BAYESOVSKÁ STATISTICKÁ ANALÝZA S PODPOROU VÝPOČETNĚ INTENZIVNÍCH PROCEDUR

Jan Hendl

Abstrakt

Vývoj počítačových technologií umožnil využít speciální algoritmy ve statistických výpočtech. Dokonce se tak změnilo nazírání na statistickou analýzu. Ačkoliv bayesovské postupy vždy tvořily alternativu k zavedeným přístupům k odhadu a ověřování statistických modelů, dnes jsme svědky v tomto směru skutečné revoluce, v které bayesovské procedury a počítače hrají rozhodující roli. V příspěvku se krátce věnujeme základům dvou hlavních škol statistického myšlení. V další části ukážeme v čem spočívá současná moderní podoba bayesovského počítání a modelování. Na rozdíl od minulosti dnešní algoritmické přístupy umožňují radikálně rozšířit využití bayesovských metod při analýze biomedicínských dat.

Klíčová slova:

Bayesovská analýza, počítačové metody, apriorní a aposteriorní pravděpodobnosti, mcmc algoritmus, modelování

Úvod

Mnoho v současnosti používaných statistických metod bylo vyvinuto ke konci 19. století, např. šlo o lineární regresi nebo metodu nejmenších čtverců. Galton navrhl v roce 1888 pojem statistické korelace. Youle a Pearson zavedli pojem měr dobré shody. Ve 20. a 30. letech 20. století došlo k velkém rozvoji statistiky pracemi R.A. Fischera o odhadování pomocí věrohodnostní funkce. Neyman a Pearson rozvinuli myšlenku testování na základě představ o rozhodování z operační analýzy. Od té doby se statistika rychle rozvíjela a dnes hraje klíčovou roli ve výzkumu v mnoha vědeckých oblastech.

Bayesovské metody tvoří část statistiky. Základ pro ně položil reverend Thomas Bayes, jehož myšlenky prezentoval před Královskou statistickou společností v roce 1763 Richard Price až několik let po smrti autora (Bayes 1763). Bayesův přístup se však dlouho příliš neprosadil, protože oboru dominoval klasický četnostní přístup Neymana a Fischera, kteří vytýkali Bayesovu přístupu subjektivitu. Někteří statistici jako Lindley nebo de Finneti tento přístup však rozvíjeli, přičemž poukazovali na určité nedostatky klasického přístupu. Teprve začátkem 90. let 20. století začal zájem o Bayesovy metody vzrůstat. Dokonce dominoval ve výzkumu statistiky. Impulsem proto byly návrhy z oblasti výpočetní statistiky, které spolu s dostupností výpočetní techniky, umožnily pracovat v rámci pružného rámce pro statistické usuzování, což vyhovovalo stále větší složitosti problémů ve vědě na konci 20, a začátku 21. století (Brooks 2003, Ashby 2006)

Rozdílnost klasického a bayesovského přístupu

Hlavní vlastností bayesovského přístupu je princip, že pravděpodobnost je interpretována jako subjektivní přesvědčení, že nastane nějaká událost,

reprezentuje nejistotu jedince. Jedním z důsledků je to, že se nepožaduje, aby se specifikovala nebo uvažovala relevantní řada opakování pokusu, v kterém událost nastává, což platí pro klasický četností rámec. V bayesovském přístupu je subjektivní pravděpodobnost upravována pomocí evidence, která je postupně shromažďována (Gelman et al. 1995).

Klasický četnostní a bayesovský přístup a jejich rozdílnost nejlépe vysvětlíme pomocí zcela jednoduchého příkladu. Půjde o parametr θ binomického rozdělení, kterým modelujeme binární náhodnou proměnnou X. Ta nabývá hodnoty 1 s pravděpodobností θ a hodnoty 0 s pravděpodobností 1 – θ. Její pravděpodobnostní f(x I θ) funkci lze zapsat ve tvaru:

$$f(x \mid \theta) = \theta^{x} (1-x)^{1-x}$$

Jestliže provedeme n pokusů s takovou náhodnou proměnnou, pak ze vyjádřit jejich společnou pravděpodobnost $f(x_1, x_2, ..., x_n \mid \theta)$ výsledků vynásobením n těchto funkcí s dosazenými hodnotami měřené veličiny. Dostaneme výraz

$$f(x_1, x_2, ..., x_n \mid \theta) = f(\mathbf{x} \mid \theta) = f(D \mid \theta) = \theta^{m} (1-x)^{n-m},$$

kde m je počet pokusů, kdy X nabyla hodnota 1. Tato pravděpodobnostní funkce tvoří základ pro rozhodování oběma přístupy. Symbol D bude označovat získaná data $x=(x_1, x_2, ..., x_n)$.

V klasickém četnostním přístupu při hledání θ postupujeme metodou maximální věrohodnosti a to tak, že maximalizujeme hodnotu funkce f(D I θ) vzhledem k parametru θ. Volíme takovou hodnotu θ, při které pravděpodobnost výsledku D by byla největší. Věrohodnostní funkcí L(θ I D) označujeme funkci f(D I θ). Za proměnou nyní považujeme parametr θ a ne jednotlivé hodnoty x_i . Hodnotu, kterou získáme maximalizací funkce L(θ I D), nazýváme maximálně věrohodným odhadem. V našem příkladu má odhad metodou maximální věrohodnosti intuitivně přirozený tvar:

$$\hat{\theta} = \frac{m}{n}$$

Věrohodnostní funkce L(θ I D) tvoří základ pro statistické usuzování o hodnotě θ v klasickém četnostním přístupu. Můžeme například porovnávat dva kandidáty na správnou hodnotu θ, přičemž tyto kandidáte jsme zjistili maximalizací věrohodnostní funkce za dvou různých omezujících podmínek na prostoru možných hodnot θ. Při maximalizaci funkce L(θ I D) používáme aparát diferenciálního počtu, který lze také použít, pokud rozložení závisí na několika různých parametrech a uvažujeme vektorový parametr θ. Maximalizace začne však být složitější a obvykle vyžaduje rafinované optimalizační numerické metody. Další obtíže nastávají, protože věrohodnostní funkce může mít mnoho lokálních maxim a úkolem je odhadnout to absolutní. Jeho hledání obvykle vyžaduje provést mnoho výpočtů. Při odhadu statistické nepřesnosti odhadu nebo získání intervalu spolehlivost vycházíme také z věrohodnostní funkce. Přitom se opíráme v jednodušších problémech o poznatek, že odhady metodou maximální věrohodnosti mají normální

rozdělení se známými hodnotami průměru a rozptylu. Přitom předpokládáme, že lze libovolně zvětšovat rozsah výběru. Tento předpoklad nemusí být vždy splněn a pak asymptotické výsledky se jeví jako zpochybnitelné. Lze teoreticky dokázat, že asymptoticky se takto zjištěný odhad blíží s pravděpodobností jedna ke správné hodnotě.

Interpretace toho, jak je odhad statisticky přesný vychází z představy opakovatelnosti celého experimentu a předpokládané chování našeho odhadu odvozujeme z možné nekonečné série vektorů výsledků **x**, které byly získány za stejných podmínek jako náš výchozí vektor dat. O intervalu spolehlivosti s hladinou spolehlivosti např. 95%, pak můžeme prohlásit, že v sérii pokusů o n pozorování (ale nikdy neuskutečněné), pokryje správný parametr θ v 95% případech. Tato představa se zdá být poněkud nepřirozená. V této souvislosti se uvádí příklad z medicínského usuzování. Jestliže diagnostický test je pozitivní u pacientů s chorobou A v 99% případů, neznamená to, že pozitivní hodnota testu T+ svědčí o tom, že daný pacient má chorobu A s pravděpodobností 99%. Ale právě tato pravděpodobnost pacienta a lékaře zajímá především. V tomto případě musíme znát pravděpodobnost v uvažované populaci, že náhodně vybraný jedinec má chorobu A. Označme ji P(A) a hledanou podmíněnou pravděpodobnost P(A | T+). Pak podle Bayesovy věty ji můžeme zjistit, jestliže také známe kromě senzitivity testu P(T+ | A) – v našem případě má hodnotu 99% - také specificitu, tedy podmíněnou pravděpodobnost P(T- | non A), že test bude negativní v případě, že jedinec chorobu nemá. Hledaná pravděpodobnost se pak vypočítá podle vzorce:

$$(4)\quad P(A|T+) = \frac{P(A)P(T+|A)}{P(A)P(T+|A) + (1-P(A))P(T-|non\ A)}$$

Jedná se o nejjednodušší formu Bayesovy formule. Na pravé straně rovnice výraz ve jmenovateli vyjadřuje pravděpodobnost jevu, že test bude pozitivní v dané populaci a jmenovatel vyjadřuje pravděpodobnost současného jevu u jedince, že má chorobu A a zároveň pozitivní test. Rovnici lze tedy přepsat do známého tvaru:

Při vysvětlování základních vztahů v bayesovské metodě vycházíme z tohoto

$$(5)\quad P(A|T+) = \frac{P(A)P(T+|A)}{P(T+)} = \frac{P(A \cap T+)}{P(T+)}$$

vzorce, ale použijeme jeho tvar pro hustoty:

$$(6)\quad f(\theta|D) = \frac{f(\theta)\,f(D|\theta)}{f(D)}$$

kde D označuje získaná data a θ parametr.

V bayesovském přístupu chceme nějakým způsobem vyjádřit naše apriorní představu o možných hodnotách parametru θ a její ovlivnění daty D. Naši apriorní představu vyjadřujeme v tomto přístupu apriorním rozložením f(θ). Toto rozložení pak modifikujeme na základě nasbíraných dat pomocí Bayesovy formule a získáme vylepšenou představu ve tvaru aposteriorního rozdělení $f(\theta \mid D)$.

Stejně jako v klasickém četnostním přístupu bayesovský přístup ústí do funkce, kde proměnnou je parametr nebo vektor parametrů při napozorovaných pevně daných datech D. V Bayesově případě funkce udává aposteriorní pravděpodobnostní rozdělení. V tom spočívá základní rozdíl. Klasičtí statistici věří, že existuje pevná hodnota parametru θ modelu a pokoušejí se odhadnout tuto hodnotu maximalizováním věrohodnostní funkce. Bayesovský přístup předpokládá, že parametr má pevné, ale neznáme rozdělení, které reprezentuje naší přibližnou představu o hodnotě parametru. Také v bayesovském přístupu lze předpokládat, že správná hodnota existuje, ale tuto hodnotu nemůžeme nikdy znát zcela jistě, proto dáváme přednost vyjádření naší nejistoty o něm pomocí pravděpodobnostního rozložení. Jak stoupá velikost informace o parametru, tak se zmenšuje rozptyl aposteriorního rozložení, v limitě přiřazuje pravděpodobnost jedna jedné hodnotě.

V praktických úlohách, kdy pracujeme s vektorovými parametry, není snadné aposteriorní rozložení vektoru parametrů interpretovat. Proto počítáme různé snadněji interpretovatelné jednorozměrné aposteriorní charakteristiky jako jeho průměry a rozptyly vybraných parametrů. Bodové odhady parametru jsou dány volbou ztrátové funkce. Jestliže zvolíme její kvadratickou formu, pak bodový odhad vede k průměrné hodnotě aposteriorní hodnoty pro daný parametr. Jiné ztrátové funkce vedou k jiným tvarům bodového odhadu. Pokud je ztráta 1 při volbě nesprávné hodnoty a 0 při volbě správné hodnoty parametru, pak jako nejlepší odhad vyjde modus aposteriorní hustoty.

Uvažujme jednoduchý příklad. Pro případ binomického rozložení lze pro parametr θ volit apriorní rozložení různým způsobem. Jestliže nemáme žádnou informaci, pak příslušná hustota má tvar f(θ) = 1, pro pružnější zachycení našich apriorních informací o parametru θ se hodí beta rozložení jejích hustota je až na konstantu rovna $U = \theta^{(a-1)} (1- \theta)^{(b-1)}$. Pro hodnoty a=b=1 tak dostaneme rovnoměrné rozložení a pro hodnoty a=b získáme rozložení symetrické kolem hodnoty 0,5. Aposterioní rozložení má tvar až na multiplikativní konstantu $\theta^{(x+a-1)} (1- \theta)^{(n-x+b-1)}$. Takže aposteriorní rozložení je také typu beta rozložení a jeho modus má hodnotu (x+a-1)/(n+a+b-2). Jestliže a=b=1, pak modus má stejnou hodnotu jako maximálně věrohodný odhad.

Poznamenejme, že obecně je možné popsat rozložení θ pomocí hyperparametrů **η**. V našem příkladu jde o hodnoty beta rozdělení **η** =(a, b). Z toho důvodu by bylo vhodnější vyjádřit apriorní rozdělení ve tvaru $f(\theta \mid \boldsymbol{\eta})$. Kvůli jednoduchosti jsme toto označení nepoužili.

Vztah mezi pozorovanou proměnnou X a parametrem θ (a někdy i hyperparametry **η**) můžeme lépe zachytit použitím grafického znázornění pomocí orientovaných acyklických grafů (directed acyclic graph) s označením DAG.

V DAG znázornění uzle kruhem označuje náhodné proměnné, šipky mezi uzly stochastické závislosti a čtverce pevně dané hodnoty. Obrázek 1a ukazuje proces generování dat. Šipka směřující od uzlu θ k uzlu y znamená, že rozložení pozorované proměnné závisí (je podmíněno) na hodnotě θ. V inferenční části zobrazené na Obrázek 1b se hodnoty y stávají evidencí, proto jsou znázorněny čtvercem, v této fázi není v y žádná nejistota. Pozorováním získaná evidence se využije proti směru původní šipky DAG (přerušovaná šipka). Tato informace slouží k revizi pravděpodobnosti θ pomocí Bayesovy formule a vede k aposteriorním rozdělení f (θ | D).

Obrázek 1 – DAG znázorňuje a) proces generování dat a b) modifikaci rozdělení θ

Bayesovský přístup lze považovat za zobecnění četnostního přístupu v tom smyslu, že při volbě ploché (tzv. neinformativní) apriorní hustoty pro odhadované parametry, aposteriorní hustota má stejný tvar jako věrohodnostní funkce. Z toho také plyne, že pokud zvolíme ztrátovou funkci typu 0 nebo 1 vedoucí k modální hodnotě, mají frekvenční a bayesovské odhady stejnou hodnotu. Proto příznivci bayesovského přístupu mohou tvrdit, že frekvenční přístup je v podstatě bayesovský s tím, že nezdůvodněně omezuje tvar apriorního rozložení a tvar ztrátové funkce.

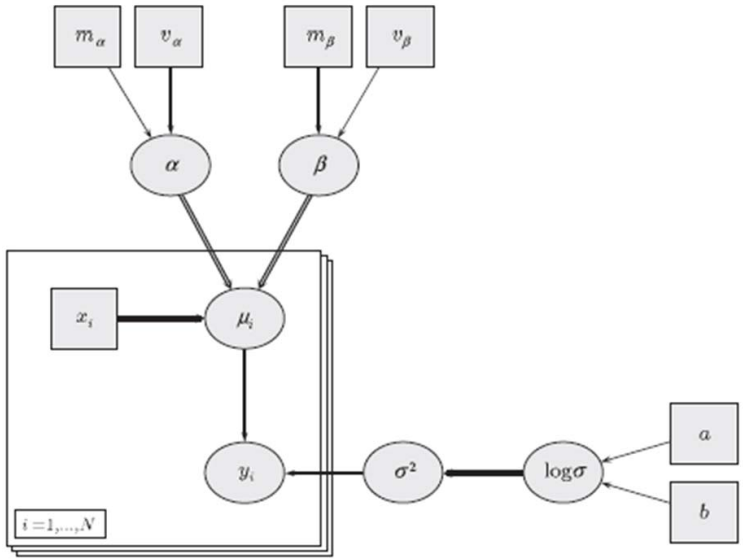
Bayesovský přístup má výhodu, že vytváří přirozený rámec pro zahrnutí nejistoty o parametru nebo parametrech. Aposteriorní rozptyl parametru v sobě nese informaci o nejistotě odhadu parametru. Pomocí aposteriorního rozložení lze také odvodit tzv. interval kredibility pro zvolený parametr, v kterém leží parametr s danou pravděpodobností. Ten má tu výhodu, že nepředpokládá hypotetickou opakovatelnost původního experimentu a jeho interpretace je přirozenější než interpretace intervalu spolehlivosti. Podobně jako u intervalu spolehlivosti, tento interval je užší s rostoucím počtem pozorování. Jeho poloha a šíře závisí na apriorním rozložení parametru. V našem jednoduchém případě má rozptyl aposteriorního rozložení tvar $(x+a)(n+b-x)/[(n+a+b)^2 (n+a+b+1)]$, z čehož je patrné, že s rostoucím n jeho hodnota klesá k nule.

Na Obrázku 2 je znázorněna situace lineární regrese s parametry, pro které volíme apriorní pravděpodobnostní rozložení. Uvažujeme jednoduchý lineární

model $E(y_i) = \mu_i = \alpha+\beta x_i$, přičemž y_i mají normální rozložení s průměr μ_i a směrodatnou odchylkou σ. Zvolíme apriorní rozložení pro neznámé parametry:

$\alpha \sim N(m_\alpha, v_\alpha)$, $\beta \sim N(m_\beta, v_\beta)$, $\ln\sigma \sim U(a,b)$.

Pomocí DAG grafu jsem znázornili tento graf na Obrázku 2. Struktura opakování (od i=1, do N) jsou symbolizovány vloženými obdélníky. Řekneme, že uzel v je rodičem dítěte w, jestliže šipka míří od v k w. Zajímáme se o stochastické uzly, o neznámé parametry a data. Výrazy pro rodiče a děti se používají pro odpovídající stochastické jevy. V našem příkladu stochastičtí rodiče pro každé y_i jsou α, β, ln(σ), přičemž μ_i a σ označujeme jako přímé rodiče. Grafy DAG slouží pro grafický popis široké třídy statistických modelů popisující lokální vztahy mezi veličinami. DAG graf vystihuje podstatnou strukturu modelu, aniž bychom museli použít mnoho rovnic. Toto znázornění se využívá pro formulaci potřebných výpočtů v nejznámějším programu pro bayesovskou analýzu WinBugs, který stručně popíšeme ve zvláštním odstavci.



Obrázek 2 – Orientovaný acyklický graf (DAG) pro příklad lineární regrese

Metody Monte Carlo

Bayesovské usuzování se soustřeďuje na aposteriorní rozdělení a některé jeho sumarizující charakteristiky (průměr, modus, rozptyl). Toto rozdělení můžeme znát v jeho explicitní funkční podobě (v našem příkladě se jednalo o beta rozdělení) nebo nemusíme. Přesto, že je známé, může mít takovou podobu, že jeho charakteristiky nelze odvodit analyticky.

Základní myšlenkou integrace metodou Monte Carlo (MC) je počítání přibližného výsledku, tím že mnohokrát simulujeme zkoumaný model. Musíme předpokládat, že známe pravděpodobnostní rozdělení, které model využívá. Předpokládejme, že známe funkční podobu aposteriorního rozdělení zkoumaného parametru θ. Pak provedeme z tohoto rozložení výběr n hodnot θ_1, θ_2, ..., θ_n. Pomocí těchto hodnot můžeme pomocí Monte Carlo integrace vypočítat zvolenou sumární charakteristiku zkoumaného rozložení θ. Například aposteriorní průměr tohoto rozložení odhadneme jako aritmetický průměr z nasimulovaných hodnot.

Zjištění aposteriorní hustoty parametru, která nám umožní generovat náhodné realizace představuje většinou složitý problém. Existuje speciální případ, kdy tento problém nenastává. jedná se o tzv. konjugované modely. Apriorní rozložení f(θ | **η**) jako funkce hyperparametrů **η** označujeme za konjugovanou pro model dat f(D | θ), jestliže aposteriorní rozložení f(θ | D, **η**) patří je stejné třídě rozdělení (příkladem je beta rozložení). Tato vlastnost je velmi užitečná, ale v praktických situacích je podmínka konjugovanosti omezující. Také se často stane, že v datovém modelu jsou rušivé proměnné. Typickým příkladem je normální rozdělení N(μ, τ). Jeho parametry μ a τ jsou obvykle korelovány, takže apriorní rozdělení má tvar f(μ, τ). Přestože máme nejistotu o parametru τ, zajímáme se však pouze o parametr μ. V tomto případě získáme aposterironí hustotu f(μ, τ | D) a rušivého parametru τ se musíme zbavit, tak že spočítáme marginální rozložení parametru μ integrací. Obvykle však taková marginalizace není jednoduchá standardní formou, protože konjugovanost je obvykle narušena korelací mezi parametry. Tento fakt byl důvodem pro to, že bayesovské aplikace se příliš neuplatňovaly v dobách s nízkou dostupností výkonných počítačů. S nástupem výkonných a levných počítačů, bylo možné řešit i složité problémy pomocí simulací. Mezi nejdůležitější simulační techniky patří metoda Monte Carlo markovovskými řetězci (MCMC).

Metody Monte Carlo markovovským řetězcem (MCMC)

Pokud řešíme praktický problém bayesovské statistické analýzy, pak vyjádření aposteriorní funkce je většinou složitým problémem. Obvykle nejsme schopni aposteriorní hustotu vyjádřit přímo jako v případě binomického rozložení. Obtížnost roste s tím, jak roste počet parametrů, které chceme odhadovat. To byl hlavní důvod, proč bayesovský přístup se neujal v běžné statistické praxi. To se změnilo navržením techniky, která využívá simulační metodu pro generování markovovského řetězce, jehož rozložení má tvar hledaného aposteriorního rozložení v bayesovském přístupu. Zkráceně se toto generování označuje MCMC metoda. Existuje několik verzí MCMC generování, proto mluvíme o MCMC metodách. Jestliže získáme simulací výběr vektoru parametrů, pak již lze snadno spočítat jeho průměr, rozptyl a další charakteristiky. Jejich hodnoty se budou blížit k správným hodnotám, pokud nasimulovaný výběr bude dostatečně veliký. To je běžná praxe metody Monte Carlo. Návrhem MCMC simulačních metod můžeme efektivně generovat náhodné výběry s libovolně složitého mnohorozměrného rozložení. MCMC metody vytvářejí

markovovský řetězec, to znamená, že následující hodnota v sérii dat je ovlivněna pouze předcházející hodnotou. Základ MCMC metod spočívá v tom, že limitní vlastnost rozložení hodnot tohoto řetězce je stejná jako požadované mnohorozměrné rozdělení. To znamená, že limitní marginální rozložení hodnot jednotlivých proměnných se bude rovnat přesně odpovídajícím aposteriorním rozložením, takže výběrový průměr simulovaných dat aproximuje správnou hodnotu aposteriorního rozložení. Důležitá je efektivita, s kterou generování dat probíhá ve srovnání s obvyklými numerickými operacemi při hledání maximálně věrohodných odhadů. Obě metody se snaží určitým způsobem mapovat mnohorozměrnou plochu. Hledání jejího vrcholu v případě metody maximální věrohodnosti je zatíženo tím, že jde o mnohem specifičtější úkol, než nalezení objemu distribuce pro určitou oblast hodnot hledaných parametrů, jak to provádějí MCMC algoritmy. Metody maximalizace mnohorozměrné funkce byly stále vylepšovány, ale nikdy zcela nepřekonaly obtíže při hledání absolutního maxima z mnoha možných lokálních maxim. Navržená pravidla a modifikace procesu maximalizace závisí na tvaru zkoumané funkce. Čím je tvar funkce složitější, tím jsou složitější pravidla, takže vždy existuje hranice složitosti, kterou daný algoritmus je schopný řešit.

Mezi nejpopulárnější metody MCMC patří tzv. Gibbsův výběr. Kroky které je potřeba udělat v tomto simulačním postupu jsou následující:

1. Definujeme počáteční hodnotu parametrů. Výběr začíná s touto hodnotu.
2. Provedeme řadu simulací tak, že markovovský řetěz konverguje ke stacionárnímu rozdělení (k požadovanému aposterionímu rozdělení). Obvykle takto získáme několik nasimulovaných řad, které startují z různých počátečních hodnot, aby bylo možné posoudit spolehlivost konvergence.
3. Pokud dosáhneme konvergence (kontrolujeme monitorováním procesu pomocí různých statistik), vybereme náhodně podmnožinu hodnot z odhadnutého rozložení, začneme počítat bodové odhady jako aposterioní průměr nebo medián, případně intervaly kredibility.

Dostupnost MCMC metod znamená, že bayesovský přístup se nemusí zabývat pouze jednoduchými modely. Máme tak prostředek ke zkoumání aposteriorních rozložení parametrů složitých modelů, které zajímají dnešní vědu (Brooks 2003).

OpenBugs a jazyk R

Program OpenBugs je určen pro bayesovskou analýzu složitých statistických modelů pomocí simulací typu MCMC použitím Gibbsova vzorkování. Je určen pro prostředí Windows a je ho možné vyvolat i z prostředí statistického programovacího jazyka R.

Jazyk R je programovatelný statistický systém, který obsahuje velkou paletu statistických prostředků včetně grafiky, prostředků pro simulaci náhodných proměnných, numerickou optimalizaci a vyrovnávání dat. Pro bayesovské výpočty jsou naprogramovány procedury pro Gibbsovo vzorkování a algoritmus Metropolise.

Uživatel programu WinBugs určuje statistický model pomocí stanovení vztahů mezi proměnnými, kde podkladem jsou grafy DAG. Program obsahuje expertní část, která určí vhodnou metodu MCMC simulování pro analýzu daného modelu. Uživatel určuje provedení schématu a rozsah možných výstupů Výhoda modelů založených na DAG grafech v tom, že se tak umožňuje rozštěpit analýzu velkých a složitých struktur na posloupnost jednoduchých výpočtů.

WinBugs vznikl začátkem 80.let minulého století, kdy výzkumníci v oblasti umělé inteligence řešili šíření nejistoty v grafických strukturách. Přitom rozpoznali, že v této souvislosti je možné použít simulaci pro statistické usuzování. Hlavním autorem programu pro oblast statistiky je David Spiegelhalter z Biostatistického institutu v Cambridgi. Shodou okolností v té době se také zabývali autoři Gelfand a Smith v nedalekém městě Nottingham algoritmy MCMC. Ty se ukázaly jako vhodné pro řešení problému simulování v programu WinBugs. S dalšími podrobnostmi vývoje programu WinBugs seznamuje publikace autorů Lunn et al. (2009). Další informace o jazyku R a WinBugs lze získat na internetových stránkách:

webpages *www.r-project.org and www.mrc-bsu.cam.ac.uk/bugs/*.

Program WinBugs používá vlastní jazyk BUGS pro textovou specifikaci grafického modelu. Má deklarativní charakter. Stochastické vztahy jsou symbolizovány znakem ‘ □ ’ , logické a deterministické vztahy symbolem ‘<-’. Opakované struktury jsou vtvořeny slovy ‘for-loops’, které mohou být vložené do jiných struktur.

Uvádíme kod pro řešení lineární regrese, jejíž zadání bylo popsáno Obrázkem 2.

```
model {
for (i in 1:N) {
y[i] ~ dnorm(mu[i], tau)
mu[i] <- alpha + beta * x[i]
}
alpha ~ dnorm(m.alpha, p.alpha)
beta ~ dnorm(m.beta, p.beta)
log.sigma ~ dunif(a, b)
sigma <- exp(log.sigma)
sigma.sq <- pow(sigma, 2)
tau <- 1 / sigma.sq
p.alpha <- 1 / v.alpha
p.beta <- 1 / v.beta
}
```

Lunn et al. (2009) k tomuto kódu poznamenávají, že funkce dnorm(.,.) používá jako parametr přesnost (1/rozptyl) místo rozptylu. Data y[1:N], x[1:N], N a hodnoty m.alpha, v.alpha, m.beta, v.beta, a a b se do programu nahrávají zvlášť. Funkce pow[.,.] umocňuje první argument druhým argumentem funkce.

Závěry

Výzkumníci by měli uvažovat alternativní statistický přístupy při řešení svých problémů. bayesovské metody ovšem vyžadují hlubší znalosti teorie pravděpodobnosti a také určité výpočetní dovednosti. Přesto se námaha spojená s osvojením těchto znalostí a dovedností vyplatí. Existují dva důvody pro tuto domněnku. Z filosofického hlediska klasický přístup je zatížen určitými filosofickými nedostatky, na některé z nich jsme upozornili. Bayesovský přístup poskytuje obecnější rámec pro statistické uvažování. Z pragmatického hlediska se výzkumníci často setkávají s dvěma typickými vlastnostmi ve svých úlohách:

1. Jsou dostupné určité apriorní znalosti zkoumaného fenoménu, které lze efektivně využít v rámci bayesovského přístupu.
2. Kromě systémové variability popsané nějakou dostupnou teorií, existuje v sociálních i biologických vědách značná individuální variabilita. Bayesovské metody mohou v případech poskytnout kvalitnější rámec pro naše usuzování.

Z těchto důvodů je oprávněné se domnívat, že porozumění problému je dokonalejší, pokud rámec modelu vychází z principů bayesovského přístupu. Výpočty jsou v tomto případě složitější a také neexistuje standardní software, kdy by bylo možné zmáčknout jednu klávesu k získání výsledku. To však nutí uživatele, aby promyslel na tři aspekty:

- Jaké jsou otázky, které ze nebo nelze zodpovědět dostupnými daty?
- Jaké máme informace o problému a jak je můžeme využít při jeho řešení?
- jaká je pravděpodobnostní struktura, která by nejlépe modelovala nejistotu spojenou s problémem?

Aplikace bayesovských metod jsou dnes mnohem přístupnější, protože jsou k dispozici volně přístupné programy a systémy jako WinBugs a R (Congdon 2003).

Literatura

[1.] Ashby D.: Bayesian statistics in medicine: a 25-year review. Stat Med 2006; 25:3589–3631.

[2.] Bayes T.: Essay towards solving a problem in the doctrine of chances. Phil. Trans. R. Soc. Londn , 1763.

[3.] Brooks S.P.: Bayesian computation: a statistical revolution. Phil. Trans. R. Soc. Lond A 2003, 361, 2681–2697

[4.] Congdon, P.: Applied Bayesian modelling. Chichester, UK: John Wiley, 2003.

[5.] Gelman, A, Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman & Hall, 1995.

[6.] Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D.: WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 2000, 10, 325-337.

Kontakt:

Doc. RNDr. Jan Hendl, CSc.
Fakulta tělesné výchovy a sportu Univerzity Karlovy,
katedra základů kinantropologie
José Martiho 31
162 52 Praha 6
e-mail: *hendl@ftvs.cuni.cz*