Tutorial and Lab Problems # 11
MATH3871/MATH5970

1. **Bayes Factors.** Consider model 1 with likelihood $\pi(x \,|\, \theta, m = 1)$ and model 2 with $\pi(x \,|\, \theta, m = 2) = \pi(x \,|\, \psi^{-1}(\theta), m = 1)$ for an invertible function ψ. If we adopt Jeffrey's priors $\pi_J(\theta)$ and $\pi_J(\psi(\theta))$ for model 1 and 2 respectively, show that the Bayes factor $B_{1\,|\,2} = B_{2\,|\,1} = 1$.

   Answer: From the definition of Jeffrey's prior we know that

$$\begin{aligned}
\pi_J(\theta) &\propto \sqrt{\mathbb{E}\left(\frac{\partial L}{\partial \theta}\right)^2} \\
&= \sqrt{\mathbb{E}\left(\frac{\partial L}{\partial \psi}\frac{\mathrm{d}\psi}{\mathrm{d}\theta}\right)^2} \\
&= \left|\frac{\mathrm{d}\psi}{\mathrm{d}\theta}\right| \sqrt{\mathbb{E}\left(\frac{\partial L}{\partial \psi}\right)^2} \\
&\propto \pi_J(\psi)|\mathrm{d}\psi/\mathrm{d}\theta| \,.
\end{aligned}$$

   Therefore, by a change of variable, we have:

$$\begin{aligned}
\pi(x \,|\, m = 1) &= \int \pi(x \,|\, \theta, m = 1)\pi_J(\theta)\mathrm{d}\theta \\
&= \int \pi(x \,|\, \psi(\theta), m = 2)\pi_J(\psi(\theta))\frac{\mathrm{d}\psi}{\mathrm{d}\theta}\mathrm{d}\theta \\
&= \int \pi(x \,|\, \psi(\theta), m = 2)\pi_J(\psi(\theta))\mathrm{d}\psi \\
&= \pi_J(x \,|\, m = 2) \,.
\end{aligned}$$

2. **Tute/Lab # 10 revisited (this question not examinable).** Assume that $\mathbf{D} = \alpha^2\mathbf{I}_p$, where $\mathbf{I}_p \in \mathbb{R}^{p\times p}$, and that σ^2 is unknown. Let $\mathbf{C} := \mathbf{I}_n + \alpha^2\mathbf{X}\mathbf{X}^\top$. Show that (use Woodbery matrix identity, see Wikipedia) an alternative formula for $\widehat{\sigma}^2$ is

$$\widehat{\sigma}^2 = \boldsymbol{y}^\top\mathbf{C}^{-1}\boldsymbol{y}/n \,.$$

Also, an alternative formula for the log-determinant of $\Sigma = \alpha^2(\mathbf{I}_p + \alpha^2 \mathbf{X}^\top \mathbf{X})^{-1}$ is

$$\ln|\Sigma| = p\ln(\alpha^2) - \ln|\mathbf{C}| \ .$$

Then, for a model with p predictors, we have

$$\begin{aligned} -2\ln g(\boldsymbol{y}\,|\,p) &= p\ln(\alpha^2) - 2\ln\Gamma(n/2) + n\ln(\pi n\widehat{\sigma}^2) - \ln|\Sigma| \\ &= n\ln(\pi \boldsymbol{y}^\top \mathbf{C}^{-1}\boldsymbol{y}) + \ln|\mathbf{C}| - 2\ln\Gamma(n/2) \ . \end{aligned}$$

Now suppose that m is the maximum number of predictors that we can include in the linear model and denote them by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m (\in \mathbb{R}^n)$. Let $\boldsymbol{z}$ be a binary vector of length m that encodes a particular model under consideration.

For example, with $m = 5$ and $\boldsymbol{z} = (0,1,0,1,1)^\top$, this means that $\mathbf{X} := [\boldsymbol{v}_2, \boldsymbol{v}_4, \boldsymbol{v}_5]$ has three columns ($p = 3$) and includes only predictors $2, 4, 5$. Thus, for a fixed n, the model (as determined by the binary vector $\boldsymbol{z}$) with the highest evidence will be the one minimizing the loss over $\boldsymbol{z} \in \{0,1\}^m$:

$$\ell(\boldsymbol{z}) := n\ln(\boldsymbol{y}^\top [\mathbf{C}(\boldsymbol{z})]^{-1}\boldsymbol{y}) + \ln|\mathbf{C}(\boldsymbol{z})| \ ,$$

where

$$\mathbf{C}(\boldsymbol{z}) := \mathbf{I}_n + \alpha^2 \sum_{k=1}^{m} \mathbb{I}\{z_k = 1\} \boldsymbol{v}_k \boldsymbol{v}_k^\top \ .$$

In fact, assuming uniform prior, the posterior density on the space of all 2^m models with m predictors can be written as:

$$g(\boldsymbol{z}) \propto \mathrm{e}^{-\ell(\boldsymbol{z})}, \qquad \boldsymbol{z} \in \{0,1\}^m \ .$$

(We may exclude the all zero vector $\boldsymbol{z} = (0, \ldots, 0)^\top$, if we want at least one predictor.)

To simulate from this posterior we can use Gibbs sampling, whereby sampling from the conditional $g(z_k \,|\, \boldsymbol{z}_{-k})$ consists of flipping a coin to decide if the k-th component is $z_k = 1$ or $z_k = 0$ (equivalent to keeping or removing the k-th predictor). The probability of success of the coin flip is:

$$g(z_k = 1 \,|\, \boldsymbol{z}_{-k}) = \frac{\mathrm{e}^{-\ell(\boldsymbol{z}; z_k = 1)}}{\mathrm{e}^{-\ell(\boldsymbol{z}; z_k = 1)} + \mathrm{e}^{-\ell(\boldsymbol{z}; z_k = 0)}} \ .$$

To be able to implement the Gibbs sampler efficiently we must be able to update the value of $\ell(\boldsymbol{z})$ after a single change to one of the z's.

In other words, given the inverse $\mathbf{C}^{-1}$ and log-determinant $\ln|\mathbf{C}|$, we want to compute very fast the inverse and determinant of the new updated matrix $\mathbf{C} \pm \boldsymbol{v}\boldsymbol{v}^\top$. This is a rank-1 update/change of $\mathbf{C}$.

Instead of computing $(\mathbf{C} \pm \boldsymbol{v}\boldsymbol{v}^\top)^{-1}$ and $\ln|\mathbf{C} \pm \boldsymbol{v}\boldsymbol{v}^\top|$ from scratch (this will take $\mathcal{O}(n^3)$ cost), we compute these using the formulas (these take $\mathcal{O}(n^2)$ cost):

$$(\mathbf{C} \pm \boldsymbol{v}\boldsymbol{v}^\top)^{-1} = \mathbf{C}^{-1} \mp \frac{\mathbf{C}^{-1}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{C}^{-1}}{1 \pm \boldsymbol{v}^\top\mathbf{C}^{-1}\boldsymbol{v}}$$
$$\ln|\mathbf{C} \pm \boldsymbol{v}\boldsymbol{v}^\top| = \ln|\mathbf{C}| + \ln(1 \pm \boldsymbol{v}^\top\mathbf{C}^{-1}\boldsymbol{v}) \ .$$

This then suggests the following $\mathcal{O}(n^2)$-cost algorithm for updating ℓ when we add a predictor (similarly for removing one).

Algorithm 1 : Updating $\ell(\boldsymbol{z})$ when we add/remove the predictor $\boldsymbol{v}$

Require: Current values in memory: $\boldsymbol{y}^\top\mathbf{C}^{-1}\boldsymbol{y}, \mathbf{C}^{-1}, \boldsymbol{v}, \ln|\mathbf{C}|$

$\boldsymbol{v} \leftarrow \alpha\boldsymbol{v}$
$\boldsymbol{a} \leftarrow \mathbf{C}^{-1}\boldsymbol{v}$
$b \leftarrow 1 \pm \boldsymbol{a}^\top\boldsymbol{v}$ (minus when removing a predictor)
$\ell \leftarrow \mp(\boldsymbol{a}^\top\boldsymbol{y})^2/b$
$\ell \leftarrow \ell + \boldsymbol{y}^\top\mathbf{C}^{-1}\boldsymbol{y}$
$\ell \leftarrow n\ln(\ell) + \ln|\mathbf{C}| + \ln(b)$
return updated value ℓ

I can give an numerical experiment/illustration of the full Bayesian model selection algorithm in the tute/lab.