Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

DISERTAČNÍ PRÁCE

Vladislav Chýna

Grafické modely pro analýzu spojitých finančních dat

Katedra ekonometrie
Školitel: Doc. RNDr. Václava Pánková, CSc.
Studijní program: Matematické metody v ekonomii

Rád bych na tomto místě poděkoval své školitelce, Doc. RNDr. Václavě Pánkové, CSc., za cenné připomínky a rady, jakož i za pomoc při řešení problémů vznikajících při psaní mé disertační práce.

Můj velký dík patří také dr. Jitce Zichové a dr. Františku Matúšovi za poskytnuté konzultace a zapůjčenou literaturu týkající se grafických modelů.

Rovněž bych chtěl poděkovat prof. Janovi Pelikánovi a prof. Petru Fialovi za přečtení práce před „malou obhajobou“ a zejména pak za jejich připomínky a návrhy.

Prohlašuji, že doktorskou práci na téma „Grafické modely pro analýzu spojitých finančních dat“ jsem vypracoval samostatně. Použitou literaturu a podkladové materiály uvádím v přiloženém seznamu literatury.

V Praze dne 22.8.2006 Vladislav Chýna

Obsah

Název práce: Grafické modely pro analýzu spojitých finančních dat
Autor: Vladislav Chýna
Katedra: Katedra ekonometrie
Školitel: Doc. RNDr. Václava Pánková, CSc.
e-mail školitele: pankova@vse.cz

Abstrakt: Grafické modely jsou vhodným nástrojem statistické analýzy. V poslední době byly aplikovány rovněž ve financích. Práce pojednává o grafických modelech určených pro zpracování dat pocházejících z normálního rozdělení. Jsou zde popsány možnosti testování shody modelů s daty a čtyři selekční algoritmy (včetně odvození vztahů mezi různými STOP pravidly) pro výběr vhodného grafu, který dobře reprezentuje konkrétní data.

Tyto dosud nepříliš známé postupy jsou pak použity k zodpovězení otázek z oblasti českého i mezinárodního kapitálového trhu (zkoumání provázanosti odvětvových indexů BCPP, globalizace světových akciových trhů, provázanosti měnových kurzů) a rovněž k testování hypotéz pro vysvětlení časové struktury úrokových sazeb v České republice.

Vytvořené programy ve formátu Mathematica 4 lze nalézt v příloze.

Klíčová slova: Orientovaný a neorientovaný graf, gaussovský grafický model, informační divergence, backward a forward selekční algoritmus, odvětvové indexy BCPP, světové akciové indexy, měnové kurzy, mechanismus determinace úrokových sazeb, SVAR grafický model, PRIBOR

ABSTRACT

Graphical models for the analysis of the continuous financial data

Graphical models are suitable tools for statistical analysis. They are commonly used in the sphere of finance. The study deals with graphical models based on the normal distribution. Some possibilities of fitting models to the data and four selection procedures (including deduction relations among various STOP rules) are described here.

These not yet much known strategies are then applied to answer questions from the area of Czech and international capital markets (investigation of cohesion of branch business indices, globalization of world share markets, cohesion of currency rates). One application tests the time structure of the interest rates in Czech Republic too.

Created programs in SW Mathematica 4 can be found in appendix.

Keywords: Directed and undirected graph, Gaussian graphical model, I-divergence, backward and forward search strategy, Branch business indices, world indices, currency rates, SVAR graphical model, PRIBOR

ABSTRAKT

Grafische Modelle für die Analyse von stetigen Daten

Grafische Modelle sind günstige Instrumente für statistische Analysen. Sie waren in der letzten Zeit auch in Finanzen angewandt. Die Arbeit behandelt sich mit grafischen Modellen vorgesehene für Daten, die aus der normalen Distribution stammen. Man beschreibt hier verschiedene Testmöglichkeiten des Einverständnisses den Modellen mit Daten und vier Selektionsalgorithmen (einschlielich der Ableitung von Beziehungen zwischen verschiedenen Stoppregeln) für die Auswahl vom geeigneten Graf, der die konkreten Daten gut repräsentiert.

Diese bisher nicht viel bekannte Vorgänge sind dann benutz um Fragen von dem Bereich der tschechischen und auch internationalen Kapitalmarkt zu antworten (die Untersuchung von Anknüpfung an börsen Branchenindexe, die Globalisierung von internationale Kapitalmarkte, die Anknüpfung an Währung). Eine Applikation testet auch die Hypothesen von zeitiger Struktur von Zinsraten in der Tschechischen Republik.

Gebildete Programme in SW Mathematica 4 kann man in der Anlage befinden.

Schlagwörter: Gerichteter und ungerichteter Graph, Gaussian grafisches Modell, Informationsdivergenz, backward und forward Selektionsalgorithmus, börsen Branchenindexe, internationale Aktienindexe, Währungskurse, SVAR grafisches Modell, PRIBOR

Úvod

Grafické modely jsou užitečným nástrojem statistické analýzy dat. Umožňují studovat strukturu závislosti v souborech proměnných a v posledních letech byly úspěšně aplikovány ve finanční analýze. Práce pojednává o grafických modelech určených pro zpracování dat pocházejících z normálního rozdělení. Jejím hlavním cílem bylo pomocí vytvořeného programového vybavení pro selekci grafických modelů zodpovědět některé otázky ze světa financí.

V první kapitole jsou popsány otázky, na které se pokusíme najít v této práci odpovědi, jakož i další použití grafických modelů, jejich výhody a nevýhody.

V kapitole 2, 3, 4 a 5 jsou shrnuty základní poznatky z teorie grafů, statistiky, některé vlastnosti náhodných vektorů a mnohorozměrného normálního rozdělení. Je zde také stručně popsána metoda maximální věrohodnosti v gaussovských grafických modelech a základní pojmy o VAR modelech časových řad.

Šestá kapitola se zabývá pojmy z teorie informace (entropie a divergence), které nachází v grafických modelech široké uplatnění.

Sedmá kapitola obsahuje popis procedur pro vstupní testy dat použitých v ilustračních příkladech, osmá kapitola se pak týká analýzy deviance. Jde o testovou statistiku pro výběr grafických modelů, které dobře reprezentují konkrétní data.

V deváté kapitole jsou popsány (a na příkladech demonstrovány) čtyři selekční algoritmy pro výběr grafických modelů. Jedná se o backward a forward algoritmy se stop pravidlem založeným na devianci vynechané/přidané hrany a se stop pravidlem založeným na celkové devianci. Jsou zde rovněž odvozeny vztahy mezi různými STOP pravidly. V závěru kapitoly je možno nalézt také krátké porovnání uvedených algoritmů.

Desátá kapitola je věnována rozšíření těchto postupů na VAR modely (tj. na spojení mnohorozměrných modelů časových řad a orientovaných grafů).

Jedenáctá kapitola je pak věnována použití těchto dosud nepříliš známých postupů na řešení konkrétních příkladů s českými i světovými finančními daty. Jedná se o zkoumání provázanosti odvětvových indexů BCPP, globalizace světových akciových trhů, provázanosti měnových kurzů a testování hypotéz pro vysvětlení časové struktury úrokových sazeb v České republice.

Důkazy vět jsou (s výjimkou důkazů vlastních) uváděny pouze odkazem na příslušnou literaturu.

V přílohách lze nalézt kromě zdrojových kódů programů také histogramy měsíčních logaritmických výnosů vybraných odvětvových indexů a výsledky testů normality a nezávislosti v příkladech použitých dat.

Kapitola 1

Proč vlastně grafické modely?

Představme si, že chceme odvodit zákonitosti platící pro studijní výsledky dosahované žáky v různých předmětech. K dispozici máme známky jednotlivých žáků ve vybraných předmětech a zajímá nás, zda lze z dobrého prospěchu žáka v jednom předmětu (například v matematice) usuzovat na dobré výsledky v předmětu jiném (například ve fyzice). Je pravda, že žáci hudebně nadaní jsou schopni logického myšlení potřebného v matematice (blíže například viz [31])?

Anebo nás zajímá otázka správného umístění investic. Vede investice do různých odvětví v České republice (do různých odvětvových indexů) k diversifikaci rizika, nebo jsou různá odvětví tak svázána, že je téměř jedno, kam investovat? Je index investičních fondů svázán s indexy vybraných průmyslových odvětví díky skutečnosti, že investiční fondy jsou vlastníky nezanedbatelného počtu akcií ostatních firem(viz kapitola 11.1.1)? Vyvíjí se bankovní sektor (index peněžnictví) v souladu s vývojem ostatních odvětví - banky přece žijí z toho, že půjčují peníze ostatním podnikům (viz kapitola 11.1.2)?

Můžeme se ptát dokonce z mezinárodního hlediska: opravdu platí, že se světové akciové trhy v posledních letech chovají stále globálněji? Promítají se změny na americké burze velmi rychle i do burz evropských (viz kapitola 11.2)?

A jak vypadá situace na poli měnových kurzů v Evropě - chovají se měny, které nejsou součástí eura, nezávisle na něm, nebo jsou s jeho vývojem stejně svázány, jako by byly jeho součástí? A co země, které by euro rády v nejbližší době zavedly (viz kapitola 11.3)?

Všechny tyto otázky jsou případem obecnější úlohy:
máme k dispozici data ve formě realizací nějakého náhodného vektoru a z těchto realizací chceme učinit závěr o struktuře vzájemné závislosti jeho složek.

První, co se nabízí, je samozřejmě **korelační koeficient**:

- snadno se spočítá (viz definice 3.14)
- má jasnou interpretaci (čím blíže je v absolutní hodnotě bližší číslu 1, tím je závislost silnější)
- existuje test jeho významnosti (viz věta 3.15)

- a možná ta nejdůležitější přednost - většině lidí je pojem korelace známý

Bohužel má i určité nevýhody:

pokud máme dvě veličiny, které nemají nic společného, ale existuje veličina třetí, která ovlivňuje obě, může se stát, že korelační koeficient vychází velmi blízko 1 (tzv. falešná korelace). K problémům pak dochází, když zmíněná závislost k třetí veličině přestává platit. Navíc nám uniká skutečná podstata vzájemných vztahů. Tak lze pomocí korelace potvrdit určité „pranostiky" - například o tom, že „když jde měsíc nahoru, více rostou houby". Vysvětlení by však mělo být takové, že „když jde měsíc nahoru, bývá více dešťových srážek a v důsledku zvýšené vlhkosti roste i více hub". Bohužel, takové vysvětlení nám korelační koeficient nenabídne.

Ještě nebezpečnější je bezmyšlenkovité použití korelací v analýze časových řad:

přestože zkoumáme vývoj 2 časových řad, které nemají nic společného, může se stát, že korelační koeficient vyjde blízko 1. To může být dáno „pouhým" společným trendem obou časových řad. Tak by se mohlo lehce stát, že například vývoj populace čápů v českých zemích a počet narozených dětí spolu úzce korelují (obojího je totiž v poslední době pomálu, i když zřejmě z úplně jiných příčin). Jak pak někomu vysvětlit, že děti opravdu nenosí čáp? V takových případech musíme před vlastním modelováním upravit časové řady tak, aby byly stacionární.

Známou nevýhodou je, že korelační koeficient (ať již výše zmíněný, nebo koeficient parciální korelace popsaný dále) je mírou pouze lineární závislosti; pokud máme nějakou nelineární závislost (byť i danou deterministickým vztahem - například: $y = x^2$), může koeficient vyjít nulový.

Tuto nevýhodu lze odstranit například použitím korelačního koeficientu ne přímo na data, ale na jejich pořadí - tzv. Spearmeanův korelační koeficient.

Zdá se, že většinu těchto problémů odstraní **koeficient parciální korelace**, a to pouze při drobných nevýhodách:

- i on se dá snadno spočítat (viz definice 3.35)
- má stejně jasnou interpretaci jako korelační koeficient (čím blíže je v absolutní hodnotě bližší číslu 1, tím je závislost silnější)
- existuje test jeho významnosti (viz věta 3.37)
- na rozdíl od korelačního koeficientu zkoumá vztah 2 zvolených veličin při pevných hodnotách ostatních uvažovaných proměnných

Problémem je samozřejmě výběr těch správných veličin, které mají nějakou souvislost.

Ve výše uvedeném příkladu s houbami by bylo zřejmě relevantní použít fázi měsíčního cyklu, počet srážek a něco, co vyjádří počet rostoucích hub. Pak bychom měli obdržet nulovou podmíněnou korelaci mezi počtem hub a fází měsíce. Naopak nenulové hodnoty by nabývala parciální korelace mezi dešťovými srážkami a počtem hub a rovněž mezi dešťovými srážkami a fází měsíce.

Zdá se tedy, že pokud máme data reprezentující několik proměnných (například akciových indexů různých odvětví), stačí spočítat vzájemné koeficienty podmíněné korelace, ty otestovat na významnost a o těch odvětvích, u kterých zamítneme hypotézu nulovosti koeficientu, prohlásit, že jsou ve vzájemném vztahu.

Bohužel problémem tohoto postupu je, že $N\times$ provedený test na pětiprocentní hladině významnosti rozhodně neznamená, že dosažená celková hladina je také pětiprocentní - viz kapitola 5.1.

Tento problém řeší právě grafické modely.

Samozřejmě lze použít i některé méně známé míry závislosti - například divergenci popsanou v kapitole 6.

Grafické modely mají ale i mnoho dalších oblastí použití - například:

- V kapitole 11.4 ukážeme spojení grafických modelů s **VAR modely** časových řad a pokusíme se zodpovědět otázku vztahu mezi velikostí úrokových měr na různou délku uložení peněz.
- Jako alternativu klasické **lineární regrese**. Jak je ukázáno na příkladu lineární regrese indexu PX50 na odvětvových indexech BCPP v [39], dává použití grafických modelů podobné výsledky jako klasická metoda nejmenších čtverců. Grafické modely však poskytují něco navíc - graf zároveň popisuje i strukturu podmíněných nezávislostí mezi vysvětlujícími proměnnými, což klasická regrese neumožňuje, a v případě potřeby tak vyžaduje použití dalších statistických metod.
- Při hodnocení žadatelů o úvěr (**credit scoring**). Grafické modely jsou zde alternativou tradičnějších metod (logistická regrese) i metod novějších (neuronové sítě). Dle některých autorů (viz [8], [30]) jsou dokonce vhodnější - umožní nejen vytvořit klasickou scoringovou kartu, ale globálněji pochopit chování žadatelů o úvěr a to využít například k predikci platební morálky u nově vznikajících produktů, k nimž zatím nemáme žádná historická data[1].
- Asi vrcholným použitím je nasazení výsledků z teorie grafických modelů v **expertních systémech** podporujících rozhodování:
  - V blízké budoucnosti[2] snad bude možno právě díky grafickým modelům poskytnout *lékařům* systém, který po zadání několika vstupních veličin sám určí s velkou pravděpodobností diagnózu pacienta a dokonce navrhne odpovídající způsob léčky.
  - Použití grafických modelů se nabízí dokonce již ve *školství*. Proč by mělo například zkoušení studentů pomocí testů probíhat klasickým způsobem, kdy každý student vypracuje celý test. Mnohem lepší by přece bylo využít test adaptivní, který by generoval další otázku podle správné/chybné odpovědi na otázku předchozí. K výsledku, který nás zajímá (tj. zda je student schopen aplikovat teoretické poznatky), bychom jednak došli mnohem rychleji a navíc by tento postup měl i psychologický efekt - student by nebyl urážen příliš jednoduchými otázkami[3].

[1] Tato vlastnost má rostoucí význam - díky zvyšující se konkurenci jsou bankovní instituce neustále nuceny nabízet nové produktu. Díky celé škále produktů se pak sama definice pojmu „špatný“ a „dobrý“ klient mění. Pokud například klient nezaplatí splátku úvěru a přitom je úvěrově pojištěn (úvěrové pojištění mu samozřejmě prodal sám věřitel), může být pro banku v globálu ještě stále klientem ziskovým.

[2] V pražském UTIA prý dokonce takový model již existuje.

[3] Příklad zabývající se testováním schopností počítat se zlomky řešený pomocí expertního systému HUGIN lze nalézt na [37].

– Známé je i použití v *troubleshootingu*, tj. jak máme postupovat, pokud se na nějakém zařízení (například počítačové tiskárně) vyskytne porucha.

Z výše uvedených aplikací by se mohlo zdát, že grafické modely nemají žádné nevýhody. Bohužel to není až tak docela pravda. Zatímco korelační koeficient spočte každý tabulkový kalkulátor (a i mnoho lepších kalkulaček), pro grafické modely je většinou nutné použít specializovaný software, případně si vytvořit vlastní (jak je tomu i v této práci). Při aplikacích grafických modelů navíc nesmíme zapomínat na psychologické pravidlo 7 ± 2, tj. skutečnost, že člověk je najednou schopen vnímat 7 ± 2 detailů.

Kapitola 2

Základní pojmy z teorie grafů

Dříve než přistoupíme k zodpovězení otázek z úvodní kapitoly, musíme zadefinovat pojmy, které budeme dále potřebovat. Protože se práce zabývá aplikací grafů ve statistice, podíváme se nejprve na několik definic z teorie grafů.

2.1 Neorientované grafy

Pokud budeme zkoumat pouze vazby mezi jednotlivými proměnnými a nebude nás zajímat příčina a důsledek, vystačíme s neorientovanými grafy:

Definice 2.1: *Neorientovaný graf*

Neorientovaný graf je uspořádaná dvojice (K, E), kde K je *množina vrcholů* (obvykle konečná podmnožina přirozených čísel) a E je podmnožinou množiny všech dvouprvkových podmnožin K. Prvky množiny E nazýváme *hrany* (ekvivalentně se používá pojmu linka). □

Úmluva 2.2:

Slovo neorientovaný budeme nadále většinou vynechávat. Pokud tedy v dalším textu píšeme o „grafu“, máme na mysli graf neorientovaný. □

Definice 2.3: *Spojené vrcholy*

Řekneme, že vrcholy v_i a v_j jsou *spojené* v grafu $G = (K, E)$, je-li mezi v_i a v_j hrana.
Tuto skutečnost budeme značit $\{v_i, v_j\} \in E$. □

Definice 2.4: *Sousedé uzlu*

Nechť $a \in K$. Množina

$$ne_G(a) = \{b \in K;\ \{a, b\} \in E\}$$

se nazývá *sousedé* uzlu a. □

Definice 2.5: *Cesta délky n, cyklus délky n*

Nechť $G = (K, E)$ je graf.
Posloupnost $\{v_0, \ldots, v_n\} \in K$ se nazývá *cesta v grafu G délky n z v_0 do v_n*, jestliže platí $\{v_{i-1}, v_i\} \in E$ pro $i = 1, \ldots, n$.
Pokud navíc platí $v_0 = v_n$, mluvíme o *cyklu délky n*. □

Definice 2.6: *Nejkratší cesta*

Nejkratší cesta spojující vrcholy v_0 a v_n je cesta s nejmenším možným počtem hran. □

Definice 2.7: *Navzájem dosažitelné vrcholy*

Řekneme, že vrcholy v_i a v_j jsou *navzájem dosažitelné* v grafu G, jestliže v grafu G existuje cesta z v_i do v_j. □

Definice 2.8: *Souvislý graf*

Řekneme, že graf je *souvislý*, jestliže jsou všechny jeho vrcholy navzájem dosažitelné. □

Definice 2.9: *Podgraf a faktor*

Nechť jsou dány dva grafy $G_1 = (K_1, E_1)$ a $G_2 = (K_2, E_2)$.
Pokud současně platí $K_1 \subseteq K_2$ a $E_1 \subseteq E_2$, pak řekneme, že graf G_1 je *podgrafem* grafu G_2.
Pokud navíc platí $K_1 = K_2$, pak řekneme, že graf G_1 je *faktorem* grafu G_2.
□

Definice 2.10: *Úplný (kompletní) graf*

Řekneme, že graf nebo podgraf je *úplný (kompletní)*, pokud je každý z jeho vrcholů spojen hranou se všemi jeho ostatními vrcholy. □

Definice 2.11: *Separační množina*

Řekneme, že množina $a \subseteq K$ *separuje* vrcholy v_i a v_j, pokud každá cesta z v_i do v_j obsahuje aspoň jeden vrchol z množiny a.
Množina a *separuje* dvě podmnožiny vrcholů $b \subseteq K$ a $c \subseteq K$, pokud separuje všechny dvojice vrcholů $v_i \in b$, $v_j \in c$. □

Definice 2.12: *Podgraf indukovaný množinou vrcholů*

Pograf G_a grafu G *indukovaný množinou* $a \subseteq K$ je graf vzniklý z G vyloučením všech vrcholů, které nepatří do a spolu se všemi hranami, jež spojují jiné vrcholy než vrcholy z a. □

Definice 2.13: *Klika*

Nechť $a \subseteq K$.
Řekneme, že množina a je *klika*, pokud indukuje úplný podgraf a pokud po přidání libovolného vrcholu indukuje podgraf, který již není úplný. □

Poznámka 2.14:

Říkáme, že klika je *maximální úplný podgraf.* □

Definice 2.15: *Komplementární graf*

Nechť $G = (K, E)$.
Komplementární graf získáme tak, že do G přidáme všechny chybějící hrany a poté původní hrany odstraníme. □

Definice 2.16: *Antiklika*

Anklika je klika komplementárního grafu. □

Pro reprezentaci v paměti počítače se nám mnohem lépe než kliky nebo antikliky hodí následující struktura:

Definice 2.17: *Matice sousednosti*

Nechť $G = (K, E)$, $K = \{v_0, \ldots, v_n\}$ je graf.
Matice sousednosti grafu G je matice $A_G = (a_{i,j})_{i,j=1}^{n}$ definovaná předpisem:

$$\begin{aligned} a_{i,j} &= 1 \; pro \; \{v_i, v_j\} \in E, \\ a_{i,j} &= 0 \; jinak. \end{aligned}$$

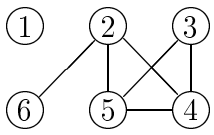
□

Poznámka 2.18:

Matice sousednosti je čtvercová matice, jejímiž prvky jsou 0 (pokud dva vrcholy nejsou spojeny hranou) a 1 (pokud dva vrcholy jsou spojeny hranou). Prvky na diagonále jsou nulové. Protože pracujeme s neorientovanými grafy, vystačíme s dolní (nebo ekvivalentně horní) trojúhelníkovou maticí. □

Příklad 2.19:

Mějme následující graf:



Množina vrcholů grafu je $K = \{1, 2, 3, 4, 5, 6\}$.
Množina hran grafu je $E = \{\{2, 4\}, \{2, 5\}, \{2, 6\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}$.
Vrcholy 3 a 6 spojují dvě *nejkratší cesty*: {3, 5, 2, 6} a {3, 4, 2, 6}.
Množina {4, 5} *separuje* vrcholy 6 a 3. Tyto vrcholy však separují rovněž všechny množiny obsahující vrchol 2.
Graf *není souvislý*, protože vrchol 1 není dostupný například z vrcholu 2.
Graf má následující *kliky*: {2, 4, 5}, {3, 4, 5}, {2, 6}, {1}.
Matice sousednosti grafu je:

$$A_G = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

□

Příklad 2.20:

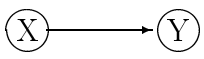
Podívejme se ještě na jeden příklad, na kterém si ukážeme 2 komplementární grafy:

První z grafů můžeme zapsat buď pomocí *klik*: {1, 2, 3}, {1, 3, 4}, nebo ekvivalentně pomocí *antiklik*: {1}, {3}, {2, 4}.
Druhý (tj. komplementární graf) má samozřejmě stejný zápis - pouze s opačným značením klik a antiklik. □

2.2 Orientované grafy

V mnoha aplikacích se vyskytují proměnné s různými rolemi (vysvětlované a vysvětlující, běžné a zpožděné,...). V takových případech již bohužel nevystačíme s neorientovanými grafy.

Situaci, kdy X ovlivňuje Y, ale Y neovlivňuje X, můžeme ale dobře vyjádřit orientovaným grafem:



Definice 2.21: *Orientovaný graf*

Orientovaný graf je uspořádaná dvojice (K, E), kde K je *množina vrcholů* (obvykle množina přirozených čísel) a E je podmnožinou množiny všech *uspořádaných* dvouprvkových podmnožin K. Prvky množiny E nazýváme *orientované hrany* (ekvivalentně se používá pojmu šipka a značení $v_i \to v_j$). □

Místo sousedů pak mluvíme o rodičích a dětech:

Definice 2.22: *Rodiče uzlu*

Nechť $a \in K$. Množina

$$pa_G(a) = \{b \in K;\ b \to a \in E\}$$

se nazývá *rodiče* uzlu a. □

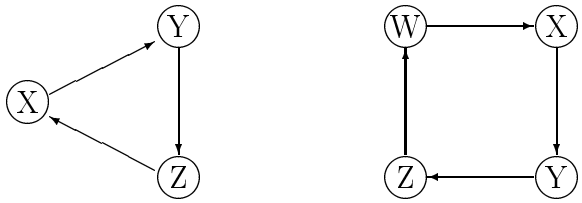
Definice 2.23: *Děti uzlu*

Nechť $a \in K$. Množina

$$ch_G(a) = \{b \in K;\ a \to b \in E\}$$

se nazývá *děti* uzlu a. □

Orientované grafy budeme používat (stejně jako neorientované) k popisu struktury podmíněných nezávislostí. Proto musíme vyloučit následující situace, které neumožňují dobře modelovat sdružené rozdělení příslušných proměnných pomocí podmíněných rozdělení.



Výše nakreslené obrázky znázorňují *orientovaný cyklus* - viz definice 2.5, ve které stačí vzít posloupnost vrcholů s ohledem na orientaci hran.

Naštěstí se ukazuje, že se bez cyklů obejdeme i v praktických aplikacích. Nadále tedy budeme pracovat s acyklickými grafy:

Definice 2.24: *Acyklický graf*

Orientovaný graf neobsahující cykly se nazývá *acyklický*. □

Podívejme se nyní na jednu z možností, jak formalizovaně zapsat, že graf neobsahuje cyklus. Na prvcích množiny K zavedeme relaci úplného uspořádání $\succ$ splňující
(pro $\forall i \in K, \forall j \in K$):

(i) buď $i \prec j$, nebo $i \succ j$,
(ii) relace $\succ$ je reflexivní,
(iii) relace $\succ$ je tranzitivní.

Pokud uvedené uspořádání aplikujeme na orientovaný graf, dostáváme, že každá hrana grafu může mít jen jednu možnou orientaci. Navíc platí následující lemma (důkaz je uveden v [31] - str. 72):

Lemma 2.25:

V orientovaném grafu jsou následující podmínky ekvivalentní:
(i) graf neobsahuje orientované cykly,
(ii) existuje úplné uspořádání vrcholů grafu. □

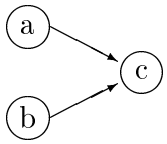
Spojení orientovaných a neorientovaných grafů se dosáhne pomocí tzv. *moralizace*. Nejde o nic jiného, než že každému orientovanému grafu přiřadíme graf neorientovaný, který je mu v jistém smyslu ekvivalentní. Nejprve však definujme asociovaný graf:

Definice 2.26: *Asociovaný neorientovaný graf*

Je-li $G^{\succ} = (K, E^{\succ})$ acyklický orientovaný graf, potom asociovaný neorientovaný graf definujeme jako $G^{asoc} = (K, E^{asoc})$ se stejnou množinou vrcholů, v němž orientované hrany jsou nahrazeny neorientovanými. □

Definice 2.27: *Imoralita, morální graf*

Buď $G^{\succ} = (K, E^{\succ})$ acyklický orientovaný graf.
Imoralitou v G budeme rozumět indukovaný podgraf G tvaru $a \to c \leftarrow b$ (tj. $a, b \in pa_G(c)$, přičemž $\{a, b\}$ ani $\{b, a\}$ není hrana v G) - tedy podgraf typu



Morálním grafem orientovaného grafu G se bude rozumět neorientovaný graf G^{mor} nad toutéž množinou vrcholů K, takový, že $\forall a, b \in K, a \neq b$ platí:
a, b je hrana v G^{mor}, pokud

- $a \to b$ nebo $a \leftarrow b$ v G
- existuje imoralita tvaru $a \to c \leftarrow b$

□

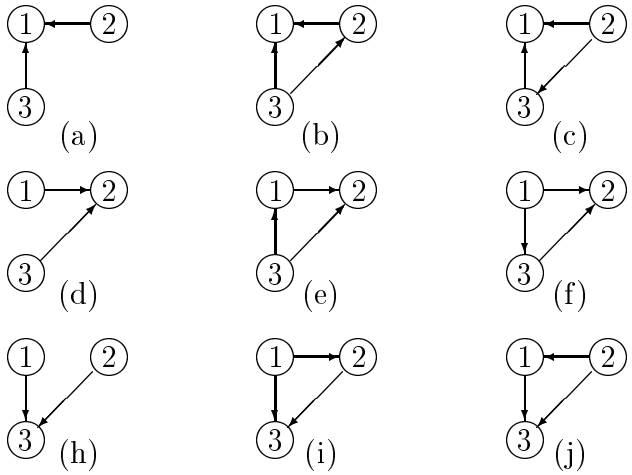
Morální graf získáme z acyklického orientovaného grafu tedy tak, že veškeré šipky nahradíme linkami (tj. vytvoříme asociovaný neorientovaný graf) a navíc přidáme hrany, které „sezdají“ rodiče daného uzlu - z toho ostatně také název morální (který zřejmě poprvé použil Lauritzen v článku [21]), protože v takovém grafu pouze sezdaní rodiče mohou mít dítě. Tento postup se nazývá *moralizace*.

Poznámka 2.28:

Místo pojmu graf neobsahuje imoralitu se můžeme setkat také s tvrzením, že graf splňuje Wermuthovu podmínku - viz např. [31]. □

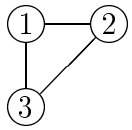
Příklad 2.29:

Podívejme se na následující třívrcholové grafy:



- U grafu (a) vrcholy $2, 3$ porušují Wermuthovu podmínku - při konstrukci morálního grafu přidáme hranu $(2, 3)$.
- U grafu (d) vrcholy $1, 3$ porušují Wermuthovu podmínku - při konstrukci morálního grafu přidáme hranu $(1, 3)$.
- U grafu (h) vrcholy $1, 2$ porušují Wermuthovu podmínku - při konstrukci morálního grafu přidáme hranu $(1, 2)$.
- Ostatní grafy neobsahují žádnou imoralitu, a tak morální graf získáme pouhým nahrazením šipek neorientovanými hranami.

Všech 9 uvedených grafů má tedy stejný morální graf:



□

Uvedený příklad dobře ilustruje zásadní problém provázání orientovaných a neorientovaných grafů přes grafy morální. Zatímco proces moralizace je jednoznačný (z jednoho orientovaného grafu dostaneme právě jeden graf neorientovaný), opačný proces - tzv. *demoralizace* - již jednoznačný není (jednomu morálnímu grafu může příslušet několik orientovaných acyklických grafů). K dohledání „správného“ orientovaného grafu je pak třeba použít dodatečných infomací/kritérií, než které jsou obsaženy ve struktuře morálního grafu - blíže si tuto problematiku přiblížíme v kapitole zabývající se aplikací grafických modelů na VAR modely.

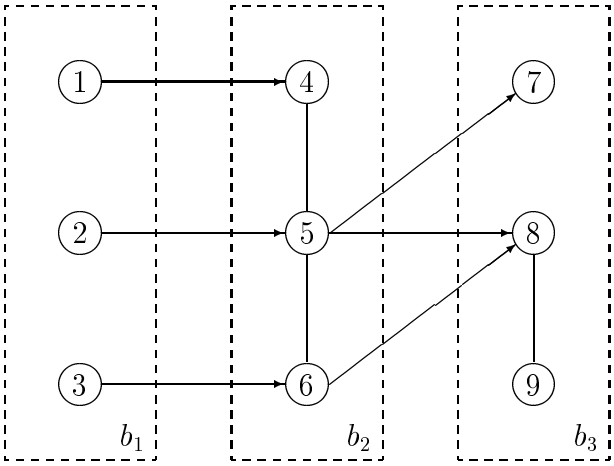
2.3 Řetězové grafy

Pro úplnost ještě zmiňme další možné zobecnění - grafy, které obsahují orientované i neorientované hrany a na kterých je navíc zavedena relace částečného uspořádání $\preceq$, jež zajistí, že jsou vrcholy uspořádané do bloků. V daném bloku se mohou vyskytovat pouze neorientované hrany, mezi bloky pak naopak pouze orientované hrany.

Protože v našich aplikacích vystačíme s neorientovanými a orientovanými grafy, omezíme se pouze na ilustrační příklad.

Příklad 2.30:

Máme graf $G = (K, E)$, $K = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ rozdělený do tří bloků:
$b_1 = \{1, 2, 3\}, b_2 = \{4, 5, 6\}, b_3 = \{7, 8, 9\}$.
Množinu hran si definujeme obrázkem:



□

Poznámka 2.31:

To, že jakákoliv hrana mezi vrcholy jednoho bloku je neorientovaná a hrany mezi vrcholy z různých bloků jsou orientované, vylučuje nejen grafy s orientovanými cykly, ale i grafy s cykly obsahujícími alespoň jednu orientovanou hranu, např.

□

Řetězové grafy je možné použít například v zobecnění lineární regrese (viz [31], aplikaci na česká finanční data je pak možno nalézt v [39]) nebo v aplikacích kreditního rizika (viz [8], [30]).

Kapitola 3

Základní pojmy ze statistiky

Nyní zadefinujeme několik (víceméně známých) statistických pojmů, které dále využijeme ve spojení s grafy a grafickými modely:

3.1 Nezávislost a podmíněná nezávislost

Definice 3.1: *Nezávislost*

Nechť X a Y jsou náhodné vektory se spojitým rozdělením.
Tyto vektory jsou *nezávislé*, právě když sdružená hustota f_{XY} splňuje vztah

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

pro všechny hodnoty x a y (tj. sdružená hustota vektorů se rovná součinu marginálních hustot).
Tuto skutečnost budeme nadále značit $X \perp Y$. □

Poznámka 3.2:

Nezávislost můžeme ekvivalentně formulovat pomocí *podmíněné hustoty*

$$X \perp Y \Leftrightarrow f_{X|Y}(x; y) = f_X(x) \quad \forall\, x.$$

□

Definice 3.3: *Podmíněná nezávislost*

Nechť X, Y, Z jsou náhodné vektory se spojitým rozdělením.
Vektory Y a Z jsou *podmíněně nezávislé* při pevné hodnotě vektoru X, právě když podmíněná hustota $f_{YZ|X}$ splňuje vztah

$$f_{YZ|X}(y, z; x) = f_{Y|X}(y; x) f_{Z|X}(z; x)$$

pro všechny hodnoty y, z a pro všechna x taková, že $f_X(x) > 0$.
Tuto skutečnost budeme nadále značit $Y \perp Z|X$. □

Podmíněnou nezávislost můžeme zapsat i mnoha jinými způsoby - blíže následující lemma:

Lemma 3.4: *Ekvivalentní definice podmíněné nezávislosti*

Nechť $Y \perp Z \mid X$. Potom jsou následující tvrzení ekvivalentní:
(i) $f_{YZ|X}(y,z;x) = f_{Y|X}(y;x) f_{Z|X}(z;x)$
(ii) $f_{Y|XZ}(y,x;z) = f_{Y|X}(y;x)$ (a analogicky $f_{Z|XY}(z,x;y) = f_{Z|X}(z;x)$)
(iii) $f_{XYZ}(x,y,z) = f_{XY}(x,y) f_{XZ}(x,z)/f_X(x)$
(iv) $f_{XYZ}(x,y,z) = f_{XY}(x,y) f_{Z|X}(z;x) = f_{Y|X}(y;x) f_{Z|X}(z;x) f_X(x)$ □

Zatímco bod (i) je definicí 3.3, bod (ii) vyjadřuje vlastnost podmíněné nezávislosti: je-li Y podmíněně nezávislé na Z, pak můžeme Z z podmiňující množiny v rozdělení $Y \mid XZ$ vynechat (symetricky pro ostatní vektory). V (iii) je podmíněná nezávislost přepsána pouze za pomoci marginálních hustot, bod (iv) vznikl triviální úpravou z (iii).

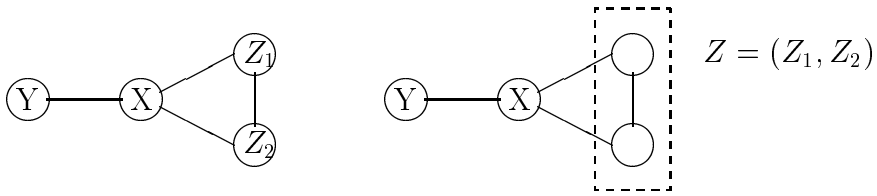
Lemma 3.5: *O blokové nezávislosti*

Nechť (X, Y, Z_1, Z_2) je náhodný vektor se spojitým rozdělením a hustotou $f > 0$.
Pak jsou následující tvrzení ekvivalentní:
(a) $Y \perp (Z_1, Z_2)|X$
(b) $Y \perp Z_1|(X, Z_2)$ a $Y \perp Z_2|(X, Z_1)$. □

Lemma je dokázáno v [31] na straně 33.

Poznámka 3.6:

Toto technické lemma je potřebné pro důkaz separační věty pro grafy podmíněných nezávislostí. Můžeme si ji přiblížit následujícím obrázkem:



□

Věta 3.7:

Nechť (X, Y, Z_1, Z_2) je náhodný vektor se spojitým rozdělením a hustotou $f > 0$.
Pak jsou následující tvrzení ekvivalentní:
(a) $Y \perp (Z_1, Z_2)|X$
(b) $Y \perp Z_2|(X, Z_1)$ a $Y \perp Z_1|X$. □

Věta je dokázána v [31] na straně 35.
Zobecněním věty 3.7 je následující tvrzení:

Věta 3.8: *Věta o separaci*

Nechť $X = (X_a, X_b, X_c)$ a nechť v grafickém modelu pro tento náhodný vektor platí, že podmnožina vrcholů a separuje podmnožiny vrcholů b a c.
Pak platí:

$$X_b \perp X_c | X_a.$$

□

Věta je dokázána v [31] na straně 67 a uvádíme ji na tomto místě s upozorněním, že pojem *grafický model* je definován až v kapitole 5.1.

3.2 Střední hodnota, rozptyl, kovariance a korelační koeficient

Úmluva 3.9:

V dalším textu budeme všechny vektory uvažovat sloupcové. □

Značení 3.10:

V dalším textu budeme používat následující značení:
Transpozici vektoru X budeme označovat X^T.
Střední hodnotu náhodné veličiny X budeme značit EX.
Rozptyl náhodné veličiny X budeme značit $Var(X) = E(X - EX)^2$.
Podmíněnou střední hodnotu v konkrétním bodě budeme značit $E_{Y|X=x}(Y)$.
Podmíněný rozptyl náhodné veličiny Y za podmínky X budeme značit $Var_{Y|X}(Y)$.
□

Lemma 3.11:

Platí následující vztahy:

$$\begin{aligned} EY &= E_X[E_{Y|X}(Y)] \\ Var(Y) &= E_X[Var_{Y|X}(Y)] + Var_X[E_{Y|X}(Y)]. \end{aligned}$$

□

Důkaz lemmatu lze nalézt např. v [1] na straně 55-56.

Definice 3.12: *Kovariance*

Kovarianci dvou náhodných vektorů X, Y definujeme předpisem

$$Cov(X, Y) = E(X - EX)(Y - EY)^T$$

□

Definice 3.13: *Korelační koeficient*

Nechť X, Y jsou náhodné veličiny s konečnými druhými momenty, pak definujeme *korelační koeficient*

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}.$$

□

Definice 3.14: *Výběrový korelační koeficient*

Nechť $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ je výběr z nějakého dvojrozměrného rozdělení, označme

$$S_{XY} = \frac{1}{n-1}\sum (X_i - \overline{X})(Y_i - \overline{Y}),$$

$$S_X^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum (Y_i - \overline{Y})^2.$$

Pokud $S_X^2 > 0$ a $S_Y^2 > 0$, definujeme *výběrový korelační koeficient* jako:

$$r = \frac{S_{XY}}{\sqrt{S_X^2\, S_Y^2}}.$$

□

Naskýtá se otázka, jak velký musí být výběrový koeficient, abychom mohli jeho teoretický protějšek považovat za nulový. Na to nám dává odpověď následující věta:

Věta 3.15:

Nechť $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ je výběr z dvojrozměrného normálního rozdělení, které má kladné rozptyly a korelační koeficient $\rho = 0$. Pak náhodná veličina

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

má rozdělení t_{n-2}. □

Věta je dokázána v [2] - str. 217.
Postup testu hypotézy nulovosti korelačního koeficientu

$$H_0\colon \rho = 0$$
$$H_1\colon \rho \neq 0$$

je tedy následující:

1. Vypočteme výběrový korelační koeficient r.
2. Vypočteme veličinu T.
3. V případě $|T| > t_{n-2}(\alpha)$ zamítáme H_0 na hladině významnosti α.

3.3 Lineární odhad metodou nejmenších čtverců

Nyní se stručně zmíníme o řešení problému nalezení lineárního odhadu náhodné veličiny Y z pozorování vektoru X.

Úmluva 3.16:

Bez újmy na obecnosti budeme nadále uvažovat střední hodnoty X a Y nulové. Pokud totiž nulové nejsou, stačí místo X uvažovat $(X - EX)$ a místo Y uvažovat $(Y - EY)$. □

Mějme náhodné veličiny $Y_1, \ldots, Y_n$ a matici daných čísel X typu $n \times k$, $k < n$. Předpokládejme, že pro náhodný vektor $Y = (Y_1, \ldots, Y_n)$ platí $Y = X\beta + e$, kde $\beta = (\beta_1, \ldots, \beta_k)^T$ je vektor neznámých parametrů a $e = (e_1, \ldots, e_n)^T$ vektor splňující $Ee = 0$, $Var(e) = \sigma^2 I$.

Jedním z možných řešení, jak odhadnout parametry $\beta_1, \ldots, \beta_k$, je *metoda nejmenších čtverců*, která minimalizuje výraz $(Y - X\beta)^T(Y - X\beta)$.

Označme odhady vektoru parametrů $\beta = (\beta_1, \ldots, \beta_k)^T$ jako $b = (b_1, \ldots, b_k)^T$. Pak platí následující věta:

Věta 3.17:

Odhady metodou nejmenších čtverců jsou $b = (X^T X)^{-1} X^T Y$. □

Důkaz lze nalézt například v [1] na straně 79.

Značení 3.18:

Odhad veličiny Y z pozorování vektoru X metodou nejmenších čtverců budeme nadále značit jako $\hat{Y}(X)$, nebo pouze $\hat{Y}$. □

Věta 3.19: *Vlastnosti lineárního odhadu $\hat{Y}$*

Reziduum $(Y - \hat{Y})$ je ortogonální k libovolné lineární transformaci vektoru X, to jest:

$$Cov(Y - \hat{Y}, X) = 0.$$

□

Věta je dokázána v [31] na straně 129.

Věta 3.20: *Rozklad rozptylu*

Rozptyl Y můžeme zapsat jako:

$$Var(Y) = Var(\hat{Y}) + Var(Y - \hat{Y}).$$

□

Věta je dokázána v [31] na straně 129.

3.4 Koeficient mnohonásobné korelace, koeficient determinace

Pomocí obyčejného korelačního koeficientu ρ se měří závislost mezi dvěma náhodnými veličinami. Velmi často je však nutné popsat závislost mezi náhodnou veličinou Y a náhodným vektorem $X = (X_1, \ldots, X_n)$. K tomu lze použít koeficient mnohonásobné korelace:

Definice 3.21: *Koeficient mnohonásobné korelace*

Nechť Y je náhodná veličina a $\hat{Y}$ její lineární odhad pomocí vektoru X metodou nejmenších čtverců.
Definujme *koeficient mnohonásobné korelace* mezi Y a X jako korelační koeficient mezi Y a $\hat{Y}$. □

Definice 3.22: *Výběrový koeficient mnohonásobné korelace*

Mějme náhodný výběr $\begin{pmatrix} Y_1 \\ \mathbf{X_1} \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \mathbf{X_n} \end{pmatrix}$ z nějakého $(p+1)$-rozměrného rozdělení. Výběrový koeficinet mezi i-tou a j-tou složkou vektorů $\mathbf{X_1}, \ldots, \mathbf{X_n}$ označme r_{ij}. Dále nechť r_{0i} značí výběrový korelační koeficient mezi Y-ovými veličinami a i-tými složkami vektorů $\mathbf{X_1}, \ldots, \mathbf{X_n}$. Zaveďme výběrové korelační matice

$$R_{\mathbf{XX}} = (r_{ij})_{i,j=1}^{p}, \quad R_{Y\mathbf{X}} = (r_{0i})_{i=1}^{p}, \quad R_{\mathbf{X}Y} = R_{Y\mathbf{X}}^{T}.$$

Pak definujeme *výběrový koeficient mnohonásobné korelace* jako:

$$r_{Y,\mathbf{X}}^2 = R_{Y\mathbf{X}} R_{\mathbf{XX}}^{-1} R_{\mathbf{X}Y}$$

pro $R_{\mathbf{XX}}$ regulární matici. □

Rovněž u koeficientu mnohonásobné korelace můžeme testovat nulovost. Využijeme přitom následující věty (viz [2] - str. 221):

Věta 3.23:

Nechť náhodné vektory $\begin{pmatrix} Y_1 \\ \mathbf{X_1} \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \mathbf{X_n} \end{pmatrix}$ jsou výběrem z regulárního normálního rozdělení. Je-li $n > p+1$ a platí-li $\rho_{Y,\mathbf{X}} = 0$, pak náhodná veličina

$$Z = \frac{n-p-1}{p} \frac{r_{Y,\mathbf{X}}^2}{1 - r_{Y,\mathbf{X}}^2}$$

má $F_{p,n-p-1}$ rozdělení. □

Postup testu hypotézy nulovosti koeficientu mnohonásobné korelace

$$H_0\text{: } \rho_{Y,\mathbf{X}} = 0$$
$$H_1\text{: } \rho_{Y,\mathbf{X}} \neq 0$$

je tedy následující:

1. Vypočteme výběrový koeficient mnohonásobné korelace $r_{Y,\mathbf{X}}$.
2. Vypočteme veličinu Z.
3. V případě $|Z| > F_{p,n-p-1}(\alpha)$ zamítáme H_0 na hladině významnosti α.

Poznámka 3.24:

Jak je patrné, jde o postup ekvivaletní s testem hypotézy v regresním modelu: $H_0 : \beta_1 = \ldots = \beta_p = 0$. □

Definice 3.25: *Koeficient determinace*

Čtverec výběrového koeficientu mnohonásobné korelace se nazývá *koeficient determinace* a značí se R^2. □

Poznámka 3.26:

Koeficient determinace vyjadřuje, do jaké míry jsou vysvětleny změny závisle proměnné (Y) působením všech nezávisle proměnných (tj. celého vektoru X). □

Následující tvrzení je důsledkem věty o rozkladu rozptylu:

Důsledek 3.27:

Koeficient determinace můžeme zapsat ve tvaru:

$$R^2 = \frac{Var\hat{Y}}{VarY}.$$

□

3.5 Parciální kovariance a rozptyl, koeficient parciální korelace

Definice 3.28:

Parciální kovarianci Y a Z při daném X definujeme jako:

$$Cov(Y - \hat{Y}(X), Z - \hat{Z}(X))$$

a značíme ji $Cov(Y, Z|X)$. □

Poznámka 3.29:

Parciální kovariance je tedy kovariancí reziduí $(Y - \hat{Y})$ a $(Z - \hat{Z})$. □

Věta 3.30: *Bilinearita parciální kovariance*

Parciální kovariance $Cov(Y, Z|X)$ je bilineární v argumentech Y a Z. □

Věta 3.31:

Parciální kovariance $Cov(Y, Z|X)$ splňuje:

$$Cov(Y, Z|X) = Cov(Y, Z) - Cov(Y, X)Var(X)^{-1}Cov(X, Z).$$

□

Definice 3.32:

Parciální rozptyl Y při daném X definujeme jako:

$$Cov(Y, Y|X)$$

a značíme $Var(Y|X)$. □

Důsledek 3.33:

Parciální rozptyl $Var(Y|X)$ splňuje následující vztahy:

$$\begin{aligned} Var(Y|X) &= Var(Y) - Var(\hat{Y}) \\ Var(Y|X) &= Var(Y) - Cov(Y, X)Var(X)^{-1}Cov(X, Y) \\ Var(Y|X) &= Var(Y - \hat{Y}). \end{aligned}$$

□

Důkazy těchto vět lze nalézt v [31] str. 135-136.

Definice 3.34: *Koeficient parciální korelace*

Nechť Y a Z jsou náhodné veličiny, pak definujeme *koeficient parciální korelace* mezi Y a Z při daném X vztahem:

$$\frac{Cov(Y, Z|X)}{\sqrt{Var(Y|X)Var(Z|X)}}$$

a značíme $corr(Y, Z|X)$, popř. jednodušeji $\rho_{Y,Z.X}$. □

Definice 3.35: *Výběrový koeficient parciální korelace*

Mějme náhodný výběr $\begin{pmatrix} Y_1 \\ \mathbf{X_1} \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \mathbf{X_n} \\ Z_n \end{pmatrix}$ z nějakého $(p+2)$-rozměrného rozdělení. Nechť $r_{p+1,i}$ je výběrový korelační koeficient mezi Z-ovými veličinami a i-tými složkami vektorů $(\mathbf{X_1}, \ldots, \mathbf{X_n})$. Označme

$$R_{Z\mathbf{X}} = (r_{p+1,i})_{i=1}^{p}, \quad R_{\mathbf{X}Z} = R_{Z\mathbf{X}}^T.$$

Pak definujeme *výběrový koeficient parciální korelace* jako:

$$r_{Y,Z.\mathbf{X}} = \frac{r_{YZ} - R_{Y\mathbf{X}} R_{\mathbf{XX}}^{-1} R_{\mathbf{X}Z}}{\sqrt{(1 - R_{Y\mathbf{X}} R_{\mathbf{XX}}^{-1} R_{\mathbf{X}Y})\,(1 - R_{Z\mathbf{X}} R_{\mathbf{XX}}^{-1} R_{\mathbf{X}Z})}}$$

pro jmenovatel různý od nuly. □

Poznámka 3.36:

Pokud máme místo vektoru pouze jednorozměrnou veličinu X, dostáváme

$$r_{Y,Z.X} = \frac{r_{YZ} - r_{YX} r_{ZX}}{\sqrt{(1 - r_{YX}^2)(1 - r_{ZX}^2)}}.$$

□

Na otázku, jak velký musí výběrový koeficient být, abychom mohli jeho teoretický protějšek považovat za nulový, dává odpověď podobná věta jako v případě jednorozměrných náhodných veličin (viz [2] - str. 223):

Věta 3.37:

Nechť náhodné vektory $\begin{pmatrix} Y_1 \\ \mathbf{X_1} \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \mathbf{X_n} \\ Z_n \end{pmatrix}$ jsou výběrem z regulárního normálního rozdělení. Platí-li $\rho_{Y,Z.\mathbf{X}} = 0$, pak při $n > p + 2$ má náhodná veličina

$$T = \frac{r_{Y,Z.\mathbf{X}}}{\sqrt{1 - r_{Y,Z.\mathbf{X}}^2}} \sqrt{n - p - 2}$$

rozdělení t_{n-p-2}. □

Postup testu hypotézy nulovosti koeficientu parciální korelace

$$H_0\text{: } \rho_{Y,Z.\mathbf{X}} = 0$$
$$H_1\text{: } \rho_{Y,Z.\mathbf{X}} \neq 0$$

je tedy následující:

1. Vypočteme výběrový koeficient parciální korelace r.
2. Vypočteme veličinu T.
3. V případě $|T| > t_{n-p-2}(\alpha)$ zamítáme H_0 na hladině významnosti α.

3.6 Varianční matice a její inverze

Značení 3.38:

Nechť X je k-rozměrný náhodný vektor. Inverzní matici k varianční matici $Var(X)$ budeme značit $D = (d_{ij})_{i,j=1}^{k}$.
Pro náhodný vektor $T = (X, Y)^T$ budeme psát D ve tvaru blokové matice:

$$D = \begin{pmatrix} D_{XX} & D_{XY} \\ D_{YX} & D_{YY} \end{pmatrix}.$$

□

Lemma 3.39: *O inverzní varianční matici*

Platí:

$$Var(X,Y)^{-1} = \begin{pmatrix} Var(X)^{-1} + B^T Var(Y|X)^{-1} B & -B^T Var(Y|X)^{-1} \\ -Var(Y|X)^{-1} B & Var(Y|X)^{-1} \end{pmatrix},$$

kde $B = Cov(Y,X)Var(X)^{-1}$. □

Důkaz lze nalézt v [31] na straně 143, kde jsou rovněž dokázány následující důsledky.

Důsledek 3.40:

Každý diagonální prvek inverzní varianční matice je převrácenou hodnotou parciálního rozptylu. □

Důsledek 3.41:

Mějme k-rozměrný náhodný vektor X, $K = \{1, 2, \ \ldots\ , k\}$. Pak každý diagonální prvek inverzní varianční matice je převrácenou hodnotou parciálního rozptylu, mimodiagonální prvek ij inverzní varianční matice škálované tak, aby měla na diagonále jednotky, je roven parciálnímu korelačnímu koeficientu mezi i-tou a j-tou složkou vektoru při pevných hodnotách ostatních složek, avšak s opačným znaménkem:

$$d_{ii} = \frac{1}{Var(X_i \mid X_{K\setminus\{i\}})} \quad \forall i \in K$$

$$\frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}} = -corr(X_i, X_j \mid X_{K\setminus\{i,j\}}) \quad \forall i \in K, \forall j \in K, i \neq j.$$

□

Důsledek 3.42:

Mimodiagonální blok D_{XY} inverzní matice k varianční matici $Var(X,Y)$ je nulový, právě když $Cov(X,Y) = 0$. □

Důsledek 3.43:

Nechť $(X, Y, Z)^T$ je náhodný vektor a D inverzní matice k jeho varianční matici. Pak platí:

$$D_{YZ} = 0 \iff Cov(Y, Z|X) = 0.$$

□

3.7 Mnohorozměrné normální rozdělení

Definice 3.44:

Nechť X je k-rozměrný náhodný vektor, μ pevně daný k-dimenzionální vektor a D symetrická pozitivně definitní matice typu $k \times k$.
Tento vektor má *mnohorozměrné normální rozdělení*, jestliže má hustotu

$$f_X(x) = (2\pi)^{-\frac{k}{2}} det(D)^{\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^T D(x-\mu)\}$$

pro všechna $x \in R^k$.
Tuto skutečnost budeme značit $X \sim N(\mu, V)$, kde $V = D^{-1}$. □

Věta 3.45:

Nechť X_a a X_b jsou náhodné vektory z normálního rozdělení.
Pak platí:

$$X_a \perp X_b \iff Cov(X_a, X_b) = 0.$$

□

Důkaz lze nalézt např. v [1] na straně 77.

Důsledek 3.46:

Nechť X_a a X_b jsou náhodné vektory z normálního rozdělení.
Pak platí:

$$X_a \perp X_b \iff D_{ab} = 0.$$

□

Věta 3.47:

Nechť X_a a X_b a X_c jsou náhodné vektory z normálního rozdělení.
Pak platí:

$$X_a \perp X_b|X_c \iff Cov(X_a, X_b|X_c) = 0.$$

□

Věta 3.48:

Nechť X_a a X_b a X_c jsou náhodné vektory z normálního rozdělení.
Pak platí:

$$X_a \perp X_b|X_c \iff D_{ab} = 0.$$

□

Důkazy lze nalézt v [31] na straně 164.

Kapitola 4

VAR modely - základní pojmy

4.1 Klasické VAR modely

VAR modely jsou jedním z přístupů k analýze vícerozměrných časových řad. I když se ukazuje (po počátečním velkém rozmachu), že nejsou většinou schopny dávat tak přesné výsledky jako velké teoretické soustavy simultálních rovnic, mají řadu výhod. Díky své ateoretičnosti je je možno lehce aplikovat i bez odpovídající ekonomické teorie, k odhadu parametrů lze použít metodu nejmenších čtverců zvlášť na každou rovnici a umožnují velmi snadné získání predikovaných hodnot.

Podívejme se nyní na základní definici:

Definice 4.1: *VAR model*

Model vektorové autoregrese (VAR(p)) je soustava rovnic tvaru:

$$\mathbf{x_t} = \mathbf{c} + \mathbf{\Phi_1 x_{t-1}} + \mathbf{\Phi_2 x_{t-2}} + \mathbf{\Phi_p x_{t-p}} + \mathbf{e_t}, \tag{4.1}$$

kde:

$\mathbf{x_t}, \ldots, \mathbf{x_{t-p}}$ jsou n-dimenzionální náhodné vektory proměnných,

$\mathbf{\Phi_1}, \ldots, \mathbf{\Phi_k}$ jsou vektory koeficientů,

$\mathbf{c}$ je vektor konstant a

$\mathbf{e_t}$ vektor náhodných chyb. □

O náhodných chybách se předpokládá, že jsou stejně rozdělené s nulovou střední hodnotou a v klasickém případě pocházejí z normálního rozdělení. (Jak je však uvedeno v [24], metodu nejmenších čtverců můžeme s úspěchem použít pro odhad parametrů i v případě širších podmínek.)

Poznámka 4.2:

Pro naše aplikace se omezíme na *stacionární časové řady* - viz další definice. Vyhneme se tak problému tzv. zdánlivé regrese. Může se totiž stát, že dané 2 řady mají společný pouze trend a jinak nic jiného. Přesto v důsledku společného trendu vykazují významnou regresní závislost. Ani vysoké hodnoty

koeficientů vícenásobné determinace R^2, resp. jeho modifikované verze $\overline{R^2}$, pak nejsou zárukou skutečné regresní závislosti těchto dvou časových řad (viz [12] - str 170). □

Definice 4.3: *Stacionární časová řada*

Časová řada $\{y_t\}$ je *(slabě) stacionární*, pokud má konstantní střední hodnotu, konstantní rozptyl a kovarianční strukturu invariantní vůči posunům v čase:

$$Cov(y_t, y_S) = Cov(y_{t+h}, y_{s+h}).$$

□

4.2 Grangerova kauzalita

V jedné z dále uvedených aplikací bude potřeba otestovat, zda změny v krátkodobých úrokových sazbách předcházejí změně sazeb dlouhodobých, zda jsou obě veličiny simultánně závislé, nebo naopak nezávislé.

K takovému testování vazeb mezi proměnnými je možné použít tzv. Grangerovy kauzality. O Grangerově kauzalitě (viz např. [12]) hovoříme tehdy, jestliže běžné a různě zpožděné hodnoty jedné proměnné (označme ji např. X_t) vysvětlují v regresi významnou měrou závislost jiné proměnné (označme ji Y_t) na zpožděných hodnotách Y_t a X_t.

Grangerovu kauzalitu nelze interpretovat jako příčinnou závislost, protože podstatou testování kauzality v Grangerově pojetí je pouze skutečnost, zda změny určité proměnné předcházejí změně jiné proměnné, nikoliv která veličina je příčinou a která následkem.

To je však v našem případě postačující, protože chceme pouze ověřit, zda se krátkodobé a dlouhodobé úrokové sazby chovají podle jedné z hypotéz uvedených dále a ne které sazby jsou primární a které odvozené.

Granger navrhuje následující testovací postupy:
Uvažujme 2 časové řady proměnných X_t (v našem případě půjde o krátkodobou úrokovou sazbu) a Y_t (v našem případě dlouhodobou úrokovou sazbu). Budeme testovat nulovou hypotézu, že proměnná X nepodmiňuje proměnnou Y. Pro tento test je potřeba odhadnout 2 formy lineární regrese.

Tzv. neomezená regrese je tvaru:

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{p} \beta_i X_{t-i} + u_t.$$

Omezená regrese se naopak pokouší vysvětlit vývoj proměnné Y pouze pomocí různě zpožděných Y_t:

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + v_t.$$

K ověření statistické významnosti zpožděných hodnot proměnné X v neomezené regresi pak využijeme F testu (ve skutečnosti nejde o nic jiného než o test podmodelu):

$$F = \frac{(e'e)_{omez} - (e'e)_{neomez}}{q \cdot (e'e)_{neomez}}(T-m) \sim F(q, T-m),$$

kde:

T - počet pozorování

m - počet odhadnutých parametrů v neomezené regresi

q - počet omezení parametrů

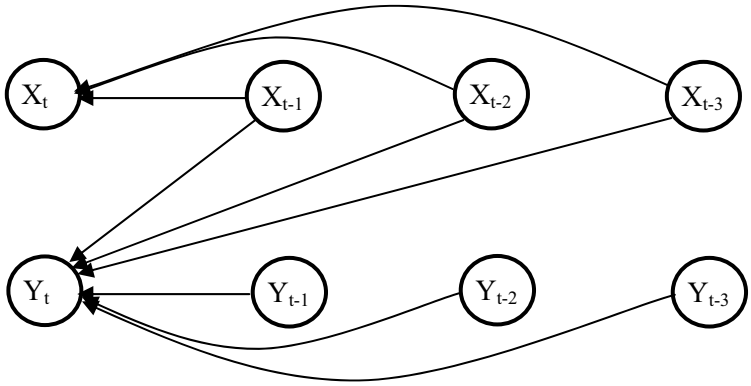
Zjistíme-li, že se parametry β_i významně odlišují od nuly, zamítneme nulovou hypotézu, že X nepodmiňuje proměnnou Y ve smyslu Grangerovy kauzality.

V dalším kroku testujeme reverzní nulovou hypotézu, že proměnná Y nepodmiňuje X.

Pokud v prvním kroku odmítneme hypotézu, že X nepodmiňuje Y, a zároveň v druhém kroku nezamítneme nulovou hypotézu, že proměnná Y nepodmiňuje X, můžeme říci, že X podmiňuje Y z hlediska Grangerovy kauzality.

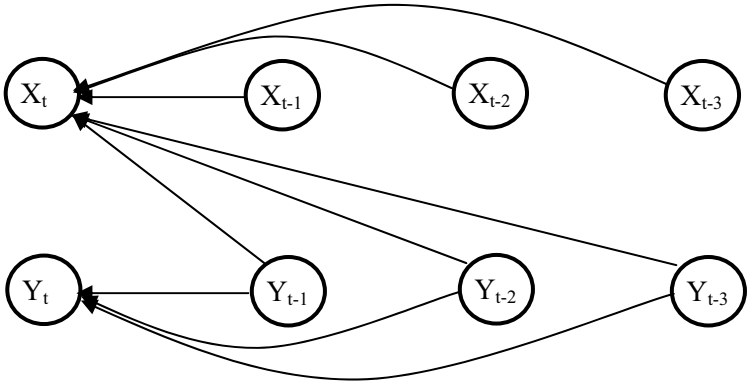
Poznámka 4.4:

Spojení VAR modelů s orientovanými grafy (viz kapitola 10) nám dává Grangerovu kauzalitu „zdarma“, tj. bez výše popsaného testu a navíc v přehledné grafické podobě[1]. Uvažujme například 2 proměnné X a Y, model VAR(3) a skutečnost, že X podmiňuje Y z hlediska Grangerovy kauzality. Pak má výsledná reprezentace VAR modelu orientovaným grafem následující podobu:

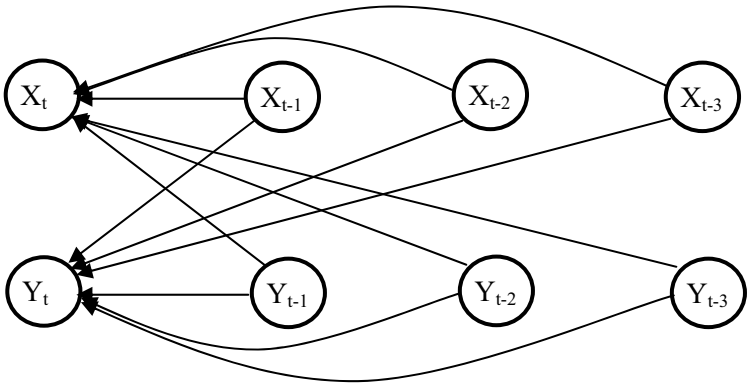


Pokud by naopak platilo, že Y podmiňuje X z hlediska Grangerovy kauzality, dostáváme následující graf:

[1] Je však třeba poznamenat, že se nejedná o nijak velkou výhodu - pokud totiž pracujeme s VAR modely v běžném ekonometrickém software, dostaneme Grangerovu kauzalitu obvykle jako jeden ze standardních výstupů spolu s dalšími testy.



Pokud neplatí ani X podmiňuje Y, ani Y podmiňuje X, dostáváme tzv. saturovaný VAR model:



□

4.3 Testy jednotkových kořenů

Námi uvažované VAR modely jsou konstruovány z časových řad, které vyhovují požadavku stacionarity. Pokud by tato vlastnost nebyla splněna, je třeba časovou řadu transformovat - nejčastěji pomocí prvních (případně i vyšších) diferencí, nebo zahrnutím trendu jako vysvětlující proměnné. Protože chceme použít VAR modely na modelování vztahů mezi krátkodobými a dlouhodobými úrokovými sazbami, je třeba stacionaritu časových řad ověřit. Jinak totiž nelze vyloučit možnost, že případná silná regresní závislost je způsobena pouze společným trendem. Skutečnou příčinou společného „směru“ vývoje úrokových sazeb pak může být závislost na nějaké jiné proměnné, která ovlivňuje vývoj jak krátkodobých, tak dlouhodobých úrokových měr.

K testování nestacionarity využijeme ADF testy v SW PC Give (protože jde o testování hypotézy, zda pro AR proces generující pozorování časové řady existuje jednotkový kořen, označují se tyto testy jako *testy jednotkového kořenu* - viz [12] - str. 173).

4.4 Strukturální VAR modely

K modelování vícerozměrných časových řad existují různé přístupy, které umožňují transformovat model 4.1 tak, aby zahrnoval i vztahy mezi běžnými proměnnými. Pro naši aplikaci (viz kapitola 11.4) bude nejvhodnější tzv. SVAR forma:

Definice 4.5: *Strukturální VAR model*

Pokud kovarianční matice H náhodných chyb $\mathbf{e_t}$ není diagonální, odpovídá VAR modelu soustava rovnic, v níž jsou vztahy mezi komponentami $\mathbf{x_t}$ skryty v kovarianční matici H (tzv. strukturální forma VAR modelu - SVAR(p)):

$$\mathbf{\Theta_0 x_t} = \mathbf{d} + \mathbf{\Theta_1 x_{t-1}} + \mathbf{\Theta_2 x_{t-2}} + \mathbf{\Theta_p x_{t-p}} + \mathbf{u_t}. \tag{4.2}$$

□

Vztah mezi SVAR a klasickým VAR modelem je dán pomocí následujících vztahů:
$\mathbf{\Theta_i} = \mathbf{\Theta_0 \Phi_i}$ pro $i = 0, \ldots, k$
$\mathbf{d} = \mathbf{\Theta_0 c}$ a
$\mathbf{u_t} = \mathbf{\Theta_0 e_t}$ s kovarianční maticí $\mathbf{\Theta_0 H \Theta_0^T} = \mathbf{D}$, která je diagonální.

Z obecných SVAR modelů se omezíme na soustavy rekurzivní (mluví se také o *jednoduše rekurzivních soustavách* - viz [5] - str. 149), protože tato struktura vyplývá z charakteru dále řešeného problému. Rekurzivnost soustavy je ekvivalentní s existencí takového přeuspořádání rovnic, které zajistí, že matice $\mathbf{\Theta_0}$ je trojúhelníková s jednotkovou diagonálou.

Poznámka 4.6:

Rekurzivní systémy (nazývané také modely řetězových příčin) neobsahují zpětné vazby mezi běžnými (endogenními) proměnnými ani vzájemně závislé náhodné složky. Endogenní proměnné jsou hierarchicky uspořádány, takže matice strukturních parametrů všech endogenních proměnných modelu je trojúhelníková, nikoli obecná, přičemž kovarianční matice náhodných složek je diagonální. V důsledku nezkorelovanosti náhodných složek rovnice je splněna pro všechny rovnice rekurzivního modelu i podmínka pro aplikaci klasické metody nejmenších čtverců, neboť stochastické vysvětlující endogenní proměnné v libovolné g-té rovnici jsou apriori nezávislé na náhodné složce této rovnice, tj. na u_g. Aplikací MNČ na jednotlivé rovnice lze dospět ke konzistentním a asymptoticky vydatným odhadům parametrů (viz [12] - str. 140). □

Definice 4.7: *Saturovaný SVAR model*

Pokud ve vektorech koeficientů nejsou žádné nulové hodnoty, je model SVAR *saturovaný* - jde o ekvivalent definice 8.1. □

V praxi se ale často stává, že některé ze zpožděných hodnot nehrají žádnou roli pro predikci běžné hodnoty x_t. V takovém případě je hodnota odpovídajícího koeficientu nulová (tj. je neodlišitelná od 0 prostřednictvím t-testu), a tudíž je SVAR „řídký“. Strukturu SVAR můžeme identifikovat pomocí koeficientů parciální korelace, jejichž významnost otestujeme například pomocí výše zmíněné deviance.

Poznámka 4.8:

Hledání „řídké“ struktury SVAR modelu (tj. snaha položit některé parametry rovny 0) má nejméně dva důvody (viz [32]):

1. Redukce počtu koeficientů může snížit chybu předpovědi, zvláště při větším počtu rovnic (a to i při malém počtu zahrnutých zpožděných proměnných).
2. Menší počet koeficientů může pomoci lépe pochopit mechanismus, kterým se řídí vývoj časové řady.

□

Výhodou je, že model SVAR může být reprezentován orientovaným acyklickým grafem, ve kterém proměnné $x_t, x_{t-1}, \ldots, x_{t-p}$ představují jednotlivé vrcholy a příčinná závislost je znázorněna orientovanými hranami, které končí v proměnné uvedené na levé straně rovnice SVAR. Způsob, jak daným datům přiřadit SVAR model, si popíšeme v kapitole 10.

Kapitola 5

Grafické modelování

Nyní již můžeme spojit teorii grafů se statistikou a ukázat nejlepší možnost, jak nalézt model, který vhodně popisuje nezávislostní strukturu konkrétních dat.

5.1 Gaussovské grafické modely

Definice 5.1: *Grafický model (orientovaný graf podmíněných nezávislostí)*

Nechť $X = (X_1, X_2, ..., X_k)$ je k-rozměrný náhodný vektor a $G = (K, E)$ graf s k vrcholy.
Grafický model pro vektor X je rodina pravděpodobnostních rozdělení vektoru X, která splňují podmíněné nezávislosti dané grafem G.
Pokud navíc platí $X \sim N(\mu, V)$, pak mluvíme o *gaussovském grafickém modelu* nebo také o *grafickém modelu pro normálně rozdělená data.* □

Poznámka 5.2:

V případě gaussovských grafických modelů platí, že podmíněné nezávislosti jsou ekvivalentní výskytu nul v inverzní varianční matici. □

V případě orientovaného grafu můžeme použít trochu jinou definici, která nepracuje v podmínce se všemi proměnnými, ale pouze s proměnnými minulými:

Definice 5.3: *Orientovaný graf podmíněných nezávislostí*

Orientovaným grafem podmíněných nezávislostí náhodného vektoru X nazýváme orientovaný graf $G^{\succ} = (K, E^{\succ})$, kde $K = \{1, 2, \ldots, k\}$, $K(j) = \{1, 2, \ldots, j\}$ a hrana (i, j), $i \prec j$ není obsažena v množině hran právě tehdy, když $X_j \perp X_i \mid X_{K(j)\setminus\{i,j\}}$. □

Jde o podobnou definici jako v případě neorientovaného grafu. V podmínce ale tentokrát nevystupují všechny proměnné (minulost i budoucnost), nýbrž pouze proměnné zpožděné (minulost). To je hlavní rozdíl mezi orientovanými a neorientovanými grafy podmíněných nezávislosti, který říká, že zatímco neorientované grafy vypovídají o nezávislosti mezi jednotlivými sdruženými distribucemi, orientované vypovídají o nezávislosti mezi sekvencí marginálních rozdělení. Minulost má dost informace pro definici sdruženého rozdělení - tuto myšlenku si formalizujme v následující větě, která je dokázána v [31] str. 73:

Věta 5.4: *Rekurzivní faktorizační identita*

$$f_{12\ldots k} = f_{k|K(k)\setminus\{k\}} f_{k-1|K(k-1)\setminus\{k-1\}} f_{2|1} f_1.$$

□

5.2 Jednoduchá selekce grafického modelu

Definice 5.5:

Mějme data ve formě N realizací k-rozměrného náhodného vektoru X:

$$\begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1N} \\ X_{21} & X_{22} & \ldots & X_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ X_{k1} & X_{k2} & \ldots & X_{kN} \end{pmatrix}.$$

Označme si $X^l = (X_{1l}, X_{2l}, ..., X_{kl})$ l-tý sloupec matice realizací, tj. l-tou realizaci náhodného vektoru X. Pak definujeme:
Vektor výběrových průměrů:

$$\bar{X} = \frac{1}{N} \sum_{l=1}^{N} X^l.$$

Výběrovou varianční matici:

$$S = \frac{1}{N} \sum_{l=1}^{N} (X^l - \bar{X})(X^l - \bar{X})^T.$$

□

Postup jednoduché selekce grafického modelu

1. Odhadneme varianční matici $V = Var(X)$ pomocí výběrové varianční matice S.
2. Spočítáme inverzní matici S^{-1}.
3. Škálujeme inverzní matici tak, aby měla na diagonále jednotky. Záporně vzaté mimodiagonální prvky jsou odhadem parciálních korelačních koeficientů $corr_N(X_i, X_j | K\setminus\{i,j\})$, kde K je množina vrcholů.
4. Odhadnuté parciální korelační koeficienty blízké nule položíme rovny nule.
5. Výsledný graf sestavíme tak, že v grafu ponecháme pouze ty hrany, kterým odpovídají nenulové prvky škálované inverzní varianční matice.

Již na první pohled je zřejmý hlavní problém tohoto přístupu: jak určit, které parciální korelační koeficienty jsou dostatečně blízké nule? Samozřejmě se naskýtá možnost použít sadu t-testů na jednotlivé korelační koeficienty. Bohužel to (jak lze lehce ukázat) není nejlepší postup:

Předpokládejme, že chceme testovat následující hypotézu:

H_0: $\rho_1 = \ldots = \rho_N = 0$

H_1: alespoň 1 korelační koeficient je nenulový

Zaveďme si náhodný jev A^i, který znamená, že zamítneme H_0^i: $\rho_i = 0$ na hladině významnosti α^*. Platí-li H_0, pak pro pravděpodobnost, že zamítneme aspoň jednu H_0^i, platí:

$$P\left(\bigcup_i A^i\right) \leq \sum_i P(A^i) = N\alpha^*.$$

Pokud bychom tedy chtěli bezpečně zajistit celkovou hladinu testu $\alpha = 5\%$, museli bychom provádět jednotlivé testy na hladině Nx menší.

Jako výhodnější se proto jeví použití metod založených na věrohodnostním přístupu, které popíšeme v další kapitole.

5.3 Maximálně věrohodné odhady

Selekce grafického modelu založená na věrohodnostním přístupu

1. Určíme systém rozdělení vektoru X: rodinu k-rozměrných normálních rozdělení.

2. Určíme množinu parametrů. Nezávislost je v mnohorozměrném normálním rozdělení charakterizována varianční maticí $V = Var(X)$, popř. její inverzí D. Díky jednoznačnému vztahu mezi D a V si můžeme zvolit, kterou použijeme. Celkem máme $\frac{k(k+1)}{2}$parametrů, z nichž k souvisí s měřítkem a $\frac{k(k-1)}{2}$ s interakcemi. Protože se nezajímáme o střední hodnotu, můžeme ji bez újmy na obecnosti považovat za nulovou.

3. Zvolíme grafický model pro test. Každá párová podmíněná nezávislost specifikovaná grafem generuje omezení na parametry. Pro mnohorozměrné normální rozdělení je to nula v inverzní varianční matici.

4. Zkonstruujeme věrohodnostní funkci (předpokládáme, že máme k dispozici náhodný výběr z mnohorozměrného normálního rozdělení).

5. Odhadneme neznámé parametry pomocí maximalizace věrohodnostní funkce za omezení vyplývajících z testovaného grafického modelu.

6. Zkontrolujeme shodu testovaného grafického modelu s daty dle testové statistiky - deviance (viz kapitola 8).

Lemma 5.6: *Věrohodnostní funkce mnohorozměrného normálního rozdělení*

Nechť $X^1, X^2, ..., X^N$ je náhodný výběr z k-rozměrného normálního rozdělení $N(0, V)$. Pak věrohodnostní funkci můžeme vyjádřit ve tvaru:

$$L(V) = \prod_{i=1}^{N} (2\pi)^{-\frac{k}{2}} det(D)^{\frac{1}{2}} \exp\{-\frac{1}{2}(X^i)^T D(X^i)\}.$$

□

Důsledek 5.7: *Logaritmická věrohodnostní funkce mnohorozměrného normálního rozdělení*

Nechť $X^1, X^2, ..., X^N$ je náhodný výběr z k-rozměrného normálního rozdělení $N(0, V)$. Pak logaritmickou věrohodnostní funkci můžeme vyjádřit ve tvaru:

$$\begin{aligned} 2l(V) &= const - \sum_{l=1}^{N} X^{l^T} V^{-1} X^l - N \log\det V \\ &= const - N tr(V^{-1} S) - N \log\det V, \end{aligned}$$

kde tr značí stopu matice a konstanta $const$ nezávisí na parametru V. □

Věta 5.8: *Věrohodnostní rovnice pro gaussovské grafické modely*

Nechť $X^1, X^2, ..., X^N$ je náhodný výběr z k-rozměrného normálního rozdělení $N(0, V)$.
Pak maximálně věrohodné odhady matice V a $D = V^{-1}$ za omezení daných grafem G splňují věrohodnostní rovnice:

$$\hat{d}_{ij} = 0$$

pro vrcholy i a j, které nejsou spojené hranou grafu G, a

$$\hat{V}_{aa} = S_{aa},$$

pro podmnožinu vrcholů a, která je klika.
Odhadnuté parametry $\hat{V}$ a $\hat{D}$ jsou jednoznačně určené s pravděpodobností jedna. Navíc platí $\hat{D} = (\hat{V})^{-1}$. □

Důkaz této věty lze nalézt v [31] na str. 176-177.

Popišme si ještě speciální případ - odhad varianční matice pro graf bez hran. Takový graf je totiž vstupem pro tzv. forward algoritmy a výpočet odhadu podle následujícího důsledku nám ušetří jedno nasazení iterační procedury.

Důsledek 5.9: *Odhad varianční matice v grafickém modelu s grafem bez hran*

Nechť $X^1, X^2, ..., X^N$ je náhodný výběr z k-rozměrného normálního rozdělení $N(0, V)$, $S = \{s_{ij}\}$ výběrová varianční matice a $G = (K, E)$ graf.
Odhad varianční matice a její inverze pro situaci $E = \emptyset$ lze psát ve tvaru:

$$\hat{V} = diag\{s_{ii}\},$$

$$\hat{D} = diag\{\frac{1}{s_{ii}}\}.$$

□

Důkaz:

Graf bez hran můžeme zapsat pomocí klik jako $\{\{1\}, \{2\}, \ldots, \{k\}\}$. Platí tedy $\hat{v}_{ii} = s_{ii}$. Protože navíc $\hat{d}_{ij} = 0$ pro nespojené vrcholy, dostáváme, že $\hat{D}$ je diagonální. Ze vztahu $\hat{V} = \hat{D}^{-1}$ pak vyplývá, že rovněž $\hat{V}$ je diagonální. □

5.4 Markovské podmínky

Jedním ze základních teoretických poznatků v grafických modelech jsou tzv. markovské podmínky (vlastnosti). Podívejme se proto nyní trochu podrobněji, o co se jedná:

5.4.1 Markovské podmínky pro neorientované grafy

Nacházíme se v následující situaci: máme náhodný vektor X a potřebujeme predikovat nějakou jeho podmnožinu. Zaveďme si tedy následující značení ($X = X_a \cup X_b \cup X_c$, X_a, X_b, X_c disjunktní):

- X_a podmnožina vrcholů, kterou chceme predikovat;
- X_b vrcholy, které jsou spojeny hranou aspoň s jedním vrcholem z X_a;
- X_c zbylé vrcholy.

Samozřejmě platí $X_a \perp X_c \mid X_b$. Markovské podmínky nám však dávají odpověď na otázku, co je podstatné pro predikci množiny X_a. Sdružené, marginální a podmíněné rozdělení indexujme pro jednoduchost množinami a, b, c namísto vektorů X_a, X_b, X_c. Pak podle lemmatu 3.4 (ii) dostáváme $f_{a|b\cup c} = f_{a|b}$. Pro predikování X_a nám tedy stačí X_b a vůbec nepotřebujeme X_c. Zapišme nyní tuto myšlenku formálněji:

Definice 5.10: *Markovské vlastnosti*

Řekneme, že náhodný vektor X s grafem podmíněných nezávislostí G má[1]

[1]Místo náhodný vektor X má příslušnou markovskou vlastnost se také říká: X splňuje příslušnou markovskou podmínku.

1. *Lokální markovskou vlastnost (LM)*, pokud platí

$$X_i \perp X_b \mid X_a,$$

kde i je libovolný vrchol v G, $a = ne_G(i)$ sousedé vrcholu i a $b = K \setminus (\{i\} \cup a)$. (Zjednodušeně $i \perp$ zbytek | sousedi.)

2. *Párovou markovskou vlastnost (PM)*, pokud pro všechny vrcholy i, j, které nejsou spojeny hranou, platí

$$X_i \perp X_j \mid X_a, \text{ kde } a = K \setminus \{i, j\}.$$

3. *Globální markovskou vlastnost (GM)*, pokud pro a, b, c po dvou disjunktní podmnožiny K platí:

$$a \text{ separuje } b, c \Rightarrow X_b \perp X_c \mid X_a.$$

□

Věta 5.11: *Ekvivalence markovských podmínek*

Platí:

$$(LM) \Leftrightarrow (PM) \Leftrightarrow (GM).$$

□

Důkaz je uveden v [31] - str. 70 - 71.

5.4.2 Markovské podmínky pro orientované grafy

Následující věta ukazuje, proč jsme v kapitole 2.2 tvrdili, že morální graf je v jistém smyslu ekvivalentní grafu orientovanému.

Věta 5.12: *Markovský teorém pro orientované grafy podmíněných nezávislostí*

Orientovaný graf $G^{\succ}$ má markovské vlastnosti svého morálního grafu G^{mor}. □

Věta je dokázána v [31] na str. 76.

Důsledek 5.13:

Pro $G^{mor} = G^{asoc}$ má orientovaný graf $G^{\succ}$ právě jenom markovské vlastnosti svého morálního grafu G^{mor}. □

Pokud má totiž morální graf více hran než asociovaný, obsahuje graf G i některé nezávislosti, které nelze z markovských vlastností morálního grafu odvodit.

Kapitola 6

Entropie a divergence

Při různých odvozeních a důkazech v grafických modelech se často setkáváme s pojmy z teorie informace. Jde nejen o užitečné nástroje pro měření rozdílnosti dvou rozdělení, ale je je možno dále zobecnit například na pozitivně definitní matice. Základní definice a vlastnosti proto uvádíme v této kapitole.

Označme:
$\mathcal{X} = \{x_1, \ldots, x_N\}$ konečný stavový prostor.
$\mathcal{P}$ pravděpodobnostní míry na $\mathcal{X}$.

Definice 6.1: *Entropie*

Buď $P \in \mathcal{P}$ pravděpodobnostní míra. Její *entropii* definujeme vztahem

$$H(P) = \sum_{x \in \mathcal{X}} -\log P(x) \cdot P(x),$$

kde dodefinujeme $\log 0 \cdot 0 = 0$. □

Poznámka 6.2:

Entropie není nic jiného, než střední hodnota funkce $-\log P(x)$. □

Lemma 6.3: *Základní vlastnosti entropie*

1. Omezení entropie zdola: $H(P) \geq 0,\ H(P) = 0 \Leftrightarrow \exists x \in \mathcal{X} : P(x) = 1.$
2. Omezení entropie shora: $H(P) \leq \log N, H(P) = \log N \Leftrightarrow P$ je rovnoměrné rozdělení $R(x) = \frac{1}{N} \forall x \in \mathcal{X}$.
3. Entropie je konkávní funkce.

□

Definice 6.4: *Divergence*

Buď $P, Q \in \mathcal{P}$ dvě pravděpodobnostní míry. Jejich *divergenci* definujeme vztahem

$$\begin{aligned} D(P \mid Q) &= \sum_{x \in \mathcal{X}} \log \frac{P(x)}{Q(x)} \cdot P(x) \; pro \; P \ll Q, \\ D(P \mid Q) &= \infty \; jinak. \end{aligned}$$

□

Poznámka 6.5:

O tom, že jde o pojem důležitý, svědčí i skutečnost, že je znám pod řadou jiných názvů. Místo jednoduchého *divergence* se můžeme setkat například s pojmy:

- Kullback - Leiblerova informační divergence
- Kullback - Leiblerova vzdálenost
- I - divergence
- Relativní entropie

□

Lemma 6.6: *Základní vlastnosti divergence*

1. Omezení divergence zdola: $D(P \mid Q) \geq 0, \; D(P \mid Q) = 0 \Leftrightarrow P = Q$.
2. Omezení entropie shora: z definice platí $D(P \mid Q) = \infty$, pokud P není absolutně spojitá vůči Q.

□

Poznámka 6.7:

Zatímco entropie je „mírou neurčitosti“ (minimální hodnoty nabývá pro degenerované rozdělení a maximální pro rozdělení rovnoměrné), divergence je „mírou nepodobnosti“ (minimální hodnoty nabývá pro stejná rozdělení, maximální pak pro absolutně nespojité míry). Je tedy dobrým nástrojem pro měření „vzdálenosti“ dvou rozdělení. □

Lemma 6.8: *Souvislost entropie a divergence*

$$\begin{aligned} D(P \mid R) &= \sum_{x \in \mathcal{X}} \log \frac{P(x)}{\frac{1}{N}} \cdot P(x) \\ &= \sum_{x \in \mathcal{X}} \log(N \cdot P(x)) \cdot P(x) \\ &= \log N \sum_{x \in \mathcal{X}} P(x) + \sum_{x \in \mathcal{X}} \log P(x) \cdot P(x) \\ &= \log N - H(P). \end{aligned}$$

„Neurčitost“ je tedy míra podobnosti s rovnoměrným rozdělením. □

Definici divergence můžeme samozřejmě zobecnit na obecné distribuce:

Definice 6.9: *Divergence pro obecné distribuce*

Buď P, Q dvě pravděpodobnostní míry. Jejich *divergenci* definujeme vztahem

$$\begin{aligned} D(P \mid Q) &= \int \log \frac{dP}{dQ}\, dP \; pro \; P \ll Q, \\ D(P \mid Q) &= \infty \; jinak. \end{aligned}$$

Speciálně: pokud existují hustoty f a g, můžeme psát

$$D(P \mid Q) = E \log \frac{f(X)}{g(X)},$$

kde střední hodnotu počítáme vzhledem k hustotě f. □

Poznámka 6.10:

Minimalizace $D(f \mid g)$ vzhledem k hustotě g odpovídá v mnohorozměrném normálním rozdělení maximalizaci věrohodnostní funkce odvozené od hustoty g.
V případě dvou náhodných vektorů X, Y z definice dostáváme

$$D(f_{XY} \mid f_X f_Y) = E \log \frac{f_{XY}}{f_X f_Y}.$$

Jsou-li vektory X, Y nezávislé, rovná se uvedený vzorec nule.
Obdobně pro tři náhodné vektory X, Y, Z dostáváme

$$D(f_{XYZ} \mid f_{Y|X} f_{Z|X} f_X) = E \log \frac{f_{XYZ}}{f_{Y|X} f_{Z|X} f_X}.$$

Tento výraz je podle lemmatu 3.4 (iv) roven nule právě tehdy, když Y a Z jsou podmíněně nezávislé při pevném X.
Na tomto speciálním případě je dobře patrné, že divergence měří vzdálenost uvedených hustot, která je v případě nezávislosti nulová. □

Jednoduše lze dokázat rovněž následující lemma o rozkladu divergence (viz např. [39] - str. 19).

Lemma 6.11:

Mějme náhodný vektor (X_a, X_b, X_c) se sdruženou hustotou f_{abc} a nenulovými marginálními hustotami. Pak platí

$$D(f_{abc} \mid f_a f_{bc}) = D(f_{abc} \mid f_{a|b} f_{c|b} f_b) + D(f_{ab} \mid f_a f_b).$$

□

Dalším zobecněním, které bylo použito pro důkaz konvergence iterativních procedur v [29], je definice divergence pro 2 pozitivně definitní matice:

Definice 6.12: *Divergence pro pozitivně definitní matice*

Buď P, R dvě čtvercové pozitivně definitní matice řádu K. Jejich *divergenci* definujeme vztahem

$$D(P \mid R) = -\frac{1}{2}\left\{\log det(PR^{-1}) + tr(I - PR^{-1})\right\}.$$

□

Tato definice vychází z vyjádření divergence pro dvě normální rozdělení s hustotami $p(x)$ a $r(x)$, která jsou charakterizována kovariančními maticemi P a R. V tomto případě se divergence chová jako norma na prostoru pravděpodobnostních měr (viz [29] - str. 143).

Vlastnosti divergence pro pozitivně definitní matice shrnuje následující lemma:

Lemma 6.13: *Vlastnosti divergence pozitivně definitních matic*

Nechť $G = (K, E)$ je graf. Označme $\mathcal{M} = |K| \times |K|$ prostor všech čtvercových pozitivně definitních matic řádu K. Pak platí:

1. Pro $P, R \in \mathcal{M}$ je $D(P \mid R) \geq 0,\ D(P \mid R) = 0 \Leftrightarrow P = R$.
2. Mějme dány 2 pozitivně definitní matice $P, R \in \mathcal{M}$. Nechť existuje matice $Q \in \mathcal{M}$ taková, že platí:
   (a) $Q(\alpha, \beta) = P(\alpha, \beta)$ pro $(\alpha, \beta) \in E$.
   (b) $Q^{-1}(\alpha, \beta) = R^{-1}(\alpha, \beta)$ pro $(\alpha, \beta) \notin E$.
   Pak platí: $D(P \mid R) = D(P \mid Q) + D(Q \mid R)$.
   Navíc, pokud taková matice Q existuje, pak je určena jednoznačně.
3. Buď $\{A_n\}, \{B_n\}$ dvě posloupnosti na kompaktním podprostoru $\mathcal{M}$. Pak $D(A_n \mid B_n) \to 0 \Rightarrow A_n - B_n \to 0$.

□

Zejména třetí vlastnost je velmi důležitá při důkazu konvergence tzv. „maticových“ algoritmů, které si popíšeme v dalším textu.

Kapitola 7

Testy vstupních dat

Hlavním cílem disertační práce je popis grafických modelů určených pro zpracování dat se spojitým rozdělením a zejména pak aplikace vybraných odhadových procedur a procedur selekce modelů na konkrétní finanční data za účelem zodpovězení otázek z první kapitoly.

Konkrétní data je sice vhodnější popsat vždy až u příslušné aplikace, ale protože budeme v dalším textu používat ilustrace s odvětvovými indexy BCPP, musíme z tohoto pravidla učinit výjimku.

7.1 Odvětvové indexy českého kapitálového trhu

Burza cenných papírů Praha, a. s., zveřejňuje hodnoty oborových a průřezových indexů. Oborové indexy jsou publikovány od 6. 4. 1995, počítány jsou od 30. 9. 1994, kdy byla jejich hodnota stanovena na počátečních 1000 bodů. Tyto indexy jsou počítány pouze tehdy, pokud je počet emisí v bázi větší než tři. Bohužel, český kapitálový trh není příliš likvidní, a tak z původních 19 oborových indexů zbývá k 31.12.2004 pouhopouhých 8. Označení indexů, odvětví, kterému index přísluší, a stav k 31.12.2004 uvádíme v následující tabulce:

Tabulka 7.1: Oborové indexy zveřejňované BCPP

Označení	Popis odvětví	Stav k 31.12.2004	Datum zrušení
BI01	Zemědělství	Zrušen	15.2.1999
BI02	Výroba potravin	Zrušen	1.3.2000
BI03	Výroba nápojů a tabáku	Zrušen	15.7.2003
BI04	Těžba a zpracování nerostů	Počítán	x
BI05	Textilní, oděvní a kožedělný průmysl	Zrušen	29.8.2002
BI06	Dřevařský a papírenský průmysl	Zrušen	2.5.2000
BI07	Chemický, farmac. a gumár. průmysl	Počítán	x
BI08	Stavebnictví a prům. stavebních hmot	Počítán	x
BI09	Hutnictví, zpracování kovů	Zrušen	21.12.2001
BI10	Strojírenství	Zrušen	30.12.2004
BI11	Elektrotechnika a elektronika	Zrušen	23.9.2004
BI12	Energetika	Počítán	x
BI13	Doprava a spoje	Počítán	x
BI14	Obchod	Počítán	x
BI15	Peněžnictví	Počítán	x
BI16	Služby	Počítán	x
BI17	Bižuterie, sklo a keramika	Zrušen	25.9.2001
BI18	Investiční fondy	Zrušen	19.9.2002
BI19	Ostatní	Zrušen	11.6.2002

Poznamenejme ještě, že „rušení“ odvětvových indexů pokračuje i po 31.12.2004[1]:

- 14. 2. 2005 zrušen BI13: Doprava a spoje.
- 29.4.2005 zrušen BI08: Stavebnictví a prům. stavebních hmot
- 12.7.2005 zrušen BI14: Obchod
- 21.7.2005 zrušen BI04: Těžba a zpracování nerostů
- 1.9.2005 zrušen BI15: Peněžnictví

BCPP zveřejňuje také 3 průřezové indexy[2]: PX 50 (od 5. 4. 1994), PX-D (od 4. 1. 1999) a PX-GLOB (od 30. 9. 1994). O nejvýznamnějším z indexů - PX 50 se zmíníme ještě dále.

[1] K 31.12.2005 jsou tak zveřejňovány pouze 3 oborové indexy:

- BI07: Chemický, farmac. a gumár. průmysl
- BI12: Energetika
- BI16: Služby

Poslední únorový den roku 2006 byly navíc zrušeny i tyto zbylé oborové indexy - viz [10].

[2] K 20.3.2006 došlo díky skutečnosti, že i když PX 50 obsahoval více emisí, na jeho pohyb měly vliv pouze emise ze SPAD a jeho vývoj byl tedy téměř totožný s vývojem indexu PX-D, ke sloučení indexů PX 50 a PX-D v jediný index s názvem PX - viz [10].

7.2 Testy předpokladů

Dříve než přistoupíme k popisu metod pro selekci grafických modelů pro spojitá data a jejich použití na vybrané indexy, musíme ověřit, zda data splňují předpoklady požadované pro dané modely. Konkrétně to znamená ověřit, zda napozorované měsíční hodnoty indexů jsou realizací nezávislých veličin s normálním rozdělením. Nejprve však provedeme výpočet tzv. *logaritmických výnosů*, u kterých můžeme spíše předpokládat výše uvedené vlastnosti:

$$\log\left(\frac{P_t}{P_{t-1}}\right) = \log P_t - \log P_{t-1}, \tag{7.1}$$

kde:
P_t = hodnota indexu na konci měsíce t,
P_{t-1} = hodnota indexu na konci měsíce $t-1$,
a log označuje přirozený logaritmus.

7.2.1 Testy nezávislosti

K ověření nezávislosti logaritmických výnosů použijeme dva z testů náhodnosti uvedených v knize [4], str. 94-99. Před samotným započetím testu nejprve z dané řady až na jednu vyškrtáme stejné sousední hodnoty.

Test založený na znaménkách diferencí

První z použitých testů je *test založený na znaménkách diferencí*:

Označme n počet pozorování (délku) testované řady dat a k počet kladných diferencí v této řadě (tj. počet bodů, v nichž daná řada roste). Pak za platnosti hypotézy nezávislosti platí (viz [4] - str. 95):

$$E(k) = \frac{n-1}{2}$$

$$Var(k) = \frac{n+1}{12}$$

a k má asymptoticky normální rozdělení. Pro větší n tedy zamítáme na hladině α hypotézu, že data jsou realizacemi nezávislých stejně rozdělených náhodných veličin, pokud platí:

$$\left|\frac{k-\frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}}\right| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ označuje $(1-\alpha/2)\cdot 100$ procentní kvantil normovaného normálního rozdělení $N(0,1)$.
V našem případě však použijeme alternativní kritérium založené na p-value, tj. na nejnižší hladině, na které daný test zamítá nulovou hypotézu. P-value pro test založený na znaménkách diferencí spočteme podle následujícího vzorce:

$$p-value = 2\left(1-\Phi\left(\left|\frac{k-\frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}}\right|\right)\right), \tag{7.2}$$

kde:
Φ značí distribuční funkci normovaného normálního rozdělení $N(0,1)$.
Hypotézu nezávislosti zamítáme na hladině α, pokud platí:

$$p - value \leq \alpha.$$

Test založený na bodech zvratu

Druhým testem je *test založený na bodech zvratu*:

Řekneme, že bod y_t je horním bodem zvratu uvažované řady, když $y_{t-1} < y_t > y_{t+1}$, a dolním bodem zvratu, pokud naopak platí $y_{t-1} > y_t < y_{t+1}$.

Nechť r označuje celkový počet horních a dolních bodů zvratu dohromady. Pak za platnosti hypotézy nezávislosti platí (viz [4] - str. 96):

$$E(r) = \frac{2(n-2)}{3}$$

$$Var(r) = \frac{16n - 29}{90}$$

a r má asymptoticky normální rozdělení. Pro větší n tedy zamítáme na hladině α hypotézu, že data jsou realizacemi nezávislých stejně rozdělených náhodných veličin, pokud platí:

$$\left| \frac{r - \frac{2(n-2)}{3}}{\sqrt{\frac{16n-29}{90}}} \right| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ označuje $(1-\alpha/2)\cdot 100$ procentní kvantil normovaného normálního rozdělení $N(0,1)$.
V našem případě opět použijeme alternativní kritérium založené na p-value. P-value pro test založený na bodech zvratu spočteme podle následujícího vzorce:

$$p - value = 2\left(1 - \Phi\left(\left| \frac{r - \frac{2(n-2)}{3}}{\sqrt{\frac{16n-29}{90}}} \right|\right)\right), \tag{7.3}$$

kde:
Φ značí distribuční funkci normovaného normálního rozdělení $N(0,1)$.
Hypotézu nezávislosti zamítáme na hladině α, pokud platí:

$$p - value \leq \alpha.$$

Poznámka 7.2:

Zatímco test založený na znaménkách diferencí se doporučuje při podezření na existenci lineárního trendu (tj. systematického posuvu nahoru nebo dolů) v předložené řadě, test založený na bodech zvratu je citlivý spíše na změny periodického charakteru - viz [4], str. 98. Protože v našem případě (tj. u měsíčních logaritmických výnosů akciových indexů) mohou teoreticky nastat obě dvě varianty, použijeme oba dva testy. □

Oba testy jsme naprogramovali v programu Mathematica 4.0 a lze je nalézt v příloze.

7.2.2 Test normality

Také k ověření normality existuje celá řada testů - například test Shapiro-Wilk, kterým byla testována zpracovávaná data[3] v [14], [15], [17], [18]. Tentokrát však použijeme test, který lze poměrně lehce implementovat do zvoleného software[4] (Mathematica 4.0) a zároveň je aplikovatelný již při malém počtu pozorování. Jde o test založený na šikmosti a_3 a špičatosti a_4, kde (viz [2]):

$$\begin{aligned} a_3 &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{s^3}, \\ a_4 &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{s^4}, \\ \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 . \end{aligned}$$

Platí-li hypotéza, že $x_1, \ldots, x_n$ je výběr z normálního rozdělení, pak a_3 a a_4 mají asymptoticky normální rozdělení s parametry:

$$E(a_3) = 0, \qquad Var(a_3) = \tfrac{6(n-2)}{(n+1)(n+3)}.$$

$$E(a_4) = 3 - \tfrac{6}{n+1}, \quad Var(a_4) = \tfrac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

K testům proti alternativě, že výběr pochází z nějakého nesymetrického rozdělení, použijeme normovanou veličinu U_3, test proti alternativám, které se liší špičatostí, založíme na U_4:

$$U_3 = \frac{a_3}{\sqrt{Var(a_3)}} \sim N(0,1), \quad U_3 = \frac{a_4 - Ea_4}{\sqrt{Var(a_4)}} \sim N(0,1).$$

Bohužel, asymptotiky lze využít teprve pro velká n (v praxi se limitních výsledků užívá v případě šikmosti pro $n \geq 200$ a v případě špičatosti dokonce až pro $n \geq 500$ - viz [2]).

V našem případě však máme k dispozici pouze desítky dat, nikoli stovky. Proto pou-

[3] Bylo využito programu Statistica, kde je test nabízen přímo v základním balíčku testů.

[4] Budeme tak mít všechny testy vstupních dat i algoritmus výpočtu v jednom programu.

žijeme vylepšení tohoto postupu, který navrhl *D'Agostino*. Položme:

$$b = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}, \qquad B = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}},$$

$$W^2 = \sqrt{2(b-1)} - 1,$$

$$\delta = \frac{1}{\sqrt{\ln W}},$$

$$a = \sqrt{\frac{2}{W^2-1}}, \qquad A = 6 + \frac{8}{B}\left(\frac{2}{B} + \sqrt{1+\frac{4}{B^2}}\right),$$

$$Z_3 = \delta \ln\left[\frac{U_3}{a} + \sqrt{\left(\frac{U_3}{a}\right)^2 + 1}\right], \quad Z_4 = \frac{1 - \frac{2}{9A} - \sqrt[3]{\frac{1-\frac{2}{A}}{1+U_4\sqrt{\frac{2}{A-4}}}}}{\sqrt{\frac{2}{9A}}}.$$

Pro veličiny Z_3 a Z_4 lze asymptotickou normality použít již pro $n \geq 9$ v případě šikmosti a pro $n \geq 20$ v případě špičatosti, což je pro naše případy zcela dostačující.

Hypotézu normality zamítneme v případě, že $|Z_3| \geq u(\alpha/2)$ nebo $|Z_4| \geq u(\alpha/2)$.

Protože nevíme, zda se námi použitá data mohou lišit od normálního rozdělení šikmostí nebo špičatostí, použijeme kombinaci obou přístupů - test založený na šikmosti a špičatosti zároveň - tzv. Omnibus test (viz [38]). Abychom mohli použít asymptotických výsledků již pro malá n ($n \geq 20$), opět využijeme veličin Z_3 a Z_4. Součet jejich kvadrátů má asymptoticky χ^2 rozdělení se 2 stupni volnosti. Normalitu tedy zamítáme v případě, že $Z_3^2 + Z_4^2 \geq \chi_2^2(\alpha)$.

V našem případě již tradičně použijeme alternativní kritérium založené na p-value. P-value pro test omnibus spočteme podle následujícího vzorce:

$$p - value = 1 - F\left(Z_3^2 + Z_4^2\right), \tag{7.4}$$

kde:
F značí distribuční funkci χ^2 rozdělení se 2 stupni volnosti.
Hypotézu nezávislosti zamítáme na hladině α, pokud platí:

$$p - value \leq \alpha.$$

Také tento test implementovali do programu Mathematica a lze jej nalézt v příloze.

7.2.3 Výsledky

Následující tabulka shrnuje výsledky provedených testů. Díky výše popsané situaci se v aplikacích musíme spokojit s daty do roku 2004[5] :

[5]Disertační práce byla zpracovávána během delšího časového období, a proto některé aplikace využívají starších dat.

Tabulka 7.3: Výsledky testů nezávislosti a normality pro odvětvové indexy BCPP v období 2001-2004

Log. výnosy indexu	**P-value testu znamének diferencí**	**P-value testu bodů zvratu**	**P-value testu OMNIBUS**
BI04	0,4579	0,1034	0,0559
BI07	0,0833	0,9074	0,4086
BI08	0,8046	0,1305	0,0863
BI12	0,4579	0,6417	0,1998
BI13	0,8046	0,8160	0,9890
BI14	0,4485	0,4051	0,0000
BI15	0,2160	0,1034	0,5683
BI16	0,4579	0,8160	0,0205

Z tabulky 7.3 je patrné, že na pětiprocentní hladině významnosti zamítáme normalitu u logaritmických výnosů indexu BI14: Obchod a BI16: Služby. Nezávislost nelze zamítnout u žádného z odvětvových indexů.

Zbývajících šest oborových indexů BCPP je tedy možno použít pro další výpočty.

Kapitola 8

Analýza deviance

Elegantním nástrojem pro test, zda daný grafický model dobře popisuje skutečnou strukturu dat, je deviance. Jak si ukážeme v této kapitole, její použití dovoluje testovat různé hypotézy a také konstruovat tabulky podobné známé analýze rozptylu.

8.1 Základní definice

Definice 8.1:

Saturovaný model je grafický model určený úplným grafem. □

Definice 8.2: *Celková deviance*

Nechť G_0 je úplný graf a G jeho faktor, ve kterém chybí f hran.
Pak definujeme *devianci* (dále budeme používat rovněž pojmu *celková deviance*) grafu G:

$$dev^{(f)} = 2 \cdot [l(S) - l(\hat{V})],$$

kde:
S je maximálně věrohodný odhad v saturovaném modelu s úplným grafem G_0, což je výběrová varianční matice,
$\hat{V}$ je maximálně věrohodný odhad v modelu s grafem G a
l značí logaritmickou věrohodnostní funkci. □

Lemma 8.3:

Nechť $X^1, X^2, ..., X^N$ je náhodný výběr z k-rozměrného normálního rozdělení $N(0, V)$.
Pak devianci můžeme vyjádřit ve tvaru:

$$dev^{(f)} = N[tr(S\hat{D}) - \log\det(S\hat{D}) - k]. \tag{8.1}$$

Navíc platí:

$$dev^{(1)} = -N\log(1 - corr(X_i, X_j|rest)^2) \sim \chi^2_1, \tag{8.2}$$

kde $rest$ označuje množinu zbývajících proměnných. □

Lemma 8.4:

Pro graf bez hran ($f = \begin{pmatrix} k \\ 2 \end{pmatrix}$) můžeme díky důsledku 5.9 psát:

$$dev^{(f)} = N[tr(S\ diag\{\frac{1}{s_{ii}}\}) - \log\det(S\ diag\{\frac{1}{s_{ii}}\}) - k]. \tag{8.3}$$

□

Věta 8.5:

Deviance $dev^{(f)}$ má asymptoticky rozdělení χ^2_f. □

Důkaz je uveden v [31] na straně 187.

Definice 8.6: *Diference deviancí*

Nechť G_0 je úplný graf, G_1 jeho faktor, ve kterém chybí f_1 hran a který má devianci $dev^{(f_1)}$, a G_2 faktor grafu G_1, ve kterém chybí f_2 hran a který má devianci $dev^{(f_2)}$. Symbolicky lze psát $G_0 \supset G_1 \supset G_2$.
Pak definujeme *diferenci deviancí* modelů s grafy $G_1 \supset G_2$ předpisem:

$$dev^* = -[dev^{(f_1)} - dev^{(f_2)}].$$

□

Věta 8.7: *Výpočetní tvar diference deviancí*

Nechť $X^1, X^2, ..., X^N$ je náhodný výběr z k-rozměrného normálního rozdělení $N(0, V)$.
Pak diferenci deviancí můžeme vyjádřit ve tvaru

$$dev^* = -N \cdot \Big\{tr[S \cdot (\hat{D}_1 - \hat{D}_2) + \log det(\hat{V}_1 \cdot \hat{D}_2)]\Big\}, \tag{8.4}$$

kde: $\hat{D}_1 = \hat{V}_1^{-1}$ je maximálně věrohodný odhad inverzní varianční matice v grafickém modelu s grafem G_1 a $\hat{D}_2 = \hat{V}_2^{-1}$ je maximálně věrohodný odhad inverzní varianční matice v grafickém modelu s grafem G_2. □

Důkaz:

Pro devianci grafu G_1 platí podle lemmatu 8.3:

$$dev^{(f_1)} = N[tr(S\hat{D}_1) - \log\det(S\hat{D}_1) - k].$$

Obdobně můžeme spočítat devianci grafu G_2:

$$dev^{(f_2)} = N[tr(S\hat{D}_2) - \log\det(S\hat{D}_2) - k].$$

S využitím vlastností stopy matice dostáváme:

$$\begin{aligned} -dev^* &= dev^{(f_1)} - dev^{(f_2)} \\ &= Ntr(S\hat{D}_1) - Ntr(S\hat{D}_2) - N\log\det(S\hat{D}_1) + N\log\det(S\hat{D}_2) \\ &= Ntr(S\hat{D}_1 - S\hat{D}_2) + N\log\frac{detS\cdot det\hat{D}_2}{detS\cdot det\hat{D}_1} \\ &= N\cdot\Big\{tr[S\cdot(\hat{D}_1-\hat{D}_2)+\log det(\hat{V}_1\cdot\hat{D}_2)]\Big\}. \end{aligned}$$

□

V dále popsaných iteračních algoritmech narazíme často na případ, kdy se 2 grafy liší pouze v jedné hraně. Vyplatí se tedy zavést ještě další speciální definice:

Definice 8.8: *Deviance vynechané hrany*

Nechť G_0 je úplný graf. Nechť G_1 je jeho faktor, ve kterém chybí f_1 hran (a má tedy devianci $dev^{(f_1)}$). Nechť je dále G_2 faktor grafu G_1, ve kterém chybí oproti G_1 pouze jedna hrana ($G_2 = G_1 \setminus \{i,j\}$). *Devianci vynechané hrany* definujeme předpisem

$$dev^*_{ij} = -(dev^{(f_1)} - dev^{(f_1-1)}) = dev(G_2) - dev(G_1).$$

□

Definice 8.9: *Deviance přidané hrany*

Nechť G_0 je úplný graf. Nechť G_1 je jeho faktor, ve kterém chybí f_1 hran (a má tedy devianci $dev^{(f_1)}$). Nechť dále G_2 má oproti G_1 jednu přidanou hranu ($G_2 = G_1 \cup \{i,j\}$). *Devianci přidané hrany* definujeme

$$dev^*_{ij} = -(dev^{(f_1+1)} - dev^{(f_1)}) = dev(G_1) - dev(G_2).$$

□

Poznámka 8.10:

Stejné značení pro devianci vynechané a přidané hrany dev^*_{ij} není zvoleno náhodně - jde totiž vždy o 2 grafy lišící se pouze o jednu hranu. □

Věta 8.11:

Diference deviancí dev^* má asymptoticky rozdělení $\chi^2_{f_2-f_1}$.
Speciálně:
Deviance vynechané/přidané hrany dev^*_{ij} má asymptoticky rozdělení χ^2_1. □

Důkaz je uveden v [31] na straně 187.

Poznámka 8.12:

Deviance je vhodným testovým kritériem pro testování shody grafického modelu s daty. Umožňuje testovat grafický model, který tvoří faktor alternativního grafu, a posoudit tak, zda lépe reprezentuje data.
Označme:

Graf	**Popis**	**Počet hran**	**Počet vynechaných hran**
G_0	Úplný graf	$\binom{k}{2}$	0
G_1	Faktor G_0	$\binom{k}{2} - f_1$	f_1
G_2	Faktor G_1	$\binom{k}{2} - f_2$	f_2

Platí tedy $G_0 \supset G_1 \supset G_2$ a $f_2 > f_1 > 0$

1. *Celková deviance* je testová statistika modelu s grafem G_1 proti saturovanému modelu s grafem G_0.
   Jako nulovou hypotézu volíme graf G_1, ve kterém jsme vynechali f_1 hran, a jako alternativní hypotézu úplný graf G_0 (tj. saturovaný model s $\binom{k}{2}$ hranami), $G_0 \subset G_1$:

   $$H_0\text{: } G_1 \qquad \binom{k}{2} - f_1 \text{ hran}$$
   $$H_1\text{: } G_0 \qquad \binom{k}{2} \text{ hran}$$

   a pokud $dev(G_1) \geq \chi^2_{f_1}(\alpha)$, zamítáme nulovou hypotézu ve prospěch alternativní. Graf, který jsme zkonstruovali vynecháním některých hran z úplného grafu, tedy nevystihuje strukturu podmíněných nezávislostí dat lépe než graf úplný (kdy všechny proměnné jsou na sobě navzájem závislé). Kritickými hodnotami jsou dle věty 8.5 hodnoty χ^2 rozdělení na zvolené hladině spolehlivosti α, počet stupňů volnosti odpovídá počtu vynechaných hran v grafu G_1.
2. *Diference deviancí* je testová statistika modelu s grafem G_2 s f_2 vynechanými hranami, proti modelu s grafem G_1 s f_1 vynechanými hranami ($G_1 \supset G_2$, $f_2 > f_1$).

   $$H_0\text{: } G_2 \qquad \binom{k}{2} - f_2 \text{ hran}$$
   $$H_1\text{: } G_1 \qquad \binom{k}{2} - f_1 \text{ hran}$$

   Nulovou hypotézu zamítáme, pokud $dev(G_2) - dev(G_1) \geq \chi^2_{f_2 - f_1}$. Graf s více vynechanými hranami tedy nevystihuje strukturu podmíněných nezávislostí dat lépe než graf, ve kterém chybí méně hran. Kritickými hodnotami jsou dle věty 8.11 hodnoty χ^2 rozdělení na zvolené hladině spolehlivosti α, počet stupňů volnosti odpovídá počtu hran, o které se 2 testované grafy liší.

3. *Deviance vynechané hrany* testuje vhodnost vynechání jedné konkrétní hrany (i, j) z grafu. Je tedy

$$H_0\text{: } G_2 = G_1 \setminus \{i, j\} \qquad \binom{k}{2} - f_1 - 1 \text{ hran}$$
$$H_1\text{: } G_1 \qquad \binom{k}{2} - f_1 \text{ hran}$$

(samozřejmě může být $f_1 = 0$ a alternativní hypotézou je tedy úplný graf). Nulovou hypotézu opět zamítáme, pokud $dev(G_2) - dev(G_1) = dev^*_{ij} \geq \chi^2_1(\alpha)$. V dále uvedených backward algoritmech příslušnou hranu z grafu G_1 vyloučíme pouze pokud $dev^*_{ij} < \chi^2_1(\alpha)$.

4. *Deviance přidané hrany* naopak testuje vhodnost přidání jedné konkrétní hrany (i, j) do grafu. Je tedy

$$H_0\text{: } G_2 \qquad \binom{k}{2} - f_2 \text{ hran}$$
$$H_1\text{: } G_1 = G_2 \cup \{i, j\} \qquad \binom{k}{2} - f_2 + 1 \text{ hran}$$

(opět může být $f_2 = 1$ a alternativní hypotézou je tedy úplný graf). Nulovou hypotézu opět zamítáme, pokud $dev(G_2) - dev(G_1) = dev^*_{ij} \geq \chi^2_1(\alpha)$. V dále uvedených forward algoritmech v tomto případě tedy do grafu G_2 hranu (i, j) přidáme.

□

8.2 Možnosti výpočtu $\hat{V}$

Podívejme se nyní na způsoby, kterými je možno z konkrétních dat získat odhad varianční matice za podmínek daných zvoleným grafickým modelem.

8.2.1 Přímý výpočet

Jednou z možností, jak určit maximálně věrohodný odhad varianční matice $\hat{V}$ za omezujících podmínek daných konkrétním grafickým modelem, je *přímý výpočet*. Tato metoda byla aplikována v diplomové práci [22], kde lze nalézt také její podrobný popis. Omezíme se proto pouze na ilustrační příklad.

Příklad 8.13:

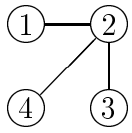
Vezměme si měsíční logaritmické výnosy následujících 4 odvětvových indexů (pro přehlednější grafické vyjádření výsledků indexy očíslujeme přirozenými čísly $1, \ldots, 4$):

Označení	Index	Odvětví
1	BI04	Těžba a zpracování nerostů a rud
2	BI07	Chemický, farmaceutický a gumozpracující průmysl
3	BI08	Stavebnictví a průmysl stavebních hmot
4	BI12	Energetika

Varianční matice těchto indexů má následující hodnoty:

$$S = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ 0,000995 & 0,000633 & 0,001765 & \\ 0,000984 & 0,001908 & 0,000776 & 0,002950 \end{pmatrix}.$$

Nechť grafický model pro tato data, za jehož omezujících podmínek chceme najít maximálně věrohodný odhad varianční matice $\hat{V}$, je určen grafem:



Graf má kliky {1, 2}, {2, 3}, {2, 4} a z věrohodnostních rovnic dostáváme

$$\begin{pmatrix} \hat{v}_{11} & & & \\ \hat{v}_{12} & \hat{v}_{22} & & \\ \hat{v}_{13} & \hat{v}_{23} & \hat{v}_{33} & \\ \hat{v}_{14} & \hat{v}_{24} & \hat{v}_{34} & \hat{v}_{44} \end{pmatrix} = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ ? & 0,000633 & 0,001765 & \\ ? & 0,001908 & ? & 0,002950 \end{pmatrix}.$$

V grafu chybí hrany {1, 3}, {1, 4}, {3, 4} a musí tedy platit $\hat{d}_{13} = 0$, $\hat{d}_{14} = 0$, $\hat{d}_{34} = 0$, kde $\hat{D} = \hat{V}^{-1}$. Tyto podmínky můžeme přepsat do tvaru:

$$det \begin{pmatrix} 0,001886 & \hat{v}_{13} & \hat{v}_{14} \\ 0,006152 & 0,000633 & 0,001908 \\ 0,001908 & \hat{v}_{34} & 0,002950 \end{pmatrix} = 0$$

$$det \begin{pmatrix} 0,001886 & \hat{v}_{13} & \hat{v}_{14} \\ 0,006152 & 0,000633 & 0,001908 \\ 0,000633 & 0,001765 & \hat{v}_{34} \end{pmatrix} = 0$$

$$det \begin{pmatrix} 0,003394 & 0,001886 & \hat{v}_{14} \\ 0,001886 & 0,006152 & 0,001908 \\ \hat{v}_{13} & 0,000633 & \hat{v}_{34} \end{pmatrix} = 0$$

K vyřešení těchto rovnic můžeme použít například následující krátký program v Mathematice 4.0.

mat1 = {{0.001886 , v13, v14}, {0.006152, 0.000633, 0.001908}, {0.001908, v34, 0.002950}};
mat2 = {{ 0.001886, v13, v14}, {0.006152, 0.000633, 0.001908}, {0.000633, 0.001765, v34}};
mat3 = {{0.003394, 0.001886, v14}, {0.001886, 0.006152, 0.001908}, {v13, 0.000633, v34}};

Solve[Det[mat1] == 0, Det[mat2] == 0, Det[mat3] == 0, {v13,v14,v34}] // N

Získali jsme však ne jedno, ale hned několik možných řešení:

Řešení 1: $(\hat{v}_{13}, \hat{v}_{14}, \hat{v}_{34}) = (-3022.22, -3559.76, 0.0021985)$

Řešení 2: $(\hat{v}_{13}, \hat{v}_{14}, \hat{v}_{34}) = (-0.384799, 0.471943, -0.00180586)$

Řešení 3: $(\hat{v}_{13}, \hat{v}_{14}, \hat{v}_{34}) = (-0.00510788, -0.000683093, 0.000196321)$

Řešení 4: $(\hat{v}_{13}, \hat{v}_{14}, \hat{v}_{34}) = (0.415831, -0.471943, -0.00180586)$

Řešení 5: $(\hat{v}_{13}, \hat{v}_{14}, \hat{v}_{34}) = (3022.15, 3559.58, 0.0021985)$

To, které zvolit jako výsledné, do značné míry závisí na subjektivní volbě, popřípadě (pokud se použije pro řešení rovnic nějaká numerická metoda) na volbě počátečních podmínek. Tento problém je ostatně zmíněn i v diplomové práci [22], kde se podobné rovnice řešily pomocí programu MATLAB. □

8.2.2 Iterační algoritmy

Řešením problému značné citlivosti řešení na počáteční podmínku (a rovněž odstraněním neúměrné pracnosti přímého výpočtu) je použít některý z iteračních algoritmů. Asi nejstarší verzí těchto algoritmů je tzv. *IPF algoritmus*, který lze nalézt v [31] na stranách 182-185. Rovněž v [29] jsou popsány 2 algoritmy založené na podobném základě. Podívejme se na vybrané postupy nyní trochu podrobněji:

IPF algoritmus

Jak již bylo řečeno, jedná se o iterační algoritmus, kde IPF je zkratkou z anglického iterative proportional fitting. Tento algoritmus je jádrem diplomové práce [26], kde lze nalézt jeho podrobné odvození. Omezíme se proto pouze na stručný popis a ilustrační příklad.

Popis algoritmu:

Mějme dány dvě hustoty, g^0 a f, pro k-rozměrný náhodný vektor X. Cílem algoritmu je nalézt hustotu g^∞, která má stejnou interakční strukturu jako g^0 a stejná marginální rozdělení jako f na podmnožinách $a_1, a_2, ..., a_m$ množiny vrcholů $K = \{v_1, \ldots, v_k\}$. Tyto podmnožiny sice nemusí být disjunktní, ale žádná nesmí být částí některé jiné. Navíc musí platit

$$\bigcup_{i=1}^{m} a_i = K.$$

Uvedené vlastnosti splňují kliky daného grafu, a proto je také budeme za tyto podmnožiny volit.

V n-tém kroku platí:

$$g_{ab}^{n+1} = g_{b|a}^{n} f_a.$$

Za a volíme postupně v cyklu kliky daného grafu, tedy:

$$\begin{aligned} a &= a_1 \; pro \; n = 1 \\ a &= a_m \; pro \; n = m \\ a &= a_1 \; pro \; n = m + 1 \end{aligned}$$

$atd.$

a b určíme jako doplněk $b = K \backslash a$.

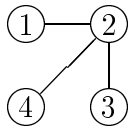
Příklad 8.14:

Pokračujme v příkladu s logaritmickými výnosy indexů BI04, BI07, BI08 a BI12.

Varianční matice těchto indexů má následující hodnoty:

$$S = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ 0,000995 & 0,000633 & 0,001765 & \\ 0,000984 & 0,001908 & 0,000776 & 0,002950 \end{pmatrix}.$$

Pomocí IPF algoritmu získáme maximálně věrohodný odhad varianční matice $\hat{V}$ za omezujících podmínek daných grafickým modelem s grafem



ve tvaru:

$$\hat{V} = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ 0,000194 & 0,000633 & 0,001765 & \\ 0,000585 & 0,001908 & 0,000196 & 0,002950 \end{pmatrix}.$$

Kliky grafu jsou $\{1, 2\}$, $\{2, 3\}$, $\{2, 4\}$ a z porovnání obou matic snadno zjistíme, že pro tyto kliky skutečně platí $S_{aa} = \hat{V}_{aa}$.

$$S - \hat{V} = \begin{pmatrix} 0,000000 & & & \\ 0,000000 & 0,000000 & & \\ 0,000801 & 0,000000 & 0,000000 & \\ 0,000399 & 0,000000 & 0,000579 & 0,000000 \end{pmatrix}.$$

Inverzní varianční matice k matici $\hat{V}$ je tvaru

$$\hat{D} = \begin{pmatrix} 355,144431 & & & \\ -108,876008 & 242,936184 & & \\ 0,000000 & -60,475978 & 588,101350 & \\ 0,000000 & -131,490557 & 0,000000 & 423,979252 \end{pmatrix}.$$

Tato matice splňuje $\hat{d}_{13} = 0$, $\hat{d}_{14} = 0$, $\hat{d}_{34} = 0$. Platí tedy, že $\hat{d}_{ij} = 0$, kdykoli nejsou vrcholy i a j v daném grafu spojeny hranou. □

„Maticové“ algoritmy

Popišme si nyní novější postup, který nám zajistí stejné výsledky, avšak ve zvoleném software bude pracovat efektivněji. Jedná se o 2 algoritmy uvedené v [29], které můžeme souhrnně označit jako *maticové*. Proč právě tento název bude patrné z následujících vzorečků.

Oba algoritmy jsou založeny na následující větě:

Věta 8.15:

Mějme dány dvě pozitivně definitní matice A a B definované na vrcholech K grafu $G = (K, E)$. Pak existuje pozitivně definitní matice V tak, že

1. $V_{\alpha,\beta} = A_{\alpha,\beta}$ když $(\alpha, \beta) \in E$ nebo $\alpha = \beta$
2. $V^{-1}_{\alpha,\beta} = B_{\alpha,\beta}$ když $(\alpha, \beta) \notin E$ nebo $\alpha \neq \beta$

Ekvivalentně můžeme toto tvrzení přeformulovat pro zápis pomocí klik:

1. $V_{c,c} = A_{c,c}$, když c je klika
2. $V^{-1}_{ac,ac} = B_{ac,ac}$, když ac je antiklika

□

I. algoritmus

První algoritmus běží v cyklu přes antikliky ac a STOP pravidlo testuje, zda jsou mimodiagonální prvky inverzní varianční matice na klikách nulové ($V^{-1}_{ac,ac} = 0$). Postup výpočtu je následující:

1. Vyjádříme graf G pomocí antiklik jako $ac_1, \ldots, ac_m$
2. Generujeme posloupnost matic $\{\hat{V}^n\}$ následujícím postupem:

- $\hat{V}^0 = S$
- $$\left(\hat{V}^{n+1}\right)^{-1} = \begin{bmatrix} B_{a,a} & B_{a,a}\left(R^n_{a,a}\right)^{-1} R^n_{a,b} \\ R^n_{b,a}\left(R^n_{a,a}\right)^{-1} B_{a,a} & R^n_{b,b} - R^n_{b,a}\left(R^n_{a,a}\right)^{-1}\left(I - B_{a,a}\left(R^n_{a,a}\right)^{-1}\right) R^n_{a,b} \end{bmatrix},$$

kde $R = \hat{V}^{-1}$, $B_{a,a} = diag\left(\left(\hat{V}^{-1}\right)^{-1}_{a,a}\right)^{-1}$,
$a = ac_1, \ldots, ac_m$, $b = K \backslash \{a\}$, $n = n' \bmod m$.

II. algoritmus

Druhý algoritmus běží v cyklu přes kliky c a STOP pravidlo testuje, zda jsou prvky varianční matice a výběrové varianční matice na klikách shodné ($V_{c,c} = S_{c,c}$). Postup výpočtu je následující:

1. Vyjádříme graf G pomocí klik jako $c_1, \ldots, c_m$

2. Generujeme posloupnost matic $\{\hat{V}^n\}$ následujícím postupem:

- $\hat{V}^0 = I$
- $\left(\hat{V}^{n+1}\right)^{-1} = \begin{bmatrix} B_{a,a} & B_{a,a}\left(R^n_{a,a}\right)^{-1} R^n_{a,b} \\ R^n_{b,a}\left(R^n_{a,a}\right)^{-1} B_{a,a} & R^n_{b,b} - R^n_{b,a}\left(R^n_{a,a}\right)^{-1}\left(I - B_{a,a}\left(R^n_{a,a}\right)^{-1}\right) R^n_{a,b} \end{bmatrix},$

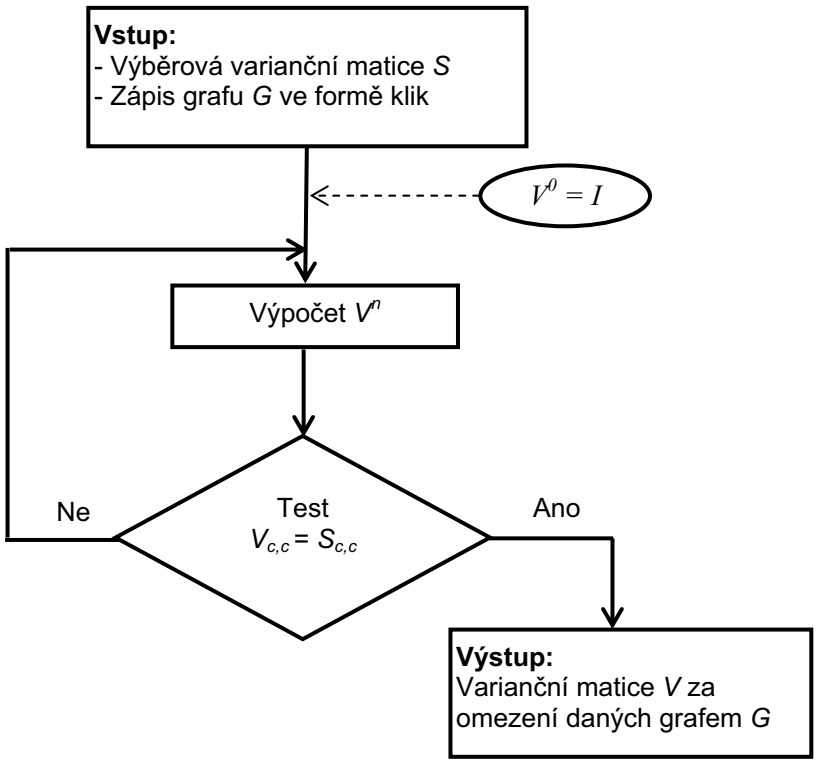
kde $R = \hat{V}$, $B_{a,a} = S_{a,a}$
$a = c_1, \ldots, c_m$, $b = K\backslash\{a\}$, $n = n' \bmod m$.

Po dosazení můžeme druhý algoritmus přepsat do tvaru:

$$\left(\hat{V}^{n+1}\right)^{-1} = \begin{bmatrix} S_{a,a} & S_{a,a}\left(\hat{V}^n_{a,a}\right)^{-1} \hat{V}^n_{a,b} \\ \hat{V}^n_{b,a}\left(\hat{V}^n_{a,a}\right)^{-1} S_{a,a} & \hat{V}^n_{b,b} - \hat{V}^n_{b,a}\left(\hat{V}^n_{a,a}\right)^{-1}\left(I - S_{a,a}\left(\hat{V}^n_{a,a}\right)^{-1}\right) \hat{V}^n_{a,b} \end{bmatrix}.$$

Druhý algoritmus jsme implementovali[1] do SW Mathematica jako jádro procedury pro odhad grafického modelu k daným datům. Celý algoritmus lze nalézt v příloze, zde se omezíme na jeho znázornění pomocí vývojového diagramu:

Obrázek 8.16: *Schéma maticového algoritmu*



Podívejme se ještě, jak vypadá odhad varianční matice a její inverze, pokud použijeme stejná data jako v příkladu 8.14 - tj. logaritmické výnosy indexů BI04, BI07, BI08 a BI12.

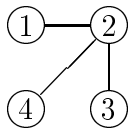
[1] K volbě právě druhého algoritmu vedla skutečnost, že pracuje s reprezentací pomocí klik a jde tak lehčeji začlenit do již hotových selekčních procedur naprogramovaných v rámci [13].

Příklad 8.17:

Varianční matice těchto indexů má následující hodnoty:

$$S = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ 0,000995 & 0,000633 & 0,001765 & \\ 0,000984 & 0,001908 & 0,000776 & 0,002950 \end{pmatrix}.$$

Pomocí maticového algoritmu získáme maximálně věrohodný odhad varianční matice $\hat{V}$ za omezujících podmínek daných grafickým modelem s grafem



ve tvaru:

$$\hat{V} = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ 0,000194 & 0,000633 & 0,001765 & \\ 0,000585 & 0,001908 & 0,000196 & 0,002950 \end{pmatrix}.$$

Kliky grafu jsou {1, 2}, {2, 3}, {2, 4} a z porovnání obou matic snadno zjistíme, že pro tyto kliky skutečně platí $S_{aa} = \hat{V}_{aa}$.

$$S - \hat{V} = \begin{pmatrix} 0,000000 & & & \\ 0,000000 & 0,000000 & & \\ 0,000801 & 0,000000 & 0,000000 & \\ 0,000399 & 0,000000 & 0,000579 & 0,000000 \end{pmatrix}.$$

Inverzní varianční matice k matici $\hat{V}$ je tvaru

$$\hat{D} = \begin{pmatrix} 355,144431 & & & \\ -108,876008 & 242,936184 & & \\ 0,000000 & -60,475978 & 588,101350 & \\ 0,000000 & -131,490557 & 0,000000 & 423,979252 \end{pmatrix}.$$

Tato matice splňuje $\hat{d}_{13} = 0,\ \hat{d}_{14} = 0,\ \hat{d}_{34} = 0$. Platí tedy, že $\hat{d}_{ij} = 0$, kdykoli nejsou vrcholy i a j v daném grafu spojeny hranou. □

Jak je na uvedeném příkladu dobře patrné, maticový algoritmus poskytuje shodné výsledky s výše popsaným algoritmem IPF. Jeho zápis v SW Mathematica je však mnohem elegantnější a algoritmus je velmi rychlý - odhad varianční matice ani v případě 7 indexů netrval nikdy déle než 0,02 vteřiny a většinou byla doba výpočtu dokonce menší než setina sekundy (doba výpočtu je samozřejmě funkcí nejen nastavené přesnosti testování shody $V_{c,c} = S_{c,c}$, ale rovněž grafu, za jehož podmínky varianční matici odhadujeme).

8.3 Rozklad deviance

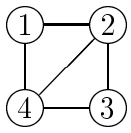
V poznámce 8.12 jsme popsali možnosti využití deviance při testování různých hypotéz. Při testování vhodnosti modelu pomocí této testové statistiky můžeme použít podobný postup jako při analýze rozptylu. Ukažme si ho nyní také na konkrétním příkladě.

Příklad 8.18:

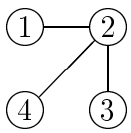
Vezměme si opět logaritmické výnosy 4 odvětvových indexů:

Označení	Index	Odvětví
1	BI04	Těžba a zpracování nerostů a rud
2	BI07	Chemický, farmaceutický a gumozpracující průmysl
3	BI08	Stavebnictví a průmysl stavebních hmot
4	BI12	Energetika

Mějme zároveň dán grafický model s grafem G_1



a druhý grafický model s grafem G_2, který je faktorem grafu G_1



Označme si navíc G_0 grafický model s úplným grafem a G grafický model s grafem bez hran. Tabulka analýzy deviance má v tomto případě následující podobu:

Test	Deviance	Stupně volnosti	Krit. hodnota
G proti G_2	32,754 - 11,244 = 21,510	6 - 3 = 3	7,81
G_1 proti G_0	5,882 - 0 = 5,882	1 - 0 = 1	3,84
G_2 proti G_1	11,244 - 5,882 = 5,362	3 - 1 = 2	5,99
G proti G_0	32,754 - 0 = 32,754	6 - 0 = 6	12,59

Z tabulky můžeme vyčíst výsledky celé řady testů. Testovou statistikou je deviance (viz definice 8.2), pokud se jedná o test proti modelu s úplným grafem, případně diference deviancí (viz definice 8.6), pokud testujeme 2 grafické modely s obecnými grafy (z nichž jeden musí být samozřejmě faktorem druhého).

Například platí:
kritická hodnota $\chi^2_2(0,05) = 5,99$, a protože $5,99 > 5,362$, nemůžeme zamítnout grafický model s grafem G_2 ve prospěch grafického modelu s grafem G_1.

Naopak $\chi^2_1(0,05) = 3,84 < 5,882$, a proto zamítáme grafický model s grafem G_1 proti saturovanému modelu s úplným grafem.

Poznamenejme ještě, že tabulku je možné lehce upravit i pro test pouze jediného grafického modelu. Otestujme tedy například grafický model s grafem G_1. Tabulka dostává následující podobu:

Test	**Deviance**	**Stupně volnosti**	**Krit. hodnota**
G proti G_1	32,754 - 5,882 = 26,872	6 - 1 = 5	11,07
G_1 proti G_0	5,882 - 0 = 5,882	1 - 0 = 1	3,84
G proti G_0	32,754 - 0 = 32,754	6 - 0 = 6	12,59

Z tabulky je opět patrný například výsledek následujícího testu:
Kritická hodnota $\chi^2_1(0,05) = 3,84$, a protože platí $3,84 < 5,882$, zamítáme shodu grafického modelu s grafem G_1 s daty. □

Kapitola 9

Selekce grafických modelů - neorientované grafy

V této kapitole se pokusíme odpovědět na otázku, jakým způsobem lze vybrat grafický model, který vykazuje dobrou shodu se zadanými daty. Ukážeme si několik různě efektivních postupů a pokusíme se rovněž podrobněji rozebrat jejich výhody a nevýhody.

9.1 Generování všech možných grafů

Nejpřímějším způsobem je testování shody všech možných grafických modelů s daty. Testovou statistikou může být deviance (tento postup je použit v [26]), ale jsou známy také metody založené na bayesovském přístupu (viz [7], [22]).

Tento přístup má však nejméně dva problémy:

(I.) Získáme celou skupinu grafů, které vhodně popisují data, a nikoli pouze 1 graf.

(II.) Tento postup je výpočetně značně náročný.

Problém číslo (II.) dobře ilustruje následující tabulka uvedená v diplomové práci [26] na straně 36:

počet vrcholů	**počet všech možných grafů**
1	1
2	2
3	8
4	64
5	1024
6	32 768
7	2 097 152
8	$2,6843456 \cdot 10^8$

V [26] na straně 36 je také uvedeno: *„výpočet algoritmu pro grafy o šesti vrcholech na Pentiu s pamětí* 96.0 *MB RAM trval přibližně* 55 *hodin, vygenerování grafů zabralo zhruba dalších* 5 *hodin“*.

Z těchto údajů je patrné, že i přes neustále se zvyšující výkonnost výpočetní techniky je tato metoda hledání „vhodných grafů“ pro grafy s větším počtem vrcholů nepoužitelná.

9.2 Backward a forward algoritmy

V knize [31] na straně 256-260 jsou popsány 4 algoritmy na vyhledání „vhodného“ grafu, který dobře reprezentuje konkrétní data. Jde o 2 *backward algoritmy*, které vycházejí z kompletního grafu (saturovaného modelu), z něhož se postupně odebírají hrany dle určitého kritéria, a o 2 *forward algoritmy*, které naopak vycházejí z grafu bez hran, do něhož se hrany dle určitého kritéria přidávají.

V dalším textu si zmíněné algoritmy popíšeme a zároveň ukážeme jejich použití na konkrétních finančních datech.

9.2.1 Backward algoritmus se stop pravidlem založeným na devianci vynechané hrany

Algoritmus si nejlépe přiblížíme pomocí následujícího příkladu:

Příklad 9.1:

Vezměme si měsíční logaritmické výnosy následujících 4 odvětvových indexů (pro přehlednější grafické vyjádření výsledků indexy očíslujeme přirozenými čísly $1, \ldots, 4$):

Označení	Index	Odvětví
1	BI04	Těžba a zpracování nerostů a rud
2	BI07	Chemický, farmaceutický a gumozpracující průmysl
3	BI08	Stavebnictví a průmysl stavebních hmot
4	BI12	Energetika

Varianční matice těchto indexů má následující hodnoty:

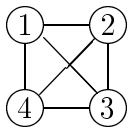
$$S = \begin{pmatrix} 0,003394 & & & \\ 0,001886 & 0,006152 & & \\ 0,000995 & 0,000633 & 0,001765 & \\ 0,000984 & 0,001908 & 0,000776 & 0,002950 \end{pmatrix}.$$

Trochu lepší přehled o struktuře dat nám poskytne korelační matice:

$$\begin{pmatrix} 1,0000 & & & \\ 0,4127 & 1,0000 & & \\ 0,4064 & 0,1920 & 1,0000 & \\ 0,3110 & 0,4478 & 0,3398 & 1,0000 \end{pmatrix}.$$

Z korelační matice je dobře patrná lineární závislost mezi jednotlivými indexy. Aplikujme nyní backward algoritmus se stop pravidlem založeným na devianci vynechané hrany.

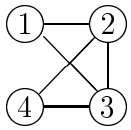
1.1. Vyjdeme z úplného grafu (saturovaného modelu). Devianci tohoto grafu položíme rovnu 0 a $\hat{V} = S$.



1.2. Z tohoto grafu postupně vynecháme jednotlivé hrany a spočteme deviance vynechaných hran, a to buď podle vzorce 8.1, nebo přímo škálováním inverzní varianční matice S^{-1} tak, aby měla na diagonále jednotky, a podle vzorce 8.2 (s využitím důsledku 3.41). Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 5,364 & * & & \\ 5,882 & 0,225 & * & \\ 0,146 & 7,211 & 3,175 & * \end{pmatrix}.$$

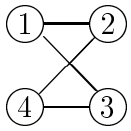
1.3. Vybereme minimální devianci vynechané hrany (tj. $0,146$ na pozici $\{1,4\}$) a porovnáme ji s kritickou hodnotou χ^2_1 pro zvolenou hladinu významnosti α (v našem případě zvolíme $\alpha = 5\%$, a tedy $\chi^2_1(0,05) = 3,84$). Protože platí $0,146 < 3,84$, vynecháme hranu $\{1,4\}$ z grafu a pokračujeme v algoritmu s novým výchozím grafem:



2.1. Z tohoto grafu postupně opět vynecháme jednotlivé hrany a spočteme deviance vynechaných hran podle vzorce 8.4. Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 7,203 & * & & \\ 6,903 & 0,320 & * & \\ \times & 9,049 & 4,195 & * \end{pmatrix}.$$

2.2. Nejmenší deviance se nachází na pozici $\{2,3\}$ a má hodnotu $0,320$. Protože platí $0,320 < 3,84$, vynecháme hranu $\{2,3\}$ z grafu a pokračujeme v algoritmu s novým výchozím grafem:

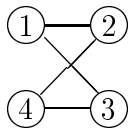


3.1. Deviance vynechaných hran z tohoto grafu jsou zaznamenány v následující matici:

$$\begin{pmatrix} * & & & \\ 6,990 & * & & \\ 6,690 & \times & * & \\ \times & 8,770 & 3,916 & * \end{pmatrix}.$$

3.2. Minimální devianci vynechané hrany má hrana {3,4}. Tato deviance však už je statisticky významná ($3,916 > 3,84$), proto hranu nevynecháme a výpočet ukončíme.

Výsledkem uvedeného postupu je jediný graf:



Modifikovaná matice sousednosti[1] výsledného grafu má následující tvar:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \end{array} \begin{array}{c} \begin{array}{cccc} BI04 & BI07 & BI08 & BI12 \end{array} \\ \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & -2 & 0 & \\ -1 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

□

Shrnutí algoritmu:

Mějme k dispozici data ve formě N realizací k-rozměrného náhodného vektoru $X \sim N(0, V)$.
Vstupem do selekční procedury jsou:

- Výběrová varianční matice S
- Saturovaný model reprezentovaný kompletním grafem G_0

Graf G s vynechanou hranou {i, j} budeme značit $G\backslash\{i, j\}$ a $rest$ označíme zbývající složky náhodného vektoru X (tj. kromě i-té a j-té složky).

Algoritmus:

1. (a) Spočítáme deviance vynechaných hran dev^*_{ij} v G_0, tzn. testujeme graf G_0 s vynechanou hranou {i,j} proti G_0. Můžeme použít dva způsoby výpočtu deviance:
   - S^{-1} škálujeme tak, aby měla na diagonále $1 \rightarrow$ mimodiagonální prvky jsou
   $$-corr(X_i, X_j|rest)$$
   a devianci spočteme podle vzorce
   $$dev^*_{ij} = -N \log(1 - corr(X_i, X_j|rest)^2) \sim \chi^2_1.$$

[1] Modifikace spočívá v tom, že namísto vynechaných hran nedáváme 0, nýbrž číslo kroku (se záporným znaménkem), ve kterém byla hrana vynechána.

- Spočteme odhady $\hat{V}_{ij}^{(0)}$ a $\hat{D}_{ij}^{(0)}$, a to buď přímým výpočtem, jak je uvedeno v subkapitole 8.2.1, nebo pomocí IPF algoritmu či jednoho z maticových algoritmů.
  Devianci spočteme podle vzorce

$$dev_{ij}^{*} = N[tr(S\hat{D}_{ij}^{(0)}) - \log det(S\hat{D}_{ij}^{(0)}) - k] \sim \chi_1^2.$$

(b) Vybereme nejmenší nevýznamnou dev_{ij}^{*} a příslušnou hranu {i,j} vyloučíme z $G_0 \rightarrow$ dostáváme nový graf $G_1 = G_0 \backslash \{i, j\}$ s odhady $\hat{V}^{(1)}$ a $\hat{D}^{(1)}$

(c) Pokud jsou všechny dev_{ij}^{*} významné $\rightarrow$ STOP: výsledkem selekce je kompletní graf G_0

2. Pro $r = 1, 2, \ldots$

(a) Spočítáme deviance vynechaných hran dev_{ij}^{*} v G_r, tzn. testujeme modely s grafem $G_r \backslash \{i, j\}$ proti modelu s grafem G_r.

- $\hat{V}^{(r)}$, $\hat{D}^{(r)}$ ...maximálně věrohodné odhady v G_r
- $\hat{V}_{ij}^{(r)}$, $\hat{D}_{ij}^{(r)}$ ...maximálně věrohodné odhady v $G_r \backslash \{i, j\}$

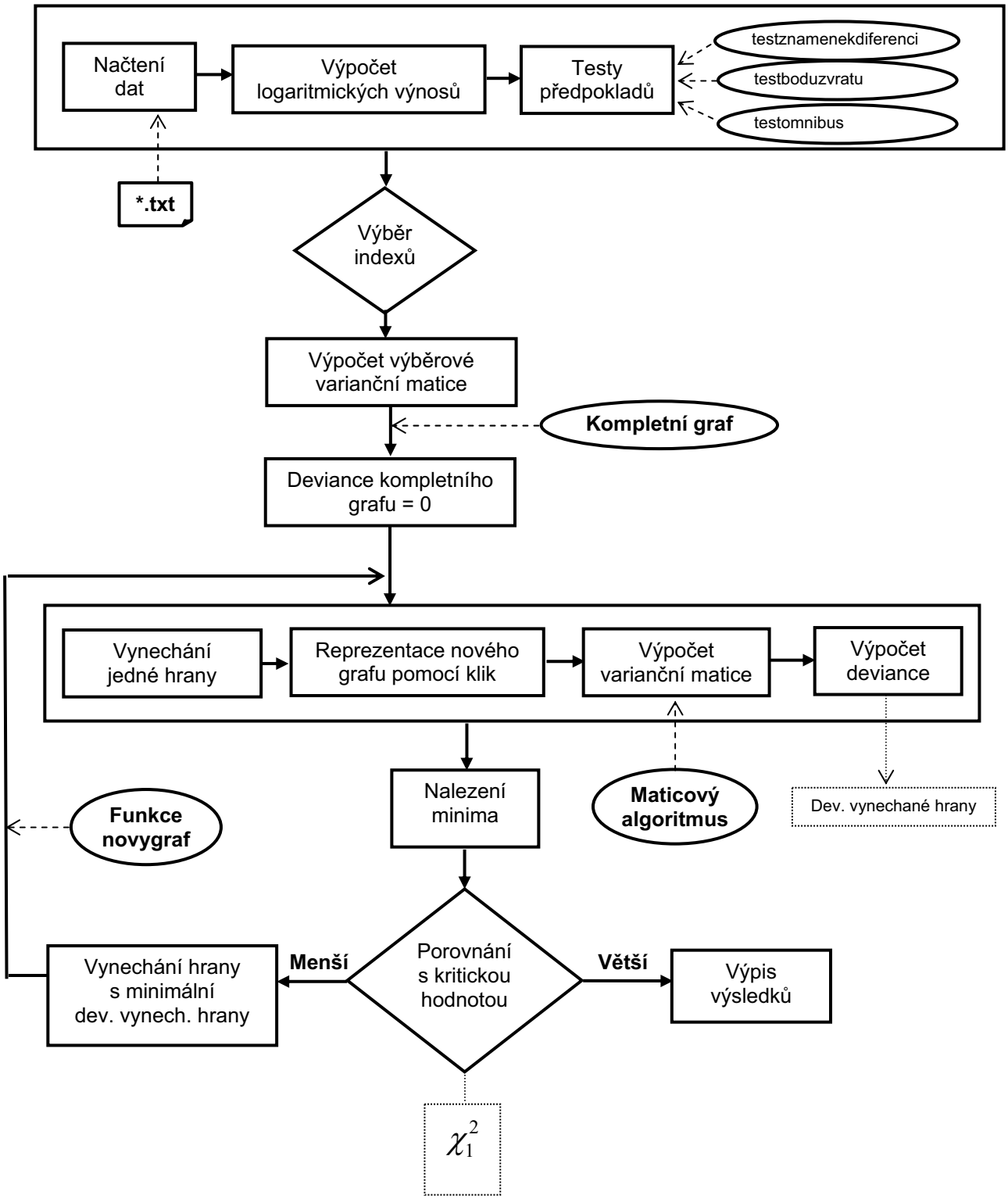
Devianci spočteme podle vzorce

$$dev_{ij}^{*} = -N \cdot \{tr[S \cdot (\hat{D}^{(r)} - \hat{D}_{ij}^{(r)}) + \log det(\hat{V}^{(r)} \cdot \hat{D}_{ij}^{(r)})]\} \sim \chi_1^2$$

(b) Vybereme nejmenší nevýznamnou dev_{ij}^{*}, příslušnou hranu {i,j} vyloučíme z G_r $\rightarrow$ dostáváme nový graf $G_{r+1} = G_r \backslash \{i, j\}$ s odhady $\hat{V}^{(r+1)}$ a $\hat{D}^{(r+1)}$.

(c) Pokud jsou všechny dev_{ij}^{*} významné $\rightarrow$ STOP: výsledkem selekce je model s grafem G_r.

Obrázek 9.2: *Schéma backward algoritmu se stop pravidlem založeným na devianci vynechané hrany*

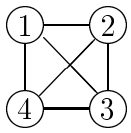


9.2.2 Backward algoritmus se stop pravidlem založeným na celkové devianci

Algoritmus si opět přiblížíme na příkladu se stejnými vstupními daty:

Příklad 9.3:

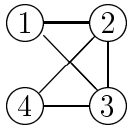
1.1. Jedná se o backward algoritmus, proto opět vyjdeme z úplného grafu (saturovaného modelu). Devianci tohoto grafu položíme rovnu 0 a $\hat{V} = S$.



1.2. Z tohoto grafu postupně vynecháme jednotlivé hrany a spočteme celkové deviance nových grafů, a to buď podle vzorce 8.1, nebo (v tomto prvním kroku) přímo škálováním inverzní varianční matice S^{-1} tak, aby měla na diagonále jednotky, a podle vzorce 8.2 (s využitím důsledku 3.41). Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 5,364 & * & & \\ 5,882 & 0,225 & * & \\ 0,146 & 7,211 & 3,175 & * \end{pmatrix}.$$

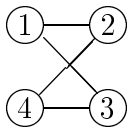
1.3. Vybereme minimální devianci (tj. $0,146$ na pozici $\{1,4\}$) a porovnáme ji s kritickou hodnotou χ^2_f, kde f značí počet chybějících hran v testovaném grafu. Opět zvolíme významnosti $\alpha = 5\%$. $\chi^2_1(0,05) = 3,84$, a protože platí $0,146 < 3,84$, vynecháme hranu $\{1,4\}$ z grafu. Získali jsme nový výchozí graf a pokračujeme v algoritmu:



2.1. Z tohoto grafu postupně opět vynecháme jednotlivé hrany a spočteme celkové deviance nových grafů podle vzorce 8.1. Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 7,349 & * & & \\ 7,048 & 0,466 & * & \\ \times & 9,195 & 4,341 & * \end{pmatrix}.$$

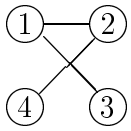
2.2. Nejmenší deviance se nachází na pozici $\{2,3\}$ a má hodnotu $0,466$. Všechny testované grafy mají 2 chybějící hrany, minimální devianci tedy porovnáme s $\chi^2_2(0,05) = 5,99$. Protože platí $0,466 < 5,99$, vynecháme hranu $\{2,3\}$ z grafu. Získali jsme nový výchozí graf a pokračujeme v algoritmu:



3.1. Celkové deviance grafů (tentokrát se třemi vynechanými hranami) jsou zaznamenány v následující matici:

$$\begin{pmatrix} * & & & \\ 7,456 & * & & \\ 7,156 & \times & * & \\ \times & 9,236 & 4,382 & * \end{pmatrix}.$$

3.2. Minimální deviance má hodnotu $4,382$ a náleží grafu s dodatečně vynechanou hranou {3,4}. Kritická hodnota $\chi^2_3(0,05)$ je rovna $7,81$. Protože platí $4,382 < 7,81$, vynecháme hranu {3,4} z grafu. Získali jsme nový výchozí graf a pokračujeme v algoritmu:

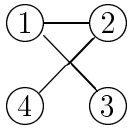


4.1. Celkové deviance grafů se čtyřmi vynechanými hranami jsou zaznamenány v následující matici:

$$\begin{pmatrix} * & & & \\ 13,346 & * & & \\ 13,046 & \times & * & \\ \times & 15,126 & \times & * \end{pmatrix}.$$

4.2. Minimální deviance má hodnotu $13,046$ a náleží grafu s dodatečně vynechanou hranou {1,3}. Kritická hodnota $\chi^2_4(0,05)$ je však rovna pouze $9,49$. Tuto devianci tedy již nemůžeme považovat za nevýznamnou a musíme algoritmus ukončit.

Výsledkem algoritmu je tedy následující graf:



Jeho reprezentace modifikovanou[2] maticí sousednosti má následující podobu:

[2]Modifikace spočívá v tom, že namísto vynechaných hran nedáváme 0, nýbrž číslo kroku (se záporným znaménkem), ve kterém byla hrana vynechána.

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \end{array} \begin{array}{c} \begin{array}{cccc} BI04 & BI07 & BI08 & BI12 \end{array} \\ \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & -2 & 0 & \\ -1 & 1 & -3 & 0 \end{pmatrix} \end{array}.$$

□

Shrnutí algoritmu:

Mějme k dispozici data ve formě N realizací k-rozměrného náhodného vektoru $X \sim N(0, V)$.

Vstupem do selekční procedury jsou:

- Výběrová varianční matice S
- Saturovaný model reprezentovaný úplným grafem G_0

Graf G s vynechanou hranou {i, j} budeme značit $G\backslash\{i, j\}$ a $rest$ označíme zbývající složky náhodného vektoru X (tj. kromě i-té a j-té složky).

Algoritmus:

1. (a) Spočítáme deviance všech grafů s jednou vynechanou hranou $dev_{ij}^{(1)}$, tzn. testujeme model s grafem $G_0\backslash\{i, j\}$ proti saturovanému modelu s grafem G_0. Můžeme použít dva způsoby výpočtu deviance:
   - S^{-1} škálujeme tak, aby měla na diagonále $1 \rightarrow$ mimodiagonální prvky jsou
   $$-corr(X_i, X_j|rest)$$
   a devianci spočteme podle vzorce
   $$dev_{ij}^{(1)} = -N \log(1 - corr(X_i, X_j|rest)^2) \sim \chi_1^2.$$
   - Spočteme odhady $\hat{V}_{ij}^{(0)}$ a $\hat{D}_{ij}^{(0)}$, a to buď přímým výpočtem, jak je uvedeno v subkapitole 8.2.1, nebo pomocí IPF algoritmu či jednoho z maticových algoritmů.
   Devianci spočteme podle vzorce
   $$dev_{ij}^{(1)} = N[tr(S\hat{D}_{ij}^{(0)}) - \log det(S\hat{D}_{ij}^{(0)}) - k] \sim \chi_1^2.$$

   (b) Vybereme nejmenší nevýznamnou $dev_{ij}^{(1)}$ a příslušnou hranu {i,j} vyloučíme z $G_0 \rightarrow$ dostáváme nový graf $G_1 = G_0\backslash\{i, j\}$ s odhady $\hat{V}^{(1)}$ a $\hat{D}^{(1)}$.

   (c) Pokud jsou všechny $dev_{ij}^{(1)}$ významné $\rightarrow$ STOP: výsledkem selekce je model s úplným grafem G_0.

2. Pro $r = 1, 2, \ldots$

   (a) Spočítáme celkové deviance $dev_{ij}^{(r+1)}$ všech *faktorů* grafu G_r, ve kterých chybí právě jedna hrana oproti grafu G_r, tzn. testujeme modely s grafem $G_r\backslash\{i, j\}$ proti saturovanému modelu s grafem G_0.

- $\hat{V}^{(r)}$, $\hat{D}^{(r)}$ ...maximálně věrohodné odhady v G_r
- $\hat{V}_{ij}^{(r)}$, $\hat{D}_{ij}^{(r)}$ ...maximálně věrohodné odhady v $G_r \backslash \{i, j\}$

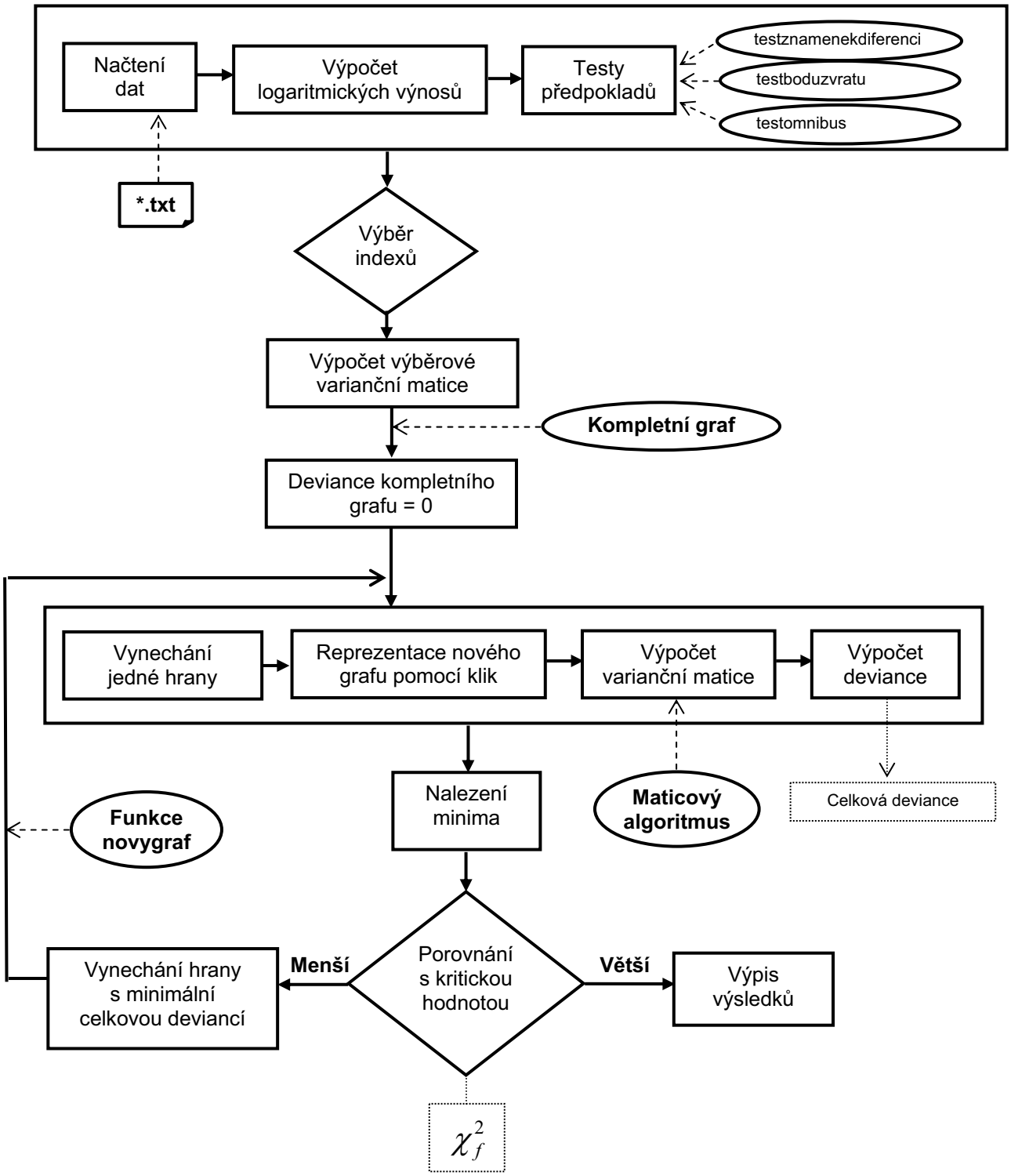
Devianci spočteme podle vzorce

$$dev_{ij}^{(r+1)} = N[tr(S\hat{D}_{ij}^{(r)}) - \log det(S\hat{D}_{ij}^{(r)}) - k] \sim \chi_{r+1}^2.$$

(b) Vybereme nejmenší nevýznamnou $dev_{ij}^{(r+1)}$, příslušnou hranu {i, j} vyloučíme z $G_r \rightarrow$ dostáváme nový graf $G_{r+1} = G_r \backslash \{i, j\}$ s odhady $\hat{V}^{(r+1)}$ a $\hat{D}^{(r+1)}$.

(c) Pokud jsou všechny $dev_{ij}^{(r+1)}$ významné $\rightarrow$ STOP: výsledkem selekce je graf G_r.

Obrázek 9.4: *Schéma backward algoritmu se stop pravidlem založeným na celkové devianci*



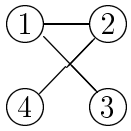
9.2.3 Porovnání výsledků backward algoritmů

Pomocí dvou různých backward algoritmů jsme získali dva různé grafy.

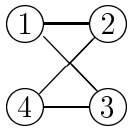
Výsledkem *backward algoritmu se stop pravidlem založeným na devianci vynechané hrany* je graf se 2 vynechanými hranami {1, 4 }, {2, 3}. Hodnota celkové deviance grafu je 0,466 a tento graf má $p - value = 0,792$.

Výsledkem *backward algoritmu se stop pravidlem založeným na celkové devianci* je graf se 3 vynechanými hranami {1, 4 }, {2, 3}, {3, 4}. Hodnota celkové deviance grafu je 4,382 a tento graf má $p - value = 0,223$.

Model s grafem

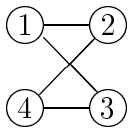


získaný podle druhého algoritmu zamítáme proti modelu s grafem

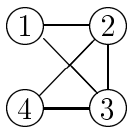


ovšem podle druhého algoritmu jej nezamítáme proti saturovanému modelu s úplným grafem.

Model s grafem



nezamítáme proti modelu s grafem



z předchozího kroku, ani proti modelu saturovanému.

Poznamenejme ještě, že spočtené hodnoty p-value se týkají testu proti alternativě saturovaného modelu s úplným grafem.

Přes rozdílné výsledné grafy se však zdá, že uvedené dva algoritmy „běží stejnou cestou“, tzn. že generují posloupnost stejných grafů, i když výsledek se poté může samozřejmě lišit. (V [13] na straně 30-37 je uveden opačný příklad, kdy *backward algoritmus se stop pravidlem založeným na devianci vynechané hrany* skončí naopak později než *backward algoritmus se stop pravidlem založeným na celkové devianci*.)

Tuto skutečnost si dokážeme v následující větě:

Věta 9.5: *Souvislost celkové deviance a deviance vynechané hrany*

Nechť G_0 je úplný graf, G jeho faktor a $\mathbf{G^{(-1)}}$ množina všech faktorů, které vzniknou z grafu G vynecháním jedné hrany.
Pak platí:
$G_{ij} \in \mathbf{G^{(-1)}}$ má nejmenší *devianci vynechané hrany* mezi všemi grafy z množiny $\mathbf{G^{(-1)}} \iff G_{ij} \in \mathbf{G^{(-1)}}$ má nejmenší *celkovou devianci* mezi všemi grafy z množiny $\mathbf{G^{(-1)}}$. □

Důkaz:

Označme si:
S = maximálně věrohodný odhad varianční matice v modelu s grafem G_0
$\hat{V}$= maximálně věrohodný odhad varianční matice v modelu s grafem G
$\hat{V}_{ij}$ = maximálně věrohodný odhad varianční matice v modelu s grafem G_{ij}
Celkovou devianci grafu G_{ij} spočteme podle vzorce:

$$2[l(S) - l(\hat{V}_{ij})].$$

Celkovou devianci grafu G spočteme podle vzorce:

$$2[l(S) - l(\hat{V})].$$

Pro devianci vynechané hrany platí:

$$\begin{aligned} dev^*_{ij} &= devG_{ij} - devG \\ &= 2[l(S) - l(\hat{V}_{ij})] - 2[l(S) - l(\hat{V})] \\ &= 2[l(\hat{V}) - l(\hat{V}_{ij})] \end{aligned}$$

Tedy: G_{ij} s nejmenší celkovou devianci (největší $l(\hat{V}_{ij})$) má nejmenší devianci vynechané hrany. □

Poznámka 9.6:

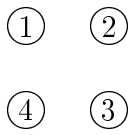
Deviance vynechané hrany je různá od deviance celkové, pokud uvažujeme vynechání hrany z grafu, který není úplný. Celková deviance je určena definicí 8.2 a deviance vynechané hrany (jakožto diference deviancí) definicí 8.6. Proto na dané hladině významnosti může být jedna ze zmíněných deviancí statistiky významná a druhá nikoliv, což vede k ukončení našich dvou backward algoritmů po různém počtu kroků. □

9.2.4 Forward algoritmus se stop pravidlem založeným na devianci přidané hrany

Také tento algoritmus si přiblížíme na příkladu se stejnými vstupními daty:

Příklad 9.7:

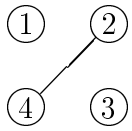
1.1. Jedná se o forward algoritmus, proto tentokrát vyjdeme z grafu bez hran. Deviance tohoto grafu je rovna $32,754$ a k jejímu výpočtu můžeme (kromě iteračních algoritmů) využít vztahu 8.3.



1.2. Do tohoto grafu postupně přidáváme jednotlivé hrany a spočteme deviance přidaných hran, a to podle vzorce 8.4. Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 8,964 & * & & \\ 8,664 & 1,802 & * & \\ 4,884 & 10,744 & 5,890 & * \end{pmatrix}.$$

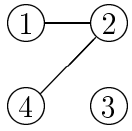
1.3. Vybereme maximální devianci přidané hrany (tj. $10,744$ na pozici $\{2,4\}$) a porovnáme ji s kritickou hodnotou χ^2_1 pro zvolenou hladinu významnosti α (v našem případě zvolíme $\alpha = 5\%$, a tedy $\chi^2_1(0,05) = 3,84$). Protože platí $10,744 > 3,84$, přidáme hranu $\{2,4\}$) do grafu a pokračujeme v algoritmu s novým výchozím grafem:



2.1. Do tohoto grafu postupně opět přidáváme jednotlivé hrany a spočteme deviance přidaných hran podle vzorce 8.4. Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 8,964 & * & & \\ 8,664 & 1,802 & * & \\ 4,884 & \times & 5,890 & * \end{pmatrix}.$$

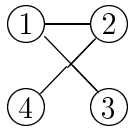
2.2. Největší deviance se nachází na pozici $\{1,2\}$ a má hodnotu $8,964$. Protože platí $8,964 > 3,84$, přidáme hranu $\{1,2\}$ do grafu a pokračujeme v algoritmu s novým výchozím grafem:



3.1. Deviance přidaných hran do tohoto grafu jsou zaznamenány v následující matici:

$$\begin{pmatrix} * & & & \\ \times & * & & \\ 8,664 & 1,802 & * & \\ 1,166 & \times & 5,890 & * \end{pmatrix}.$$

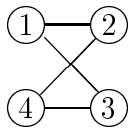
3.2. Maximální devianci přidané hrany má hrana {1,3}. Protože platí $8,664 > 3,84$, přidáme hranu {1,3} do grafu a pokračujeme v algoritmu s novým výchozím grafem:



4.1. Deviance přidaných hran do tohoto grafu opět pro přehlednost zaznamenáme v matici:

$$\begin{pmatrix} * & & & \\ \times & * & & \\ \times & 0,041 & * & \\ 1,166 & \times & 3,916 & * \end{pmatrix}.$$

4.2. Maximální devianci přidané hrany má hrana {3,4}. Protože i v tomto případě platí $3,916 > 3,84$, přidáme hranu {3,4} do grafu a pokračujeme v algoritmu s novým výchozím grafem:

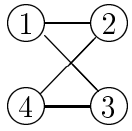


5.1. Deviance přidaných hran do tohoto grafu jsou zaznamenány v následující matici:

$$\begin{pmatrix} * & & & \\ \times & * & & \\ \times & 0,320 & * & \\ 0,241 & \times & \times & * \end{pmatrix}.$$

5.2. Maximální devianci přidané hrany má hrana {2,3}. Tato deviance však už je statisticky nevýznamná ($0,320 < 3,84$), proto hranu nepřidáme a výpočet ukončíme.

Výsledkem uvedeného postupu je jediný graf:



Modifikovaná matice sousednosti[3] výsledného grafu má následující tvar:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \end{array} \begin{array}{c} \begin{array}{cccc} BI04 & BI07 & BI08 & BI12 \end{array} \\ \begin{pmatrix} 0 & & & \\ 2 & 0 & & \\ 3 & 0 & 0 & \\ 0 & 1 & 4 & 0 \end{pmatrix} \end{array}.$$

□

Shrnutí algoritmu:

Mějme k dispozici data ve formě N realizací k-rozměrného náhodného vektoru $X \sim N(0, V)$.

Vstupem do selekční procedury jsou:

- Výběrová varianční matice S
- Model vzájemné nezávislosti s grafem bez hran G_0

Graf G s přidanou hranou {i, j} budeme značit $G \cup \{i, j\}$ a *rest* označíme zbývající složky náhodného vektoru X (tj. kromě i-té a j-té složky).

Algoritmus:

1. Pro $r = 0, 1, 2 \ldots, \begin{pmatrix} k \\ 2 \end{pmatrix} - 2$

   (a) Spočítáme deviance přidaných hran dev^*_{ij} v grafu G_r, tzn. testujeme graf G_r proti grafu s přidanou hranou $G \cup \{i, j\}$.

   - $\hat{V}^{(r)}$, $\hat{D}^{(r)}$ ...maximálně věrohodné odhady v G_r
   - $\hat{V}^{(r)}_{ij}$, $\hat{D}^{(r)}_{ij}$ ...maximálně věrohodné odhady v $G_r \cup \{i, j\}$

   Devianci spočteme podle vzorce

   $$dev^*_{ij} = +N \cdot \{tr[S \cdot (\hat{D}^{(r)} - \hat{D}^{(r)}_{ij}) + \log det(\hat{V}^{(r)} \cdot \hat{D}^{(r)}_{ij})]\} \sim \chi^2_1$$

   (b) Vybereme největší významnou dev^*_{ij}, příslušnou hranu {i,j} přidáme do $G_r \rightarrow$ dostáváme nový graf $G_{r+1} = G_r \cup \{i, j\}$ s odhady $\hat{V}^{(r+1)}$ a $\hat{D}^{(r+1)}$.

   (c) Pokud jsou všechny dev^*_{ij} nevýznamné $\rightarrow$ STOP: výsledkem selekce je model s grafem G_r.

2. Pro $r = \begin{pmatrix} k \\ 2 \end{pmatrix} - 1$ (tzn. v G_r chybí pouze 1 hrana oproti kompletnímu grafu)

   (a) Spočítáme devianci přidané hrany dev^*_{ij} v grafu G_r, tzn. testujeme graf G_r proti grafu s přidanou hranou $G \cup \{i, j\}$ (tj. proti kompletnímu grafu).

   Můžeme použít dva způsoby výpočtu deviance:

[3]Modifikace spočívá v tom, že namísto přidaných hran nedáváme automaticky 1, nýbrž číslo kroku (tentokrát s kladným znaménkem), ve kterém byla hrana do grafu přidána.

- S^{-1} škálujeme tak, aby měla na diagonále 1 $\rightarrow$ mimodiagonální prvky jsou

$$-corr(X_i, X_j|rest)$$

a devianci spočteme podle vzorce

$$dev^*_{ij} = -N \log(1 - corr(X_i, X_j|rest)^2) \sim \chi^2_1.$$

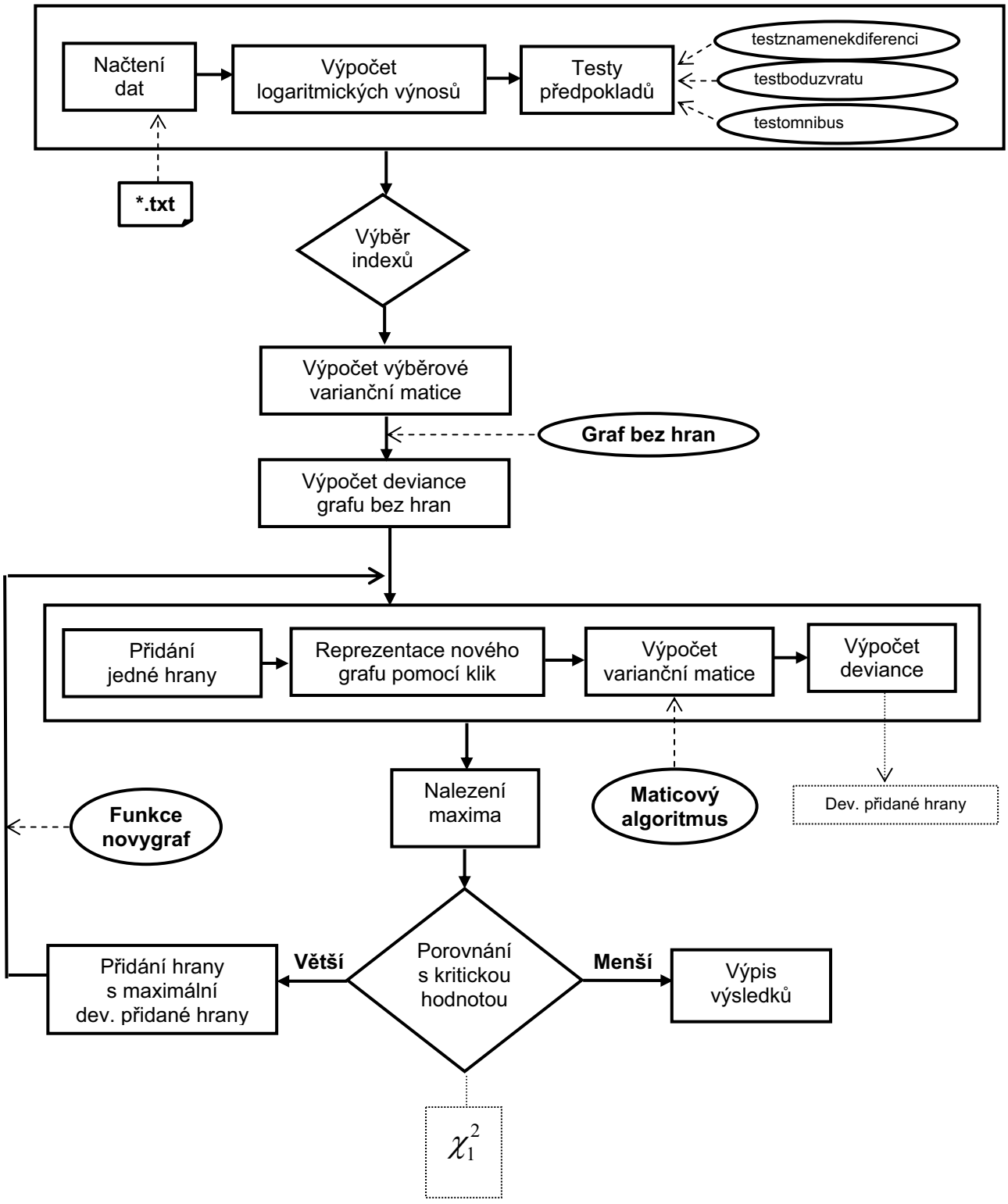
- Označme $\hat{V}^{(r)}$ a $\hat{D}^{(r)}$ odhady v grafu G_r, pak devianci spočteme podle vzorce

$$dev^*_{ij} = N[tr(S\hat{D}^{(r)}) - \log det(S\hat{D}^{(r)}) - k] \sim \chi^2_1.$$

(b) Pokud je $dev^*_{ij} > \chi^2_1(0,05) \rightarrow$ výsledkem selekce je kompletní graf.

(c) Pokud je $dev^*_{ij} \leq \chi^2_1(0,05) \rightarrow$ výsledkem selekce je graf G_r.

Obrázek 9.8: *Schéma forward algoritmu se stop pravidlem založeným na devianci přidané hrany*

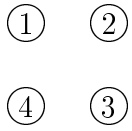


9.2.5 Forward algoritmus se stop pravidlem založeným na celkové devianci

Opět využijeme stejných vstupních dat pro 4 zvolené odvětvové indexy, abychom si tento algoritmus blíže přiblížili:

Příklad 9.9:

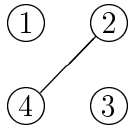
1.1. Jedná se o forward algoritmus, proto vyjdeme z grafu bez hran. Deviance tohoto grafu je rovna $32,754$ a k jejímu výpočtu můžeme (kromě iteračních algoritmů) využít vztahu 8.3.



1.2. Do tohoto grafu postupně přidáváme jednotlivé hrany a spočteme celkové deviance pro nové grafy, a to podle vzorce 8.1. Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 23,790 & * & & \\ 24,090 & 30,952 & * & \\ 27,870 & 22,010 & 26,864 & * \end{pmatrix}.$$

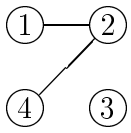
1.3. Vybereme minimální celkovou devianci (tj. $22,010$ na pozici $\{2,4\}$), a protože v grafu chybí tentokrát 5 hran, porovnáme ji s kritickou hodnotou χ^2_5 pro zvolenou hladinu významnosti α (v našem případě zvolíme $\alpha = 5\%$, a tedy $\chi^2_5(0,05) = 11,07$). Protože platí $22,010 > 11,07$, přidáme hranu $\{2,4\}$ do grafu a pokračujeme v algoritmu s novým výchozím grafem:



2.1. Do tohoto grafu postupně opět přidáváme jednotlivé hrany a spočteme celkové deviance nových grafů podle vzorce 8.1. Tyto deviance jsou shrnuty v následující matici:

$$\begin{pmatrix} * & & & \\ 13,046 & * & & \\ 13,346 & 20,208 & * & \\ 17,127 & \times & 16,120 & * \end{pmatrix}.$$

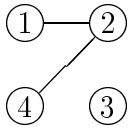
2.2. Nejmenší celková deviance se nachází na pozici $\{1,2\}$ a má hodnotu $13,046$. Protože v tomto grafu chybí 4 hrany, porovnáme ji s $\chi^2_4(0,05) = 9,49$. Platí $13,046 > 9,49$, přidáme tedy hranu $\{1,2\}$ do grafu a pokračujeme v algoritmu s novým výchozím grafem:



3.1. Celkové deviance grafů s další přidanou hranou jsou zaznamenány v následující matici:

$$\begin{pmatrix} * & & & \\ \times & * & & \\ 4,382 & 11,244 & * & \\ 11,880 & \times & 7,156 & * \end{pmatrix}.$$

3.2. Minimální celkovou devianci má graf s přidanou hranou {1,3}. V tomto novém grafu chybí 3 hrany oproti úplnému grafu, a protože platí $4,382 < 7,81 = \chi^2_3(0,05)$, nemůžeme již přidat hranu {1,3} do grafu a algoritmus ukončíme. Výsledkem je tedy minulý grafický model:



Modifikovaná matice[4] výsledného grafu má následující tvar:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \end{array} \begin{array}{c} \begin{array}{cccc} BI04 & BI07 & BI08 & BI12 \end{array} \\ \begin{pmatrix} 0 & & & \\ 2 & 0 & & \\ 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}.$$

□

Shrnutí algoritmu:

Mějme k dispozici data ve formě N realizací k-rozměrného náhodného vektoru $X \sim N(0, V)$.

Vstupem do selekční procedury jsou:

- Výběrová varianční matice S
- Model vzájemné nezávislosti s grafem bez hran G_0

Graf G s přidanou hranou {i, j} budeme značit $G \cup \{i, j\}$ a *rest* označíme zbývající složky náhodného vektoru X (tj. kromě i-té a j-té složky).

Algoritmus:

1. Pro $r = 0, 1, 2 \ldots, \begin{pmatrix} k \\ 2 \end{pmatrix} - 2$

[4]Modifikace spočívá v tom, že namísto přidaných hran nedáváme automaticky 1, nýbrž číslo kroku (tentokrát s kladným znaménkem), ve kterém byla hrana do grafu přidána.

(a) Spočítáme celkové deviance grafů S přidanou hranou $dev_{ij}^{(r+1)}$ v grafu G_r, tzn. testujeme graf G_r proti grafu s přidanou hranou $G \cup \{i,j\}$.

- $\hat{V}^{(r)}$, $\hat{D}^{(r)}$ ...maximálně věrohodné odhady v G_r
- $\hat{V}_{ij}^{(r)}$, $\hat{D}_{ij}^{(r)}$ ...maximálně věrohodné odhady v $G_r \cup \{i,j\}$

Devianci spočteme podle vzorce

$$dev_{ij}^{(r)} = N[tr(S\hat{D}_{ij}^{(r)}) - \log det(S\hat{D}_{ij}^{(r)}) - k] \sim \chi_{r+1}^2.$$

(b) Vybereme minimální významnou celkovou devianci $dev_{ij}^{(r+1)}$, příslušnou hranu {i,j} přidáme do $G_r \rightarrow$ dostáváme nový graf $G_{r+1} = G_r \cup \{i,j\}$ s odhady $\hat{V}^{(r+1)}$ a $\hat{D}^{(r+1)}$.

(c) Pokud jsou všechny $dev_{ij}^{(r+1)}$ nevýznamné $\rightarrow$ STOP: výsledkem selekce je model s grafem G_r.

2. Pro $r = \begin{pmatrix} k \\ 2 \end{pmatrix} - 1$ (tzn. v G_r chybí pouze 1 hrana oproti kompletnímu grafu)

(a) Spočítáme devianci přidané hrany dev_{ij}^{*} v grafu G_r, tzn. testujeme graf G_r proti grafu s přidanou hranou $G \cup \{i,j\}$ (tj. proti kompletnímu grafu).
Můžeme použít dva způsoby výpočtu deviance:

- S^{-1} škálujeme tak, aby měla na diagonále 1 $\rightarrow$ mimodiagonální prvky jsou

$$-corr(X_i, X_j|rest)$$

a devianci spočteme podle vzorce

$$dev_{ij}^{*} = -N \log(1 - corr(X_i, X_j|rest)^2) \sim \chi_1^2.$$

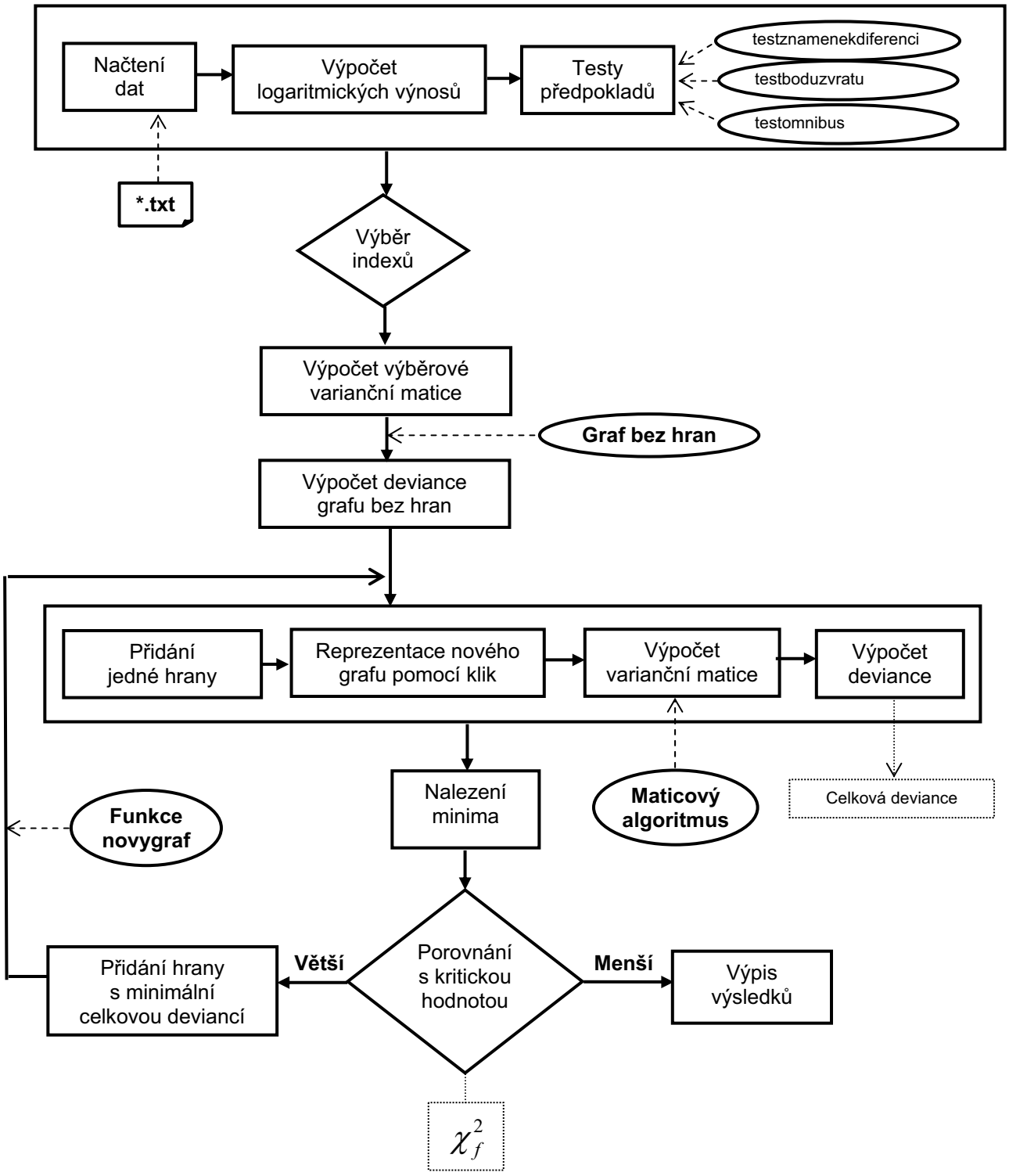
- Označme $\hat{V}^{(r)}$ a $\hat{D}^{(r)}$ odhady v grafu G_r, pak devianci spočteme podle vzorce

$$dev_{ij}^{*} = N[tr(S\hat{D}^{(r)}) - \log det(S\hat{D}^{(r)}) - k] \sim \chi_1^2.$$

(b) Pokud je $dev_{ij}^{*} > \chi_1^2(0,05) \rightarrow$ výsledkem selekce je kompletní graf.

(c) Pokud je $dev_{ij}^{*} \leq \chi_1^2(0,05) \rightarrow$ výsledkem selekce je graf G_r.

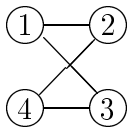
Obrázek 9.10: *Schéma forward algoritmu se stop pravidlem založeným na celkové devianci*



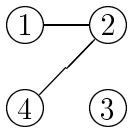
9.2.6 Porovnání výsledků forward algoritmů

Pomocí dvou různých forward algoritmů jsme opět získali dva různé grafy.

Výsledkem *forward algoritmu se stop pravidlem založeným na devianci přidané hrany* je graf se 4 přidanými hranami $\{1, 2\}$, $\{1, 3\}$, $\{2, 4\}$, $\{3, 4\}$.



Výsledkem *forward algoritmu se stop pravidlem založeným na celkové devianci* je graf s pouze 2 přidanými hranami $\{1, 2\}$, $\{2, 4\}$.



Přes rozdílné výsledné grafy se však zdá, že i tyto dva algoritmy „běží stejnou cestou“, tzn. že generují posloupnost stejných grafů, i když výsledek se poté může samozřejmě lišit, a to i přes to, že zatímco v jednom algoritmu pracujeme s *maximální* deviancí přidané hrany, v druhém naopak s *minimální* celkovou deviancí.

Tuto skutečnost si dokážeme v následující větě, která je v mnohém podobná větě pro backward algoritmy (věta 9.5):

Věta 9.11: *Souvislost celkové deviance a deviance přidané hrany*

Nechť G_0 je úplný graf, G jeho faktor a $\mathbf{G^{(+1)}}$ množina všech grafů, které vzniknou z grafu G přidáním jedné hrany (G je samozřejmě faktorem $\mathbf{G^{(+1)}}$).
Pak platí:
$G_{ij} \in \mathbf{G^{(+1)}}$ má největší *devianci přidané hrany* mezi všemi grafy z množiny $\mathbf{G^{(+1)}} \iff G_{ij} \in \mathbf{G^{(+1)}}$ má nejmenší *celkovou devianci* mezi všemi grafy z množiny $\mathbf{G^{(+1)}}$. □

Důkaz:

Označme si:
S = maximálně věrohodný odhad varianční matice v modelu s grafem G_0
$\hat{V}$= maximálně věrohodný odhad varianční matice v modelu s grafem G
$\hat{V}_{ij}$ = maximálně věrohodný odhad varianční matice v modelu s grafem G_{ij}
Celkovou devianci grafu G_{ij} spočteme podle vzorce:

$$2[l(S) - l(\hat{V}_{ij})].$$

Celkovou devianci grafu G spočteme podle vzorce:

$$2[l(S) - l(\hat{V})].$$

Pro devianci přidané hrany platí:

$$\begin{aligned} dev^*_{ij} &= devG - devG_{ij} \\ &= 2[l(S) - l(\hat{V})] - 2[l(S) - l(\hat{V}_{ij})] \\ &= 2[l(\hat{V}_{ij}) - l(\hat{V})]. \end{aligned}$$

Tedy: G_{ij} s nejmenší celkovou deviancí (největší $l(\hat{V}_{ij})$) má největší devianci přidané hrany. □

Poznámka 9.12:

Deviance přidané hrany je různá od deviance celkové, pokud uvažujeme přidání hrany do grafu, který se liší od grafu úplného o více než 1 hranu. Celková deviance je určena definicí 8.2 a deviance přidané hrany (jakožto diference deviancí) definicí 8.6. Proto na dané hladině významnosti může být jedna ze zmíněných deviancí statistiky významná a druhá nikoliv, což vede k ukončení našich dvou forward algoritmů po různém počtu kroků. □

9.2.7 Forward, nebo backward algoritmy?

Forward algoritmus použil poprvé Dempster v roce 1972, backward algoritmus pak jeho žák Wermut (v roce 1976). Ve skutečnosti existuje ještě mnoho dalších variant než naše 4 uvedené. Některé dají výsledky dříve než jiné, ale vždy záleží na konkrétních datech. V ilustračních příkladech jsme navíc ukázali značnou citlivost na zvolené STOP pravidlo. Nejlepším řešením je zřejmě použít všechny 4 algoritmy a (v případě rozdílných výsledků) vybrat finální graf například s ohledem na dobrou interpretaci výsledků.

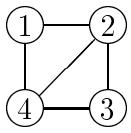
9.3 Programové řešení - hlavní myšlenky a problémy

Backward ani forward algoritmy nejsou nijak složité, jediným problémem může být skutečnost, že pro iterační výpočet odhadu varianční matice potřebujeme reprezentaci grafu pomocí klik. V této kapitole se pokusíme nastínit základní ideu řešení tohoto problému. Celý zdrojový kód (v programu Mathematica 4) spolu s komentářem je pak uveden v příloze.

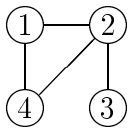
Vlastní algoritmus vychází z jednoduché myšlenky: jestliže z kliky vynecháme jednu hranu, získáváme 2 nové kliky, v jejichž zápisu vždy chybí jeden z vrcholů vynechané hrany. Jediným problémem může být skutečnost, že je-li graf zapsán pomocí více klik s neprázdným průnikem, nově vzniklé kliky mohou být obsaženy v některé z větších klik grafu a nejsou tedy ve skutečnosti klikami. Ukažme si tuto skutečnost na příkladě.

Příklad 9.13:

Mějme grafický model s následujícím grafem:



Tento graf můžeme zapsat pomocí klik {1, 2, 4} a {2, 3, 4}.
Po vyloučení hrany {3, 4} získáváme následující graf:



Z kliky $\{2, 3, 4\}$ vznikly vynecháním hrany $\{3, 4\}$ nové „kliky“ $\{2, 3\}$, $\{2, 4\}$, avšak $\{2, 4\} \subseteq \{1, 2, 4\}$. Tedy $\{2, 4\}$ není klikou grafu, jak je ostatně patrné i z výše uvedeného obrázku. □

Tyto případy jsou v programu ošetřeny funkcí *novygraf*, která přebytečné podmnožiny odstraní a vrátí reprezentaci grafu pomocí klik. Poznamenejme ještě, že graf potřebujeme reprezentovat pomocí klik kvůli maticovému algoritmu. Jako vhodnější reprezentace pro záznam vynechaných hran se nám však jeví matice sousednosti (protože pracujeme s neorientovanými grafy, stačí nám dokonce pouze její dolní trojúhelníková podmatice). Tato dvojí reprezentace sice zabírá trochu více paměti počítače, ušetří nám ale mnoho operací. Matici sousednosti lze navíc použít i jako jeden z výstupů - pokud v kroku k vynecháme (v příslušném backward algoritmu) z grafu hranu $\{i, j\}$, dosadíme na pozici prvek $(-k)$, pokud naopak (v příslušném forward algoritmu) hranu do grafu přidáme, dosadíme na pozici prvek $(+k)$. Získáme tak dobrý přehled o tom, jak algoritmus „běžel“, tj. v jakém pořadí za sebou byly jednotlivé hrany vynechány/přidány.

Kapitola 10

Selekční algoritmus pro VAR modely - orientované grafy

Podívejme se nyní blíže na algoritmus, který daným datům přiřadí konkrétní VAR model a jemu odpovídající orientovaný graf:

Algoritmus pro přidělení modelu datům

0. Dříve než přistoupíme k vlastní identifikaci VAR modelu, je třeba otestovat, zda jsou zpracovávané časové řady stacionární a případně provést jejich transformaci pomocí prvních (vyšších) diferencí[1].

1. V prvním kroku je potřeba určit maximální počet zpoždění (p) pro VAR model popisující dané časové řady. Optimální je taková volba zpoždění, která zajistí nekorelovanost náhodných složek. Pro volbu zpoždění v AR modelech existuje řada návodů - velmi jednoduché je například kritérium založené na autokorelační a parciální autokorelační funkci (viz např. [4]). V [24], [27], [28], [32], [33] je doporučován následující postup, který bere v úvahu rovněž počet odhadovaných parametrů VAR modelu:

A. Pro zvolená možná p odhadneme pomocí metody nejmenších čtverců jednotlivé rovnice modelu SVAR.

B. Pro každou rovnici (i=1,...,m) modelu spočteme reziduální součet čtverců S_i.

C. Optimální délku zpoždění určíme pomocí minimalizace Akeikeho kritéria (AIC), které spočteme pro dané p pomocí vzorce:

$$AIC = n \sum \log S_i + 2k,$$

kde:

$k = pm^2 + m(m-1)/2$.

Další kroky již souvisí s využitím grafického modelování:

2. Spočteme výběrovou kovarianční matici V odpovídající datové matici X s použitím zvoleného zpoždění p.

[1] V článku [33] je pomocí simulací dokázáno, že dále zmíněné procedury platí pro integrované řady prvního řádu - I(1), v takovém případě je tedy transformace řady do určité míry volitelná.

3. Odhadneme matici výběrových koeficientů parciální korelace mezi všemi (běžnými i zpožděnými) proměnnými.

4. Určíme, které koeficienty parciální korelace je možné považovat za nulové, a přiřadíme graf podmíněných nezávislostí (neorientovaný graf), jenž nejlépe odpovídá daným datům.

V tomto kroku využijeme výše popsané backward algoritmy. Nebudou však „startovat“ z úplného grafu, ale z grafu, ve kterém již vynecháme hrany, které by odpovídaly vazbám mezi zpožděnými proměnnými. Ponecháme tedy pouze hrany mezi běžnými proměnnými a mezi běžnými a zpožděnými proměnnými.

5. Pomocí procesu tzv. „demoralizace“ (jde o opačnou proceduru k moralizaci) zorientujeme hrany. Samozřejmě postupujeme tak, že šipka vede od zpožděných proměnných k běžným. Problémem však zůstávají případné hrany mezi běžnými proměnnými.

6. Pokud se stane, že dostaneme více přípustných modelů (proces „demoralizace“ není na rozdíl od „moralizace“ jednoznačný), porovnáme je pomocí některé z metod maximální věrohodnosti - konkrétně použijeme opět AIC.

Kapitola 11

Zpracování konkrétních finančních dat

Nyní můžeme konečně přistoupit k řešení rozsáhlejších příkladů, které nám snad poskytnou odpovědi na otázky z úvodní kapitoly.

11.1 Provázanost odvětví na českém kapitálovém trhu

11.1.1 Situace v letech 1994 - 2000[1]

Podívejme se nejprve, jak vypadala situace „v minulém tisíciletí".

Vezměme si nyní logaritmické výnosy šesti odvětvových indexů (pro přehlednější grafické vyjádření výsledků indexy očíslujeme přirozenými čísly $1, \ldots, 6$):

Tabulka 11.1: Odvětvové indexy zvolené pro analýzu v letech 1994 - 2000

Označení	Index	Odvětví
1	BI03	Výroba nápojů a tabáku
2	BI05	Textilní, oděvní a kožedělný průmysl
3	BI09	Hutnictví a průmysl zpracování kovů
4	BI11	Elektrotechnický a elektronický průmysl
5	BI16	Služby
6	BI18	Investiční fondy

Varianční matice těchto indexů má následující hodnoty:

$$S = \begin{pmatrix} 0,003856 & & & & & \\ 0,001510 & 0,005387 & & & & \\ 0,002634 & 0,002214 & 0,009566 & & & \\ 0,001626 & 0,002510 & 0,002172 & 0,004989 & & \\ 0,001625 & 0,001941 & 0,003190 & 0,002376 & 0,006322 & \\ 0,001404 & 0,001513 & 0,001731 & 0,001703 & 0,001511 & 0,002969 \end{pmatrix}.$$

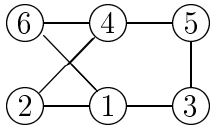
[1] Závěry byly publikovány v článcích [14], [15].

Trochu lepší přehled o struktuře dat nám poskytne korelační matice:

$$\begin{pmatrix} 1 & & & & & \\ 0,3313 & 1 & & & & \\ 0,4338 & 0,3084 & 1 & & & \\ 0,3708 & 0,4842 & 0,3144 & 1 & & \\ 0,3291 & 0,3326 & 0,4103 & 0,4231 & 1 & \\ 0,4151 & 0,3783 & 0,3249 & 0,4426 & 0,3488 & 1 \end{pmatrix}.$$

Z korelační matice je dobře patrná lineární závislost mezi jednotlivými indexy.

Aplikací backward algoritmu se stop pravidlem založeným na devianci vynechané hrany získáváme následující graf a jeho matici sousednosti:

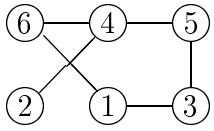


$$\begin{array}{c} \\ BI03 \\ BI05 \\ BI09 \\ BI11 \\ BI16 \\ BI18 \end{array} \begin{array}{c} \begin{array}{cccccc} BI03 & BI05 & BI09 & BI11 & BI16 & BI18 \end{array} \\ \begin{pmatrix} & & & & & \\ 1 & & & & & \\ 1 & -7 & & & & \\ -6 & 1 & -1 & & & \\ -2 & -5 & 1 & 1 & & \\ 1 & -4 & -3 & 1 & -8 & \end{pmatrix} \end{array}.$$

V tomto grafu je vynecháno 8 hran, má celkovou devianci 8,22 a p-value 0,412.

Připomeňme ještě, že vynechané hrany jsou v matici sousednosti označeny záporným znaménkem a číslem kroku algoritmu, v němž byly vynechány, existující hrany jsou označeny 1.

Pomocí backward algoritmu se stop pravidlem založeným na celkové devianci získáme i v tomto případě odlišný graf:



$$\begin{array}{c} \\ BI03 \\ BI05 \\ BI09 \\ BI11 \\ BI16 \\ BI18 \end{array} \begin{array}{c} \begin{array}{cccccc} BI03 & BI05 & BI09 & BI11 & BI16 & BI18 \end{array} \\ \begin{pmatrix} & & & & & \\ -9 & & & & & \\ 1 & -7 & & & & \\ -6 & 1 & -1 & & & \\ -2 & -5 & 1 & 1 & & \\ 1 & -4 & -3 & 1 & -8 & \end{pmatrix} \end{array}.$$

V tomto grafu je vynecháno 9 hran, má celkovou devianci 13,09 a p-value 0,159.

Žádný z výsledných modelů tedy nelze zamítnout proti alternativě saturovaného modelu s úplným grafem, která představuje podmíněnou závislost všech dvojic indexů.

Z obou uvedených grafů jsou patrné následující podmíněné nezávislosti mezi logaritmickými výnosy odvětvových indexů při pevných hodnotách zbývajících indexů:

$$BI03 \perp BI11$$

$$BI03 \perp BI16$$

$$BI05 \perp BI09$$

$$BI05 \perp BI16$$

$$BI05 \perp BI18$$

$$BI09 \perp BI11$$

$$BI09 \perp BI18$$

$$BI16 \perp BI18$$

Zjištěné výsledky jsou dosti překvapující. U finančních dat, jimiž údaje o odvětvových indexech jsou, bych očekával mnohem větší provázanost. Například v diplomové práci [26] je na straně 38-41 uvedena podobná analýza pro čtyři kurzy měn. Ze dvou tam uvedených grafů je v jednom vynechána 1 hrana a druhý graf je dokonce úplný.

Osobně jsem očekával, že vrchol grafu reprezentující index investiční fondy (BI18) bude spojen se všemi ostatními. Ve skutečnosti je vrchol 6 odpovídající indexu investičních fondů BI18 spojen pouze s vrcholem 1 reprezentujícím index výroby nápojů a tabáku BI03 a s vrcholem 4 reprezentujícím elektroprůmysl BI11. To odpovídá struktuře korelační matice, kde proměnná BI18 vykazuje největší korelaci právě s proměnnými BI03 a BI11.

Index elektroprůmyslu BI11 vykazuje největší provázanost s ostatními indexy, konkrétně s textilním, oděvním a kožedělným průmyslem, se službami a s investičními fondy, jak ukazují hrany vycházející z vrcholu 4 v obou grafech.

Důsledkem věty o separaci (věta 3.8), který můžeme pozorovat na výsledném grafu z backward algoritmu se stop pravidlem založeným na celkové devianci, je například existence dvou podmíněně nezávislých skupin indexů {BI05} a {BI03, BI09, BI16, BI18} při pevných hodnotách indexů BI11. Tedy index textilního, oděvního a kožedělného průmyslu se zdá být téměř neovlivněn ostatními indexy ze sledované skupiny.

Na druhou stranu je poměrně překvapivé, že vrchol 5 reprezentující index služeb (BI16) je spojen s vrcholem 3 reprezentujícím odvětvový index hutního průmyslu a průmyslu zpracování kovů (BI09).

Tyto závěry zřejmě spíše souvisí s nepříliš vysokou průhledností českého kapitálové trhu jako takového, než se skutečnými vztahy mezi vývojem jednotlivých odvětví.

11.1.2 Situace v letech 2001 - 2004

Nyní se podíváme, jak vypadá situace v časově bližší době - konkrétně budeme zkoumat strukturu podmíněných závislostí mezi zvolenými měsíčními logaritmickými výnosy odvětvových indexů BCCP v letech 2001 - 2004.

Zajímavé by jistě bylo zvolit stejných 6 indexů jako ve výše popsané aplikaci. Bohužel, zatímco v letech 1994 - 2000 jsme měli ještě na výběr z dostatečného množství indexů, v dalším období se tato situace razantně zhoršila. Například index BI18: investiční fondy se od 11.6.2002 (jistě k nemalé radosti majitelů akcií investičních fondů z kuponové privatizace) již nepočítá - blíže viz tabulka 7.1.

Přesto i koncem roku 2004 zbývají některé indexy, které nám snad dají přehled o situaci na českém kapitálovém trhu - podívejme se na měsíční logaritmické výnosy šesti z nich (pro přehlednější grafické vyjádření výsledků indexy opět očíslujeme přirozenými čísly $1, \ldots, 6$):

Tabulka 11.2: Odvětvové indexy zvolené pro analýzu v letech 2001 - 2004

Označení	**Index**	**Odvětví**
1	BI04	Těžba a zpracování nerostů
2	BI07	Chemický, farmac. a gumár. průmysl
3	BI08	Stavebnictví a prům. stavebních hmot
4	BI12	Energetika
5	BI13	Doprava a spoje
6	BI15	Peněžnictví

Varianční matice těchto indexů má následující hodnoty:

$$S = \begin{pmatrix} 0,003394 & & & & & \\ 0,001886 & 0,006152 & & & & \\ 0,000995 & 0,000633 & 0,001765 & & & \\ 0,000984 & 0,001908 & 0,000776 & 0,002950 & & \\ 0,000170 & 0,002444 & 0,001256 & 0,002684 & 0,013185 & \\ 0,000667 & 0,000783 & 0,000154 & 0,000985 & 0,000000 & 0,003001 \end{pmatrix}.$$

Trochu lepší přehled o struktuře dat nám poskytne korelační matice:

$$\begin{pmatrix} 1 & & & & & \\ 0,4127 & 1 & & & & \\ 0,4064 & 0,1920 & 1 & & & \\ 0,3110 & 0,4478 & 0,3398 & 1 & & \\ 0,0254 & 0,2714 & 0,2603 & 0,4304 & 1 & \\ 0,2089 & 0,1822 & 0,0670 & 0,3311 & 0,0015 & 1 \end{pmatrix}.$$

Z korelační matice je patrná lineární závislost mezi většinou indexů.

Přesnější výsledky nám opět poskytnou grafické modely. Aplikací 4 výše uvedených procedur získáme následující (upravené) matice sousednosti (testy nezávislosti a normality logaritmických výnosů jsme provedli v programu Mathematica a jsou uvedeny v příloze):

Pomocí backward algoritmu se STOP pravidlem založeném na celková devianci dostaneme tuto matici sousednosti:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \\ BI13 \\ BI15 \end{array} \begin{array}{c} \begin{array}{cccccc} BI04 & BI07 & BI08 & BI12 & BI13 & BI15 \end{array} \\ \begin{pmatrix} & & & & & \\ 1 & & & & & \\ 1 & -1 & & & & \\ -2 & 1 & -3 & & & \\ -7 & -5 & -8 & 1 & & \\ -6 & -9 & -10 & 1 & -4 & \end{pmatrix} \end{array}.$$

Backward algoritmus se STOP pravidlem založeném na devianci vynechané hrany má tento výstup:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \\ BI13 \\ BI15 \end{array} \begin{array}{c} \begin{array}{cccccc} BI04 & BI07 & BI08 & BI12 & BI13 & BI15 \end{array} \\ \begin{pmatrix} & & & & & \\ 1 & & & & & \\ 1 & -1 & & & & \\ -2 & 1 & -3 & & & \\ -7 & -5 & -8 & 1 & & \\ -6 & -9 & 1 & 1 & -4 & \end{pmatrix} \end{array}.$$

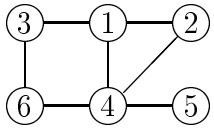
Forward algoritmu se STOP pravidlem založeném na celkové devianci vygeneruje následující matici:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \\ BI13 \\ BI15 \end{array} \begin{array}{c} \begin{array}{cccccc} BI04 & BI07 & BI08 & BI12 & BI13 & BI15 \end{array} \\ \begin{pmatrix} & & & & & \\ 4 & & & & & \\ 5 & 0 & & & & \\ 0 & 2 & 0 & & & \\ 0 & 0 & 0 & 3 & & \\ 0 & 0 & 0 & 1 & 0 & \end{pmatrix} \end{array}.$$

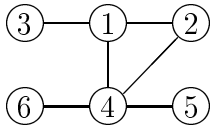
A konečně forward algoritmus se STOP pravidlem založeným na devianci vynechané hrany poskytne výstup ve formě:

$$\begin{array}{c} \\ BI04 \\ BI07 \\ BI08 \\ BI12 \\ BI13 \\ BI15 \end{array} \begin{array}{c} \begin{array}{cccccc} BI04 & BI07 & BI08 & BI12 & BI13 & BI15 \end{array} \\ \begin{pmatrix} & & & & & \\ 4 & & & & & \\ 5 & 0 & & & & \\ 0 & 2 & 0 & & & \\ 0 & 0 & 0 & 3 & & \\ 0 & 0 & 6 & 1 & 0 & \end{pmatrix} \end{array}.$$

Algoritmy se stop pravidlem založeným na devianci vynechané hrany (backward i forward) vybraly shodně graf se 6 hranami (tj. s 9 chybějícími hranami oproti saturovanému modelu):



Algoritmy se stop pravidlem založeným na celkové devianci vybraly (opět shodně) graf, který obsahuje dokonce pouze 5 hran (tj. oproti kompletnímu grafu v něm chybí 10 hran):



Zdá se tedy, že i přes skutečnost, že odvětvových indexů zveřejňovaných BCPP značně ubylo, rozhodně nedošlo k jejich většímu provázání. Navíc stále platí určité závěry, které jsou v jistém smyslu „proti logice“. Očekával jsem, že index peněžnictví bude mít úzkou vazbu na vývoj všech ostatních indexů. Vždyť banky přece „žijí“ z toho, že poskytují úvěry podnikům. Když se těmto společnostem daří, nehrozí potíže se splácením úvěrů a tato situace je pozitivní i pro bankovní sektor (a samozřejmě naopak). V uvedených grafech je index BI15: peněžnictví (reprezentovaný vrcholem 6) ale spojen hranou pouze s indexy BI08: Stavebnictví a prům. stavebních hmot (vrchol 3) a BI12: Energetika (vrchol 4), případně pouze s vrcholem 4 v grafu vybraném pomocí algoritmů se STOP pravidlem založeným na celkové devianci. Znamená to snad, že banky půjčují peníze pouze stavebním a energetickým firmám? Podle mě leží příčina uvedeného výsledku někde jinde.

Odpověď na otázku, proč se index peněžnictví chová značně nezávisle na ostatních indexech, lze dle mého názoru nalézt například v článku [11]: „za vysokými zisky českých bank stojí především rekordní výnosy z poplatků za bankovní služby, které činí téměř 40 procent všech příjmů“ (i když se jedná o popis situace v roce 2005, jde o dlouhodobější vývojový trend, kterému zřejmě nezabrání ani snahy antimonopolního úřadu). Hledat souvislost mezi tím, kolik zaplatí majitelé účtu za jeho vedení, výběry z účtu a dokonce i vklady peněz na účet a vývojem průmyslových odvětví, je zřejmě opravdu zbytečné.

11.2 Globalizace světových akciových trhů[2]

Podívejme se nyní na provázanost akciových indexů z širšího hlediska.

V dnešní době se stále častěji mluví o globalizaci světových akciových trhů. Jakýkoli pohyb na jednom akciovém trhu se prý během velmi krátké doby projeví i na trzích ostatních. Snad nejlépe situaci vystihuje přísloví finančníků: „Když Amerika kýchne, Evropa dostane rýmu.“ V následující aplikace se pomocí grafických modelů pokusíme na tuto situaci podívat poněkud podrobněji - popíšeme strukturu závislosti mezi několika americkými a několika evropskými akciovými indexy, a to jak v současnosti, tak v nedávné minulosti.

Pro zodpovězení otázky týkající se globalizace světových akciových trhů jsme vybrali 3 americké a 3 evropské indexy ve dvou čtyřletých obdobích (1992-1995 a 2000-2003). Stručný popis jednotlivých indexů obsahuje následující tabulka:

[2]Závěry byly publikovány v článcích [17], [18].

Tabulka 11.3: Světové akciové indexy

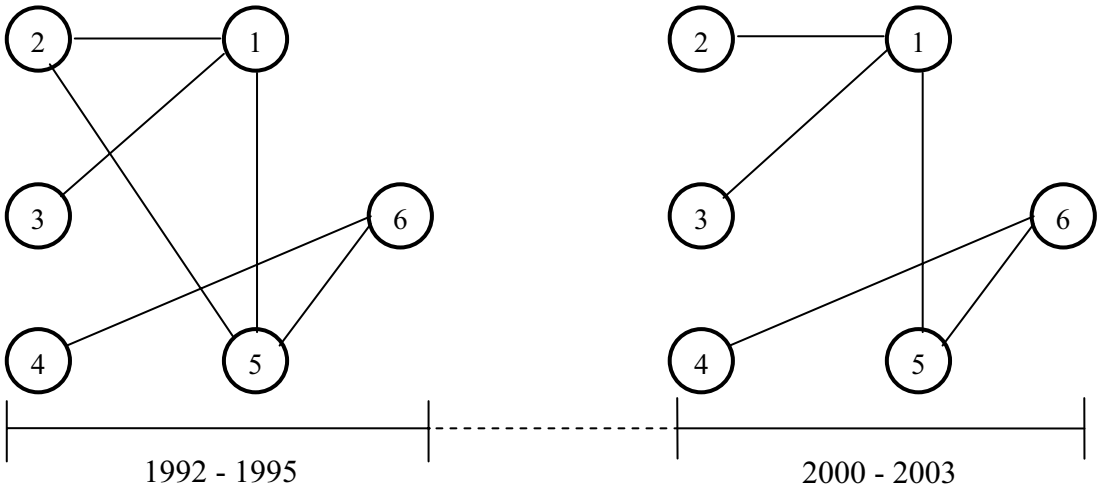
Označení	Region	Popis indexu
DJIA	Amerika - USA	Dow Jones Industrial Average je snad nejznámějším akciovým indexem. Byl založen Charlesem Dowem již v roce 1896[3] a může tak být právem považován za „otce“ všech moderních burzovních indexů. V současnosti je založen na 30 akciích nejvýznamnějších (z hlediska obratu) průmyslových firem (tzv. „blue chips“, které jsou obchodovány na NYSE (New York Stock Exchange). Můžeme zde najít například akcie firem Boeing, IBM, Procter & Gamble, Coca-Cola, Eastman Kodak, General Motors, Philip Morris, Texaco, Westinghouse Electric atd. Co se týká výpočtu indexu, je naprosto nestandardní a lze ho označit „jednoduše za velmi složitý“ - viz [3] - str. 21.
DJTA	Amerika - USA	Dow Jones Transportation Average sestává z 20 akcií největších dopravních a přepravních společností (např. Canadian Pacific, Delta Air Lines, Union Pacific atd.)
DJUA	Amerika - USA	Dow Jones Utilities Average je počítán z 15 akcií veřejně prospěšných společností (tj. především elektrárenských a plynárenských společností).
DAX30	Evropa - Německo	DAX (Deutsche Aktienindex) je indexem německé akciové burzy ve Frankfurtu. Sestává z 30 nejčastěji obchodovaných akcií.
FTSE100	Evropa - Velká Británie	Financial Times Stock Exchange (zvaný „footsie“) je konvexní lineární kombinací 100 společností s největší tržní kapitalizací z Velké Británie. Tyto společnosti tvoří zároveň přibližně 70 % celkové kapitalizace britského akciového trhu. Výchozí hodnota 1000 byla nastavena 3.1.1984.
CAC40	Evropa - Francie	CAC 40 index je hlavním indexem akciové burzy Euronext Paris. Obsahuje 40 akciových titulů, které jsou vybrány mezi stem akcií s největší tržní kapitalizací.

U finančních dat (jimiž údaje o hodnotě akciových indexů bezesporu jsou) bývá často

[3]Ve skutečnosti ještě o trochu dříve - již v červenci 1884 začali novináři Charles Dow a Edward Jones sledovat průměrný vývoj kursů jedenácti největších průmyslových firem, od roku 1896 pak začali pravidelně zveřejňovat DJIA - viz [9]. V historických dobách ale tento index obsahoval různý počet akcií jednotlivých firem. Původně šlo o 11 významných firem (především z oblasti železnic), roku 1916 se počet zvýšil na 20 firem a konečně v roce 1928 na současný počet 30 firem.

problém s nezávislostí a normalitou. Proto jsme i v tomto případě použili transformaci na logaritmické výnosy (vlastní testy nezávislosti a normality jsou uvedeny v příloze).

Za pomoci backward algoritmu se STOP pravidlem založeným na celkové devianci jsme získali následující 2 grafy:



Jak je dobře patrné z obrázků, vztahy mezi indexy nejsou zdaleka tak těsné, jak by se dalo očekávat z výše uvedeného přísloví. V prvním grafu (který reprezentuje indexy v období let 1992-1995) je vynecháno 9 hran a ve druhém dokonce 10 hran.

Zajímavá je absence hrany mezi vrcholy 2 a 3, což znamená, že $DJTA \perp DJUA|DJIA$. Také evropské indexy DAX30 a FTSE100 jsou podmíněně nezávislé při pevných hodnotách indexu CAC40. Jediná vazba mezi Amerikou a Evropou je přes indexy DJIA a FTSE100.

Dle mého názoru lze důvod v menší provázanosti amerického a evropských akciových trhů hledat především v mezinárodních událostech, které mohly vést k nestabilitě finančních trhů v poslední době. Jedná se zejména o válku v Iráku, teroristické útoky na USA a povodně v Evropě.

11.2.1 Souvislost světových akciových trhů a trhu českého

Pokusme se nyní zodpovědět otázku, zda je také český akciový trh (reprezentovaný indexem PX 50) ovlivňován zahraničními akciovými trhy (zahraničními akciovými indexy). Nejprve si však něco řekněme o hlavním indexu Pražské burzy - PX 50[4].

Index PX 50

Burza zavedla svůj oficiální index PX 50 při příležitosti prvního výročí zahájení obchodování. Byl zvolen standardní výpočet indexu ve shodě s metodologií IFC (International Finance Corporation) doporučenou pro tvorbu indexů na vznikajících trzích. Na základě rozborů bylo rozhodnuto vytvořit bázi složenou z 50 emisí. V současné době je počet

[4]Tato věta přestala k 20.3.2006 platit - ten den totiž došlo k sloučení indexu PX50 a indexu PX-D v jeden index s názvem PX - viz [10].

bazických emisí variabilní. V souladu se Zásadami aktualizace báze indexu PX 50 schválenými v prosinci 2001 však nemůže převýšit padesát. Do báze indexu se nezařazují emise oboru č.18 (investiční fondy) a holdingových společností vzniklých transformací z investičních fondů, neboť v jejich kursech se již promítají cenové pohyby bazických emisí. Za výchozí burzovní den byl zvolen 5. 4. 1994, výchozí hodnotou indexu PX 50 se stalo 1 000 bodů.

K 31.12.2003 sestává index PX 50 nikoli z 50 akciových titulů (jak by snad mohl mylně napovídat název), ale pouze z 15[5]. Jedná se o nejvýznamnější společnosti na českém trhu, což je dobře patrné z následující tabulky (viz [35]).

Tabulka 11.4: Báze indexu PX 50 v roce 2003

Pořadí	ISBN	Název CP	Tržní kapitalizace	Váha %
1	AT0000652011	ERSTE BANK	103659,4	23,38
2	CZ0008019106	KOMERČNÍ BANKA	91907,8	20,73
3	CZ0005112300	ČEZ	86285,1	19,46
4	CZ0009093209	ČESKÝ TELECOM	61777,3	13,93
5	CS0008418869	PHILIP MORRIS ČR	30098,6	6,79
6	CZ0008002755	ČESKÁ POJIŠŤOVNA	17085,1	3,85
7	CZ0009091500	UNIPETROL	12047,9	2,72
8	CZ0009054607	ČESKÉ RADIOKOMUN.	10660,5	2,40
9	CZ0005102350	SEVEROČESKÉ DOLY	7025,6	1,58
10	CZ0005098251	ISPAT NOVÁ HUŤ	5575,9	1,26
11	CZ0005077354	ZČ ENERGETIKA	4678,6	1,06
12	CZ0005078253	STČ ENERGETICKÁ	4538,1	1,02
13	CZ0005103952	SOKOLOVSKÁ UHELNÁ	3510,1	0,79
14	CZ0005077057	JČ ENERGETIKA	2910,4	0,66
15	CS0008416251	RM-S HOLDING	1607,3	0,36

Vlastní index se pak počítá podle následujícího vzorce (viz [35]):

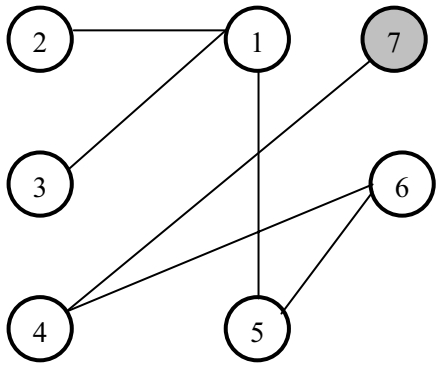
$$\mathrm{PX50}(t) = K(t)\frac{M(t)}{M(0)}1000,$$

kde $M(t)$ je tržní kapitalizace báze v čase t,
$M(0)$ je tržní kapitalizace v základním (výchozím) období,
$K(t)$ je faktor zřetězení v čase t (zohledňuje změny provedené v bázi indexu) .

Poznamenejme ještě, že ke změně hodnoty báze ve výše uvedeném vztahu dochází nejen při výměně bazické emise, ale i při všech operacích, které mění tržní kapitalizaci bazické emise (zvýšení, resp. snížení počtu cenných papírů v emisi, splynutí emisí, emise předkupních práv). Speciální operace výplata dividendy nezpůsobuje v případě indexu PX 50 změnu hodnoty báze, dividendové výnosy se nezohledňují. Jedná se tedy o index cenový, nikoliv o „return“ index. Transformace hodnoty báze je založena na principu spojitosti indexu v okamžiku změny báze.

[5]K 31.12.2004 ze 16 emisí a k 31.12.2005 dokonce pouze ze 14 emisí, přičemž dochází rovněž k významným změnám váhy jednotlivých společností.

Odpověď na otázku provázání českého a zahraničních trhů nám poskytuje následující graf (použili jsme logaritmické výnosy za období 2000-2003 a index PX50 je reprezentován vrcholem 7 a test nezávislosti a normality je opět uveden v příloze):



Z grafu je patrné, že PX50 je podmíněně nezávislý se všemi ostatními indexy při pevných hodnotách německého akciového indexu DAX30. To si zdůvodňuji velmi úzkou provázaností české a německé ekonomiky.

11.3 Provázanost měnových kurzů

Jiná otázka, která může být zodpovězena za pomoci grafických modelů pro finanční data, je, do jaké míry jsou provázány kurzy evropských měn. Oproti aplikacím uvedeným v [22] a [26] máme sice k dispozici díky zavedení eura na výběr z mnohem menšího počtu kurzů, ale zato můžeme zkoumat, jak se chovají kurzy měn zemí, které si vlastní měnu dosud ponechaly (a zřejmě ani v budoucnu nemají zájem na této situaci nic měnit) - Velká Británie, Švýcarsko, severské státy. Na druhé straně stojí země, které by euro rády v nejbližší době zavedly - Polsko a Slovensko.

Protože se jedná o měnové kurzy, které jsou vždy dvoustranné, budeme danou problematiku zkoumat z hlediska české koruny jako referenční měny.

Země vybrané pro analýzu, jejich měny a označení uvádí následující tabulka (pro vlastní výpočet jsme použili týdenní logaritmické výnosy těchto měn v období let 2003-2005 a testy nezávislosti a normality lze nalézt opět v příloze):

Tabulka 11.5: Kurzy měn zvolené pro analýzu v letech 2003 - 2005

Označení	Kód	Množství	Země	Měna
1	EUR	1	Evropská unie	euro
2	DKK	1	Dánsko	koruna
3	NOK	1	Norsko	koruna
4	SEK	1	Švédsko	koruna
5	CHF	1	Švýcarsko	frank
6	GBP	1	Velká Británie	libra
7	PLN	1	Polsko	zlotý
8	SKK	100	Slovensko	koruna

Z korelační matice je patrná poměrně silná lineární závislost mezi jednotlivými měnovými kurzy:

$$\begin{pmatrix} 1 & & & & & & & \\ 0,9988 & 1 & & & & & & \\ 0,5433 & 0,5472 & 1 & & & & & \\ 0,7484 & 0,7533 & 0,5338 & 1 & & & & \\ 0,8508 & 0,8594 & 0,5787 & 0,6273 & 1 & & & \\ 0,5280 & 0,5374 & 0,2347 & 0,3854 & 0,5293 & 1 & & \\ 0,3505 & 0,3511 & 0,2267 & 0,3273 & 0,2128 & 0,2225 & 1 & \\ 0,6426 & 0,6460 & 0,3713 & 0,5283 & 0,5091 & 0,3595 & 0,5061 & 1 \end{pmatrix}.$$

Aplikujme nyní naše 4 selekční algoritmy a pokusme se intepretovat výsledky:

- Matice sousednosti získaná pomocí backward algoritmu se STOP pravidlem založeným na celkové devianci

$$\begin{array}{c} \\ EUR \\ DKK \\ NOK \\ SEK \\ CHF \\ GBP \\ PLN \\ SKK \end{array} \begin{array}{c} \begin{array}{cccccccc} EUR & DKK & NOK & SEK & CHF & GBP & PLN & SKK \end{array} \\ \begin{pmatrix} & & & & & & & \\ 1 & & & & & & & \\ -3 & -6 & & & & & & \\ -15 & 1 & 1 & & & & & \\ 1 & 1 & 1 & -16 & & & & \\ -19 & 1 & -14 & -5 & -13 & & & \\ -2 & -17 & -11 & -8 & -18 & -9 & & \\ -12 & 1 & -4 & -7 & -10 & -1 & 1 & \end{pmatrix} \end{array}.$$

- Matice sousednosti získaná pomocí backward algoritmu se STOP pravidlem založeným na devianci vynechané hrany

$$\begin{array}{c} \\ EUR \\ DKK \\ NOK \\ SEK \\ CHF \\ GBP \\ PLN \\ SKK \end{array} \begin{array}{c} \begin{array}{cccccccc} EUR & DKK & NOK & SEK & CHF & GBP & PLN & SKK \end{array} \\ \begin{pmatrix} & & & & & & & \\ 1 & & & & & & & \\ -3 & -6 & & & & & & \\ -15 & 1 & 1 & & & & & \\ 1 & 1 & 1 & -16 & & & & \\ 1 & 1 & -14 & -5 & -13 & & & \\ -2 & 1 & -11 & -8 & 1 & -9 & & \\ -12 & 1 & -4 & -7 & -10 & -1 & 1 & \end{pmatrix} \end{array}.$$

- Matice sousednosti získaná pomocí forward algoritmu se STOP pravidlem založe-

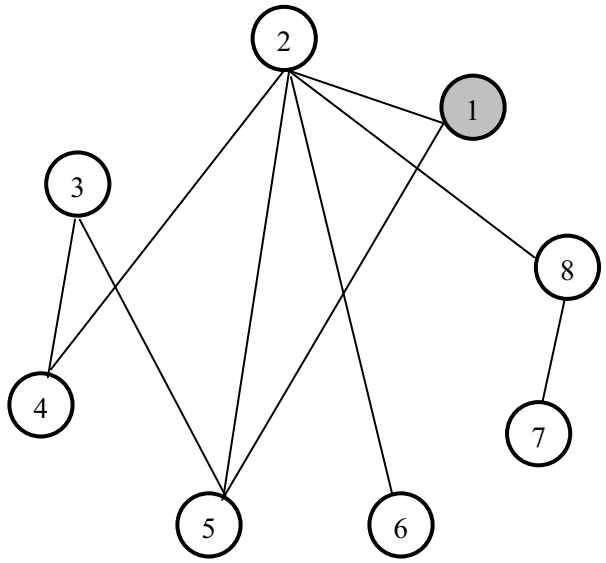
ným na celkové devianci

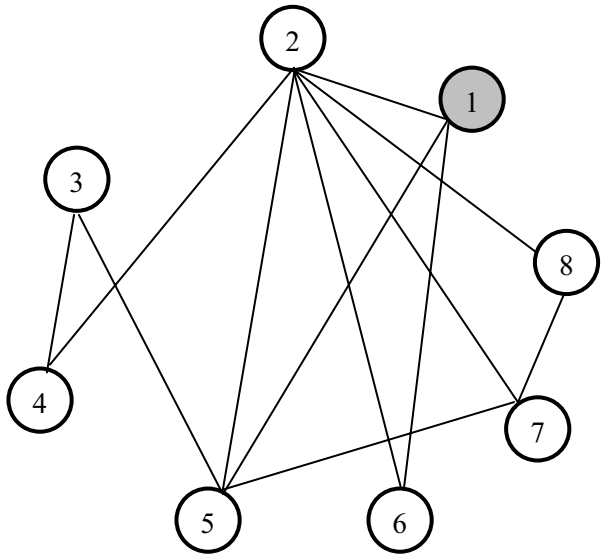
$$\begin{array}{c} \\ EUR \\ DKK \\ NOK \\ SEK \\ CHF \\ GBP \\ PLN \\ SKK \end{array} \begin{array}{c} \begin{array}{cccccccc} EUR & DKK & NOK & SEK & CHF & GBP & PLN & SKK \end{array} \\ \begin{pmatrix} & & & & & & & \\ 1 & & & & & & & \\ 0 & 0 & & & & & & \\ 0 & 3 & 9 & & & & & \\ 8 & 2 & 5 & 0 & & & & \\ 10 & 6 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & \\ 0 & 4 & 0 & 0 & 0 & 0 & 7 & \end{pmatrix} \end{array}.$$

- Matice sousednosti získaná pomocí forward algoritmu se STOP pravidlem založeným na devianci přidané hrany

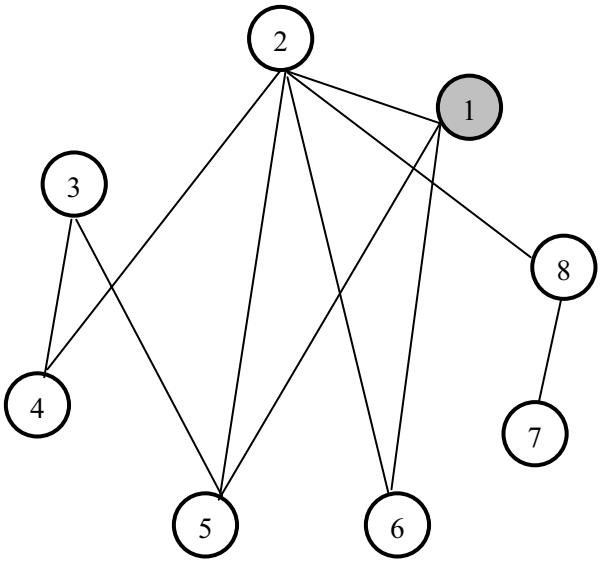
$$\begin{array}{c} \\ EUR \\ DKK \\ NOK \\ SEK \\ CHF \\ GBP \\ PLN \\ SKK \end{array} \begin{array}{c} \begin{array}{cccccccc} EUR & DKK & NOK & SEK & CHF & GBP & PLN & SKK \end{array} \\ \begin{pmatrix} & & & & & & & \\ 1 & & & & & & & \\ 0 & 0 & & & & & & \\ 0 & 3 & 9 & & & & & \\ 8 & 2 & 5 & 0 & & & & \\ 10 & 6 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & \\ 0 & 4 & 0 & 0 & 0 & 0 & 7 & \end{pmatrix} \end{array}.$$

Pomocí backward algoritmů jsme tedy získali 2 různé grafy (v jednom je vynecháno 16 a v druhém dokonce 19 hran):





Výstupem obou forward algoritmů je jeden graf s 10 přidanými hranami:



Zatímco vývoj měn Dánska (vrchol 2) a Švýcarska (vrchol 5) a ve dvou z uvedených grafických modelů rovněž Velké Británie (vrchol 6) je (aspoň z pohledu české koruny) přímo svázán s vývojem eura (vrchol 1), Polsko (vrchol 7) a Slovensko (vrchol 8) jsou s eurem podmíněně nezávislé. Nikoho jistě nepřekvapí hrana mezi vrcholy 3 a 4 (Norsko, Švédsko) a rovněž provázanost měn Polska a Slovenska. Poměrným překvapením je pak silná provázanost měnového kurzu dánské koruny s většinou ostatních měn.

Zatímco země, které nezavedly euro (a často ani nejsou součástí evropské unie), si z hlediska vývoje kurzu příliš „nepomohly“, tak země, které by euro rády zavedly, jsou ve sledovaném období s touto měnou podmíněně nezávislé. Je ale třeba mít na paměti, že toto se týká pouze pohledu ze strany české koruny jako referenční měny. Z hlediska jiných zemí může být pohled značně odlišný.

Velký počet vynechaných hran, resp. malý počet přidaných hran oproti poměrně silné provázanosti měnových kurzů uvedené v [22] a [26], lze vysvětlit skutečností, že zatímco zmiňovaní autoři používali převážně měny, které se nedlouho poté staly součástí eura, v našem případě byl zvolen opačný přístup. Pro analýzu jsme vybrali kurzy měn, které s eurem „nechtějí mít nic společného“ - i když se jim to ne vždy daří, případně země, které by euro rády zavedly, ale zřejmě na tom bude potřeba ještě „odvést pořádný kus práce“.

Poznámka 11.6:

Pro analýzu jsme zvolili celkem 8 měn. Tato volba je už na hranici pravidla 7 ± 2, což je ostatně patrné i na uvedených grafech, které začínají být mírně nepřehledné. □

11.4 Hypotézy pro vysvětlení časové struktury úrokových sazeb a jejich test pomocí grafických VAR modelů

Poslední aplikace se již bude týkat orientovaných grafů - pokusíme se aplikovat zmíněné teoretické poznatky o grafických VAR modelech na finanční data - konkrétně se blíže podíváme na strukturu úrokových sazeb[6]. Je možné odvodit dlouhodobé úrokové sazby z krátkodobých, krátkodobé z dlouhodobých, nebo mezi nimi není žádný vztah?

Poznámka 11.7:

Vztah mezi výnosností do splatnosti a dobou do splatnosti dluhopisů se obecně označuje jako výnosová křivka. Výnosová křivka vypovídá věrohodně o závislosti výnosností na době do splatnosti jen tehdy, pokud je konstruována na základě dluhopisů lišících se pouze dobou do splatnosti, ale jinak přibližně shodných vlastností (typ emitenta, riziko ve formě ratingu, kupónová sazba, zdanění, podmínky svolatelnosti atd.). Proto se výnosové křivky konstruují především pro státní dluhopisy - viz [25]. □

Pro vysvětlení struktury úrokových sazeb (tj. pro vztahy úrokových sazeb pro instrumenty s různou dobou splatnosti) se vyvinulo několik teorií. Mezi nejznámější patří:

1. hypotéza očekávání
2. hypotéza oddělených trhů
3. hypotéza preferovaného umístění

Hypotéza očekávání vychází z představy, že jednotlivé dluhopisy jsou dokonalé substituty. Časová struktura úrokových sazeb je pak tedy ovlivňována pouze očekáváním o vývoji budoucích úrokových sazeb. Dlouhodobé sazby jsou dány očekávanými budoucími krátkodobými úrokovými sazbami dle vztahu:

$$(1 + i_{0;\,n}) = \sqrt[n]{(1 + i_{0;\,1}) \cdot (1 + i_{1;\,2}) \ldots (1 + i_{n-1;\,n})},$$

kde:

$i_{0;\,n}$ je dlouhodobá úroková sazba (sazba z dluhopisu se splatností n let),
$i_{0;\,1}$ je běžná jednoroční úroková sazba,
$i_{t-1;\,t}$ je očekávaná úroková sazba z dluhopisu se splatností 1 rok za t-1 let.

Poznámka 11.8:

Pro kratší období je nutno použít jednoduché úročení[7] - pak má výše uvedený vztah následující podobu:

$$i_{0;\,t_n} = \frac{i_{0;\,t_1} + i_{t_1;\,t_2} + i_{t_{n-1};\,t_n}}{t_n},$$

[6] Jde o „klasický“ případ použití - viz např. [24], [27], [32].

[7] Jednoduché úročení se používá při uložení kapitálu na kratší dobu než je 1 úrokovací období a od složeného úročení se liší tím, že se úroky nepřipisují k vloženému kapitálu a dále se spolu s ním neúročí - tj. nepočítají se úroky z úroků.

kde:

$i_{t_{i-1};\,t_i}$ jsou očekávané roční úrokové sazby při uložení kapitálu v čase t_{i-1} na dobu $t_i - t_{i-1}$. □

Proti této hypotéze hovoří skutečnost, že ve vyspělých státech mají výnosové křivky zpravidla stoupající strukturu (viz [23] - str. 400). To by však znamenalo, že se očekává budoucí růst krátkodobých úrokových sazeb. Není však žádný důvod, aby krátkodobé úrokové sazby neustále rostly.

Hypotéza oddělených trhů naopak vychází z představy, že dluhopisy s rozdílnou dobou splatnosti nejsou substituty. Na dluhopisových trzích totiž investují různí investoři s různými cíli. Zatímco krátkodobé cenné papíry preferují banky z důvodu řízení likvidity, dlouhodobé dluhopisy jsou nástrojem uložení peněz pro pojišťovny životního pojištění a penzijní fondy. Časová struktura úrokových sazeb je podle této hypotézy způsobena odlišnou poptávkou a nabídkou na jednotlivých segmentech trhů dluhopisů s rozdílnou dobou splatnosti.

Nevýhodou této hypotézy je skutečnost, že nedokáže vysvětlit, proč se na vyspělých finančních trzích úrokové sazby z instrumentů s různou dobou splatnosti nepohybují nezávisle.

Za nejvíce přijímané (viz [23] - str. 401) vysvětlení chování časové struktury úrokových sazeb je považována **hypotéza preferovaného umístění.**

Tato hypotéza je založena na dvou předpokladech

1. Dluhopisy jsou poměrně dobrými substituty
2. Investoři mají určité preference

Podle této hypotézy úrokové sazby sice reflektují současné a očekávané budoucí krátkodobé úrokové sazby (stejně jako v hypotéze očekávání), je však třeba uvažovat také s tzv. prémií za riziko. Protože nejistota vzrůstá s prodlužováním doby splatnosti, investoři preferují zapůjčovat peníze v krátkém období. Vypůjčovatelé však naopak preferují získávat dlouhodobé zdroje. Investor tedy požaduje určitou odměnu (prémii) za ochotu investovat dlouhodobě. Dlouhodobá úroková sazba se tak rovná průměru krátkodobých úrokových sazeb, které se očekávají během doby splatnosti dlouhodobého dluhopisu, a prémii za riziko.

11.4.1 Struktura úrokových sazeb na českém finančním a kapitálovém trhu

Podívejme se nyní na vývoj úrokových sazeb v České republice. Použití státních obligací (jak je uvedeno v poznámce 11.7) by bylo jistě nejlepším řešením, bohužel narážíme na značný problém s daty. Proto jsme k analýze zvolili týdenní vývoj úrokové míry PRIBOR, který zveřejňuje na svých stránkách ČNB (viz [34]) a který osobně považujeme za mnohem transparentnější, než je vývoj na trhu (zejména dlouhodobých) státních dluhopisů.

Zvolili jsme 3 různé doby splatnosti peněžních prostředků - 1, 6 a 12 měsíců (PRIBOR 1M, PRIBOR 6M, PRIBOR 12M) a budeme se zajímat o jejich vzájemné vazby. Protože VAR modely pracují často s velkým počtem odhadnutých parametrů, tyto sazby budeme zkoumat za pětileté období (2001-2005). K dispozici tedy máme více než 250 pozorování.

Postupovat budeme přesně podle návodu uvedeného v 11.4.

0. Test stacionarity použitých časových řad.
Výsledky získané ze SW PCGive shrnuje následující tabulka:

Úroková míra	**ADF test**	**odhad β_{-1}**
PRIBOR 1M	-2.047	0.99298
PRIBOR 6M	-2.218	0.99177
PRIBOR 12M	-2.264	0.99060

Kritická hodnota pro ADF test je na 5% hladině významnosti = -2,87. Ani u jedné časové řady tedy nemůžeme zamítnou hypotézu, že $\beta_{-1} = 1$ (tj. že proces obsahuje jednotkový kořen). Přistoupil jsem proto k transformaci řad pomocí prvních diferencí a test zopakoval:

Úroková míra	**ADF test**	**odhad β_{-1}**
PRIBOR 1M	-14.47	0.098321
PRIBOR 6M	-13.11	0.19569
PRIBOR 12M	-12.31	0.25919

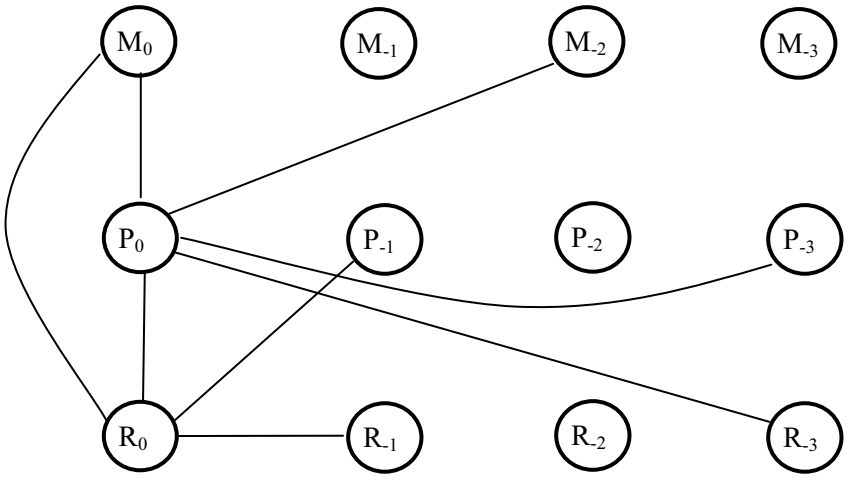
Nyní již můžeme považovat časové řady za stacionární (hypotézu $\beta_{-1} = 1$ zamítáme i na 1% hladině významnosti - kritická hodnota testu je -3.46).

1. Určení maximálního počtu zpoždění (p). Opět jsme využili PCGive a spočetli hodnoty Akeikeho kritéria (AIC) pro různé volby p.

Zpoždění	**AIC**
1	-7.5284
2	-7.5808
3	-7.6941
4	-7.6833

AIC tedy nabývá své minimální hodnoty pro délku zpoždění $p = 3$.

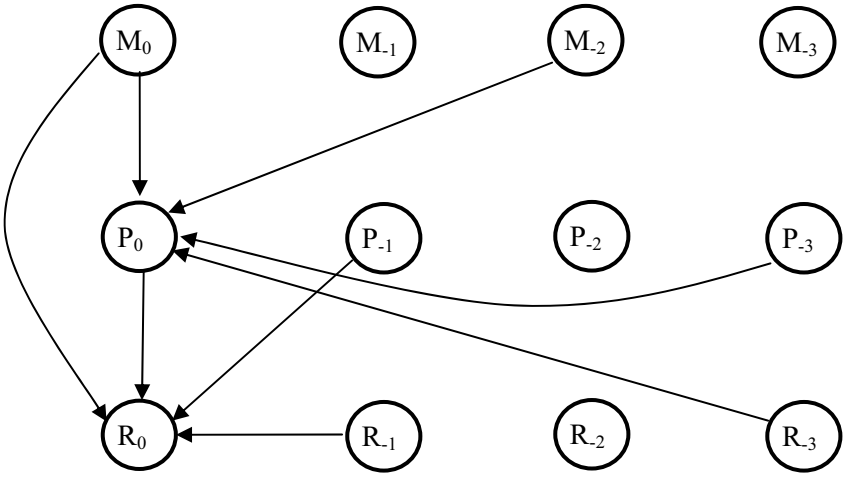
2, 3, 4 Spočteme výběrovou kovarianční matici S, odhadneme koeficienty parciální korelace a určíme, které je možno považovat za nulové.
V těchto krocích jsme využili naprogramované selekční procedury backward algoritmu se STOP pravidlem založeným na devianci vynechané hrany a získali jsme následující neorientovaný graf podmíněných nezávislostí - pro jednoduchost zaveďme následující značení proměnných:
Pribor 1M = M (jako měsíční)
Pribor 6M = P (jako pololetní)
Pribor 12M = R (jako roční):



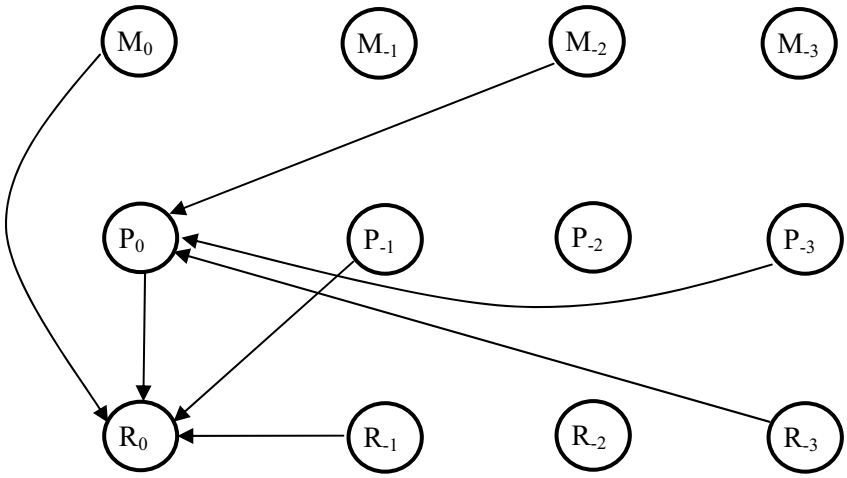
5, 6 Zorientování neorientovaných hran, případné porovnání konkurenčních modelů.

Zatímco orientace hran od minulých proměnných k proměnným běžným je jednoznačná, hrany mezi běžnými proměnnými navzájem mohou mít několik možných orientací (navíc některé z nich mohou být morální). Získali jsme tedy následující konkurenční modely, které jsme (opět v PCGive) porovnali pomocí AIC.

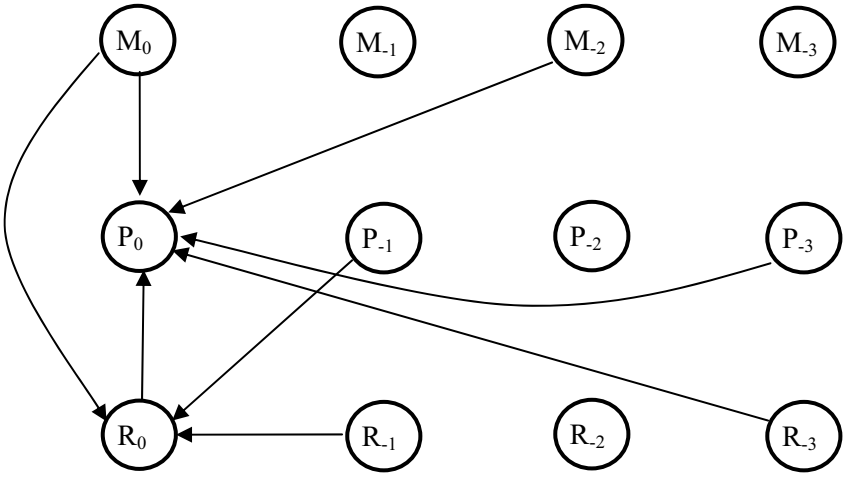
Model 1: AIC = -10.484



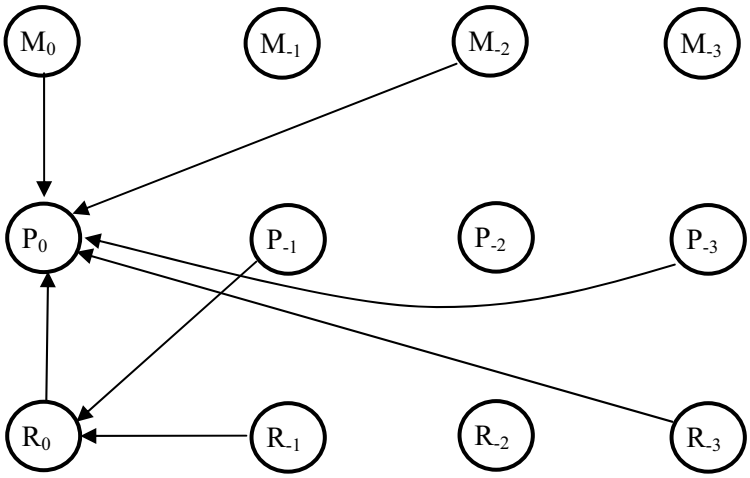
Model 2: AIC = -9.6810



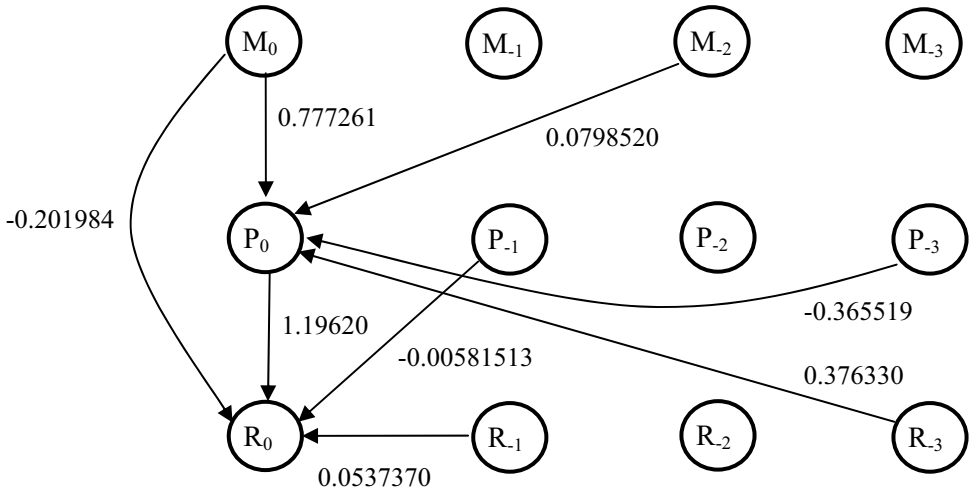
Model 3: AIC = -10.474



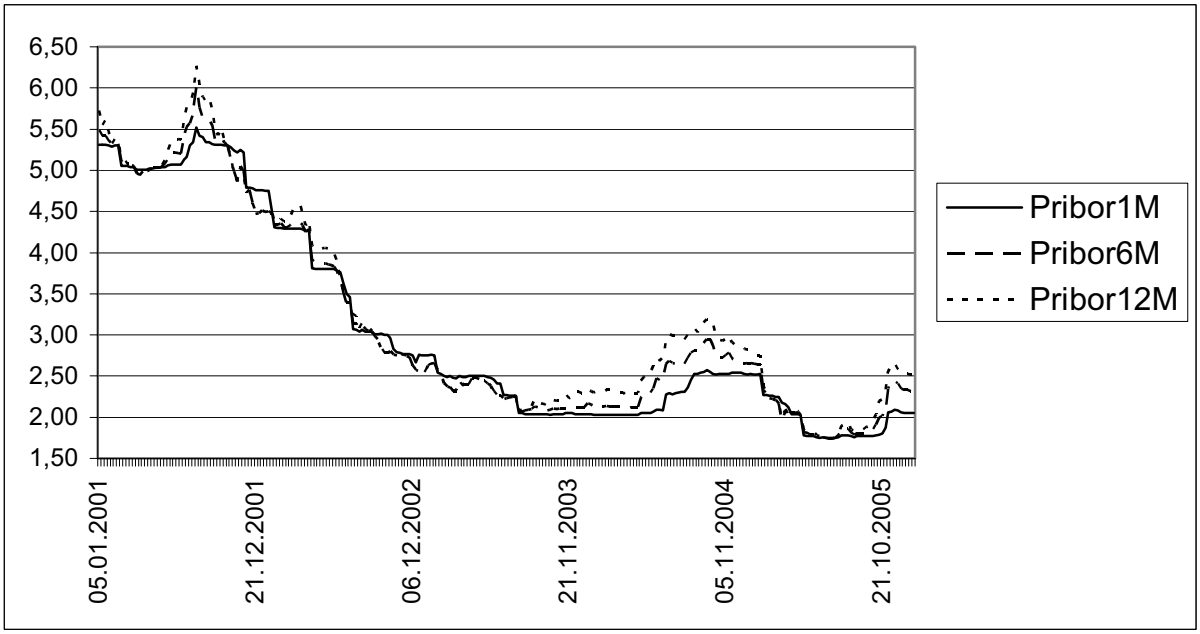
Model 4: AIC = -9.9437



Jako nejlepší se tedy jeví model 1 (i když model 3 za ním příliš nezaostává). V obou těchto modelech je však jasně vidět dominantní postavení úrokové míry PRIBOR 1M (proměnná M). Platí, že M podmiňuje z hlediska Grangerovy kauzality jak P, tak R.



Grafické modelování nám tedy potvrdilo, že existuje vztah mezi krátkodobými a dlouhodobými úrokovými mírami, a sice ve směru od měr krátkodobých k dlouhodobým. Co se týče odpovědi na otázku, zda platí hypotéza očekávání, nebo hypotéza preferovaného umístění, již teorie grafických modelů odpověď nedává. Z grafu vývoje úrokových měr je však patrné, že skutečně existuje jakási prémie za delší splatnost peněžních prostředků.



Závěr

Hlavním cílem práce je popis problematiky grafických modelů určených pro zpracování dat se spojitým rozdělením s ohledem na odhadové procedury a selekci modelu a zejména pak použití těchto procedur na řešení otázek z oblasti financí.

V práci jsou podrobně popsány čtyři selekční iterační algoritmy (a dokázány vztahy mezi různými STOP pravidly) určené pro vyhledání grafického modelu s grafem, který dobře reprezentuje konkrétní data. Tyto algoritmy, které jsou založeny na maximálně věrohodných odhadech varianční matice a testové statistice devianci, lze použít i pro grafy s velkým množstvím vrcholů. Omezení plyne pouze z horší interpretace dosažených výsledků u rozsáhlých grafických modelů. V kapitole 8 lze rovněž nalézt popis několika procedur, které slouží k vlastními odhadu varianční matice a které je možno použít jako „jádro“ do libovolného z selekčních algoritmů. Zdrojové kódy všech 4 algoritmů (byly naprogramovány v SW Matematica) se nalézají v příloze.

Kapitola 11 se zabývá vlastním řešením problémů z oblasti kapitálových a měnových trhů.

První aplikace se pokouší zodpovědět otázku, jaké jsou vztahy mezi odvětvími (odvětvovými indexy) na BCPP. V období let 1994-2000 jsou výsledky dosti překvapivé. Graf reprezentující vztahy podmíněné nezávislosti mezi šesti zvolenými odvětvovými indexy (z důvodu splnění podmínek normality a nezávislosti byly k výpočtům použity měsíční logaritmické výnosy) má velké množství vynechaných hran. Chybějící hrana mezi 2 vrcholy přitom znamená, že proměnné, které vrcholy reprezentují (v našem případě odvětvové indexy), jsou podmíněně nezávislé. Osobně jsem u finanční dat, jimiž údaje o odvětvových indexech jsou, očekával mnohem větší provázanost. Například vrchol grafu reprezentující index investiční fondy (BI18) je spojen pouze s vrcholem indexu výroby nápojů a tabáku BI03 a s vrcholem reprezentujícím elektroprůmysl BI11. Na druhou stranu je poměrně překvapivé, že vrchol reprezentující index služeb (BI16) je spojen s vrcholem reprezentujícím odvětvový index hutního průmyslu a průmyslu zpracování kovů (BI09). Tyto závěry zřejmě spíše souvisí s nepříliš vysokou průhledností českého kapitálové trhu jako takového, než se skutečnými vztahy mezi vývojem jednotlivých odvětví.

V období let 2001-2004 testuji provázanost na případě jiných 6 odvětvových indexů (některé z původních indexů byly totiž zrušeny, což potvrzuje názor malé průhlednosti a likvidnosti českého kapitálového trhu - hodnoty odvětvových indexů BCPP zveřejňuje pouze v případě, že počet emisí v indexu obsažených překračuje hodnotu 3). Z výsledného grafu je patrné, že i přes skutečnost, že odvětvových indexů zveřejňovaných BCPP značně ubylo, rozhodně nedošlo k jejich většímu provázání. Navíc stále platí určité závěry, které jsou v jistém smyslu „proti logice“. Například index peněžnictví nemá úzkou

vazbu na vývoj všech ostatních indexů. V uvedených grafech je index BI15: peněžnictví spojen hranou pouze s indexy BI08: Stavebnictví a prům. stavebních hmot a BI12: Energetika Znamená to snad, že banky půjčují peníze pouze stavebním a energetickým firmám? Podle mého názoru leží příčina uvedeného výsledku někde jinde - v neustále rostoucím podílu poplatků na výnosech českých bank. Hledat souvislost mezi tím, kolik zaplatí majitelé účtu za jeho vedení, výběry z účtu a dokonce i vklady peněz na účet, a vývojem průmyslových odvětví, je zřejmě opravdu zbytečné.

Druhá aplikace se týká otázky provázanosti světových akciových trhů. V dnešní době se stále častěji mluví o globalizaci světových akciových trhů. Jakýkoli pohyb na jednom akciovém trhu se prý během velmi krátké doby projeví i na trzích ostatních. Snad nejlépe situaci vystihuje přísloví finančníků: „Když Amerika kýchne, Evropa dostane rýmu." Závěry analýzy však tuto skutečnost nepotvrdily. Aplikací selekčních algoritmů totiž obdržíme grafy, z nichž je patrné, že vztahy mezi indexy nejsou zdaleka tak těsné, jak by se dalo očekávat z výše uvedeného přísloví. V mezidobí (pro analýzu byly zvoleny měsíční logaritmické výnosy 6 indexů světových akciových burz v období let 1992-1995 a 2000-2003) se provázanost nezvýšila, ba právě naopak (ve výsledném grafu z let 2000-2003 chybí oproti grafu z let 1992-1995 navíc jedna hrana). Jediná vazba mezi Amerikou a Evropou je přes indexy DJIA a FTSE100. Dle mého názoru lze důvod v menší provázanosti amerického a evropských akciových trhů hledat především v mezinárodních událostech, které mohly vést k nestabilitě finančních trhů v poslední době. Jedná se zejména o válku v Iráku, teroristické útoky na USA a povodně v Evropě.

Rovněž český kapitálový trh (reprezentovaný indexem PX50) nejeví v letech 2000-2003 příliš velkou provázanost s vývojem na světových burzách (s vývojem 6 světových burzovních indexů). Analýza ukázala, že PX50 je podmíněně nezávislý s 5 zvolenými zahraničními indexy při pevných hodnotách německého akciového indexu DAX30. To si zdůvodňuji velmi úzkou provázaností české a německé ekonomiky.

V třetí aplikaci je za pomoci grafických modelů pro finanční data řešena otázka, do jaké míry jsou provázány kurzy evropských měn. Konkrétně, jak se chovají kurzy měn zemí, které si i po zavedení eura vlastní měnu ponechaly (a zřejmě ani v budoucnu nemají zájem na této situaci nic měnit) - Velká Británie, Švýcarsko, severské státy - oproti zemím, které by euro rády v nejbližší době zavedly - Polsko a Slovensko.

Protože se jedná o měnové kurzy, které jsou vždy dvoustranné, je tato problematika zkoumána z hlediska české koruny jako referenční měny.

Analýza ukázala, že zatímco země, které nezavedly euro (a často ani nejsou součástí evropské unie), si z hlediska vývoje kurzu příliš „nepomohly" - vrcholy reprezentující jejich měny jsou spojeny hranou s vrcholem představujícím euro, tak země, které by euro rády zavedly, jsou ve sledovaném období s touto měnou podmíněně nezávislé.

Poslední aplikace je z trochu jiné oblasti - zabývá se použitím grafických modelů k zodpovězení otázky, jakým mechanismem se řídí časová struktura úrokových sazeb na českém finančním trhu (jednotlivé hypotézy byly testovány pomocí grafických VAR modelů, z důvodu zajištění dostatečného počtu pozorování byly proto tentokrát použity týdenní hodnoty úrokových sazeb z let 2001-2005). Analýza ukázala dominantní vliv krátkodobé úrokové míry (Pripor 1M) na míry dlouhodobější (Pribor 6M, Pribor 12M) a navíc existenci určité prémie za nelikviditu. Závěry tak hovoří pro platnost hypotézy

preferovaného umístění.

Jak je patrné, grafické modely mohou být nápomocny k zodpovězení mnohých otázek z oblasti financí. Výše uvedené aplikace však zároveň dokládají, že i přes skutečnost, že jde o mocný nástroj statistické analýzy, nemohou vysvětlit zdaleka vše. O příčinách některých skutečností (a zajisté nejen na českém kapitálovém trhu) se můžeme stále jenom dohadovat. A to je snad dobře, protože čím by se měl člověk zabývat, kdyby bylo vše známé?

Literatura

[1] Anděl, J. (1985): Matematická statistika. SNTL, Praha.

[2] Anděl, J. (1998): Statistické metody. Matfyzpress, Praha.

[3] Brada, J. (1994): Materiály ke kurzu Peníze, banky a finanční trhy, ZS 94-95.

[4] Cipra, T. (1986): Analýza časových řad s aplikacemi v ekonomii, SNTL, Praha.

[5] Cipra, T. (1984): Ekonometrie, Státní pedagogické nakladatelství, Praha.

[6] finance.yahoo.com

[7] Giudici,P. (1996): Learning in graphical Gaussian models. Bayesian Statistics 5, 621-628.

[8] Hand, D. J., McConway, K. J., Stanghellini, E. (1997): Graphical models of applicants for credit, IMA Journal of Mathematics Applied in Business and Industry, 143 - 155.

[9] Hospodářské noviny, 26.4.2005: Dow Jones tíží staromódnost

[10] Hospodářské noviny 21.3.2006: Pražská burza vítala nový index posílením, Sloučení indexů zlepší orientaci investorů.

[11] Hospodářské noviny 22.3.2006: ČSOB si upevnila pozici bankovní jedničky

[12] Hušek, R. (1999): Ekonometrická analýza, Ekopress, s.r.o..

[13] Chýna, V. (2002): Grafické modely pro analýzu spojitých finančních dat. Diplomová práce, MFF UK.

[14] Chýna, V. (2003): Grafické modely pro spojitá finanční data - aplikace na odvětvové indexy BCPP, Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze, Oeconomica Praha, 191-204.

[15] Chýna, V. (2003): Graphical Models for the Analysis of the Continuous Financial Data - an Application on the Branch Business Indices (PSE), Mathematical Methods in Economics 2003, ČZU Praha, 139-146.

[16] Chýna, V. (2003): Iterative algorithms for Gaussian Markov distributions over finite graphs (Implementation in SW Mathematica), Week of Doctoral Students 2003, Matfyzpress Praha, 184-189.

[17] Chýna, V. (2004): Globalizace světových akciových trhů. 1. MEDZINÁRODNY SEMINÁR DOKTORANDOV Katedry operačného výskumu a ekonometrie FHI EU v Bratislave a Katedry ekonometrie FIS VŠE v Prahe, VŠE Praha, 45-51.

[18] Chýna, V. (2004): The interconnection of stock indices (by the help of iterative algorithms for Gaussian Markov distributions over finite graphs), Mathematical Methods in Economics 2004, Masarykova univerzita Brno, 140-145.

[19] Johnston, J., DiNardo, J. : Econometric methods - Fourth Edition, The McGraw-Hill Companies, Inc., 287 - 326.

[20] Lauritzen, S. L. (1996): Graphical models. Clarendon Press, Oxford.

[21] Lauritzen, S. L., Spiegehalter, D. J. (1988): Local computations with probabilities on graphical structures and their application to expert system (with discussion). Journal of the Royal Statistical Society, Series B, 50, 157 - 224.

[22] Lněnička, R. (2001): Bayesovské metody pro zpracování finančních dat, Diplomová práce, MFF UK.

[23] Musílek, P. (1996): Finanční trhy: instrumenty, instituce a manegement - II.díl, Ediční oddělení VŠE Praha

[24] Oxley, L., Reale, M., Wilson, G. T. (2004): Finding directed acyclic graphs for vector autoregressions, Compstat 2004 - section: Time series analysis.

[25] Radová J., Chýna V., Málek J. (2005): Finanční matematika v příkladech, Professional Publishing, Praha.

[26] Randáková, G. (2001): Implementace grafických modelů pro finanční analýzu v programu Mathematica, Diplomová práce, MFF UK.

[27] Reale, M., Wilson, G. T. (2001): Identification of vector AR models with recursive structural errors using conditional independence graphs, Statistitcal Methods and Applications 10, 49 - 65.

[28] Reale, M., Wilson, G. T. (2002): The sampling properties of conditional independence graphs for structural vector autoregressions, Biometrika 89, 457 - 461.

[29] Speed, T. P., Kiivery, H. T. (1986): Gaussian Markov Distributions over Finite Graph, The Annals of Statistics 1986, Vol. 14, 138 - 150.

[30] Stanghellini, E. (1999): A discretge variable chain graph for applicants for crecit, Appl. Statistic, 239 - 251.

[31] Whittaker,J. (1990): Graphical models in Applied multivariate Statistics. John Wiley, New York.

[32] Wilson, G. T., Reale, M. (2001): Developments in multivariate time series modeling, Estadística 53, 353 - 395.

[33] Wilson, G. T., Reale, M. (2003): Directed acyclic graphs for I(1) structural VAR models, dosud nepublikováno.

[34] www.cnb.cz

[35] www.pse.cz

[36] www.economagic.com

[37] www.utia.cas.cz/vomlel

[38] Zvára, K. (2004): R & Regrese, Elektronické materiály ke kurzu Regrese, MFF UK.

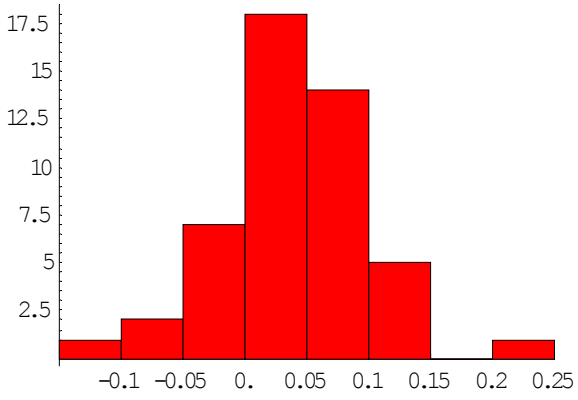
[39] Zelinková, J. (2003): Regrese a grafické modely pro finanční data, Diplomová práce, MFF UK.

PŘÍLOHY

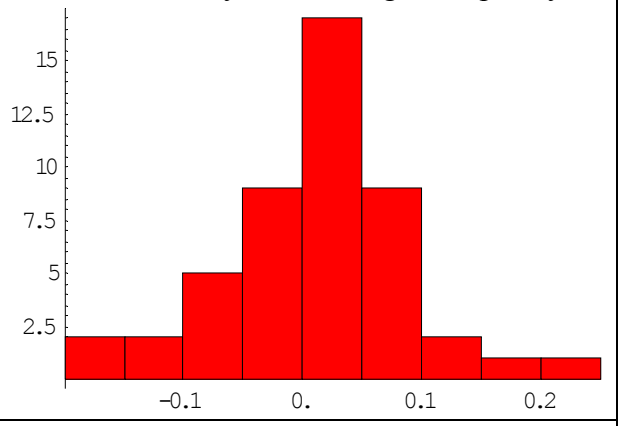
Příloha č. 1: Histogramy vybraných odvětvových indexů

V této příloze jsou uvedeny histogramy měsíčních logaritmických výnosů šesti odvětvových indexů českého kapitálové trhu, které byly použity v ilustračních příkladech.

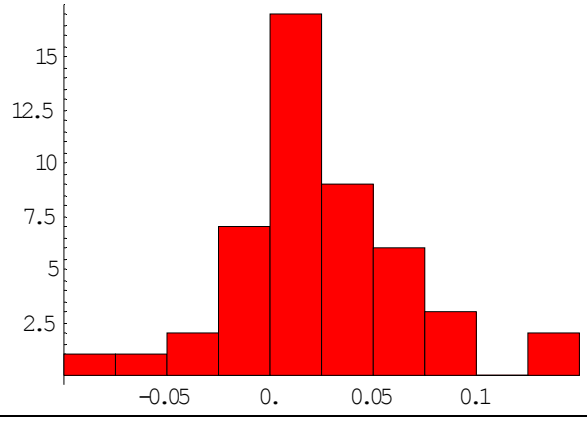
BI04: Těžba a zpracování nerostů



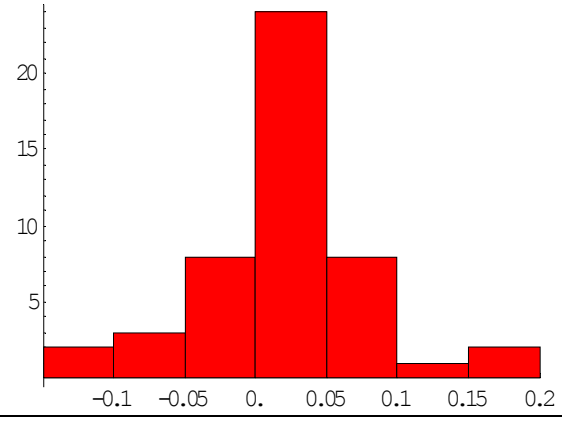
BI07: Chemický, farmac. a gumár. průmysl



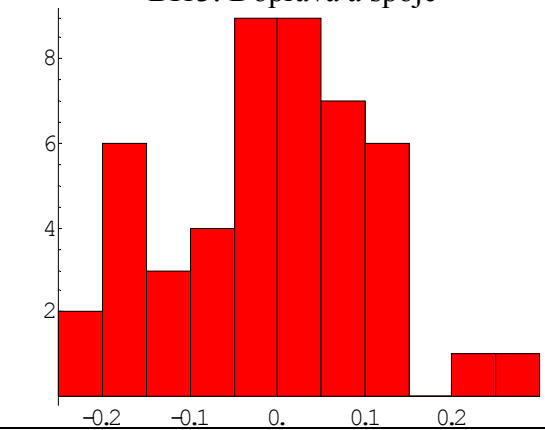
BI08: Stavebnictví a prům. stavebních hmot



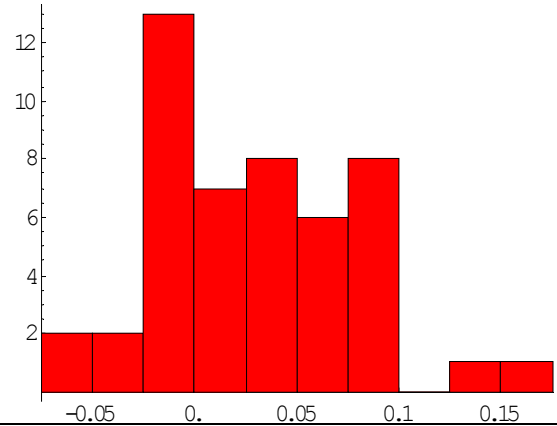
BI12: Energetika



BI13: Doprava a spoje



BI15: Peněžnictví



Příloha č. 2: Výsledky testů nezávislosti a normality

V této příloze jsou uvedeny výsledky testů nezávislosti a normality pro měsíční logaritmické výnosy oborových indexů BCPP v období 30.9.1994 - 31.7.2001, které byly použity v jedné z aplikací.

Index	Test založený na znaménkách diferencí (p-value)	Test Shapiro-Wilk (p-value)
BI03	0,849	0,054
BI05	0,342	0,398
BI09	0,568	0,670
BI11	0,849	0,098
BI16	0,568	0,986
BI18	0,568	0,232

Příloha č. 3: Výsledky testů nezávislosti a normality

V této příloze jsou uvedeny výsledky testů nezávislosti a normality pro měsíční logaritmické výnosy světových akciových indexů a indexu PX50 v období 1992 - 1995 a 2000 - 2003, které byly použity v jedné z aplikací.

Index	**Test založený na znaménkách diferencí (p-value)**		**Test Shapiro-Wilk (p-value)**	
	1992-1995	***2000-2003***	***1992-1995***	***2000-2003***
DJIA	0.80	0.80	0,23	0,49
DJTA	0.80	0.80	0,28	0,22
DJUA	0.46	0.46	0,09	0,41
DAX	0.80	0.46	0,48	0,20
FTSE100	0.46	0.46	0,54	0,58
CAC40	0.46	0.80	0,24	0,13
PX50	xxx	0.80	xxx	0,72

Příloha č. 4: Výsledky testů nezávislosti a normality

V této příloze jsou uvedeny výsledky testů nezávislosti a normality pro týdenní logaritmické výnosy měnových kurzů v období 2003 - 2005, které byly použity v jedné z aplikací.

.

Měna	Test založený na znaménkách diferencí (p-value)	Test založený na bodech zvratu (p-value)	Test omnibus (p-value)
EUR	0,213	0,750	0,903
DKK	0,333	0,484	0,875
NOK	0,213	0,226	0,313
SEK	0,489	0,750	0,605
CHF	0,128	0,143	0,816
GBP	0,678	0,656	0,957
PLN	0,678	0,899	0,307
SKK	0,890	0,524	0,210

Příloha č. 5: Zdrojový kód backward algoritmu se stop pravidlem založeným na devianci vynechané hrany

V této příloze je uveden zdrojový kód (realizovaný v systému Mathematica 4.0) selekčního backward algoritmu se stop pravidlem založeným na devianci vynechané hrany. Program je demonstrován na příkladu šesti odvětvových indexů BCPP.

```
(*   BACKWARD ALGORITMUS SE STOP
  PRAVIDLEM ZALOZENYM NA DEVIANCI VYNECHANE HRANY   *)
(*   nacteni analyzovanych dat   *)
(* nacteni dat z textoveho souboru
   Formát dat:
     1. řádek = číslo udávající počet sloupců bez prvního sloupce popisků
   2. řádek = popisky sloupců
   1. sloupec = popisky měsíců *)
stream = OpenRead@"C:êChynaêDataêCR_Odvetvi_Mesic_2001_2004.txt"D;
n = Read@stream, NumberD;(* n=pocet sloupcu indexu, tj. bez sloupce mesicu *)
nazvy = Read@stream, Table@Word, 8n + 1<DD;
data = ReadList@stream, Table@Number, 8n + 1<DD;
Close@streamD;
(* ========================================================================== *)

(*   uprava dat   *)
(* vymazani sloupce mesicu z dat *)
data2 = Transpose@Drop@Transpose@dataD, 1DD;
(* spocteni logaritmickych vynosu *)
logvynosy@cena_D := Log@Drop@cena, 1D ê Drop@cena, -1DD;
data3 = Transpose@Map@logvynosy@#D &, Transpose@data2DDD;

(* ========================================================================== *)

Needs@"Statistics`Master`"D
(* vypocet hodnoty p-value z normalniho rozdeleni N(μ,σ), μ=0,σ =1 *)
pvaluenormal@hodnota_D := (1 - CDF@NormalDistribution@0, 1D, hodnotaD)  2
pvn = pvaluenormal; (* definice alias *)
(* ------------------------------------------------------------------------- *)
(* vypocet hodnoty p-value z ch2 rozdeleni  *)
pvaluech2@stupnevolnosti_, hodnota_D :=
 1 - CDF@ChiSquareDistribution@stupnevolnostiD, hodnotaD
pvch2 = pvaluech2 (* definice alias *);
(* ========================================================================== *)

(*   testy nezavislosti   *)
(* test zalozeny na znamenkach diferenci *)
(* d = vektor diferenci po vyskrtani duplicit, k = pocet kladnych diferenci,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testznamenekdiferenci@list_D :=
  Module@8d, k, n<, d = DeleteCases@(Drop@list, 1D - Drop@list, -1D) êê N, 0.D ; k =
    Length@Select@d, (# > 0) &DD; n = Length@dD + 1; Abs@k - (n - 1) ê 2D ê Sqrt@(n + 1) ê 12DD;
tzd = testznamenekdiferenci; (* definice alias *)
(* ------------------------------------------------------------------------- *)
(* test zalozeny na bodech zvratu *)
(* d = vektor diferenci po vyskrtani duplicit,
 r = pocet bodu zvratu = pocet zapornych prvku po
    pronasobeni vektoru diferenci a vektoru diferenci posunuteho,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testboduzvratu@list_D :=
  Module@8d, k, n<, d = DeleteCases@(Drop@list, 1D - Drop@list, -1D) êê N, 0.D ;
   r = Length@Select@Drop@d, 1D  Drop@d, -1D, (# < 0) &DD;
   n = Length@dD + 1; Abs@r - 2  (n - 2) ê 3D ê Sqrt@(16  n - 29) ê 90DD;
```

```
tbz = testboduzvratu; (* definice alias *)
(* ========================================================================== *)

(*   testy normality   *)
(* test zalozeny na korigovane sikmosti *)
(* n = delka dat,
 b = koeficient,
 w2 = koeficient,
 delta = koeficient,
 a = koeficient,
 u3 = normovana sikmost *)
testsikmosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b = H3  Hn^2 + 27  n - 70L  Hn + 1L  Hn + 3LL ê HHn - 2L  Hn + 5L  Hn + 7L  Hn + 9LL;
   w2 = Sqrt@2  Hb - 1LD - 1;
   delta = 1 ê Sqrt@Log@Sqrt@w2DDD;
   a = Sqrt@2 ê Hw2 - 1LD;
   u3 = Skewness@listD ê Sqrt@6  Hn - 2L ê HHn + 1L  Hn + 3LLD;
   delta  Log@u3 ê a + Sqrt@Hu3 ê aL^2 + 1DDD;
(* ------------------------------------------------------------------------- *)
(* test zalozeny na korigovane spicatosti *)
(* n = delka dat,
 b = koeficient,
 a = koeficient,
 u4 = normovana spicatost *)
testspicatosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b =
    6  Hn^2 - 5  n + 2L ê HHn + 7L  Hn + 9LL  Sqrt@6  Hn + 3L  Hn + 5L ê Hn  Hn - 2L  Hn - 3LLD;
   a = 6 + 8 ê b  H2 ê b + Sqrt@1 + 4 ê b^2DL;
   u4 = HKurtosis@listD - H3 - 6 ê Hn + 1LLL ê
     Sqrt@24  n  Hn - 2L  Hn - 3L ê HHn + 1L^2  Hn + 3L  Hn + 5LLD;
   H1 - 2 ê H9  aL - HH1 - 2 ê aL ê H1 + u4  Sqrt@2 ê Ha - 4LDLL^H1 ê 3LL ê Sqrt@2 ê H9  aLDD;
(* ------------------------------------------------------------------------- *)
(* test omnibus *)
testomnibus@list_D := testsikmosti@listD^2 + testspicatosti@listD^2;

(* ========================================================================== *)

(*   vypis vysledku testu normality a nezavislosti   *)
pvaluetestutzd := Map@pvn@tzd@#DD &, Transpose@data3DD;
(* vektor p-value testu znamenek diferenci *)
pvaluetestutbz := Map@pvn@tbz@#DD &, Transpose@data3DD;
(* vektor p-value testu bodu zvratu *)
pvaluetestuomnibus := Map@pvch2@2, testomnibus@#DD &, Transpose@data3DD;
(* vektor p-value testu omnibus *)
vysledkytestu := Transpose@8Join@8"index", "-----"<, Drop@nazvy, 1DD,
    Join@8"diference", "-----"<, pvaluetestutzdD êê N,
    Join@8"body zvratu", "-----"<, pvaluetestutbzD êê N,
    Join@8"omnibus", "-----"<, pvaluetestuomnibusD êê N
   <D;
(* vypis hodnot p-value pro jednotlive indexy *)
Print@"Vysledky testu:   ", vysledkytestu êê TableFormD;

(* ========================================================================== *)
```

```
H   vlastni algoritmus   L
Needs@"Statistics`MultiDescriptiveStatistics`"D
vybraneindexy = 81, 2, 3, 4, 5, 7<;H konkretni vybrane indexy L

H Vypis zvolenych indexu L
Print@"Zvolene indexy: "D;
Print@
  Transpose@8Range@Length@vybraneindexyDD, Map@nazvy@@# + 1DD &, vybraneindexyD<D êê
   TableFormD;
H Pricitam +1, protoze v nazvech je jeste polozka mesic L

H data4 = konkretni log vynosy vybranych indexu L
data4 = Transpose@Map@Flatten@Take@Transpose@data3D, 8#, #<DD &, vybraneindexyDD;

kk = pocetvrcholu = Length@vybraneindexyD;
nn = pocetrealizaci = Length@Transpose@data4D@@1DDD;H delka dat = pocet realizaci L
s = kovariancnimatice =
   CovarianceMatrix@data4D.DiagonalMatrix@Table@Hnn - 1L ê nn, 8kk<DD;
korelacnimatice = CorrelationMatrix@data4D;

Print@"Pocet realizaci = delka log vynosu =", nnD
Print@"pocet vrcholu = ", kkD
Print@"variancni matice=", kovariancnimatice êê MatrixFormD
Print@"korel matice=", korelacnimatice êê MatrixFormD
H ------------------------------------------------------------------------ L

H pomocne funkce L
podmnozina@mnozina_, prvek_D :=
  Module@8vp<, vp = Map@HIntersection@prvek, #DL &, mnozinaD; MemberQ@vp, prvekDD;

neobsazena@m1_, m2_D :=
  Complement@m2, Select@m2, Hpodmnozina@m1, #DL &DD;

novygraf@graf_, hrana_D := Module@8klikys, klikybez, bez1, bez2, bezduplicit<,
   klikys = Select@graf, MemberQ@#, hrana@@1DDD && MemberQ@#, hrana@@2DDD &D;
   klikybez = Complement@graf, klikysD;
   bez1 = DeleteCases@klikys, hrana@@1DD, 2D;
   bez2 = DeleteCases@klikys, hrana@@2DD, 2D;
   bezduplicit = neobsazena@klikybez, Union@bez1, bez2DD;
   Union@klikybez, bezduplicitDD;
H ------------------------------------------------------------------------ L

H Maticový algoritmus pro odhad varianční matice L
odhadmatice@s_, clique_D := Module@8kk, knew, kold, iter, a, b<,
   kk = Length@sD;
   knew = IdentityMatrix@kkD;
kold = IdentityMatrix@kkD;
stop = 0;
iter = 0;
While@ stop < Length@Flatten@cliqueDD,
    a = clique@@Mod@iter, Length@cliqueDD + 1DD;
    iter = iter + 1;
    b = Complement@Range@kkD, aD;
    knew@@a, aDD = s@@a, aDD;
```

```
    knew@@a, bDD = s@@a, aDD.Inverse@kold@@a, aDDD.kold@@a, bDD;
    knew@@b, aDD = kold@@b, aDD.Inverse@kold@@a, aDDD.s@@a, aDD;
    knew@@b, bDD = kold@@b, bDD - kold@@b, aDD.Inverse@kold@@a, aDDD.
       HIdentityMatrix@Length@aDD - s@@a, aDD.
       Inverse@kold@@a, aDDDL.kold@@a, bDD;
    kold = knew;

      H testing the stopping rule L
    Map@HIf@Max@Abs@Flatten@knew@@clique@@#DD, clique@@#DDDD - s@@
  clique@@#DD, clique@@#DDDDDDD < 0.000001  , stop =
  stop + 1, stop = 0DL &, Range@Length@cliqueDDD;D;
knewD;
H ------------------------------------------------------------------------ L

H iteracni procedura L
clique = 8Range@kkD<;
maticesouslednosti = Array@If@H#1 > #2L, 1, 0D &, 8kk, kk<D;
krok = H-1L;
konec = False;
minminulehografu = 0;
klikaposlednihografu = clique;

While@konec == False,
 minimum = 1000;
 vektorhran = Position@maticesouslednosti, 1D;
 vektorklik = Map@Hnovygraf@klikaposlednihografu, #DL &, vektorhranD;

 For@ii = 1, ii ≤ Length@vektorklikD,
  H výpočet deviance pro jednotlivé hrany L
  v = odhadmatice@s, vektorklik@@iiDDD ;
  d =  Inverse@vD;
  deviance = pocetrealizaci HTr@s . dD - Log@Det@s . dDD -  pocetvrcholuL ;

  If@deviance < minimum, minimum = deviance; poziceminima = iiD;

  ii++D; H konec for L
 H konec výpočtu deviance pro jednotlivé hrany L

 If@minimum - minminulehografu < 3.84, minminulehografu = minimum;
        maticesouslednosti@@vektorhran@@poziceminima, 1DD,
     vektorhran@@poziceminima, 2DDDD = krok;
  klikaposlednihografu = vektorklik@@poziceminimaDD,
        konec = TrueD;


krok = krok - 1;

 If@-krok >= Binomial@kk, 2D + 1, konec = TrueD;

D H konec while L
H ========================================================================== L

H   vypis vysledku a vykresleni grafu   L
Needs@"DiscreteMath`Combinatorica`"D
Print@"matice souslednosti = ", maticesouslednosti êê MatrixFormD;
```

```
Print@"Pocet vynechanych hran = ",
  If@-krok >= Binomial@kk, 2D + 1, Binomial@kk, 2D, -krok - 2DD;
ShowLabeledGraph@g = MakeGraph@Range@kkD, HMemberQ@Position@
        maticesouslednosti + Transpose@maticesouslednostiD, 1D, 8#1, #2<DL &DD;
```

Vysledky testu:

index	diference	body zvratu	omnibus
BI04	0.457901	0.103405	0.0559108
BI07	0.0832645	0.907394	0.408593
BI08	0.804571	0.130472	0.0862969
BI12	0.457901	0.641713	0.199821
BI13	0.804571	0.816031	0.989007
BI14	0.448489	0.405122	9.26277×10^{-7}
BI15	0.216021	0.103405	0.568274
BI16	0.457901	0.816031	0.0204966

Zvolene indexy:

1 BI04
2 BI07
3 BI08
4 BI12
5 BI13
6 BI15

Pocet realizaci = delka log vynosu =48

pocet vrcholu = 6

variancni matice=

0.00339391	0.00188588	0.000994744	0.000984176	0.00017004	0.000687797
0.00188588	0.00615159	0.000632584	0.00190782	0.00244384	0.000183418
0.000994744	0.000632584	0.00176544	0.000775545	0.00125598	0.000709797
0.000984176	0.00190782	0.000775545	0.00295029	0.00268428	0.00115512
0.00017004	0.00244384	0.00125598	0.00268428	0.0131847	0.00149201
0.000687797	0.000183418	0.000709797	0.00115512	0.00149201	0.00222446

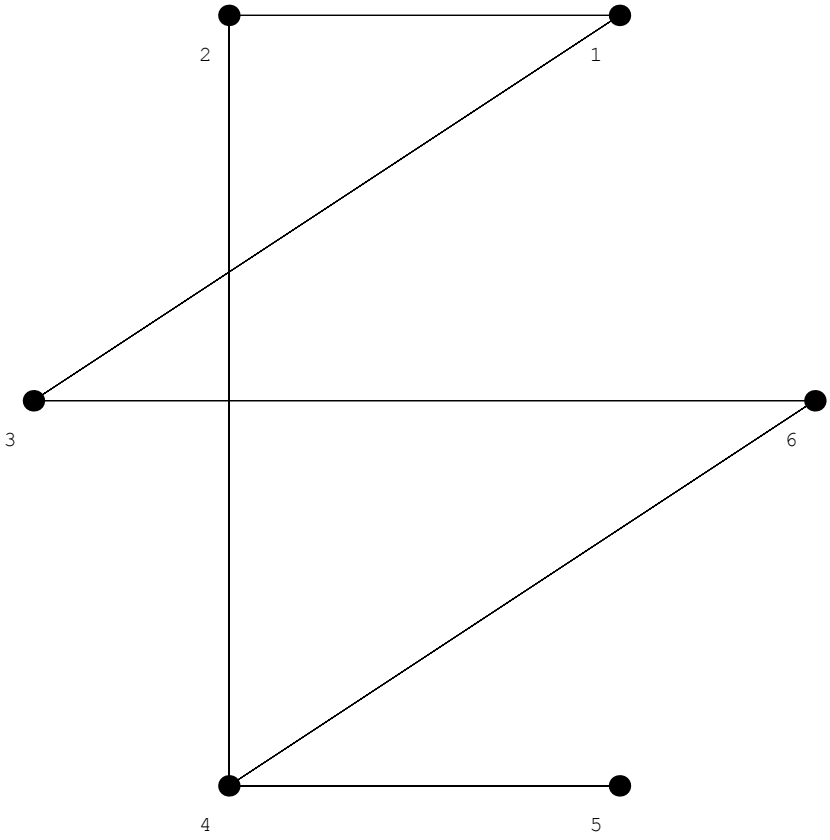
korel matice=

1.	0.412735	0.406383	0.311022	0.0254194	0.250321
0.412735	1.	0.191954	0.447828	0.271359	0.0495833
0.406383	0.191954	1.	0.33982	0.260328	0.358175
0.311022	0.447828	0.33982	1.	0.430388	0.450902
0.0254194	0.271359	0.260328	0.430388	1.	0.275501
0.250321	0.0495833	0.358175	0.450902	0.275501	1.

matice souslednosti =

0	0	0	0	0	0
1	0	0	0	0	0
1	-1	0	0	0	0
-2	1	-3	0	0	0
-7	-5	-8	1	0	0
-6	-9	1	1	-4	0

Pocet vynechanych hran = 9


2
1
3
6
4
5

Příloha č. 6: Zdrojový kód backward algoritmu se stop pravidlem založeným na celkové devianci

V této příloze je uveden zdrojový kód (realizovaný v systému Mathematica 4.0) selekčního backward algoritmu se stop pravidlem založeným na celkové devianci. Program je demonstrován na příkladu šesti odvětvových indexů BCPP.

```
(*   BACKWARD ALGORITMUS SE STOP PRAVIDLEM ZALOZENYM NA CELKOVE DEVIANCI   *)
(*   nacteni analyzovanych dat   *)
(* nacteni dat z textoveho souboru
   Formát dat:
     1. řádek = číslo udávající počet sloupců bez prvního sloupce popisků
   2. řádek = popisky sloupců
   1. sloupec = popisky měsíců *)
stream = OpenRead@"C:êChynaêDataêCR_Odvetvi_Mesic_2001_2004.txt"D;
n = Read@stream, NumberD;(* n=pocet sloupcu indexu, tj. bez sloupce mesicu *)
nazvy = Read@stream, Table@Word, 8n + 1<DD;
data = ReadList@stream, Table@Number, 8n + 1<DD;
Close@streamD;
(* ========================================================================== *)

(*   uprava dat   *)
(* vymazani sloupce mesicu z dat *)
data2 = Transpose@Drop@Transpose@dataD, 1DD;
(* spocteni logaritmickych vynosu *)
logvynosy@cena_D := Log@Drop@cena, 1D ê Drop@cena, -1DD;
data3 = Transpose@Map@logvynosy@#D &, Transpose@data2DDD;

(* ========================================================================== *)

Needs@"Statistics`Master`"D
(* vypocet hodnoty p-value z normalniho rozdeleni N(μ,σ), μ=0,σ =1 *)
pvaluenormal@hodnota_D := H1 - CDF@NormalDistribution@0, 1D, hodnotaDL  2
pvn = pvaluenormal; (* definice alias *)
(* -------------------------------------------------------------------------- *)
(* vypocet hodnoty p-value z ch2 rozdeleni  *)
pvaluech2@stupnevolnosti_, hodnota_D :=
 1 - CDF@ChiSquareDistribution@stupnevolnostiD, hodnotaD
pvch2 = pvaluech2 (* definice alias *);
(* ========================================================================== *)

(*   testy nezavislosti   *)
(* test zalozeny na znamenkach diferenci *)
(* d = vektor diferenci po vyskrtani duplicit, k = pocet kladnych diferenci,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testznamenekdiferenci@list_D :=
  Module@8d, k, n<, d = DeleteCases@HDrop@list, 1D - Drop@list, -1DL êê N, 0.D ; k =
    Length@Select@d, H# > 0L &DD; n = Length@dD + 1; Abs@k - Hn - 1L ê 2D ê Sqrt@Hn + 1L ê 12DD;
tzd = testznamenekdiferenci; (* definice alias *)
(* -------------------------------------------------------------------------- *)
(* test zalozeny na bodech zvratu *)
(* d = vektor diferenci po vyskrtani duplicit,
 r = pocet bodu zvratu = pocet zapornych prvku po
    pronasobeni vektoru diferenci a vektoru diferenci posunuteho,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testboduzvratu@list_D :=
  Module@8d, k, n<, d = DeleteCases@HDrop@list, 1D - Drop@list, -1DL êê N, 0.D ;
   r = Length@Select@Drop@d, 1D  Drop@d, -1D, H# < 0L &DD;
   n = Length@dD + 1; Abs@r - 2  Hn - 2L ê 3D ê Sqrt@H16  n - 29L ê 90DD;
tbz = testboduzvratu; (* definice alias *)
```

```
H ========================================================================== L

H   testy normality   L
H test zalozeny na korigovane sikmosti L
H n = delka dat,
 b = koeficient,
 w2 = koeficient,
 delta = koeficient,
 a = koeficient,
 u3 = normovana sikmost L
testsikmosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b = H3  Hn^2 + 27  n - 70L  Hn + 1L  Hn + 3LL ê HHn - 2L  Hn + 5L  Hn + 7L  Hn + 9LL;
   w2 = Sqrt@2  Hb - 1LD - 1;
   delta = 1 ê Sqrt@Log@Sqrt@w2DDD;
   a = Sqrt@2 ê Hw2 - 1LD;
   u3 = Skewness@listD ê Sqrt@6  Hn - 2L ê HHn + 1L  Hn + 3LLD;
   delta  Log@u3 ê a + Sqrt@Hu3 ê aL^2 + 1DDD;
H ------------------------------------------------------------------------ L
H test zalozeny na korigovane spicatosti L
H n = delka dat,
 b = koeficient,
 a = koeficient,
 u4 = normovana spicatost L
testspicatosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b =
    6  Hn^2 - 5  n + 2L ê HHn + 7L  Hn + 9LL  Sqrt@6  Hn + 3L  Hn + 5L ê Hn  Hn - 2L  Hn - 3LLD;
   a = 6 + 8 ê b  H2 ê b + Sqrt@1 + 4 ê b^2DL;
   u4 = HKurtosis@listD - H3 - 6 ê Hn + 1LLL ê
     Sqrt@24  n  Hn - 2L  Hn - 3L ê HHn + 1L^2  Hn + 3L  Hn + 5LLD;
   H1 - 2 ê H9  aL - HH1 - 2 ê aL ê H1 + u4  Sqrt@2 ê Ha - 4LDLL^H1 ê 3LL ê Sqrt@2 ê H9  aLDD;
H ------------------------------------------------------------------------ L
H test omnibus L
testomnibus@list_D := testsikmosti@listD^2 + testspicatosti@listD^2;

H ========================================================================== L

H   vypis vysledku testu normality a nezavislosti   L
pvaluetestutzd := Map@pvn@tzd@#DD &, Transpose@data3DD;
H vektor p-value testu znamenek diferenci L
pvaluetestutbz := Map@pvn@tbz@#DD &, Transpose@data3DD;
H vektor p-value testu bodu zvratu L
pvaluetestuomnibus := Map@pvch2@2, testomnibus@#DD &, Transpose@data3DD;
H vektor p-value testu omnibus L
vysledkytestu := Transpose@8Join@8"index", "-----"<, Drop@nazvy, 1DD,
    Join@8"diference", "-----"<, pvaluetestutzdD êê N,
    Join@8"body zvratu", "-----"<, pvaluetestutbzD êê N,
    Join@8"omnibus", "-----"<, pvaluetestuomnibusD êê N
   <D;
H vypis hodnot p-value pro jednotlive indexy L
Print@"Vysledky testu:   ", vysledkytestu êê TableFormD;

H ========================================================================== L
```

```
(*   vlastni algoritmus   *)
Needs@"Statistics`MultiDescriptiveStatistics`"D
vybraneindexy = 81, 2, 3, 4, 5, 7<;(* konkretni vybrane indexy *)

(* Vypis zvolenych indexu *)
Print@"Zvolene indexy: " D;
Print@
  Transpose@8Range@Length@vybraneindexyDD, Map@nazvy@@# + 1DD &, vybraneindexyD<D êê
   TableFormD;
(* Pricitam +1, protoze v nazvech je jeste polozka mesic *)

(* data4 = konkretni log vynosy vybranych indexu *)
data4 = Transpose@Map@Flatten@Take@Transpose@data3D, 8#, #<DD &, vybraneindexyDD;

kk = pocetvrcholu = Length@vybraneindexyD;
nn = pocetrealizaci = Length@Transpose@data4D@@1DDD;(* delka dat = pocet realizaci *)
s = kovariancnimatice =
   CovarianceMatrix@data4D.DiagonalMatrix@Table@Hnn - 1L ê nn, 8kk<DD;
korelacnimatice = CorrelationMatrix@data4D;

Print@"Pocet realizaci = delka log vynosu =", nnD
Print@"pocet vrcholu = ", kkD
Print@"variancni matice=", kovariancnimatice êê MatrixFormD
Print@"korel matice=", korelacnimatice êê MatrixFormD

(* ---------------------------------------------------------------------------- *)
(* pomocne funkce *)
podmnozina@mnozina_, prvek_D :=
  Module@8vp<, vp = Map@HIntersection@prvek, #DL &, mnozinaD; MemberQ@vp, prvekDD;

neobsazena@m1_, m2_D :=
  Complement@m2, Select@m2, Hpodmnozina@m1, #DL &DD;

novygraf@graf_, hrana_D := Module@8klikys, klikybez, bez1, bez2, bezduplicit<,
   klikys = Select@graf, MemberQ@#, hrana@@1DDD && MemberQ@#, hrana@@2DDD &D;
   klikybez = Complement@graf, klikysD;
   bez1 = DeleteCases@klikys, hrana@@1DD, 2D;
   bez2 = DeleteCases@klikys, hrana@@2DD, 2D;
   bezduplicit = neobsazena@klikybez, Union@bez1, bez2DD;
   Union@klikybez, bezduplicitDD;
(* ---------------------------------------------------------------------------- *)
(* Maticový algoritmus pro odhad varianční matice *)
odhadmatice@s_, clique_D := Module@8kk, knew, kold, iter, a, b<,
   kk = Length@sD;
   knew = IdentityMatrix@kkD;
kold = IdentityMatrix@kkD;
stop = 0;
iter = 0;
While@ stop < Length@Flatten@cliqueDD,
    a = clique@@Mod@iter, Length@cliqueDD + 1DD;
    iter = iter + 1;
    b = Complement@Range@kkD, aD;
    knew@@a, aDD = s@@a, aDD;
    knew@@a, bDD = s@@a, aDD.Inverse@kold@@a, aDDD.kold@@a, bDD;
    knew@@b, aDD = kold@@b, aDD.Inverse@kold@@a, aDDD.s@@a, aDD;
```

```
    knew@@b, bDD = kold@@b, bDD - kold@@b, aDD.Inverse@kold@@a, aDDD.
        HIdentityMatrix@Length@aDD - s@@a, aDD.
        Inverse@kold@@a, aDDDL.kold@@a, bDD;
    kold = knew;

      H testing the stopping rule L
    Map@HIf@Max@Abs@Flatten@knew@@clique@@#DD, clique@@#DDDD - s@@
  clique@@#DD, clique@@#DDDDDDD < 0.000001  , stop =
  stop + 1, stop = 0DL &, Range@Length@cliqueDDD;D;
knewD;
H ------------------------------------------------------------------------ L

H iteracni algoritmus L
clique = 8Range@kkD<;
maticesouslednosti = Array@If@H#1 > #2L, 1, 0D &, 8kk, kk<D;
krok = H-1L;
konec = False;
minminulehografu = 0;
klikaposlednihografu = clique;

While@konec == False,
 minimum = 1000;
 vektorhran = Position@maticesouslednosti, 1D;
 vektorklik = Map@Hnovygraf@klikaposlednihografu, #DL &, vektorhranD;
 For@ii = 1, ii ≤ Length@vektorklikD,
  H výpočet deviance pro jednotlivé hrany L
  v = odhadmatice@s, vektorklik@@iiDDD ;
   d =  Inverse@vD;
  deviance = pocetrealizaci HTr@s . dD - Log@Det@s . dDD -  pocetvrcholuL ;
  If@deviance < minimum, minimum = deviance; poziceminima = iiD;

  ii++D; H konec for L
 H konec výpočtu deviance pro jednotlivé hrany L

 If@minimum < Quantile@ChiSquareDistribution@-krokD, 0.95D,
        maticesouslednosti@@vektorhran@@poziceminima, 1DD,
     vektorhran@@poziceminima, 2DDDD = krok;
  klikaposlednihografu = vektorklik@@poziceminimaDD; minminulehografu = minimum,
        konec = TrueD;

 krok = krok - 1;

 If@-krok >= Binomial@kk, 2D + 1, konec = TrueD;

D H konec while L

H ========================================================================== L
H   vypis vysledku a vykresleni grafu   L
Needs@"DiscreteMath`Combinatorica`"D
Print@"matice souslednosti = ", maticesouslednosti êê MatrixFormD;
Print@"Pocet vynechanych hran = ",
  If@-krok >= Binomial@kk, 2D + 1, Binomial@kk, 2D, -krok - 2DD;
ShowLabeledGraph@g = MakeGraph@Range@kkD, HMemberQ@@Position@
        maticesouslednosti + Transpose@maticesouslednostiD, 1D, 8#1, #2<DL &DD;
```

Vysledky testu:

index	diference	body zvratu	omnibus
BI04	0.457901	0.103405	0.0559108
BI07	0.0832645	0.907394	0.408593
BI08	0.804571	0.130472	0.0862969
BI12	0.457901	0.641713	0.199821
BI13	0.804571	0.816031	0.989007
BI14	0.448489	0.405122	9.26277×10^{-7}
BI15	0.216021	0.103405	0.568274
BI16	0.457901	0.816031	0.0204966

Zvolene indexy:

1 BI04
2 BI07
3 BI08
4 BI12
5 BI13
6 BI15

Pocet realizaci = delka log vynosu =48

pocet vrcholu = 6

variancni matice=

0.00339391	0.00188588	0.000994744	0.000984176	0.00017004	0.000687797
0.00188588	0.00615159	0.000632584	0.00190782	0.00244384	0.000183418
0.000994744	0.000632584	0.00176544	0.000775545	0.00125598	0.000709797
0.000984176	0.00190782	0.000775545	0.00295029	0.00268428	0.00115512
0.00017004	0.00244384	0.00125598	0.00268428	0.0131847	0.00149201
0.000687797	0.000183418	0.000709797	0.00115512	0.00149201	0.00222446

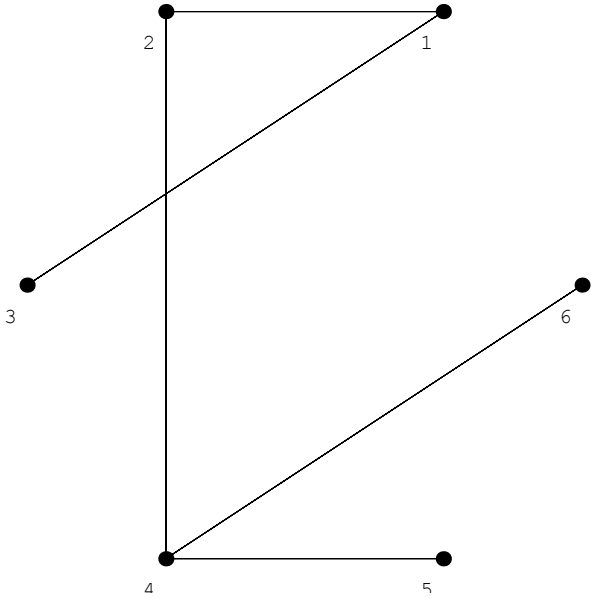
korel matice=

1.	0.412735	0.406383	0.311022	0.0254194	0.250321
0.412735	1.	0.191954	0.447828	0.271359	0.0495833
0.406383	0.191954	1.	0.33982	0.260328	0.358175
0.311022	0.447828	0.33982	1.	0.430388	0.450902
0.0254194	0.271359	0.260328	0.430388	1.	0.275501
0.250321	0.0495833	0.358175	0.450902	0.275501	1.

matice souslednosti =

0	0	0	0	0	0
1	0	0	0	0	0
1	−1	0	0	0	0
−2	1	−3	0	0	0
−7	−5	−8	1	0	0
−6	−9	−10	1	−4	0

Pocet vynechanych hran = 10


2
1
3
6
4
5

Příloha č. 7: Zdrojový kód forward algoritmu se strop pravidlem založeným na devianci přidané hrany

V této příloze je uveden zdrojový kód (realizovaný v systému Mathematica 4.0) selekčního forward algoritmu se stop pravidlem založeným na devianci přidané hrany. Program je demonstrován na příkladu šesti odvětvových indexů BCPP.

```
(*   FORWARD ALGORITMUS SE STOP PRAVIDLEM ZALOZENYM NA DEVIANCI PRIDANE HRANY   *)
(*   nacteni analyzovanych dat   *)
(* nacteni dat z textoveho souboru
   Formát dat:
     1. řádek = číslo udávající počet sloupců bez prvního sloupce popisků
   2. řádek = popisky sloupců
   1. sloupec = popisky měsíců *)
stream = OpenRead@"C:êChynaêDataêCR_Odvetvi_Mesic_2001_2004.txt"D;
n = Read@stream, NumberD;(* n=pocet sloupcu indexu, tj. bez sloupce mesicu *)
nazvy = Read@stream, Table@Word, 8n + 1<DD;
data = ReadList@stream, Table@Number, 8n + 1<DD;
Close@streamD;
(* ========================================================================== *)

(*   uprava dat   *)
(* vymazani sloupce mesicu z dat *)
data2 = Transpose@Drop@Transpose@dataD, 1DD;
(* spocteni logaritmickych vynosu *)
logvynosy@cena_D := Log@Drop@cena, 1D ê Drop@cena, -1DD;
data3 = Transpose@Map@logvynosy@#D &, Transpose@data2DDD;

(* ========================================================================== *)

Needs@"Statistics`Master`"D
(* vypocet hodnoty p-value z normalniho rozdeleni N(μ,σ), μ=0,σ =1 *)
pvaluenormal@hodnota_D := H1 - CDF@NormalDistribution@0, 1D, hodnotaDL  2
pvn = pvaluenormal; (* definice alias *)
(* ------------------------------------------------------------------------- *)
(* vypocet hodnoty p-value z ch2 rozdeleni  *)
pvaluech2@stupnevolnosti_, hodnota_D :=
 1 - CDF@ChiSquareDistribution@stupnevolnostiD, hodnotaD
pvch2 = pvaluech2 (* definice alias *);
(* ========================================================================== *)

(*   testy nezavislosti   *)
(* test zalozeny na znamenkach diferenci *)
(* d = vektor diferenci po vyskrtani duplicit, k = pocet kladnych diferenci,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testznamenekdiferenci@list_D :=
  Module@8d, k, n<, d = DeleteCases@HDrop@list, 1D - Drop@list, -1DL êê N, 0.D ; k =
    Length@Select@d, H# > 0L &DD; n = Length@dD + 1; Abs@k - Hn - 1L ê 2D ê Sqrt@Hn + 1L ê 12DD;
tzd = testznamenekdiferenci; (* definice alias *)
(* ------------------------------------------------------------------------- *)
(* test zalozeny na bodech zvratu *)
(* d = vektor diferenci po vyskrtani duplicit,
 r = pocet bodu zvratu = pocet zapornych prvku po
    pronasobeni vektoru diferenci a vektoru diferenci posunuteho,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testboduzvratu@list_D :=
  Module@8d, k, n<, d = DeleteCases@HDrop@list, 1D - Drop@list, -1DL êê N, 0.D ;
   r = Length@Select@Drop@d, 1D  Drop@d, -1D, H# < 0L &DD;
   n = Length@dD + 1; Abs@r - 2  Hn - 2L ê 3D ê Sqrt@H16  n - 29L ê 90DD;
tbz = testboduzvratu; (* definice alias *)
```

```
H ============================================================================== L

H   testy normality   L
H test zalozeny na korigovane sikmosti L
H n = delka dat,
 b = koeficient,
 w2 = koeficient,
 delta = koeficient,
 a = koeficient,
 u3 = normovana sikmost L
testsikmosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b = H3  Hn^2 + 27  n - 70L  Hn + 1L  Hn + 3LL ê HHn - 2L  Hn + 5L  Hn + 7L  Hn + 9LL;
   w2 = Sqrt@2  Hb - 1LD - 1;
   delta = 1 ê Sqrt@Log@Sqrt@w2DDD;
   a = Sqrt@2 ê Hw2 - 1LD;
   u3 = Skewness@listD ê Sqrt@6  Hn - 2L ê HHn + 1L  Hn + 3LLD;
   delta  Log@u3 ê a + Sqrt@Hu3 ê aL^2 + 1DDD;
H ------------------------------------------------------------------------------ L
H test zalozeny na korigovane spicatosti L
H n = delka dat,
 b = koeficient,
 a = koeficient,
 u4 = normovana spicatost L
testspicatosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b =
    6  Hn^2 - 5  n + 2L ê HHn + 7L  Hn + 9LL  Sqrt@6  Hn + 3L  Hn + 5L ê Hn  Hn - 2L  Hn - 3LLD;
   a = 6 + 8 ê b  H2 ê b + Sqrt@1 + 4 ê b^2DL;
   u4 = HKurtosis@listD - H3 - 6 ê Hn + 1LLL ê
     Sqrt@24  n  Hn - 2L  Hn - 3L ê HHn + 1L^2  Hn + 3L  Hn + 5LLD;
   H1 - 2 ê H9  aL - HH1 - 2 ê aL ê H1 + u4  Sqrt@2 ê Ha - 4LDLL^H1 ê 3LL ê Sqrt@2 ê H9  aLDD;
H ------------------------------------------------------------------------------ L
H test omnibus L
testomnibus@list_D := testsikmosti@listD^2 + testspicatosti@listD^2;

H ============================================================================== L

H   vypis vysledku testu normality a nezavislosti   L
pvaluetestutzd := Map@pvn@tzd@#DD &, Transpose@data3DD;
H vektor p-value testu znamenek diferenci L
pvaluetestutbz := Map@pvn@tbz@#DD &, Transpose@data3DD;
H vektor p-value testu bodu zvratu L
pvaluetestuomnibus := Map@pvch2@2, testomnibus@#DD &, Transpose@data3DD;
H vektor p-value testu omnibus L
vysledkytestu := Transpose@8Join@8"index", "-----"<, Drop@nazvy, 1DD,
    Join@8"diference", "-----"<, pvaluetestutzdD êê N,
    Join@8"body zvratu", "-----"<, pvaluetestutbzD êê N,
    Join@8"omnibus", "-----"<, pvaluetestuomnibusD êê N
   <D;
H vypis hodnot p-value pro jednotlive indexy L
Print@"Vysledky testu:   ", vysledkytestu êê TableFormD;

H ============================================================================== L
```

```
H   vlastni algoritmus   L
Needs@"Statistics`MultiDescriptiveStatistics`"D
vybraneindexy = 81, 2, 3, 4, 5, 7<;H konkretni vybrane indexy L
H Vypis zvolenych indexu L
Print@"Zvolene indexy: "D;
Print@
  Transpose@8Range@Length@vybraneindexyDD, Map@nazvy@@# + 1DD &, vybraneindexyD<D êê
   TableFormD;
H Pricitam +1, protoze v nazvech je jeste polozka mesic L

H data4 = konkretni log vynosy vybranych indexu L
data4 = Transpose@Map@Flatten@Take@Transpose@data3D, 8#, #<DD &, vybraneindexyDD;

kk = pocetvrcholu = Length@vybraneindexyD;
nn = pocetrealizaci = Length@Transpose@data4D@@1DDD;H delka dat = pocet realizaci L
s = kovariancnimatice =
   CovarianceMatrix@data4D.DiagonalMatrix@Table@Hnn - 1L ê nn, 8kk<DD;
korelacnimatice = CorrelationMatrix@data4D;

Print@"Pocet realizaci = delka log vynosu =", nnD
Print@"pocet vrcholu = ", kkD
Print@"variancni matice=", kovariancnimatice êê MatrixFormD
Print@"korel matice=", korelacnimatice êê MatrixFormD
H ------------------------------------------------------------------------- L

H pomocne funkce L
podmnozina@mnozina_, prvek_D :=
  Module@8vp<, vp = Map@HIntersection@prvek, #DL &, mnozinaD; MemberQ@vp, prvekDD;

neobsazena@m1_, m2_D :=
  Complement@m2, Select@m2, Hpodmnozina@m1, #DL &DD;

novygraf@graf_, hrana_D := Module@8klikys, klikybez, bez1, bez2, bezduplicit<,
   klikys = Select@graf, MemberQ@#, hrana@@1DDD && MemberQ@#, hrana@@2DDD &D;
   klikybez = Complement@graf, klikysD;
   bez1 = DeleteCases@klikys, hrana@@1DD, 2D;
   bez2 = DeleteCases@klikys, hrana@@2DD, 2D;
   bezduplicit = neobsazena@klikybez, Union@bez1, bez2DD;
   Union@klikybez, bezduplicitDD;
H ------------------------------------------------------------------------- L

H Maticový algoritmus pro odhad varianční matice L
odhadmatice@s_, clique_D := Module@8kk, knew, kold, iter, a, b<,
   kk = Length@sD;
   knew = IdentityMatrix@kkD;
kold = IdentityMatrix@kkD;
stop = 0;
iter = 0;
While@ stop < Length@Flatten@cliqueDD,
    a = clique@@Mod@iter, Length@cliqueDD + 1DD;
    iter = iter + 1;
    b = Complement@Range@kkD, aD;
    knew@@a, aDD = s@@a, aDD;
    knew@@a, bDD = s@@a, aDD.Inverse@kold@@a, aDDD.kold@@a, bDD;
    knew@@b, aDD = kold@@b, aDD.Inverse@kold@@a, aDDD.s@@a, aDD;
```

```
    knew@@b, bDD = kold@@b, bDD - kold@@b, aDD.Inverse@kold@@a, aDDD.
        HIdentityMatrix@Length@aDD - s@@a, aDD.
        Inverse@kold@@a, aDDDL.kold@@a, bDD;
    kold = knew;

      H testing the stopping rule L
    Map@HIf@Max@Abs@Flatten@knew@@clique@@#DD, clique@@#DDDD - s@@
  clique@@#DD, clique@@#DDDDDDD < 0.000001 , stop =
  stop + 1, stop = 0DL &, Range@Length@cliqueDDD;D;
knewD;
H ------------------------------------------------------------------------ L

H První výpočet deviance pro graf bez hran L
maticesousednosti = Table@0, 8kk<, 8kk<D;
v = DiagonalMatrix@Map@Hs@@#, #DDL &, Range@Length@sDDDD;
d = DiagonalMatrix@1 ê Map@Hs@@#, #DDL &, Range@Length@sDDDD;
deviance = pocetrealizaci HTr@s.dD - Log@Det@s.dDD - pocetvrcholuL;
H ------------------------------------------------------------------------ L

H hrany, které chybí v aktuální matici sousednosti L
vektorhran = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;
vektorhranpropridani = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;
H přidáme do vektoru hran 1 hranu tak,
 že vezmene vektor vynechaných hran a uděláme doplněk s jednou hranou L
vektorhran = Map@HComplement@vektorhran, 8vektorhran@@#DD<DL &,
   Range@Length@vektorhranDDD;
H vygenerování vektoru klik pro všechny přidané hrany L
vektorklik = 8<;
For@jj = 1, jj ≤ Length@vektorhranD,
  clique = 8Range@kkD<;
  odstranithrany = vektorhran@@jjDD;
  Map@Hclique = novygraf@clique, odstranithrany@@#DDDL &,
   Range@Length@odstranithranyDDD;
  vektorklik = Append@vektorklik, cliqueD;
  jj++D;


minminulehografu = deviance;
krok = 1;
konec = False;

H ------------------------------------------------------------------------ L
H iteracni procedura L
While@konec == False,
 minimum = 1000;

 For@ii = 1, ii ≤ Length@vektorklikD,
  v = odhadmatice@s, vektorklik@@iiDDD ;
  d = Inverse@vD;
  deviance = pocetrealizaci HTr@s . dD - Log@Det@s . dDD - pocetvrcholuL ;


   If@deviance < minimum, minimum = deviance; poziceminima = iiD;

  ii++D; H konec for L
```

```
 If@minminulehografu - minimum > 3.84, Hminminulehografu = minimum;
        maticesousednosti@@vektorhranpropridani@@poziceminima, 1DD,
      vektorhranpropridani@@poziceminima, 2DDDD = krokL,
        konec = TrueD;

 H hrany, které chybí v aktuální matici sousednosti L
 vektorhran = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;

 vektorhranpropridani = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;
 H přidáme do vektoru hran 1 hranu tak,
  že vezmene vektor vynechaných hran a uděláme doplněk s jednou hranou L
 vektorhran = Map@HComplement@vektorhran, 8vektorhran@@#DD<DL &,
   Range@Length@vektorhranDDD;

 H vygenerování vektoru klik pro všechny přidané hrany L
 vektorklik = 8<;
 For@jj = 1, jj ≤ Length@vektorhranD,
  clique = 8Range@kkD<;
  odstranithrany = vektorhran@@jjDD;
  Map@Hclique = novygraf@clique, odstranithrany@@#DDDL &,
   Range@Length@odstranithranyDDD;

  vektorklik = Append@vektorklik, cliqueD;
  jj++D;

 krok = krok + 1;

 If@krok >= Binomial@kk, 2D + 1, konec = TrueD;

D H konec while L
H ============================================================================== L

H   vypis vysledku a vykresleni grafu   L
Needs@"DiscreteMath`Combinatorica`"D
H vypocet hodnoty p-value z ch2 rozdeleni  L
pvaluech2@stupnevolnosti_, hodnota_D :=
 1 - CDF@ChiSquareDistribution@stupnevolnostiD, hodnotaD
pvch2 = pvaluech2 H definice alias L;


Print@"Pocet pridanych hran = ",
  If@krok >= Binomial@kk, 2D + 1, Binomial@kk, 2D, krok - 2DD;
Print@"matice sousednosti: ", maticesousednosti êê MatrixFormD;
ShowLabeledGraph@g = MakeGraph@Range@kkD, HNot@MemberQ@Position@
         maticesousednosti + Transpose@maticesousednostiD, 0D, 8#1, #2<DDL &DD;
```

Vysledky testu:

index	diference	body zvratu	omnibus
BI04	0.457901	0.103405	0.0559108
BI07	0.0832645	0.907394	0.408593
BI08	0.804571	0.130472	0.0862969
BI12	0.457901	0.641713	0.199821
BI13	0.804571	0.816031	0.989007
BI14	0.448489	0.405122	9.26277×10^{-7}
BI15	0.216021	0.103405	0.568274
BI16	0.457901	0.816031	0.0204966

Zvolene indexy:

1 BI04
2 BI07
3 BI08
4 BI12
5 BI13
6 BI15

Pocet realizaci = delka log vynosu =48

pocet vrcholu = 6

variancni matice=

0.00339391	0.00188588	0.000994744	0.000984176	0.00017004	0.000687797
0.00188588	0.00615159	0.000632584	0.00190782	0.00244384	0.000183418
0.000994744	0.000632584	0.00176544	0.000775545	0.00125598	0.000709797
0.000984176	0.00190782	0.000775545	0.00295029	0.00268428	0.00115512
0.00017004	0.00244384	0.00125598	0.00268428	0.0131847	0.00149201
0.000687797	0.000183418	0.000709797	0.00115512	0.00149201	0.00222446

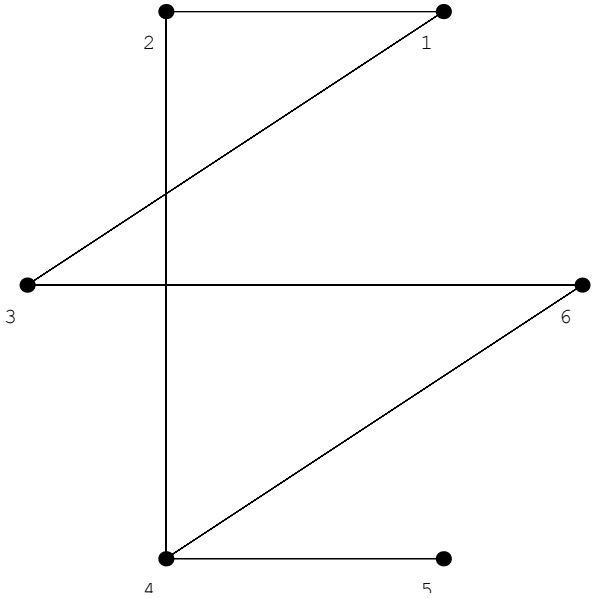
korel matice=

1.	0.412735	0.406383	0.311022	0.0254194	0.250321
0.412735	1.	0.191954	0.447828	0.271359	0.0495833
0.406383	0.191954	1.	0.33982	0.260328	0.358175
0.311022	0.447828	0.33982	1.	0.430388	0.450902
0.0254194	0.271359	0.260328	0.430388	1.	0.275501
0.250321	0.0495833	0.358175	0.450902	0.275501	1.

Pocet pridanych hran = 6

matice sousednosti:

0	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
0	2	0	0	0	0
0	0	0	3	0	0
0	0	6	1	0	0


2
1
3
6
4
5

Příloha č. 8: Zdrojový kód forward algoritmu se strop pravidlem založeným na celkové devianci

V této příloze je uveden zdrojový kód (realizovaný v systému Mathematica 4.0) selekčního forward algoritmu se stop pravidlem založeným na celkové devianci. Program je demonstrován na příkladu šesti odvětvových indexů BCPP.

```
(*   FORWARD ALGORITMUS SE STOP PRAVIDLEM ZALOZENYM NA CELKOVE DEVIANCI   *)
(*   nacteni analyzovanych dat   *)
(* nacteni dat z textoveho souboru
   Formát dat:
     1. řádek = číslo udávající počet sloupců bez prvního sloupce popisků
   2. řádek = popisky sloupců
   1. sloupec = popisky měsíců *)
stream = OpenRead@"C:êChynaêDataêCR_Odvetvi_Mesic_2001_2004.txt"D;
n = Read@stream, NumberD;(* n=pocet sloupcu indexu, tj. bez sloupce mesicu *)
nazvy = Read@stream, Table@Word, 8n + 1<DD;
data = ReadList@stream, Table@Number, 8n + 1<DD;
Close@streamD;
(* ========================================================================== *)

(*   uprava dat   *)
(* vymazani sloupce mesicu z dat *)
data2 = Transpose@Drop@Transpose@dataD, 1DD;
(* spocteni logaritmickych vynosu *)
logvynosy@cena_D := Log@Drop@cena, 1D ê Drop@cena, -1DD;
data3 = Transpose@Map@logvynosy@#D &, Transpose@data2DDD;

(* ========================================================================== *)

Needs@"Statistics`Master`"D
(* vypocet hodnoty p-value z normalniho rozdeleni N(μ,σ), μ=0,σ =1 *)
pvaluenormal@hodnota_D := H1 - CDF@NormalDistribution@0, 1D, hodnotaDL  2
pvn = pvaluenormal; (* definice alias *)
(* ------------------------------------------------------------------------- *)
(* vypocet hodnoty p-value z ch2 rozdeleni  *)
pvaluech2@stupnevolnosti_, hodnota_D :=
 1 - CDF@ChiSquareDistribution@stupnevolnostiD, hodnotaD
pvch2 = pvaluech2 (* definice alias *);
(* ========================================================================== *)

(*   testy nezavislosti   *)
(* test zalozeny na znamenkach diferenci *)
(* d = vektor diferenci po vyskrtani duplicit, k = pocet kladnych diferenci,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testznamenekdiferenci@list_D :=
  Module@8d, k, n<, d = DeleteCases@HDrop@list, 1D - Drop@list, -1DL êê N, 0.D ; k =
    Length@Select@d, H# > 0L &DD; n = Length@dD + 1; Abs@k - Hn - 1L ê 2D ê Sqrt@Hn + 1L ê 12DD;
tzd = testznamenekdiferenci; (* definice alias *)
(* ------------------------------------------------------------------------- *)
(* test zalozeny na bodech zvratu *)
(* d = vektor diferenci po vyskrtani duplicit,
 r = pocet bodu zvratu = pocet zapornych prvku po
    pronasobeni vektoru diferenci a vektoru diferenci posunuteho,
 n = pocet clenu rady po vyskrtani duplicit = delka vektoru diferenci + 1 *)
testboduzvratu@list_D :=
  Module@8d, k, n<, d = DeleteCases@HDrop@list, 1D - Drop@list, -1DL êê N, 0.D ;
   r = Length@Select@Drop@d, 1D  Drop@d, -1D, H# < 0L &DD;
   n = Length@dD + 1; Abs@r - 2  Hn - 2L ê 3D ê Sqrt@H16  n - 29L ê 90DD;
tbz = testboduzvratu; (* definice alias *)
```

```
(* ========================================================================== *)

(*   testy normality   *)
(* test zalozeny na korigovane sikmosti *)
(* n = delka dat,
 b = koeficient,
 w2 = koeficient,
 delta = koeficient,
 a = koeficient,
 u3 = normovana sikmost *)
testsikmosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b = H3  Hn^2 + 27  n - 70L  Hn + 1L  Hn + 3LL ê HHn - 2L  Hn + 5L  Hn + 7L  Hn + 9LL;
   w2 = Sqrt@2  Hb - 1LD - 1;
   delta = 1 ê Sqrt@Log@Sqrt@w2DDD;
   a = Sqrt@2 ê Hw2 - 1LD;
   u3 = Skewness@listD ê Sqrt@6  Hn - 2L ê HHn + 1L  Hn + 3LLD;
   delta  Log@u3 ê a + Sqrt@Hu3 ê aL^2 + 1DDD;
(* ---------------------------------------------------------------------------- *)
(* test zalozeny na korigovane spicatosti *)
(* n = delka dat,
 b = koeficient,
 a = koeficient,
 u4 = normovana spicatost *)
testspicatosti@list_D := Module@8n, b, w, delta, a, u3<,
   n = Length@listD;
   b =
    6  Hn^2 - 5  n + 2L ê HHn + 7L  Hn + 9LL  Sqrt@6  Hn + 3L  Hn + 5L ê Hn  Hn - 2L  Hn - 3LLD;
   a = 6 + 8 ê b  H2 ê b + Sqrt@1 + 4 ê b^2DL;
   u4 = HKurtosis@listD - H3 - 6 ê Hn + 1LLL ê
     Sqrt@24  n  Hn - 2L  Hn - 3L ê HHn + 1L^2  Hn + 3L  Hn + 5LLD;
   H1 - 2 ê H9  aL - HH1 - 2 ê aL ê H1 + u4  Sqrt@2 ê Ha - 4LDLL^H1 ê 3LL ê Sqrt@2 ê H9  aLDD;
(* ---------------------------------------------------------------------------- *)
(* test omnibus *)
testomnibus@list_D := testsikmosti@listD^2 + testspicatosti@listD^2;

(* ========================================================================== *)

(*   vypis vysledku testu normality a nezavislosti   *)
pvaluetestutzd := Map@pvn@tzd@#DD &, Transpose@data3DD;
(* vektor p-value testu znamenek diferenci *)
pvaluetestutbz := Map@pvn@tbz@#DD &, Transpose@data3DD;
(* vektor p-value testu bodu zvratu *)
pvaluetestuomnibus := Map@pvch2@2, testomnibus@#DD &, Transpose@data3DD;
(* vektor p-value testu omnibus *)
vysledkytestu := Transpose@8Join@8"index", "-----"<, Drop@nazvy, 1DD,
    Join@8"diference", "-----"<, pvaluetestutzdD êê N,
    Join@8"body zvratu", "-----"<, pvaluetestutbzD êê N,
    Join@8"omnibus", "-----"<, pvaluetestuomnibusD êê N
   <D;
(* vypis hodnot p-value pro jednotlive indexy *)
Print@"Vysledky testu:   ", vysledkytestu êê TableFormD;

(* ========================================================================== *)
```

```
H   vlastni algoritmus   L
Needs@"Statistics`MultiDescriptiveStatistics`"D
vybraneindexy = 81, 2, 3, 4, 5, 7<;H konkretni vybrane indexy L

H Vypis zvolenych indexu L
Print@"Zvolene indexy: " D;
Print@
  Transpose@8Range@Length@vybraneindexyDD, Map@nazvy@@# + 1DD &, vybraneindexyD<D êê
   TableFormD;
H Pricitam +1, protoze v nazvech je jeste polozka mesic L

H data4 = konkretni log vynosy vybranych indexu L
data4 = Transpose@Map@Flatten@Take@Transpose@data3D, 8#, #<DD &, vybraneindexyDD;

kk = pocetvrcholu = Length@vybraneindexyD;
nn = pocetrealizaci = Length@Transpose@data4D@@1DDD;H delka dat = pocet realizaci L
s = kovariancnimatice =
   CovarianceMatrix@data4D.DiagonalMatrix@Table@Hnn - 1L ê nn, 8kk<DD;
korelacnimatice = CorrelationMatrix@data4D;

Print@"Pocet realizaci = delka log vynosu =", nnD
Print@"pocet vrcholu = ", kkD
Print@"variancni matice=", kovariancnimatice êê MatrixFormD
Print@"korel matice=", korelacnimatice êê MatrixFormD
H ------------------------------------------------------------------------------ L

H pomocne funkce L
podmnozina@mnozina_, prvek_D :=
  Module@8vp<, vp = Map@HIntersection@prvek, #DL &, mnozinaD; MemberQ@vp, prvekDD;

neobsazena@m1_, m2_D :=
  Complement@m2, Select@m2, Hpodmnozina@m1, #DL &DD;

novygraf@graf_, hrana_D := Module@8klikys, klikybez, bez1, bez2, bezduplicit<,
   klikys = Select@graf, MemberQ@#, hrana@@1DDD && MemberQ@#, hrana@@2DDD &D;
   klikybez = Complement@graf, klikysD;
   bez1 = DeleteCases@klikys, hrana@@1DD, 2D;
   bez2 = DeleteCases@klikys, hrana@@2DD, 2D;
   bezduplicit = neobsazena@klikybez, Union@bez1, bez2DD;
   Union@klikybez, bezduplicitDD;
H ------------------------------------------------------------------------------ L

H Maticový algoritmus pro odhad varianční matice L
odhadmatice@s_, clique_D := Module@8kk, knew, kold, iter, a, b<,
   kk = Length@sD;
   knew = IdentityMatrix@kkD;
kold = IdentityMatrix@kkD;
stop = 0;
iter = 0;
While@ stop < Length@Flatten@cliqueDD,
    a = clique@@Mod@iter, Length@cliqueDD + 1DD;
    iter = iter + 1;
    b = Complement@Range@kkD, aD;
    knew@@a, aDD = s@@a, aDD;
    knew@@a, bDD = s@@a, aDD.Inverse@kold@@a, aDDD.kold@@a, bDD;
```

```
    knew@@b, aDD = kold@@b, aDD.Inverse@kold@@a, aDDD.s@@a, aDD;
    knew@@b, bDD = kold@@b, bDD - kold@@b, aDD.Inverse@kold@@a, aDDD.
       HIdentityMatrix@Length@aDD - s@@a, aDD.
       Inverse@kold@@a, aDDDL.kold@@a, bDD;
    kold = knew;

      H testing the stopping rule L
    Map@HIf@Max@Abs@Flatten@knew@@clique@@#DD, clique@@#DDDD - s@@
  clique@@#DD, clique@@#DDDDDDD < 0.000001  , stop =
  stop + 1, stop = 0DL &, Range@Length@cliqueDDD;D;
knewD;
H ------------------------------------------------------------------------ L

H První výpočet deviance pro graf bez hran L
maticesousednosti = Table@0, 8kk<, 8kk<D;
v = DiagonalMatrix@Map@Hs@@#, #DDL &, Range@Length@sDDDD;
d = DiagonalMatrix@1 ê Map@Hs@@#, #DDL &, Range@Length@sDDDD;
deviance = pocetrealizaci HTr@s.dD - Log@Det@s.dDD - pocetvrcholuL;
H ------------------------------------------------------------------------ L

H hrany, které chybí v aktuální matici sousednosti L
vektorhran = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;
vektorhranpropridani = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;
H přidáme do vektoru hran 1 hranu tak,
 že vezmene vektor vynechaných hran a uděláme doplněk s jednou hranou L
vektorhran = Map@HComplement@vektorhran, 8vektorhran@@#DD<DL &,
   Range@Length@vektorhranDDD;
H vygenerování vektoru klik pro všechny přidané hrany L
vektorklik = 8<;
For@jj = 1, jj ≤ Length@vektorhranD,
  clique = 8Range@kkD<;
  odstranithrany = vektorhran@@jjDD;
  Map@Hclique = novygraf@clique, odstranithrany@@#DDDL &,
   Range@Length@odstranithranyDDD;
  vektorklik = Append@vektorklik, cliqueD;
  jj++D;

minminulehografu = deviance;
krok = 1;
konec = False;
H ------------------------------------------------------------------------ L

H iteracni procedura L
While@konec == False,
 minimum = 1000;

 For@ii = 1, ii ≤ Length@vektorklikD,
  v = odhadmatice@s, vektorklik@@iiDDD ;
  d =  Inverse@vD;
  deviance = pocetrealizaci HTr@s . dD - Log@Det@s . dDD -  pocetvrcholuL ;

   If@deviance < minimum, minimum = deviance; poziceminima = iiD;

  ii++D; H konec for L
```

```
 If@minimum > Quantile@ChiSquareDistribution@krokD, 0.95D,
  Hminminulehografu = minimum;
        maticesousednosti@@vektorhranpropridani@@poziceminima, 1DD,
      vektorhranpropridani@@poziceminima, 2DDDD = krokL,
        konec = TrueD;

 H hrany, které chybí v aktuální matici sousednosti L
 vektorhran = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;

 vektorhranpropridani = Select@Position@maticesousednosti, 0D, H#@@1DD > #@@2DDL &D;
 H přidáme do vektoru hran 1 hranu tak,
  že vezmene vektor vynechaných hran a uděláme doplněk s jednou hranou L
 vektorhran = Map@HComplement@vektorhran, 8vektorhran@@#DD<DL &,
   Range@Length@vektorhranDDD;

 H vygenerování vektoru klik pro všechny přidané hrany L
 vektorklik = 8<;
 For@jj = 1, jj ≤ Length@vektorhranD,
  clique = 8Range@kkD<;
  odstranithrany = vektorhran@@jjDD;
  Map@Hclique = novygraf@clique, odstranithrany@@#DDDL &,
   Range@Length@odstranithranyDDD;

  vektorklik = Append@vektorklik, cliqueD;
  jj++D;

 krok = krok + 1;

 If@krok >= Binomial@kk, 2D + 1, konec = TrueD;

D H konec while L
H ============================================================================== L

H   vypis vysledku a vykresleni grafu   L
Needs@"DiscreteMath`Combinatorica`"D
H vypocet hodnoty p-value z ch2 rozdeleni  L
pvaluech2@stupnevolnosti_, hodnota_D :=
 1 - CDF@ChiSquareDistribution@stupnevolnostiD, hodnotaD
pvch2 = pvaluech2 H definice alias L;

Print@"Pocet pridanych hran = ",
  If@krok >= Binomial@kk, 2D + 1, Binomial@kk, 2D, krok - 2DD;
Print@"matice sousednosti: ", maticesousednosti êê MatrixFormD;
ShowLabeledGraph@g = MakeGraph@Range@kkD, HNot@MemberQ@Position@
         maticesousednosti + Transpose@maticesousednostiD, 0D, 8#1, #2<DDL &DD;
```

Vysledky testu:

index	diference	body zvratu	omnibus
BI04	0.457901	0.103405	0.0559108
BI07	0.0832645	0.907394	0.408593
BI08	0.804571	0.130472	0.0862969
BI12	0.457901	0.641713	0.199821
BI13	0.804571	0.816031	0.989007
BI14	0.448489	0.405122	9.26277×10^{-7}
BI15	0.216021	0.103405	0.568274
BI16	0.457901	0.816031	0.0204966

Zvolene indexy:

1 BI04
2 BI07
3 BI08
4 BI12
5 BI13
6 BI15

Pocet realizaci = delka log vynosu =48

pocet vrcholu = 6

variancni matice=

0.00339391	0.00188588	0.000994744	0.000984176	0.00017004	0.000687797
0.00188588	0.00615159	0.000632584	0.00190782	0.00244384	0.000183418
0.000994744	0.000632584	0.00176544	0.000775545	0.00125598	0.000709797
0.000984176	0.00190782	0.000775545	0.00295029	0.00268428	0.00115512
0.00017004	0.00244384	0.00125598	0.00268428	0.0131847	0.00149201
0.000687797	0.000183418	0.000709797	0.00115512	0.00149201	0.00222446

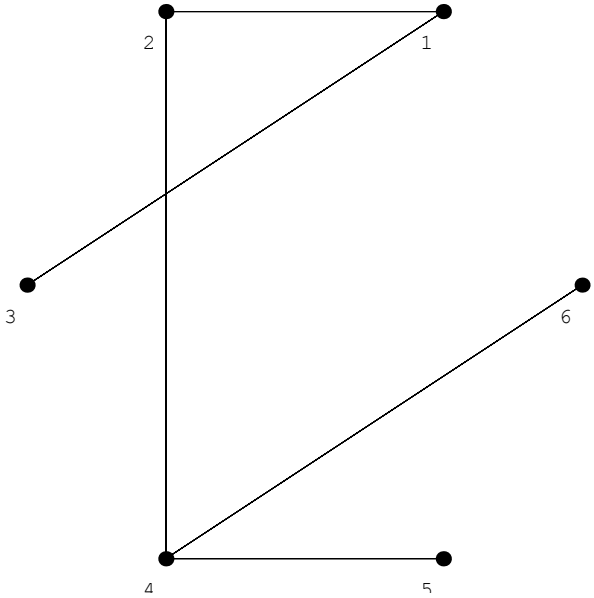
korel matice=

1.	0.412735	0.406383	0.311022	0.0254194	0.250321
0.412735	1.	0.191954	0.447828	0.271359	0.0495833
0.406383	0.191954	1.	0.33982	0.260328	0.358175
0.311022	0.447828	0.33982	1.	0.430388	0.450902
0.0254194	0.271359	0.260328	0.430388	1.	0.275501
0.250321	0.0495833	0.358175	0.450902	0.275501	1.

Pocet pridanych hran = 5

matice sousednosti:

0	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
0	2	0	0	0	0
0	0	0	3	0	0
0	0	0	1	0	0


2
1
3
6
4
5