Exam 2 Practice Problems

Math 315, Fall 2019

1. Explain why we prefer to use DIC or WAIC to compare models instead of in-sample deviance.

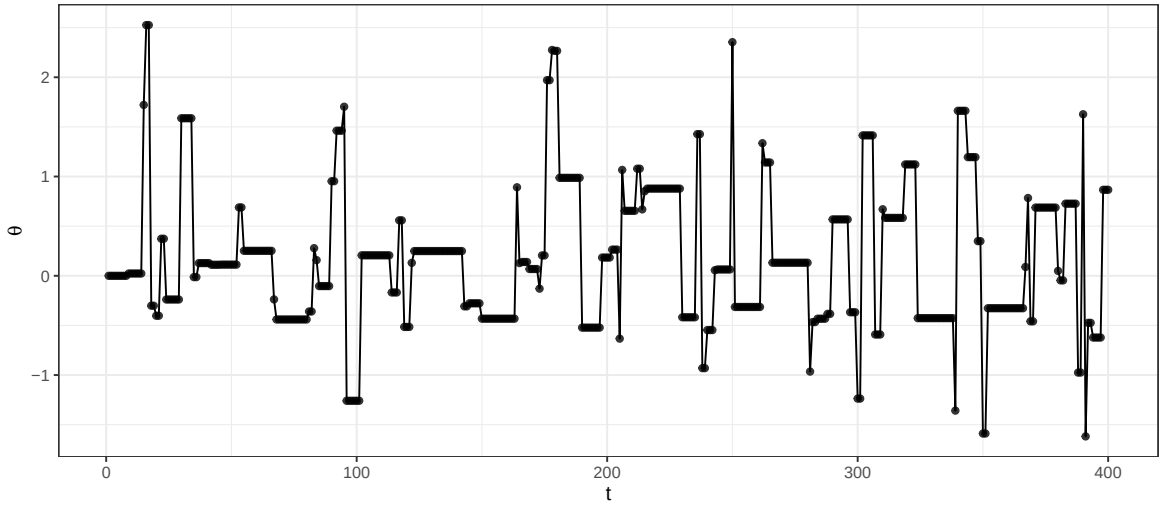
2. Explain what overfitting is and describe one strategy to avoid it.

3. Suppose $y_1, \ldots, y_n$ form a random sample from $\mathcal{N}(\mu, \sigma^2)$. The joint posterior distribution that results from the reference prior is

$$p(\mu, \sigma^2 | y_1, \ldots, y_n) \propto \left(\sigma^2\right)^{-n/2-1} \exp\left\{\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$$

   Find $p(\sigma^2 | \mu, y_1, \ldots, y_n)$, the conditional posterior of σ^2 given μ and the data. If it is a member of a named family of distributions, be sure to specify this, along with its parameter values.

4. The figure below is a trace plot from 400 steps of an MCMC (Metropolis) run.



   (a) Is the acceptance rate: too high, too low, or just right? Briefly explain your reasoning.

   (b) If the acceptance rate for a random walk Metropolis algorithm using a normal proposal (jump) density is too high, how should the standard deviation be adjusted?

5. Suppose that you have a random sample, $x_1, \ldots, x_n$, from a Galenshore distribution with PDF

$$f(x_i|\theta) = \frac{2}{\Gamma(a)}\theta^{2a}x_i^{2a-1}e^{-\theta^2 x_i^2}$$

where $x_i, \theta > 0$ and a is a known constant. Further, suppose that you put a Gamma$(3, 1)$ prior on θ.

   (a) Derive the unnormalized posterior distribution for θ.

   (b) Describe a method for obtaining draws, $\theta^{(1)}, \ldots, \theta^{(m)}$, from the posterior distribution. If helpful, you may use R function names, but you need to also describe the process.

6. Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program. The maximum oxygen uptake of each subject was measured (in liters per minute) while running on an inclined treadmill, both before and after the 12-week program. Of interest is how a subjects change in maximal oxygen uptake may depend on which program they were assigned to. However, other factors, such as age, are expected to affect the change in maximal uptake as well.

   The researchers considered the following five models:

Model	μ_i
m1	$\mu_i = \alpha$
m2	$\mu_i = \alpha + \beta_1 \texttt{group}_i$
m3	$\mu_i = \alpha + \beta_2 \texttt{age}_i$
m4	$\mu_i = \alpha + \beta_1 \texttt{group}_i + \beta_2 \texttt{age}_i$
m5	$\mu_i = \alpha + \beta_1 \texttt{group}_i + \beta_2 \texttt{age}_i + \beta_3 \texttt{group}_i \times \texttt{age}_i$

   (a) Below is (slightly modified) output from the `compare(m1, m2, m3, m4, m5)`. Based on this information, which model do you prefer? Why?

	WAIC	SE
m1	97.41	9.63
m2	89.23	7.96
m3	75.20	9.99
m4	70.78	11.99
m5	78.72	12.68

7. Suppose that you have been recruited to create a regression model to predict the price of dinner in New York City to help set prices at a new restaurant. Using data from a recent Zagat survey, you fit a multiple linear regression model with the following mean function:

$$\mu(\texttt{price}_i|\texttt{X}_i) = \alpha + \beta_1\texttt{food} + \beta_2\texttt{decor} + \beta_3\texttt{service},$$

where the `food`, `decor`, and `service` variables are average customer ratings out of 30 points, and `price` is recorded in dollars. You fit this multiple linear regression model using the reference prior distribution for multiple linear regression and obtain the following results:

Parameter	Mean	StdDev	5.5%	94.5%
β_0	-24.642	4.697	-32.148	-17.136
β_1	1.556	0.369	0.967	2.145
β_2	1.847	0.215	1.504	2.191
β_3	0.135	0.391	-0.490	0.760
σ	5.734	0.313	5.234	6.234

(a) Give a careful interpretation of the *maximum a posteriori* estimate of β_3, in the context of the problem.

(b) Does `service` appear to be an important predictor of price, after controlling for `food` and `decor`? Justify your answer.