BAYESIÁNSKÁ ANALÝZA – CVIČENÍ 3

Toto cvičení je založeno na znalosti čtvrté kapitoly z učebnice Koop (2003): *Bayesian econometrice*, případně na odpovídající kapitole podkladového učebního textu *Bayesiánská analýza*.

Co bude náplní cvičení?

- ✎ Odhad a posteriorní analýza normálního lineárního regresního modelu s nezávislou normální-gama apriorní hustotou a s omezeními ve tvaru nerovnosti.
- ✎ Osvojení si Gibbsova vzorkovače.

Zadání příkladů

1. *Gibbsův vzorkovač a jeho vlastnosti*: Předpokládejme elementární příklad modelu, kdy při odhadu jeho parametrů získáváme posteriorní hustotu odpovídající dvourozměrnému normálnímu rozdělení (to, že máme známou posteriorní hustotu bude sloužit k tomu, že výsledky simulátorů budeme schopni porovnat s výsledky analytickými):
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$
kde $|\rho| < 1$ je známá posteriorní korelace mezi parametry θ_1 a θ_2.
   (a) Začněte si vytvářet skript, kdy si nejprve zadefinujete počet generovaných vzorků S (např. $S = 10000$), vytvořte vektor středních hodnot μ jakožto nulový vektor (později ho můžete změnit), korelační koeficient ρ a kovarianční matici Σ, kdy na diagonále budou jedničky a mimo ní koeficienty ρ (díky tomu odpovídá kovarianční matice přímo i korelační matici).
   (b) Vytvořte si vlastní funkci, která využívá Monte Carlo integraci k výpočtu posteriorní střední hodnoty, směrodatné odchylky parametrů θ_1 a θ_2 a numerické standardní chyby (NSE) pro střední hodnotu a rozptyl odhadu parametrů, popřípadě využijte dodanou funkci `MC_int.m`. V rámci této funkce můžete využít generátor náhodných čísel z vícerozměrného rozdělení buď přímo dostupný v rámci statistického toolboxu Matlabu (funkce `mvnrnd.m`) nebo funkci LeSageho ekonometrického toolboxu `norm_rnd.m`. *Poznámka:* Je třeba si uvědomit, že Monte Carlo integrace nevyužívá nic jiného než zákon velkých čísel, který říká, že výběrový průměr nějaké funkce parametrů bude konvergovat ke střední hodnotě této funkce parametrů pro rostoucí velikost vygenerovaných vzorků rozdělení, z něhož tyto parametry pocházejí. NSE je definováno jako $\sqrt{\frac{\sigma_g^2}{S}}$, kde σ_g^2 je rozptyl funkce parametrů (nahrazován odhadem), který nás zajímá.
   (c) Vytvořte funkci, která budeý využívat Gibbsův vzorkovač k výpočtu střední hodnoty a směrodatné odchylky parametrů θ_1 a θ_2 (využijte vlastností vícerozměrného normálního rozdělení pro výpočet odpovídajících podmíněných hustot pravděpodobnosti).
      i. Vyjděte z funkce sdružené hustoty normálního rozdělení (dvojrozměrné), se střední hodnotou $\mu = (\mu_1, \mu_2)'$ a kovarianční maticí Σ definovanou v úvodu příkladu:
$$\frac{1}{(2\pi)^{\frac{2}{2}}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\theta - \mu)' \Sigma^{-1} (\theta - \mu) \right]$$
      ii. Odvoďtě podmíněné hustoty pro $\theta_1|\theta_2$ a $\theta_2|\theta_1$
      iii. Pokud jste šikovnější, odvoďte podmíněné hustoty pro obecné k-rozměrné normální rozdělení s arbitrárním dělením vektoru středních hodnot a kovarianční matice (zde je dobré využít teorém o inverzi a determinantu dělené matice z přílohy učebního textu či Koopovy učebnice). Důkaz inverze dělené matice lze nalézt např. zde a odvození podmíněných hustot zde.

iv. Výsledné funkce podmíněných hustot pravděpodobnosti jsou $\theta_1|\theta_2 \sim N(\mu_{(1|2)}, \Sigma_{(1|2)})$ a $\theta_2|\theta_1 \sim N(\mu_{(2|1)}, \Sigma_{(2|1)})$:

$$\begin{aligned}\mu_{(1|2)} &= \mu_1 + \rho * (\theta_2 - \mu_2)\\ \Sigma_{(1|2)} &= 1 - \rho^2\\ \mu_{(2|1)} &= \mu_2 + \rho * (\theta_1 - \mu_1)\\ \Sigma_{(2|1)} &= 1 - \rho^2\end{aligned}$$

(d) Nastavte $\rho = 0$ pro porovnání výsledků z části (a) a (b). Kolik replikací je nutných k odhadu středních hodnot a směrodatných odchylek parametrů θ_1 a θ_2 s přesností na dvě desetinná místa?

(e) Zopakujte část (c) pro $\rho = 0.5, 0.9, 0.95, 0.99, 0.999$. Jak velikost korelace ovlivní výkonnost Gibbsova vzorkovače? Pro srovnání vykreslete průběh Gibbsova vzorkovače pro prvních 50 a 1000 iterací a rovněž i výslednou sekvenci vzorků po odstranění počátečních S_0 vzorků, a to pro odlehlé počáteční hodnoty parametrů.

2. Soubor `data_cocaine.mat` obsahuje 56 pozorování proměnných vztahujících se k prodeji kokainu v severovýchodní Kalifornii v období 1984-1991. Data jsou podmnožinou dat použitých ve studii Culkins, J.P. a Padman, R. (1993): „Quantity Discounts and Quality Premia for Illicit Drugs," *Journal of the American Statistical Association*, 88, 748-757. Proměnné jsou načteny v rámci výchozího skriptu `cocaine.m` a mají následující význam:

- $price$ = cena za gram kokainu v rámci dané transakce;
- $quant$ = počet gramů kokainu prodaných v dané transakci;
- $qual$ = kvalita kokainu vyjádřená jako procento čistoty;
- $trend$ = časová proměnná s hodnotami od 1984=1 až po 1991=8.

Předpokládejme regresní model

$$price = \beta_0 + \beta_1 quant + \beta_2 qual + \beta_3 trend + \epsilon.$$

(a) Jaká znaménka koeficientů byste očekávali u parametrů β_1, β_2 a β_3?

(b) Odhadněte daný model (předpokládáme, že se jedná o NLRM s nezávislou normální gama aprioní hustotou). Zvolte si vhodné hyperparametry dle vašich zkušeností. Jsou znaménka parametrů v souladu s vašim očekáváním?

(c) Říká se, že čím větší objem obchodů, tím větší riziko, že vás dostihne ruka zákona. Prodejci tak jsou ochotni akceptovat nižší cenu, pokud prodávají větší množství. Pokuste se testovat tuto hypotézu.

(d) Ověřte hypotézu, že kvalita kokainu nemá vliv na jeho cenu.

(e) Jaká je průměrná roční změna ceny kokainu? Zamyslete se nad tím, proč by se měla cena takto měnit.

(f) Odhadněte výchozí model se zahrnutím o předpokladech na znaménka parametrů dle vašich očekávání a otestujte znovu vybrané hypotézy.

3. Každé ráno mezi 6:30 a 8:00 opouští Bill Melbournské předměstí Carnegie, aby se dostal do práce na University of Melbourne. Čas, který Bill stráví cestou do práce, $time$, závisí na času odjezdu, $depart$, počtu červených světel na semaforech, $reds$ a počtu vlaků, kvůli kterým musí čekat na Murrumbeenském přejezdu, $trains$. Pozorování těchto proměných je celkem získáno za 231 pracovních dní v roce 2006 a jsou obsahem souboru `data_commute.mat`, resp. jsou jíž načtena v rámci skriptu `commute.m`. Proměnná $time$ je měřena v minutách, $depart$ je počet minut po 6:30, které uplynou než Bill vyrazí z domu.

   (a) Odhadněte rovnici (v kontextu NLRM s nezávislou normální gama apriorní hustotou)

   $$time = \beta_0 + \beta_1 depart + \beta_2 reds + \beta_3 trains + \epsilon.$$

   (b) Jaká znaménka koeficientů byste očekávali u parametrů β_1, β_2 a β_3?

   (c) Otestujte hypotézu, že každé červené světlo zpozdí Billa nejméně o 2 minuty.

   (d) Testujte hypotézu, že čas odjezdu nemá vliv na čas strávený cestováním.

   (e) Otestujte hypotézu, čas cestování navíc díky čekání na jednom semaforu je stejný jako čas čekání průjezdu jednoho vlaku.

   (f) Odhadněte výchozí model se zahrnutím o předpokladech na znaménka parametrů dle vašich očekávání a otestujte znovu vybrané hypotézy.