komponent. Jedná se o prvních pět sloupců matice $\boldsymbol{V}^T$. Vezmeme-li pro určení důležitosti jako hraniční mez hodnotu koeficientu rovnu $0,33$ (v absolutní hodnotě), vidíme, že na první hlavní komponentu mají zásadní vliv znaky "Splatnost", "Vyse pujcky", "Nejc aktiva", "Zamestnani" a "Telefon". O těchto znacích lze říci, že mají první pořadí důležitosti pro příspěvek k celkovému rozptylu. Na druhou komponentu mají zásadní vliv "Kredit", "Stav konta" a "Moralka". V případě

dalších hlavních komponent se již některé znaky opakují (byly "důležité" již pro nějaké hlavní komponenty s nižším pořadovým číslem) a nebo již nemá daná hlavní komponenta takovou váhu (reprezentuje jen malou část celkového rozptylu). O ostatních znacích by se chtělo říct, že nemají výrazný vliv na celkový rozptyl. Tento závěr by však nebyl příliš podložený, neboť první hlavní komponenty nám nereprezentují tak velkou část celkového rozptylu, jakou bychom si přáli.

Proměnná	v_1	v_2	v_3	v_4	v_5
Kredit	0.095826	**-0.388367**	-0.306861	0.103056	0.216386
Stav konta	-0.020192	**-0.342291**	**-0.387899**	-0.047905	0.205011
Splatnost	**0.417210**	-0.208413	0.046805	-0.118998	0.064385
Moralka	-0.013146	**-0.423892**	-0.082691	0.074007	**-0.492304**
Ucel	-0.100946	0.059085	0.016404	-0.132887	0.013359
Vyse pujcky	**0.424141**	-0.173375	0.211208	**-0.402470**	0.067997
Vklady	-0.086463	-0.210975	-0.248016	-0.000799	**0.414917**
Delka zam	-0.167394	-0.320869	0.190643	-0.064017	0.110537
Pomer k prijmu	-0.039490	-0.099664	0.253968	**-0.562948**	0.016756
Rodinny stav	-0.023480	-0.152595	0.197373	0.087840	0.207815
Garance	0.120723	0.065962	0.180654	0.279721	-0.020716
Doba byd	-0.145337	-0.174857	0.191591	-0.062591	-0.044762
Nejc aktiva	**-0.411973**	0.077520	0.086742	-0.085534	0.059814
Vek	-0.168076	-0.320441	0.248096	0.047434	0.095462
Dalsi uvery	0.096622	-0.103206	-0.296142	-0.095842	-0.210783
Typ bydleni	-0.283489	-0.125832	0.266870	-0.025189	0.170076
Uver zde	-0.071667	-0.316426	0.143874	0.140765	**-0.543143**
Zamestnani	**-0.349033**	0.033372	-0.211925	-0.132743	-0.060434
Vel domacnosti	0.030292	0.125333	-0.312403	**-0.371069**	-0.167624
Telefon	**-0.334355**	-0.071389	-0.194603	-0.015025	-0.039314
Cizinec	-0.159492	0.050047	-0.041645	**-0.426824**	-0.123355

Závěr: K variabilitě souboru výrazně přispívají znaky "Splatnost", "Vyse pujcky", "Nejc aktiva" a "Zamestnani", které silně ovlivňují hodnotu první hlavní komponenty. Za významné považujeme ještě znaky "Kredit", "Stav konta" a "Moralka" silně ovlivňující hodnotu druhé hlavní komponenty. Vzhledem k nízkému procentu vysvětlení variability prvními hlavními komponentami nelze přímo říci, že by některé znaky byly pro celkový rozptyl souboru nevýznamné. Možnými kandidáty na takové méně významné znaky jsou "Ucel" a "Rodinny stav", které jsou jen málo korelované s většinou ostatních znaků a zároveň mají nízké koeficienty významnosti pro hodnoty prvních pěti hlavních komponent.

Kapitola 2

Vyšetřování struktury závislosti v množině proměnných

2.1 Korelační matice nejvýznamnějších znaků

Jak plyne ze závěrů analýzy hlavních komponent, významný podíl na variabilitě dat mají znaky "Splatnost", "Vyse pujcky", "Nejc aktiva", "Zamestnani", "Kredit", "Stav konta" a "Moralka". Abychom lépe porozumněli vzájemným vztahům těchto znaků, je vhodné podívat se na jejich výběrovou korelační matici.
Poznámka: V tuto chvíli tedy uvažujeme náhodný výběr o $n = 1000$ sedmirozměrných pozorováních.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1.000	**0.351**	0.208	0.229	0.100	-0.143	-0.033
X_2	**0.351**	1.000	0.070	0.192	0.015	-0.032	0.041
X_3	0.208	0.070	1.000	0.075	**0.632**	-0.300	-0.212
X_4	0.229	0.192	0.075	1.000	0.037	-0.054	0.010
X_5	0.100	0.015	**0.632**	0.037	1.000	**-0.320**	-0.300
X_6	-0.143	-0.032	-0.300	-0.054	**-0.320**	1.000	0.276
X_7	-0.033	0.041	-0.212	0.010	-0.300	0.2761	1.000

Legenda: X_1 = Kredit, X_2 = Stav konta, X_3 = Splatnost, X_4 = Moralka, X_5 = Vyse pujcky, X_6 = Nejc aktiva, X_7 = Zamestnani

Jak vidíme z korelační matice, výrazně korelované jsou znaky "Splatnost" a "Vyse pujcky". Oba tyto znaky mají vysoký násobící koeficient v první hlavní komponentě. Pro potřeby případného regresního modelu bychom tedy mohli uvažovat pouze jeden z těchto znaků. Hodnotou větší než 30% jsou dále korelované znaky "Vyse pujcky" a "Nejc aktiva" v první hlavní komponentě a "Kredit" a "Stav konta" v druhé.

2.2 Kontingenční tabulky

Všechny znaky databáze, se kterou pracujeme, jsou kategoriálního charakteru, tzn. jedná se o veličiny nabývající konečného počtu hodnot, které mají význam kódování určitých kategorií.
Nyní zmíníme metodu umožňující testování nezávislosti kategoriálních proměnných.

Nechť X nabývá hodnot $x_1, x_2, ..., x_k$ s pravděpodobnostmi $p_{i.} = P(X = x_i),\ i = 1, 2, ..., k$ a Y nabývá hodnot $y_1, y_2, ..., y_r$ s pravděpodobnostmi $p_{.j} = P(Y = y_j),\ j = 1, 2, ..., r$. Označme $p_{ij} = P(X = x_i, Y = y_j),\quad p_{ij}$ nazveme sdružené pravděpodobnosti, $p_{i.}$ a $p_{.j}$ pravděpodobnosti marginální. Platí:

$$p_{i.} = \sum_{j=1}^{r} p_{ij},\ p_{.j} = \sum_{i=1}^{k} p_{ij}$$

Kontingenční tabulkou rozumíme tabulku sdružených a marginálních pravděpodobností ve tvaru:

i/j	y_1	y_2	$\cdots$	y_r	**Součet**
x_1	p_{11}	p_{12}	$\cdots$	p_{1r}	$p_{1.}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2r}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	p_{k1}	p_{k2}	$\cdots$	p_{kr}	$p_{k.}$
Součet	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.r}$	1

Naší snahou je vyšetřit, zda jsou veličiny X a Y nezávislé. Testujeme tedy hypotézu

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j},\ \forall i, j,\ i = 1, 2, ...k,\ j = 1, 2, ..., r$$

Jelikož jednotlivé pravděpodobnosti v praxi neznáme, musíme je odhadnout pomocí četností.
Označme n_{ij} četnost jevu $X = x_i \wedge Y = y_j$, $n_{i.}$, $n_{.j}$ odpovídající marginální četnosti a N celkový počet pozorování. Pro N platí:

$$N = \sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij} = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{r} n_{.j}$$

Pravděpodobnosti $p_{i.}$ a $p_{.j}$ odhadneme následujícím způsobem:

$$\hat{p}_{i.} = \frac{n_{i.}}{N},\ \hat{p}_{.j} = \frac{n_{.j}}{N}$$

Odhadem p_{ij} za platnosti hypotézy H_0 jsou hodnoty

$$\hat{p}_{ij} = \frac{n_{i.} \cdot n_{.j}}{N^2}$$

Protějškem naměřených četností n_{ij} jsou teoretické četnosti Np_{ij}. Ty jsou za platnosti hypotézy odhadnuty jako

$$\frac{n_{i.} \cdot n_{.j}}{N}$$

Testovou statistikou pro test hypotézy o nezávislosti X a Y je veličina C:

$$C = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n.j/N}$$

Veličina C má pro dostatečně velké hodnoty teoretický četností Np_{ij} (doporučeno $Np_{ij} > 5$) rozdělení $\chi^2_{(k-1)(r-1)}$.

Hypotézu tedy zamítáme, pokud C je větší než $(1 - \alpha)$-tý kvantil daného rozdělení ($C > \chi^2_{(k-1)(r-1)}(1 - \alpha)$), kde α je hladina testu.

Poznámka: Nejsou-li odhady teoretických četností $n_{i.}n_{.j}/N > 5$, doporučuje se sloučení řádků nebo sloupců kontingenční tabulky četností (odpovídá sloučení kategorií veličiny X) a test se pro takové dvojice provede znovu.

2.2.1 Zpracování dat programem NCSS

Jak bylo řečeno výše, jako nejvýznamnější z hlediska variability souboru se jeví znaky "Splatnost", "Vyse pujcky", "Nejc aktiva", "Zamestnani", "Kredit", "Stav konta" a "Moralka". Určitou představu o jejich vzájemné nezávislosti jsme si udělali již z výběrové korelační matice. Nyní se podívejme, jak dopadlo testování hypotéz o nezávislosti jednotlivých dvojic na základě kontingenčních tabulek.

Na příkladu dvojice "Kredit" a "Stav uctu" ukažme celý výstup programu NCSS. Jako první uvádíme kontingenční tabulku četností, tedy tabulku ukazující četnosti n_{ij} respektive $n_{i.}$ a $n_{.j}$:

kredit **Stav uctu**	0	1	**Součet**
1	135	139	274
2	105	164	269
3	14	49	63
4	46	348	394
Součet	300	700	1000

Druhým výstupem je kontingenční tabulka relativních četností:

kredit **Stav uctu**	0	1	**Součet**
1	13.5%	13.9%	27.4%
2	10.5%	16.4%	26.9%
3	1.4%	4.9%	6.3%
4	4.6%	34.8%	39.4%
Součet	30%	70%	100%

Důležitým ukazatelem je kontingenční tabulka očekávaných četností zobrazující prvky $\frac{n_{i.} \cdot n_{.j}}{N}$, respektive $n_{i.}/N$ a $n_{.j}/N$:

kredit **Stav uctu**	0	1	**Součet**
1	82.2	191.8	274.0
2	80.7	188.3	269.0
3	18.9	44.1	63.0
4	118.2	275.8	394.0
Součet	300.0	700.0	1000.0

Posledním a zároveň nejdůležitějším výstupem je výsledek testování hypotézy o nezávislosti těchto dvou znaků. Kolonka "Závěr" uvádí rozhodnotí zamítnutí nebo nezamítnutí hypotézy o nezávislosti daných dvou znaků na hladině 5%:

Chí kvadrát	123, 720944
Stupně volnosti	3
P-hodnota	0.000000
Závěr	Zamítnout Ho

Z prostorových důvodů zde nebudeme uvádět kontingenční tabulky pro všechny dvojice nejvýznamnějších znaků. Pro představu si uveďme alespoň tabulku výsledků testů nezávislosti jednotlivých dvojic. Z znamená zamítnutí hypotézy na hladině 5%, N její nezámítnutí.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1		Z	Z	Z	Z	Z	N
X_2	Z		N	Z	N	Z	Z
X_3	Z	N		Z	Z	Z	Z
X_4	Z	Z	Z		Z	N	N
X_5	Z	N	Z	Z		Z	Z
X_6	Z	Z	Z	N	Z		Z
X_7	N	Z	Z	N	Z	Z	

Legenda: X_1 = Kredit, X_2 = Stav konta, X_3 = Splatnost, X_4 = Moralka, X_5 = Vyse pujcky, X_6 = Nejc aktiva, X_7 = Zamestnani

U některých dvojic znaků bylo z důvodu nesplnění požadavku $n_{i.}n_{.j}/N > 5,\ \forall i,j$ nutno sloučit některé kategorie tak, aby test splňoval daný předpoklad. Jak vidíme hypotézu o nezávislosti nezamítáme u dvojic Kredit-Zamestnani, Zamestnani-Moralka, Stav konta-Splatnost, Moralka-Nejc aktiva a Vyse pujcky-Stav konta , které mají zároveň i velmi nízkou výběrovou korelaci.

2.3 Grafické modely

2.3.1 Teoretické zázemí

Další ukázkou metod vícerozměrné statistické analýzy jsou grafické modely. Protože se jedná o metodu založenou na teorii grafů, uveďme si nejprve několik pojmů používaných v tomto odvětví matematiky.

Grafem rozumíme dvojici (K, E), tedy množinu vrcholů a hran spojujících vrcholy z K. Grafy dělíme na orientované (kde bereme v potaz také směr hrany) a neorientované. My s orientovanými grafy pracovat nebudeme. Vrcholy, mezi kterými je hrana, nazýváme spojené. Graf je úplný, pokud v něm je každý vrchol spojen s každým. Graf G_1 nazveme podgrafem grafu G_2, pokud $K_1 \subset K_2$ a zároveň $E_1 \subset E_2$. Podgraf indukovaný množinou $K_0 \subset K$ je graf, který vznikne z G vynecháním vrcholů z $(K - K_0)$ a všech hran do nich vedoucích. Klika je taková podmnožina vrcholů $K_{kl} \subset K$, která indukuje úplný podgraf a zároveň je maximální v tom smyslu, že $(K_0 + k)$, kde k je libovolný vrchol z $K - K_0$, již úplný podgraf neindukuje.

My budeme nadále pracovat zejména s tzv. grafy podmíněné nezávislosti. Pro úplnost dodejme, že náhodné veličiny X a Y jsou při pevném Z podmíněně nezávislé (značíme $X \perp Y|Z$), pokud se jejich sdružená podmíněná hustota rovná součinu marginálních podmíněných hustot X a Y. Tedy:

$$f_{XY|Z}(x, y; z) = f_{X|Z}(x; z) f_{Y|Z}(y; z),\ \forall x, y;\ z : f_Z(z) > 0$$

Pro kategoriální veličiny je podmíněná nezávislost definována vztahem. (při konvenci $p_X(x) = P(X = x), p_{XY|Z}(x, y; z) = P(X = x, Y = y|Z = z)$ atp.):

$$X \perp Y|Z \iff p_{XY|Z}(x, y, z) = p_{X|Z}(x; z) p_{Y|Z}(y; z),\ \forall x, y, z.$$

Předpokládáme, že $P_Z(z) \neq 0$ pro všechna z z množiny hodnot veličiny Z.

Grafem podmíněné nezávislosti rozumíme graf G, kde za K vezmeme množinu složek náhodného vektoru $\boldsymbol{X} = X^1, X^2, ..., X^p$ a za množinu E hrany spojující vrcholy, které nejsou podmíněně nezávislé. Pro dané vrcholi $i, j \in K$ tedy neplatí $\boldsymbol{X}^i \perp \boldsymbol{X}^j | \boldsymbol{X}^{K-\{i,j\}}$. Zde $\boldsymbol{X}^a$ představuje podvektor vzniklý z $\boldsymbol{X}$ ponecháním pouze těch složek, které mají indexy v množině a.

Grafický model je rodina rozdělení vektoru $\boldsymbol{X}$ splňujících podmíněné nezávislosti dané grafem G. Je-li graf G úplný, nazveme grafický model saturovaný.
Cílem této metody je nalezení takového grafického modelu, který nejlépe popisuje strukturu podmíněných nezávislostí v naměřených datech. K těmto účelům nám poslouží deviance - testová statistika pro ověření shody uvažovaného modelu s daty.

Nechť $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ je realizace náhodného vektoru $\boldsymbol{X}$. Označme $p(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x})$. Logaritmickou věrohodnostní funkcí nazveme funkci:

$$l(p; \boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n) = \sum_x n(x) \dot{l}n(p(x)), \;\; n(x) = \sum_{i=1}^{n} \delta(x, \boldsymbol{X}_i),$$

$n(x)$ je četnost hodnoty x v datech a zápis logaritmické věrohodnostní funkce můžeme zkrátit: $l(p; X_1, X_2, ..., X_n) = l(p; n)$. Pravděpodobnosti $p(x)$, četnosti $n(x)$ pro všechna možná x reprezentuje p-rozměrná kontingenční tabulka.

Deviance

Mějme graf podmíněné nezávislosti G_m, který je podgrafem úplného grafu G_s se shodnou množinou vrcholů ($K_m = K_s$). V grafu G_m chybí oproti G_s k hran. Deviancí grafického modelu M určeného grafem G_m nazveme veličinu:

$$dev^{(k)} = 2[l(\hat{p}^S; n) - l(\hat{p}^M; n)] = 2[l(\frac{n}{N}; n) - l(\hat{p}^M; n)] = 2 \sum_x n(x) \frac{\dot{n(x)}}{N \hat{p}^M(x)},$$

kde $\hat{p}^S$ je maximálně věrohodný odhad $p(x)$ v saturovaném modelu určeném G_s a $\hat{p}^M$ je maximálně věrohodný odhad $p(x)$ v modelu M.
Odhadem $p(x)$ v saturovaném modelu je relativní četnost $\frac{n(x)}{N}$. Odhad parametrů p v modelu M se provádí pomocí tzv. IPF iteračního algoritmu,

který je blíže popsán v [6, str. 27]. Tento algoritmus pracuje s klikami grafu, kterým je definován model M.

Jsou-li G_1 a G_2 podgrafy G_s ve stejném smyslu jako výše, přičemž platí $k_2 > k_1$, $G_2 \subset G_1 \subset G_s$ definujeme diferenci deviancí modelů M_2 a M_1 náležících grafům G_2, resp. G_1 jako:

$$dev^* = dev^{k_2} - dev^{k_1}.$$

V případě, že G_2 vznikne z G_1 vynecháním jedné hrany $\{i, j\}$, nazýváme diferenci deviancí $dev^*_{ij} = dev^{k_2} - dev^{k_1} = dev^{k_1+1} - dev^{k_1}$ deviancí vynechané hrany.
Naopak, pokud G_2 vznikne z G_1 přidáním hrany, nazýváme diferenci deviancí $dev^{k_1} - dev^{k_2} = dev^{k_1} - dev^{k_1-1}$ deviancí přidané hrany.

Z pohledu výpočtů je dobré uvést ještě jiný vzorec pro diferenci deviancí:

$$dev^* = 2 \sum_x n(x) \frac{\hat{p}^1\dot{(}x)}{\hat{p}^2(x)}$$

kde kde $\hat{p}^1$ je maximálně věrohodný odhad $p(x)$ v modelu určeném G_1 a $\hat{p}^2$ je maximálně věrohodný odhad $p(x)$ v modelu určeném $G_2 \subset G_1$.

Jak bylo řečeno výše, deviance je testovou statistikou pro testování shody grafického modelu s daty. Za platnosti modelu M odpovídajícímu případu $X^a \perp X^b | X^c$, kde r_a, r_b, r_c jsou počty kategorií vektoru X^a, respektive X^b a X^c, má asymptoticky rozdělení $\chi^2_{r_c(r_a-1)(r_b-1)}$ a platí:

$$dev(X^a \perp X^b | X^c) = 2 \sum n_{abc} \cdot ln \frac{n_{abc} n_c}{n_{ac} n_{bc}}.$$

[6, str. 24]

Diference deviancí $dev^* = dev^{k_2} - dev^{k_1}$ má asymptoticky rozdělení $\chi^2_{s_2-s_1}$, kde s_2, s_1 jsou počty stupňů volnosti v rozdělení dev^{k_2}, respektive dev^{k_1}.
[6, str. 25]

Volba vhodného grafického modelu

Způsobů volby vhodného grafického modelu je mnoho. Jednou z možností je testování všech možných grafických modelů, které pro danou množinu K připadají v úvahu. Tento způsob je sice pro mohutnější množiny K příliš časově náročný, pro malé množiny K ale přesto použitelný.

2.3.2 Zpracování dat

Zpracovávaná data jsou shodná s těmi, která byla analyzována v předchozích kapitolách. V návaznosti na výsledky analýzy hlavních komponent jsme se nejprve zaměřili na znaky, které nám vyšly jako nejvýznamnější pro první hlavní komponentu. Jsou to znaky "Splatnost", "Vyse pujcky", "Nejc aktiva" a "Zamestnani".

Vzhledem k velkému počtu kategorií těchto proměnných a související složitosti výpočtů jsme přistoupili ke sloučení kategorií podle následujícího schématu:

Náhodná veličina X_1 odpovídá znaku "Nejc aktiva"
$X_1 = 0$, pokud Nejc aktiva = 1
$X_1 = 1$, jinak

Náhodná veličina X_2 odpovídá znaku "Zamestnani"
$X_3 = 0$, pokud Zamestnani = 1, 2
$X_3 = 1$, jinak

Náhodná veličina X_3 odpovídá znaku "Splatnost"
$X_2 = 0$, pokud Splatnost = 9, 10
$X_2 = 1$, pokud Splatnost = 7, 8
$X_2 = 2$, jinak

Náhodná veličina X_4 odpovídá znaku "Vyse pujcky"
$X_4 = 0$, pokud Vyse pujcky = 9, 10
$X_4 = 1$, pokud Vyse pujcky = 7, 8
$X_4 = 2$, pokud Vyse pujcky = 5, 6
$X_4 = 3$, jinak

Máme tedy prostor kategorií:

$$H_4 = \{0,1\} \times \{0,1\} \times \{0,1,2\} \times \{0,1,2,3\}$$

Sestavme tabulku četností prvků z prostoru H_4:

		x_1	0			1			
$n_{1234}(x)$		x_2	0	1	2	0	1	2	
X_1 0	X_4 0		26	2	0	17	4	0	49
	1		35	16	0	50	48	4	153
	2		12	11	4	15	18	16	76
	3		0	0	1	0	0	3	4
1	0		17	4	0	33	12	1	67
	1		21	29	1	96	102	19	268
	2		6	17	14	28	134	102	301
	3		1	0	5	2	14	60	82
			118	79	25	241	332	205	1000

Jak je z této tabulky patrné, některé sdružené kategorie nejsou v naměřených datech vůbec zastoupené, to svědčí o silné vzájemné korelovanosti jednotlivých proměnných. Protože ani po snížení počtu kategorií nejsou splněny podmínky pro počítání s kontingenčními tabulkami (viz. minulá kapitola), mají následující výsledky jen informativní charakter.

Jelikož budeme zkoumat graf o čtyřech vrcholech, nabízí se celkem 64 grafických modelů, z nichž jeden je saturovaný. V případě zbylých 63 budeme vždy testovat hypotézu o dobré shodě daného grafu s daty oproti alternativě saturovaného modelu.
Pro výpočet konkrétních hodnot testových statistik a kritických hodnot příslušných rozdělení bylo použito softwaru přiloženého k práci [5] naprogamovaného v prostředí Excel.

Výsledkem bylo zamítnutí všech hypotéz o vhodnosti jednotlivých nesaturovaných grafů. To svědčí o vzájemné závislosti všech prvků, které jsou významné pro první hlavní komponentu.
Poznamenejme, že závislost proměnných “Splatnost”, “Vyse pujcky”, “Nejc aktiva” a “Zamestnani” se prokázala i v analýze kontingenčních tabulek

(hypotéza nezávislosti byla zamítnuta pro všechny dvojice) a že jednotlivé dvojice vykazují i nezanedbatelnou korelaci.

Na závěr si uveďme ještě výsledky z práce [5], kde byla zkoumána vhodnost grafických modelů čtveřice proměnných "Kredit"(1), "Vyse pujcky"(2), "Splatnost"(3) a "Vek"(4).

Výsledkem bylo nezamítnutí hypotéz o vhodnosti grafů s hranami (postupně):

1-2, 1-3, 1-4, 2-3 a 2-4;
1-2, 1-3, 1-4, 2-3 a 3-4;
1-3, 1-4, 2-3, 2-4 a 3-4;
1-2, 1-3, 1-4, 2-3;
1-3, 1-4, 2-3, 3-4;
1-3, 1-4, 2-3;

V žádném nezamítnutém grafickém modelu tedy nechybí hrany 1-3, 1-4 a 2-3 značící závislost znaků Kredit-Splatnost, Kredit-Vek a Vyse pujcky-Splatnost.

Závěr

Cílem této práce bylo popsat některé metody vícerozměrné statistické analýzy a ty pak aplikovat na finanční data z oblasti kreditscoringu.

Na základě analýzy hlavních komponent jsme usoudili, že nejvýznamnějšími znaky z hlediska jejich vlivu na celkovou variabilitu souboru jsou "Splatnost", "Vyse pujcky", "Nejc aktiva", "Zamestnani" a "Telefon", které mají největší vliv na hodnotu první hlavní komponenty, a dále 'Kredit", "Stav konta" a "Moralka".
Když jsme zkoumali výběrovou korelační matici těchto znaků (s výjimkou znaku "Telefon"), nevýrazně korelované se jevily dvojice znaků "Kredit" - "Zamestnani", "Stav konta" - "Splatnost", "Stav konta" - "Vyse pujcky", "Stav konta - Nejc Aktiva", "Splatnost" - "Moralka", "Moralka" - "Vyse pujcky", "Moralka" - "Nejc Aktiva", "Stav konta" - "Zamestnani" a "Moralka" - "Zamestnani". Při testování metodou kontingenčních tabulek jsme hypotézu o vzájemné nezávislosti nezamítli u dvojic "Kredit-Zamestnani", "Stav konta"-"Splatnost", "Stav konta"-"Vyse pujcky", "Moralka"-"Nejc aktiva" a "Moralka"-"Zamestnani".
Výsledky grafických metod ukázaly na vzájemnou závislost znaků "Splatnost", "Vyse pujcky", "Nejc aktiva" a "Zamestnani" významných pro hodnotu první hlavní komponenty, což potvrdila i korelační analýza a testy nezávislosti v kontingenčních tabulkách.

Literatura

[1] Ajvazjan S., Bežajevová Z., Staroverov O.: *Metody vícerozměrné analýzy*, SNTL, Praha, 1981.

[2] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2005

[3] Bican L.: *Lineární algebra a geometrie*, ACADEMIA, Praha, 2000.

[4] Hebák P., Hustopecký J.: *Vícerozměrné statistické metody s aplikacemi*, SNTL-Alfa, Praha, 1987.

[5] Kýpeť M.: *Diplomová práce: Bayesovská analýza diskrétních finančních dat*, MFF UK, Praha, 2005.

[6] Svobodová B.: *Diplomová práce: Analýza kategoriálních finančních dat*, MFF UK, Praha, 2003.

[7] Universita Mnichov: *www.stat.uni-muenchen.de/service/datenarchiv*.