Cvičení č. 4, od 6.5.2024:

Datový soubor `Cars2004.RData` (k přímému načtení do R) obsahuje údaje o automobilech dostupných v roce 2004 na trhu v USA. V dalším se budeme zabývat problémem vyhodnocení závislosti prodejní ceny (`price.retail`) na následujících charakteristikách jednotlivých aut: spotřeba ve městě (`cons.city`), spotřeba na dálnici (`cons.highway`), objem motoru (`engine.size`), koňská síla (`horsepower`), hmotnost (`weight`), obvod kola (`wheel.base`), délka (`length`), šířka (`width`), počet válců (`ncylinder`). Ze souboru předem vyřaďte vozy s hybridním, respektive rotačním motorem (`fhybrid` roven `"Yes"`, respektive `ncylinder` roven `−1`). Výsledná data by měla obsahovat údaje o $n = 423$ vozech.

Jako Y_i označme prodejní cenu $i$tého vozu v 1 000 USD (tj. `price.retail` vydělené 1 000), jako $\mathbf{X}_i = \big(X_{i,1}, \ldots, X_{i,p}\big)^\top$ $(p = 9)$ výše zmíněné charakteristiky. Uvažujte následující (hierarchický) model:

$$\begin{aligned} \mathbf{X}_i &\sim \mathcal{N}_p\big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big), && i = 1, \ldots, n, \\ Y_i \,|\, \mathbf{X}_i &\sim \mathcal{N}\big(\beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta},\, \sigma^2\big), && i = 1, \ldots, n, \end{aligned}$$

$\big(Y_i,\, \mathbf{X}_i^\top\big)^\top$ nezávislé pro $i = 1, \ldots, n$. Primárním parametrům tedy odpovídají: β_0, $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_p\big)^\top$, σ^2, $\boldsymbol{\mu} = \big(\mu_1, \ldots, \mu_p\big)^\top$, $\boldsymbol{\Sigma}$.

Jako apriorní rozdělení uvažte rozdělení založené na následujícím rozkladu (apriorní nezávislost):

$$p\big(\beta_0,\, \boldsymbol{\beta},\, \sigma^2,\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big) = p(\beta_0)\, p(\boldsymbol{\beta})\, p(\sigma^2)\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma}),$$

kde (náhodný hyperparametr λ je uveden explicitně v podmínce):

$$\beta_0 \sim \mathcal{N}(80,\, 100^2),$$

$$\boldsymbol{\beta} \,|\, \lambda \sim \mathcal{N}_p\big(\mathbf{0},\, \lambda^{-1}\, \mathrm{I}_p\big),$$
$$\lambda \sim \mathcal{G}(1,\, 0.005).$$

$$\mu_{\texttt{cons.city}} \sim \mathcal{N}(10,\, 10^2), \quad \mu_{\texttt{cons.highway}} \sim \mathcal{N}(10,\, 10^2), \quad \mu_{\texttt{engine.size}} \sim \mathcal{N}(3,\, 10^2),$$
$$\mu_{\texttt{horsepower}} \sim \mathcal{N}(200,\, 1\,000^2), \quad \mu_{\texttt{weight}} \sim \mathcal{N}(1\,500,\, 1\,000^2), \quad \mu_{\texttt{wheel.base}} \sim \mathcal{N}(250,\, 100^2),$$
$$\mu_{\texttt{length}} \sim \mathcal{N}(500,\, 100^2), \quad \mu_{\texttt{width}} \sim \mathcal{N}(10,\, 100^2), \quad \mu_{\texttt{ncylinder}} \sim \mathcal{N}(5,\, 10^2),$$

$$\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_p(9,\, \mathrm{diag}(0.001, \ldots, 0.001)).$$

Pro reziduální rozptyl (či jeho přímou transformaci) uvažte postupně dvě možné apriorní volby:

(i) $\sigma^{-2} \sim \mathcal{G}(1,\, 0.005)$ (klasické gama rozdělení pro inverzní rozptyl);

(ii) $\sigma \sim \mathcal{U}(0.1,\, 100)$ (rovnoměrné rozdělení pro směrodatnou odchylku).

Pro každou z apriorních voleb proveďte následující kroky (pokud řešení některého z kroků nezávisí na volbě apriorního rozdělení pro parametr σ^2, stačí ho uvést jednou…).

1. Nakreslete (stačí rukou na papír) orientovaný acyklický graf (DAG) popisující uvažovaný model včetně apriorního rozdělení.

2. Odvoďte (stačí rukou na papír) plně podmíněnou hustotu (stačí tvar hustoty známý až na multiplikativní konstantu) pro vektorový parametr $\boldsymbol{\beta}$ (regresní koeficienty kromě absolutního členu), která by byla použita v rámci Gibbsova algoritmu, jestliže by vektor regresních koeficientů $\boldsymbol{\beta}$ byl generován najednou v rámci jednoho z kroků Gibbsova algoritmu. Je potřeba provádět dvě různá odvození při různých apriorních rozděleních pro reziduální rozptyl σ^2?

   Na základě odvozeného podiskutujte o roli parametru λ v uvažovaném modelu.

3. Implementujte výše uvedený model v JAGSu a vygenerujte dva markovské řetězce (pro každou z voleb apriorního rozdělení parametru σ^2), jejichž limitním rozdělením bude aposteriorní rozdělení pro uvažovaný model.

4. Nakreslete trajektorie (`traceplots`) pro parametry β_0, $\boldsymbol{\beta}$, σ, $\boldsymbol{\mu}$ a dále pro parametr λ a pro devianci modelu (kreslete oba řetězce do jednoho obrázku dvěma různými barvami). Nakreslete odhady autokorelačních funkcí pro regresní koeficienty β_0 a $\boldsymbol{\beta}$ (pro alespoň jeden z vygenerovaných řetězců).

   Posuďte, zda lze předpokládat použitelnost markovského řetězce pro aposteriorní inferenci.

5. Pro parametry β_0, všechny složky vektoru $\boldsymbol{\beta}$, parametr σ a parametr λ spočtěte základní charakteristiky aposteriorního rozdělení, 95% HPD věrohodnostní intervaly a nakreslete odhady aposteriorních hustot. Číselné hodnoty uveďte ve formě vhodné tabulky, ze které bude možné snadno porovnat výsledky při dvou apriorních volbách pro parametr σ^2. Taktéž aposteriorní hustoty kreslete tak, aby bylo možné snadno porovnávat výsledky získané při různých apriorních volbách.

   Které charakteristiky auta ovlivňují statisticky významně prodejní cenu auta, po očištění od možného vlivu zbývajících charakteristik?

6. Pro složky vektoru $\boldsymbol{\beta}$ uveďte klasické 95% intervaly spolehlivosti založené na odhadu normálního lineárního modelu metodou nejmenších čtverců. Liší se výrazně šířky některých/všech těchto intervalů od šířek HPD věrohodnostních intervalů? Jste schopni nalézt vysvětlení možných odlišností? Liší se nějak též závěry týkající se statistické významnosti vlivu jednotlivých charakteristik auta na prodejní cenu?

7. Spočtěte bayesovský bodový i intervalový (95%) odhad korelačního koeficientu mezi spotřebou ve městě a na dálnici, resp. mezi obvodem kola a délkou auta. Nakreslete odhady aposteriorních hustot pro oba tyto korelační koeficienty. Aposteriorní hustoty opět kreslete tak, aby bylo možné snadno porovnávat výsledky získané při různých apriorních volbách.

8. Pomocí metody Monte Carlo nakreslete marginální, tj. po vyintegrování charakteristik auta reprezentovaných náhodným vektorem $\mathbf{X}$, prediktivní hustotu prodejní ceny auta (reprezentované náhodnou veličinou Y). Spočtěte též související 95% HPD věrohodnostní interval (v tomto kontextu se jedná o analogii predikčního intervalu). Na základě posouzení prediktivní hustoty komentujte, zda uvažovaný model netrpí nějakými (zjevnými) nedostatky.

Deadline pro odevzdání vypracovaného úkolu (e-mailem na `komarek[AT]karlin.mff...`) je pondělí 20.5. v 08:05 CEST.

Exercise #4, since 06/05/2024:

The data file `Cars2004.RData` (to be read directly in R) contains data on cars available in the U.S. market in 2004. In the following, we will address the problem of evaluating the dependence of the retail price (`price.retail`) on the following characteristics of individual cars: city consumption (`cons.city`), highway consumption (`cons. highway`), engine size (`engine.size`), horsepower (`horsepower`), weight (`weight`), wheel circumference (`wheel.base`), length (`length`), width (`width`), number of cylinders (`ncylinder`). Before starting the analyzis, exclude cars with a hybrid or rotary engine from the dataset (`fhybrid` equals `"Yes"` or `ncylinder` equals -1). The resulting data should contain data on $n = 423$ cars.

Let Y_i be the retail price of the ith car in 1 000 USD (that is, `price.retail` divided by 1 000). Let $\mathbf{X}_i = \big(X_{i,1}, \ldots, X_{i,p}\big)^\top$ ($p = 9$) be the above characteristics. Consider the following (hierarchical) model:

$$\begin{aligned} \mathbf{X}_i &\sim \mathcal{N}_p\big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big), & i &= 1, \ldots, n, \\ Y_i \,|\, \mathbf{X}_i &\sim \mathcal{N}\big(\beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta},\, \sigma^2\big), & i &= 1, \ldots, n, \end{aligned}$$

$\big(Y_i,\, \mathbf{X}_i^\top\big)^\top$ independent for $i = 1, \ldots, n$. Hence the primary parameters: β_0, $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_p\big)^\top$, σ^2, $\boldsymbol{\mu} = \big(\mu_1, \ldots, \mu_p\big)^\top$, $\boldsymbol{\Sigma}$.

As a prior distribution, consider the distribution based on the following decomposition (prior independence):

$$p\big(\beta_0,\, \boldsymbol{\beta},\, \sigma^2,\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big) = p(\beta_0)\, p(\boldsymbol{\beta})\, p(\sigma^2)\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma}),$$

where (the random hyperparameter λ is explicitely mentioned in the condition):

$$\beta_0 \sim \mathcal{N}(80,\, 100^2),$$

$$\boldsymbol{\beta} \,|\, \lambda \sim \mathcal{N}_p\big(\mathbf{0},\, \lambda^{-1}\, \mathrm{I}_p\big),$$

$$\lambda \sim \mathcal{G}(1,\, 0.005).$$

$$\mu_{\texttt{cons.city}} \sim \mathcal{N}(10,\, 10^2), \quad \mu_{\texttt{cons.highway}} \sim \mathcal{N}(10,\, 10^2), \quad \mu_{\texttt{engine.size}} \sim \mathcal{N}(3,\, 10^2),$$

$$\mu_{\texttt{horsepower}} \sim \mathcal{N}(200,\, 1\,000^2), \quad \mu_{\texttt{weight}} \sim \mathcal{N}(1\,500,\, 1\,000^2), \quad \mu_{\texttt{wheel.base}} \sim \mathcal{N}(250,\, 100^2),$$

$$\mu_{\texttt{length}} \sim \mathcal{N}(500,\, 100^2), \quad \mu_{\texttt{width}} \sim \mathcal{N}(10,\, 100^2), \quad \mu_{\texttt{ncylinder}} \sim \mathcal{N}(5,\, 10^2),$$

$$\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_p(9,\, \mathrm{diag}(0.001, \ldots, 0.001)).$$

For the residual variance (or its direct transformation), consider consecutively two possible prior choices:

(i) $\sigma^{-2} \sim \mathcal{G}(1,\, 0.005)$ (classical gamma distribution for the inverted variance);

(ii) $\sigma \sim \mathcal{U}(0.1,\, 100)$ (uniform distribution for the standard deviation).

For each of the the prior choices, perform the following steps (if the solution to any of the steps does not depend on the choice of the prior distribution for the parameter σ^2, do that just once...).

1. Draw (just by hand) a Directed Acyclic Graph (DAG) describing the model under consideration.

2. Derive (just by hand on paper) the full conditional density (just the form of the density known up to a multiplicative constant) for the vector $\boldsymbol{\beta}$ (regression coefficients except the intercept term) that would be used in the Gibbs algorithm if the vector of regression coefficients $\boldsymbol{\beta}$ were to be generated jointly within one step of the Gibbs algorithm. Is it necessary to perform two different derivations with different prior distributions for the residual variance of σ^2?

   Based on the derivation, discuss the role of the parameter λ in the model under consideration.

3. Implement the above model in JAGS and generate two Markov chains (for each of the choices of the prior distribution of the parameter σ^2) whose limiting distribution is the posterior distribution for the model under consideration.

4. Draw trajectories (`traceplots`) for the parameters β_0, $\boldsymbol{\beta}$, σ, $\boldsymbol{\mu}$ and also for the parameter λ and for the deviance of the model (draw both chains in one plot with two different colors). Draw estimates of the autocorrelation functions for the regression coefficients β_0 and $\boldsymbol{\beta}$ (for at least one of the generated chains).

   Assess whether it can be assume that generated chains are applicable for the posterior inference.

5. For parameter β_0, all components of the vector $\boldsymbol{\beta}$, the parameter σ, and the parameter λ, compute the basic characteristics of the posterior distribution, 95% HPD credible intervals, and plot the estimates of the posterior densities. Report the numerical values in a form of a suitable table from which the results for the two prior choices for the σ^2 parameter can be easily compared. Also, draw the aposterior densities in such a way that the results obtained at different prior choices can be easily compared.

   Which car characteristics have a statistically significant effect on the retail price of a car, after adjusting for the possible effect of the remaining characteristics?

6. For elements of the vector $\boldsymbol{\beta}$, provide classical 95% confidence intervals based on the least squares estimation in the normal linear model. Do the widths of any/all of these intervals differ significantly from the widths of the HPD confidence intervals? Are you able to find explanations for the possible differences? Do the conclusions regarding the statistical significance of the effect of individual car characteristics on the retail price also differ in any way?

7. Calculate the Bayesian point as well as interval (95%) estimate of the correlation coefficient between city and highway speeds, or between wheel circumference and car length. Plot the estimates of the posterior densities for both of these correlation coefficients. Again, plot the posterior densities in such a way that it is easy to compare the results obtained with different prior choices.

8. Use the Monte Carlo method and plot the marginal, i.e., after integrating out the characteristics of the car represented by the random vector $\mathbf{X}$, predictive density of the retail price of the car (represented by the random variable Y). Also calculate the associated 95% HPD credible interval (in this context, it is an analogy of the prediction interval). Based on the assessment of the predictive density, comment on whether the model under consideration suffers from any (obvious) deficiencies (or not).

Deadline to deliver the report (e-mail to `komarek[AT]karlin.mff...`): Monday 20 May at 08:05 CEST.