FORUM STATISTICUM SLOVACUM Vol XVIII (2022) No 2 pp 1-12

Bayesovské modelování dat v prostředí programového systému SAS
Bayesian data modelling in the SAS software system environment

Roman Pavelka
Štatistický úrad Slovenskej republiky, Odbor metód štatistických zisťovaní, Lamačská cesta 3, 840 05 Bratislava, Slovenská republika
Statistical Office of the Slovak Republic, Statistical Surveys and Methodology Department, Lamačská cesta 3, 840 05 Bratislava, Slovak Republic
roman.pavelka@statistics.sk

Abstrakt: *Bayesovské metody se stávají stále populárnějšími v moderní statistické analýze a jsou aplikovány na širokou škálu spektra vědeckých oborů a oblastí výzkumu. Předkládaný příspěvek představuje principy postupů procedur v modulu STAT programového sytému SAS. Je zde uvedena procedura MCMC, která je určena pro obecně použitelná pro bayesovské výpočty. Procedura MCMC poskytuje výzkumníkům provádění analýzy dat na široké škále komplexních bayesovských statistických modelů. Procedura využívá Markovský řetězec Monte Carlo (MCMC) algoritmus pro náhodné výběry vzorků z libovolného aposteriorního rozdělení, které je definováno apriorními rozděleními parametrů a pravděpodobnostními funkcemi pro sledovaná zadaná data. Tento článek popisuje, jak používat proceduru MCMC v postupech pro odhad, inferenci a predikci a okrajově se zmiňuje i o dalších procedurách s možností bayesovského přístupu k analýze dat.*

Abstract: *Bayesian methods are becoming increasingly popular in modern statistical analysis and they are applied to a wide range of scientific disciplines and fields of research. The submitted contribution presents the principles of this procedure in the STAT module of the SAS programming system. There MCMC procedure is designed for generally applicable Bayesian calculations. The MCMC procedure provides researchers the data analysis on a wide range of complex Bayesian statistical models. The procedure uses the Monte Carlo Markov Chains (MCMC) algorithm for random sampling of any posterior distribution, which is defined by any prior distributions with parameters and probability functions for the data to be analysed. This article describes how to use the MCMC procedure in estimations, inferences and predictions, and marginally mentions other procedures with the possibility of a Bayesian approach to data analysis.*

Klíčové slova: *apriorní rozdělení, aposteriorní rozdělení, MCMC, pravděpodobnostní funkce.*

Key words: *prior distribution, posterior distribution, MCMC, probability function.*

1 Úvod

Nejčastěji používané statistické metody jsou známé jako četnostní (nebo-li tzv. klasické) metody. Tyto metody předpokládají, že neznámé parametry jsou fixními konstantami, a definují pravděpodobnost pomocí omezujících relativních frekvencí. Z těchto předpokladů vyplývá, že pravděpodobnosti jsou objektivní a že nelze dělat pravděpodobnostní úsudky o parametrech, protože jsou pevně

dané. Bayesovské metody nabízejí alternativní přístup; zacházejí s parametry jako s náhodnými proměnnými a definují pravděpodobnost jako „stupně víry“[1]. Z těchto postulátů vyplývá, že pravděpodobnosti jsou subjektivní a že je možné o parametrech uvádět pravděpodobnostní výroky. Termín „bayesovský“ pochází z převládajícího používání Bayesovy věty, která byla pojmenována po reverendu Thomasi Bayesovi, presbyteriánském duchovním z osmnáctého století. Bayes se zajímal o řešení otázky inverzní pravděpodobnosti: jaká je po pozorování množiny událostí pravděpodobnost jedné události?

2 Bayesův vzorec pro náhodné jevy a hypotézy

Předpokládejme, že se má realizovat odhad parametru θ z dat $\boldsymbol{y} = \{y_1, \cdots y_n\}$ pomocí statistického modelu popsaného hustotou pravděpodobnosti $p(\boldsymbol{y}|\theta)$ za platnosti parametru θ. Bayesovská statistika postuluje, že tento odhad přesně určit nelze, přičemž nejistota ohledně parametru θ je vyjádřena pomocí pravděpodobnostních úsudků a rozdělení. Například, pokud pravděpodobnostní rozdělení parametru θ se řídí normálním rozdělením s průměrem *0* a rozptylem *1*, v duchu bayesovské statistiky se má za to, že toto rozdělení nejlépe vystihuje nejistotu spojenou s parametrem θ. Následující kroky popisují základní prvky bayesovské dedukce:

- Rozdělení pravděpodobnosti pro parametr θ je formulováno jako $\pi(\theta)$ a je známé jako tzv. rozdělení apriorní[2].
- Na základě pozorovaných dat $\boldsymbol{y}$ se zvolí statistický model $p(\boldsymbol{y}|\theta)$ pro popis rozdělení daného $\boldsymbol{y}$ za předpokládané platnosti parametru θ.
- Víra (přesvědčení) o parametru θ se upraví kombinací informace z apriorního rozdělení $\pi(\theta)$ a dat $\boldsymbol{y}$ výpočtem aposteriorního rozdělení $p(\theta|\boldsymbol{y})$.

Úpravy víry (přesvědčení) o parametru θ se provádí pomocí tzv. Bayesovy věty, která odpovídá kombinaci apriorního rozdělení $\pi(\theta)$ a statistického modelu $p(\boldsymbol{y}|\theta)$ následujícím způsobem:

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta,\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\theta)\pi(\theta)}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\theta)\pi(\theta)}{\int p(\boldsymbol{y}|\theta)\pi(\theta)\mathrm{d}(\theta)}\,, \tag{1}$$

kde

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\theta)\pi(\theta)\mathrm{d}(\theta). \tag{2}$$

Veličina $p(\boldsymbol{y})$ daná výrazem (2) vyjadřuje marginální (okrajové) rozdělení pozorovaných dat $\boldsymbol{y}$ a sehrává úlohu normalizační konstanty aposteriorního rozdělení. Pokud existuje marginální rozdělení ve tvaru konečného (vlastního)

[1] Pravděpodobnost události je míra, s jakou je možné se domnívat, že je událost pravdivá.

[2] Vyjadřuje víru (přesvědčení) o parametru θ ještě před analýzou dat $\boldsymbol{y}$.

integrálu, neposkytuje žádnou dodatečnou informaci o aposteriorním rozdělení $p(\theta|\boldsymbol{y})$. Proto je aposteriorní rozdělení $p(\theta|\boldsymbol{y})$ úměrné statistickému modelu $p(\boldsymbol{y}|\theta)$, který specifikuje podmíněné rozdělení pozorovaných dat za daných parametrů, a apriorní pravděpodobnosti $\pi(\theta)$ pro parametr θ. Způsob zápisu Bayesovy věty (1) pomocí věrohodnostní funkce $L(\theta)$ je následující:

$$p(\theta|\boldsymbol{y}) = \frac{L(\Theta)\pi(\Theta)}{\int L(\Theta)\pi(\Theta)\mathrm{d}(\Theta)}. \tag{3}$$

V závislosti na vlivu apriorního rozdělení může vaše předchozí přesvědčení ovlivnit generované aposteriorní rozdělení buď silně (subjektivní nebo informativní apriorní rozdělení), nebo minimálně (objektivní nebo neinformativní apriorní rozdělení). Na apriorní hustotu se lze v jistém smyslu dívat jako na míru našeho přesvědčení, že θ nabývá právě této hodnoty. Tato víra může být podložena historickými daty či pouze statistikovým subjektivním názorem (Vávra, 2018). Bayesovu vzorci (1) se také říká zákon o inverzní pravděpodobnosti (Hebák, 2012). Pojem inverzní pravděpodobnosti se především používá v souvislosti s úsudky o neznámém modelu na základě známých dat, na rozdíl od přímých pravděpodobností, které se naopak používají pro úsudky o neznámých skutečnostech na základě známého modelu.

Bayesovská analýza dat se zabývá výlučně aposteriorním rozdělením $p(\theta|\boldsymbol{y})$. Na rozdíl od klasické analýzy dat jsou parametry považovány za náhodné veličiny mající svoje pravděpodobnostní rozdělení. Veškerá statistická inference v bayesovském pojetí vychází ze souhrnných charakteristik aposteriorního rozdělení $p(\theta|\boldsymbol{y})$ včetně bodových a intervalových odhadů. Například průměr nebo medián aposteriorního rozdělení slouží jako bodové odhady a kvantily rozdělení vedou k bayesovskému intervalu spolehlivosti, tzv. "credibility" intervalu. Podobně jako u klasické statistiky tyto intervaly mohou být obou- nebo jednostranné. Každý interval (d, h) podle (Samaniego, 2010), pro který platí

$$P(\theta_d \leq \theta \leq \theta_h) = \int_{\theta_d}^{\theta_h} p(\theta|\boldsymbol{y})d\theta = 1 - \alpha. \tag{4}$$

je $(1-\alpha)100\%$ bayesovským intervalem spolehlivosti odhadovaného parametru. Interval s vyhovujícími vlastnostmi je znám jako „*highest posteriori density*“ (HPD) interval, který pro jednovrcholová rozdělení je nejkratším, neboli má nejmenší rozdíl $(\theta_h - \theta_d)$.

3 Ilustrativní příklad použití bayesovské analýzy dat

Nechť urna obsahuje *n* míčků, z nichž n_1 je bílých a n_2 černých. Z urny se náhodně vytáhne 1 míček a poté se vrací do urny nazpět; jedná se tedy o výběr s vracením. Každý míček má stejnou pravděpodobnost vytažení rovnou $1/n$. Pravděpodobnost vytažení bílého míčku je θ , černého pak $1-\theta$. Tento proces

se bude opakovat *n*-krát. Náhodná veličina udávající počet vytažených bílých míčků se označí jako *X*. Ze všech *n* pokusů je možné vybrat přesně *k* bílých míčků právě $C(k,n)$ způsoby (kde $C(k,n)$ je počet kombinací k–té třídy spomezi n prvků). Náhodná veličina *X* podléhá binomickému rozdělení, a proto statistický model (věrohodnostní funkce) bude vyjádřen ve tvaru

$$L(\theta) = C(k,n)\,\theta^k(1-\theta)^{n-k}. \tag{5}$$

Na rozdíl od četnostní statistiky, bayesovská statistika předpokládá, že parametr θ může nabývat více hodnot s apriorní vírou (přesvědčením) o pravděpodobnostech uvedených hodnot parametru θ. Příklad apriorního rozdělení parametru udávajícího pravděpodobnost θ je uveden v tabulce 1. Pro *n = 20* a *k = 2* jsou v prvním, resp. druhém sloupci všechny možné hodnoty parametru θ, resp. jejich apriorní pravděpodobnosti (rozdělení). Na základě hodnot parametru θ v prvním a hodnot apriorního rozdělení ve druhém sloupci Tabulky 1 lze spočítat hodnotu věrohodnosti do třetího sloupce pro každou z hodnot parametru θ podle vztahu (5). Výsledné aposteriorní rozdělení parametru θ (pátý sloupec tabulky 1) je určeno pomocí Bayesovy věty (3), přičemž hodnota normalizační konstanty představuje součet hodnot věrohodnostní funkce vážených apriorními pravděpodobnostmi parametru θ (ve čtvrtém sloupci tabulky 1).

Tab. 1 Empirické aposteriorní rozdělení *(Zdroj: Dohnálek, 2016)*

Hodnoty θ	Apriorní rozdělení	Funkce věrohodnosti	Funkce věrohodnosti vážená apriórním rozdělením	Aposteriorní rozdělení
0,10	0,125	0,2852	0,0356	0,1600
0,12	0,150	0,2740	0,0411	0,1845
0,14	0,225	0,2466	0,0555	0,2491
0,16	0,225	0,2109	0,0474	0,2130
0,18	0,175	0,1730	0,0259	0,1165
0,20	0,125	0,1369	0,0171	0,0768
Celkem	**1,000**		**0,2501**	**1,0000**

Takovým způsobem dochází k úpravám apriorních pravděpodobností využitím informace získané z dat na hodnoty aposteriorních pravděpodobností. Hodnoty věrohodnostní funkce pro pozorovaná data vážené apriorním rozdělením určují rozdělení aposteriorní pravděpodobnosti, které představují zkorigované prvotní, počáteční přesvědčení (víru) o parametru θ pomocí informace ze získaných dat.

4 Funkcionality bayesovské statistiky v systému SAS

Programový systém SAS nabízí 2 způsoby analýzy dat bayesovskou metodou. Ve vybraných procedurách pro statistické modelování se bayesovská analýza volá příkazem BAYES a univerzální proceduru MCMC pro obecné statistické

modelování pomocí Bayesovy věty (SAS Institute Inc., 2018a). Statistická inference – bodové i intervalové odhady - vychází z aposteriorního rozdělení $p(\theta|\boldsymbol{y})$.

Software SAS/STAT poskytuje bayesovské možnosti v sedmi procedurách: BCHOICE, BGLIMM, FMM, GENMOD, LIFEREG, PHREG a MCMC. Procedury FMM, GENMOD, LIFEREG a PHREG umožňují bayesovské analýzy jako doplněk ke standardním četnostním analýzám (Chen, 2014). Procedury tak poskytují pohodlný přístup k bayesovskému modelování a inferenci pro modely konečných směsí rozdělení, zobecněné lineární modely, modely dožití, Coxovy regresní modely a po částech exponenciální modely. Procedura BCHOICE poskytuje Bayesovu analýzu pro diskrétní modely výběru. Procedura BGLIMM dovoluje bayesovskou analýzu pro generalizované lineární modely se smíšenými efekty.

Bayesovské modelování také nabízí alternativní modelové řešení v analýze chybějících hodnot procedurou MI, ve kterém se s chybějícími hodnotami zachází jako s neznámými parametry a podle toho se odhadují. Tyto chybějící hodnoty jsou jednoduše imputovány přijatelnými hodnotami prostřednictvím simulací aposteriorního rozdělení neúplných údajů.

Jádro procedury MCMC pro generování pseudonáhodných čísel z pravděpodobnostních rozdělení tvoří metody založené na simulaci pomocí Markovských řetězců. Podle (SAS Institute Inc., 2018b) je markovský řetězec sekvence náhodných veličin, u nichž rozdělení každého prvku závisí výlučně na hodnotě prvku předchozím. Pro účely simulace metodou Markovských řetěců jsou nejpopulárnějšími algoritmy Gibbsovské vzorkování (Gibbs Sampling) a Metropolis-Hastings algoritmus. Metropolis-Hastings algoritmus simuluje výběr z aposteriorního pravděpodobnostního rozdělení za účelem aproximace požadovaného rozdělení a provádí rozhodování o přijetí nebo o odmítnutí vygenerovaného náhodného vzorku. Pokud je nově generovaný vzorek akceptován, algoritmus generuje vzorek nový, pokud je náhodně generovaný vzorek odmítnut, simulace stávajícího výběru se opakuje. Generování náhodných vzorků končí v okamžiku dosažení konvergence k cílovému aposteriornímu rozdělení. Po dosažení konvergence se vybere náhodně z několika nasimulovaných řad a vybírá se řada podle spolehlivosti konvergence k cílovému rozdělení. Gibbsovo vzorkování (jako speciální případ předchozího algoritmu) rozděluje simultánní aposteriorní rozdělení vícerozměrné náhodné veličiny na podmíněná rozdělení pro každý parametr, ze kterých náhodně generuje řetězce a sleduje konvergenci k požadovanému aposteriornímu rozdělení.

5 Ukázky bayesovského modelování procedurou MCMC

Procedura MCMC je flexibilní, simulačně založená procedura, která je vhodná pro modelování podle široké škály bayesovských modelů. Pro úspěšné modelování je třeba zadat tvar podmíněného rozdělení pozorovaných dat za daných parametrů (statistický model) a apriorní rozdělení pro odhadované parametry. Pro modelování hierarchických modelů se vkládá také i apriorní rozdělení pro parametry náhodných efektů. Simulací procedura MCMC pak získává vzorky z požadovaného aposteriorního rozdělení, vytváří souhrnné a diagnostické statistiky a ukládá generované vzorky do výstupního datového souboru, který lze použít pro další analýzu. I když procedura MCMC podporuje celou sadu standardních rozdělení, je možné procedurou modelovat data s jakýmkoliv podmíněným rozdělením (statistickým modelem), apriorním či aposteriorním rozdělením, jsou-li jejich matematické tvary programovatelné pomocí příkazů DATA kroku. Na parametry vstupující do modelu neexistují žádná omezení, parametry do procedury mohou být zadávány buď v lineární, nebo v jakékoliv nelineární funkční podobě.

Procedura MCMC může modelovat automaticky proměnné závislé (response variable), proměnné nezávislé (kovariáty) i chybějící hodnoty. Chybějící hodnoty procedura považuje za neznámé parametry a zahrnuje chybějící hodnoty do procesu simulace. Pro každý parametr, resp. blok parametrů procedura vybírá nejvhodnější metody vzorkování. Například je-li generované rozdělení konjugované[3], vzorky jsou generovány přímo z podmíněných aposteriorních rozdělení využitím standardních číselných generátorů. V jiných případech procedura používá Metropolis-Hastings algoritmus, Gibbsovo vzorkování, případně a jiné.

Nejjednodušším příkladem bayesovského modelování procedurou MCMC je simulace vzorků z normálního (aposteriorního) rozdělení pravděpodobnosti, které se realizuje pomocí následující příkazové syntaxe:

```
data X;
     stop;
run;

ods graphics on;
ods exclude nobs;
proc mcmc data=X outpost=simout seed=23 nmc=10000 diagnostics=none
     plots=density;
     parm alpha 0;
     prior alpha ~ normal(0, sd=1);
     model general(0);
run;
```

[3] Je-li aposteriorní rozdělení členem stejné rodiny rozdělení jako rozdělení apriorní, pak obě tato rozdělení jsou konjugovaná, neboli obecně, jsou stejného typu.

Vstupem do procedury MCMC je prázdný datový soubor nazvaný X. Vstupní datový soubor X nemusí obsahovat datové věty (DATA krok s příkazem STOP), protože úlohou procedury je generování nových vzorků. Příkazem NMC=*10 000* se nastaví počet simulovaných vzorků, příkaz PARM slouží k označení parametru, jehož aposteriorní rozdělení bude procedura simulovat včetně přiřazení jeho počáteční hodnoty. Příkazem PLOTS se určí, které grafy má procedura generovat. Apriorní rozdělení simulovaného parametru se nastaví příkazem PRIOR, když v tomto případě se jedná o parametr ALPHA s aposteriorním normálním rozdělením $N(0,1)$. Příkaz MODEL, který vkládá do procedury tvar podmíněného rozdělení pozorovaných dat za daných parametrů (statistický model), musí mít každá příkazová syntaxe pro proceduru MCMC. Pro generování parametru ALPHA, kde neexistuje závislá proměnná, nabývá příkaz MODEL hodnoty GENERAL(0). Simulovaná data se ukládají do výstupního souboru SIMOUT (nastavení parametru OUTPOST). Vzhledem k jednoduchosti úlohy, není potřebná jakákoliv diagnostika - např. počet iterací do stacionarity MCMC řetězců, rychlosti konvergence generování řetězců, atd.

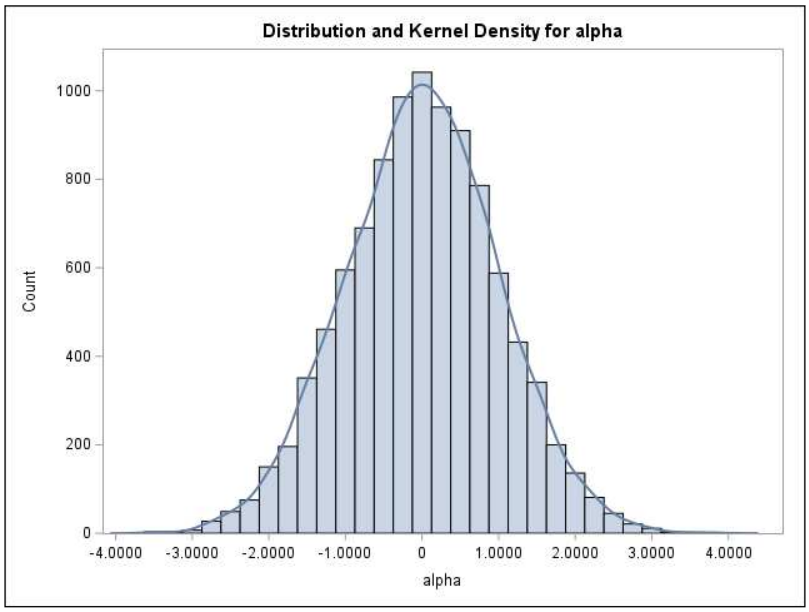
Po provedené simulaci je výstupem činnosti procedury MCMC soubor SIMOUT obsahující *10 000* vygenerovaných hodnot parametru ALPHA, jehož rozdělení pravděpodobnosti odpovídá požadovanému normálnímu rozdělení $N(0,1)$. Generování hodnot procedurou bylo implicitně realizováno přímo z podmíněného aposteriorního rozdělení (normálního) využitím standardních číselných generátorů, protože apriorní i podmíněné aposteriorní rozdělení parametru ALPHA patří do stejné rodiny rozdělení, a tedy jsou konjugované. Výstupní report v systému SAS obsahuje tabulku odhadovaných parametrů „Parameters“ s názvem použité metody vzorkování, názvem a počáteční hodnotou parametru a popisem apriorního rozdělení parametru. Sumární statistiky a inference (vycházející výlučně z aposteriorního rozdělení) jsou zachyceny v tabulce 2 „Posterior Summaries and Intervals“, kde je uveden název parametru, jeho střední hodnota, standardní odchylka a 95% HPD interval.

Tab. 2 Sumární statistiky a inference simulovaného parametru *(Zdroj: vlastní zpracování)*

Parameters				
Block	Parameter	Sampling Method	Initial Value	Prior Distribution
1	alpha	Direct	0	normal(0, sd=1)

Posterior Summaries and Intervals					
Parameter	N	Mean	Standard Deviation	95% HPD Interval	
alpha	10000	0.00195	0.9949	-1.9664	1.9302

Na obrázku 1 byl zaznamenán průběh rozdělení parametru ALPHA.



Obr. 1 Aposteriorní rozdělení parametru ALPHA *(Zdroj: vlastní zpracování)*

Následující příklad použití bayesovského modelování se bude týkat problému s chybějícími údaji. Poprvé se uvedený příklad objevil v diskusním příspěvku Murraya (1977), který představuje datový soubor s dvojrozměrným normálním rozdělením s chybějícími daty. Jedná se o uměle vytvořený problém, kterým však později zabývali i mnozí jeho následovníci jako Tanner & Wong (1987) a Tan et al. (2010).

Tab. 3 Struktura souboru s normálním rozdělením s chybějícími daty *(Zdroj: Murray, 1977)*

Proměnná	Hodnoty dat s dvojrozměrným normálním rozdělením											
x_1	1	1	-1	-1	2	2	-2	-2	.	.	.	.
x_2	1	-1	1	-1	.	.	.	.	2	2	-2	-2

Soubor dat v tabulke 3 k analýze chybějících hodnot se vytvoří pomocí příkazů:

```
data binorm;
     input x1 x2 @@;
     datalines;
1 1 1 -1 -1 1 -1 -1 2 . 2 . -2 . -2 . . 2 . 2 . -2 . -2;
run;
```

Soubor dat (nazvaný BINORM) se skládá z 12 pozorování z dvojrozměrného normálního rozdělení. U obou komponent tohoto vícerozměrného rozdělení proměnných chybí hodnoty. Strukturu souboru s chybějícími daty zobrazuje tabulka 3. Dvojrozměrné normální rozdělení má obě komponenty se stejnými nulovými průměry μ, s různými rozptyly σ_1^2 a σ_2^2 a korelačním koeficientem ρ.

Tyto předpoklady se promítají do následující kovarianční matice Σ:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{6}$$

Apriorní rozdělení kovarianční matice Σ je $\pi(\Sigma)$ a odpovídá podle Boxa & Tiaa (1973) inverznímu Wishartovu rozdělení s determinantem $\sigma_1^2\sigma_2^2(1-\rho^2)$, tj.

$$\pi(\Sigma) \propto [\Sigma]^{-\frac{p+1}{2}} = \left(\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-(p+1)}. \tag{7}$$

Při analýze kompletních pozorování, kdy se do úvahy berou pouze úplná pozorování, by se bayesovským modelováním vytvořil poměrně jednoduchý aposteriorní odhad: 2 datové dvojice ((1, 1) a (-1, -1)) mají korelaci rovnou 1 a 2 dvojice dat ((1, -1) a (-1, 1)) mají korelaci -1. V případě, že se budou chybějící hodnoty považovat za zcela náhodně chybějící, lze podle Chena (2009) považovat nepozorované hodnoty buď (téměř) dokonale pozitivně nebo dokonale negativně korelující s pozorovanými hodnotami 2 a –2 a že chybějící hodnoty mohou nabývat pouze hodnot, které se od nich příliš neliší. Navíc vyřazením částečně pozorovaných hodnot se ztrácí podstatná část při odhadu rozptylu obou komponent dvojrozměrného normálního rozdělení. Bayesovské modelování chybějících hodnot dvojrozměrného normálního rozdělení vykonávají následující příkazy:

```
proc mcmc data=binorm nmc=20000 seed=17 outpost=postout
     diag=none plots=none monitor=(rho sig1 sig2);
     array x[2] x1 x2;
     array mu[2] (0 0);
     array sigma[2, 2];
     parms rho 0.5 / slice;
     parms sig1 1 sig2 1;
     if (sig1 > 0 and sig2 > 0) then
          lprior = -3 * log(sig1 * sig2 * sqrt(1-rho*rho));
     else lprior = .;
     prior rho sig1 sig2 ~ general(lprior);
     sigma[1,1] = sig1*sig1;
     sigma[1,2] = rho * sig1 * sig2;
     sigma[2,1] = sigma[1,2];
     sigma[2,2] = sig2*sig2;
     model x ~ mvn(mu, sigma);
run;
```

Příkazy PARMS specifikují tři modelované parametry modelu. Volba SLICE v prvním příkazu PARMS vybere vzorkovací algoritmus, kterým se generuje parametr ρ. I když výpočetně náročnější (Neal, 2003), tento typ algoritmu je velmi vhodný pro simulaci dat z nenormálních rozdělení (kam náleží i rozdělení korelačního koeficientu ρ). Příkazy IF-ELSE zajišťují, aby simulované hodnoty parametrů směrodatné odchylky σ_1 a σ_2 byly zahrnovány do aposteriorních odhadů kovarianční matice pouze při své kladné hodnotě. Příkazem LPRIOR a příkazem PRIOR se do procedury zadávají simultánní apriorní rozdělení

parametrů σ_1, σ_2 a ρ na logaritmické stupnici[4]. Prvky kovarianční matice Σ jsou konstruovány použitím dalších programových příkazů jako funkcí parametrů modelu (v syntaxi příkazů označené jako `sigma[1,1]` až `sigma[2,2]`). Příkazem MODEL se nastavuje specifikace požadované funkce věrohodnosti nutná pro simulaci a následnou aposteriorní inferenci.

Po provedené simulaci na základě výše uvedené syntaxe je výstupem činnosti procedury MCMC soubor POSTOUT obsahující *20 000* vygenerovaných chybějících hodnot v obou proměnných x_1 a x_2 ve dvojrozměrném normálním rozdělení a vygenerované rozdělení parametrů σ_1, σ_2 a ρ podle zadání. Analýza chybějících hodnot je ilustrována v části „Missing Data Information Table" tabulky 4. Proměnná x_1 obsahuje *4* chybějící hodnoty, proměnná x_2 obsahuje také *4* chybějící hodnoty. Tabulka 4 zaznamenává indexy chybějících hodnot za jednotlivé proměnné. Aposteriorní inference o odhadovaných parametrech σ_1, σ_2 a ρ je zobrazena v části „Posterior Summaries and Intervals" tabulky 4.

Tab. 4 Sumární statistiky a inference dvojrozměrného rozdělení *(Zdroj: vlastní zpracování)*

Missing Data Information Table			
Variable	Number of Missing Obs	Observation Indices	Sampling Method
x1	4	9 10 11 12	Direct
x2	4	5 6 7 8	Direct

Posterior Summaries and Intervals					
Parameter	N	Mean	Standard Deviation	95% HPD Interval	
rho	20000	-0.00184	0.5582	-0.8591	0.8519
sig1	20000	1.6555	0.4458	0.9828	2.6014
sig2	20000	1.6306	0.4444	0.9478	2.4860

Průběh aposteriorního rozdělení parametru ρ ilustruje obrázek 2. Průměrný odhad ρ se blíží 0, se směrodatnou odchylkou 0,56. Tento odhad je nepřesný, protože aposteriorní rozdělení není unimodální. Obrázek 3 ilustruje aposteriorní rozdělení odhadovaných parametrů σ_1, σ_2.

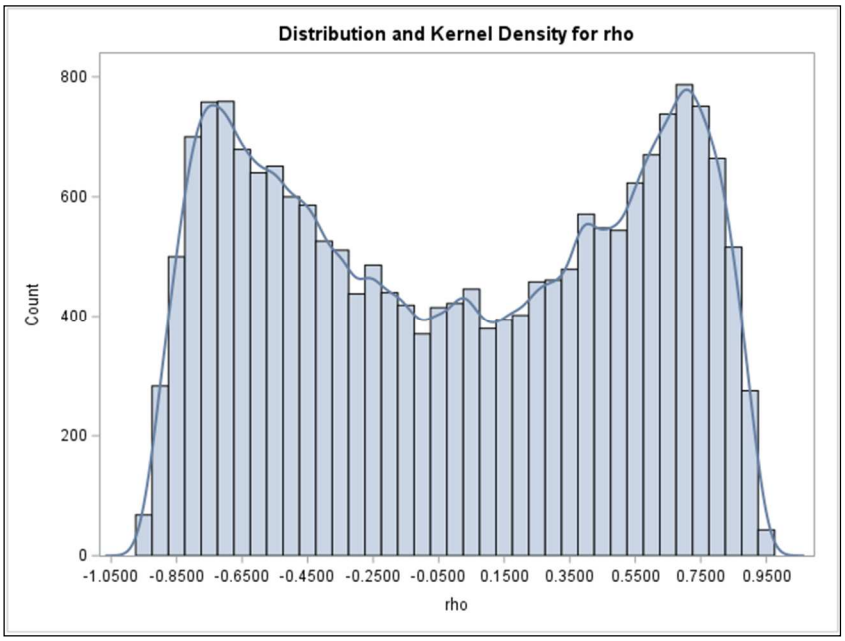
Podobně jako v předchozím případě jsou apriorní i podmíněné aposteriorní rozdělení chybějících údajů konjugované. Podmíněná aposteriorní rozdělení chybějících hodnot $\pi(x_1|.)$ a $\pi(x_2|.)$ pro jednotlivé proměnné x_1 a x_2 jsou jednorozměrná normální rozdělení a jsou dána výrazem:

$$\pi(x_1|\rho,\sigma_1,\sigma_2,x_2)\sim N\left(\rho\frac{\sigma_1}{\sigma_2}x_2,\sigma_1^2(1-\rho^2)\right). \quad (8)$$

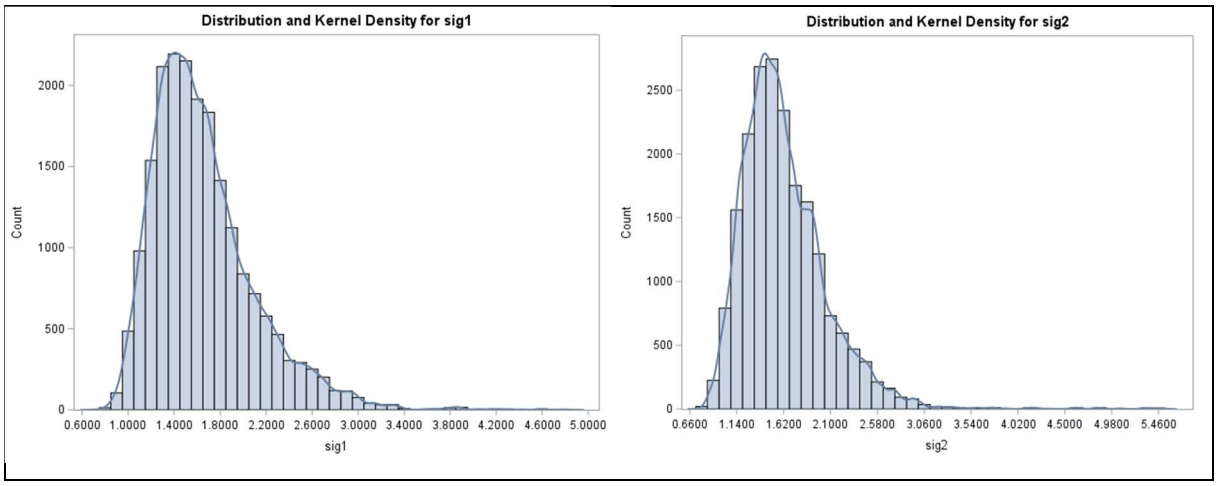
$$\pi(x_2|\rho,\sigma_1,\sigma_2,x_1)\sim N\left(\rho\frac{\sigma_2}{\sigma_1}x_1,\sigma_2^2(1-\rho^2)\right). \quad (9)$$

Podmíněná aposteriorní rozdělení chybějících hodnot (8) a (9) pro jednotlivé proměnné x_1 a x_2 jsou unimodální. Výstup zde ale není graficky zobrazen.

[4] Použitím výpočtů na logaritmické stupnici se dosahuje větší stabilita a rychlejší konvergence v číselném výpočtu.



Obr. 2 Aposteriorní rozdělení parametru ρ *(Zdroj: vlastní zpracování)*



Obr. 3 Aposteriorní rozdělení parametrů σ_1, σ_2 *(Zdroj: vlastní zpracování)*

6 Závěr

Při řešení analytických problémů by výzkumníci měli zvažovat alternativní statistické přístupy. I když bayesovské metody ve srovnání s klasickými postupy vyžadují hlubší znalosti teorie současně s určitými výpočetními dovednostmi, jejich správná aplikace ve výzkumné činnosti se rozhodně vyplatí. Bayesovská statistika poskytuje obecnější přístup pro statistickou inferenci. Na rozdíl od četnostní statistiky bayesovský přístup považuje odhadované parametry jako náhodné, jejichž rozdělení je možné upřesňovat informacemi získaných z dat.

Navíc v rámci bayesovského přístupu jsou znalosti zkoumaného fenoménu, které lze efektivně využít, zpravidla vždy dostupné.

V současné době je aplikace bayesovských metod mnohem přístupnější, protože už existují vhodné volně přístupné programy i placené programové systémy, například právě SAS.

7 Literatura

Box, G. E. P.,& Tiao, G. C. (1973). Bayesian inference in statistical analysis. New York: Wiley, 1973.

Dohnálek, R. (2016). Rozdělení pravděpodobnosti odvozená z urnových modelů. Bakalářská práce obhájena na Přirodovědecké univerzity Masarykovy univerzity v Brně.

Hebák, P. (2012). Srovnání klasické a bayesovské pravděpodobnosti a statistiky (1). Acta Oeconomica Pragensia 20(1):69-87.

Chen, F. (2014). Practical Bayesian computation using SAS. ASA Conference on Statistical Practices. February 20, 2014.

Murray, G. D. (1977). Contribution to discussion of paper by Dempster, Laird, and Rubin. Journal of the Royal Statistical Society, Series B, 39, 27–28.

Neal, R. M. (2003). Slice sampling. Technical report No. 2005. Department of Statistics and Department of Computer Science. University of Toronto, Toronto, Ontario, Canada, https://www.cs.toronto.edu/~radford/ftp/slc-samp.pdf.

Samaniego, F. J. (2010). A comparison of the Bayesian and frequentist approaches to estimation. New York: Springer, 2010.

SAS Institute Inc. (2018a). SAS/STAT® 15.1 user's guide. Chapter 7: Introduction to Bayesian analysis procedures. Cary, NC: SAS Institute Inc., 2018, s.129.

SAS Institute Inc. (2018b). SAS/STAT® 15.1 user's guide. Chapter 77: The MCMC procedure. Cary, NC: SAS Institute Inc., 2018, s.6011.

Tan, M. T., Tian, G. L., & Ng, K. W. (2010). Bayesian missing data problems: EM, data augmentation, and non-iterative computation. New York: Chapman & Hall/CRC, 2010.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.

Vávra, Jan. (2018). Bayesovská faktorová analýza. Univerzita Karlova. Matematicko-fyzikální fakulta. Praha, 2018.