Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 15: Inference under model uncertainty [a]

Lecturer: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim: To explain, design, and apply model selection and model determination Bayesian procedures.

References:

- Robert, C. (2007; Section 7.1 & 7.2). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- O'Hagan, A., & Forster, J. J. (2004; Section 7 Model comparison). Kendall's advanced theory of statistics, volume 2B: Bayesian inference (Vol. 2). Arnold.

[a] Author: Georgios P. Karagiannis.

1 Model uncertainty problem

Note 1. All (statistical/mathematical) models are approximations/simplifications of the reality. They are arguable. Their purpose is to allow You draw conclusions about reality and take decisions.

- "All models are wrong, but some are useful". –George Box's quote

Note 2. Let $y = (y_1, ..., y_n)$ be observables generated from the real but unknown data-generating process $y \sim R(y)$. Let $z = (y_{n+1}, ..., y_{n+m})$ be future outcomes from the same process $R(\cdot)$.

Note 3. Statistician may be unsure of the proper way to formulate the statistical model $\mathscr{M} = \{F(y|\theta);\, \theta \in \Theta\}$ in order to acceptably represent the data-generating process $R(y)$. We call this situation 'model uncertainty'.

Note 4. To address problems under model uncertainty, often

- statistician initially considers a restricted set $\{\mathscr{M}_k;\, k \in \mathcal{K}\}$ of available - reasonable - potential - competing - alternative statistical models $\mathscr{M}_k = \{F_k(y|\theta_k);\, \theta_k \in \Theta_k\}$. Then,
- interest lies in learning a model $\mathscr{M}$ which 'best' represents $R(y)$, performing inference (predictive or parametric), and drawing conclusions under the pretense of model uncertainty.

Example 5. Consider the Normal linear regression problem. Let $\{\phi_0, \phi_1, ..., \phi_{d-1}\}$ be a set of d possible regressor variables, which may affect the values of the response variable $y \in \mathcal{Y} \subseteq \mathbb{R}$. We are interested in learning the mapping

$$\{\phi_0, ..., \phi_{d-1}\} \overset{?}{\longmapsto} y$$

Assume there are available n pairs of observations: (Φ_1, y_1), (Φ_2, y_2), ..., (Φ_n, y_n).

Consider the following approximations of the real mapping $\{\phi_0, ..., \phi_{d-1}\} \longmapsto y$: Assume that observables y_i are Normally distributed and independent for $i = 1, ..., n$, with unknown variance σ^2, and unknown mean which can be parametrised as a linear combination of some of the regressors $\{\phi_0, ..., \phi_{d-1}\}$ with unknown coefficients β. Finally, consider, we are uncertain which set of $\{\phi_0, ..., \phi_{d-1}\}$ actually affects values of y.

2 Standard framework

Note 6. Consider a collection of statistical models $\mathcal{M} = \{\mathscr{M}_k;\ k \in \mathcal{K}\}$, where each model

$$\mathscr{M}_k = \{F_k(y|\theta_k);\ \theta_k \in \Theta_k\},$$

is labeled by an index $k \in \mathcal{K}$, and associated with the sampling distribution $F_k(y|\theta_k)$ with pdf/pmf $f_k(y|\theta_k)$, unknown parameter $\theta_k \in \Theta_k$, and parametric space Θ_k.

Definition 7. The encompassing model $\mathcal{M}$ is defined as the collection of statistical models

$$\begin{aligned}\mathcal{M} &= \{\mathscr{M}_k;\ k \in \mathcal{K}\} \\ &= \{\{F_k(y|\theta_k);\ \theta_k \in \Theta_k\};\ k \in \mathcal{K}\}\end{aligned}$$

labeled by parameter $\vartheta = (k, \theta_k) \in \Theta$ which is defined on the joint parameter space

$$\Theta = \cup_{k\in\mathcal{K}}\{k\} \times \Theta_k$$

Example 8. (Cont...) We set-up the encompassing model $\mathcal{M} = \{\mathscr{M}_k;\ k \in \mathcal{K}\}$ as the collection of statistical models

$$\mathscr{M}_k = \left\{y|\beta_k, \sigma^2, k \sim \mathrm{N}\left(\Phi_k\beta_k, I\sigma^2\right)\ ;\ \beta_k \in \mathbb{R}^{d_k},\ \sigma^2 \in \mathbb{R}_+\right\}, \tag{1}$$

where $k \subseteq \{0, 1, ..., d-1\}$ is the label of the model indicating the sub-set of the regressors included in the statistical model $\mathscr{M}_k$, and $\mathcal{K}$ is the collection of all sub-sets of $\{0, 1, ..., d-1\}$ so that $k \in \mathcal{K}$. Given $\mathscr{M}_k$, Φ_k is the design matrix consisting of the regressors with indexes in k, $\beta_k \in \mathbb{R}^{d_k}$ is the associated vector of regression coefficients, and d_k is the number of regressors. Notice that σ^2 is kept to be common among all the models, however one could have set it to be different between different models (e.g.; σ_k^2). The within model parameters are $\theta_k = \left(\beta_k, \sigma^2\right)$ defined on space $\Theta_k = \mathbb{R}^{d_k} \times \mathbb{R}$. The parameter of $\mathcal{M}$ is $\vartheta = (k, \beta_k, \sigma^2)$ and defined on space $\Theta = \cup_{k\in\mathcal{K}}\{k\} \times \mathbb{R}^{d_k} \times \mathbb{R}$.

Note 9. To complete the Bayesian model we specify prior distributions on the joint parametric space Θ. Consider a collection of conditional priors on θ_k given model $\mathscr{M}_k$ as

$$\theta_k|\mathscr{M}_k \sim \Pi(\theta_k|\mathscr{M}_k)$$

with pdf/pmf $\pi(\theta_k|\mathscr{M}_k)$; and consider a marginal prior distribution on $\mathscr{M}_k$

$$\mathscr{M}_k \sim \Pi(\mathscr{M}_k)$$

with pdf/pmf $\pi(\mathscr{M}_k)$, for all $k \in \mathcal{K}$. The encompassing prior model has distribution $\Pi(\mathscr{M}_k, \theta_k)$ with pdf/pmf

$$\pi(\mathscr{M}_k, \theta_k) = \pi(\theta_k|\mathscr{M}_k)\pi(\mathscr{M}_k),\ \ \forall(k, \theta_k) \in \Theta$$

Note 10. Specifying the prior distribution $\Pi(\mathscr{M}_k, \theta_k)$ is a delicate matter.

1. Within model priors $\theta_k|\mathscr{M}_k \sim \Pi(\theta_k|\mathscr{M}_k)$ should not be improper because the normalizing constants are not canceled in the joint posterior (7), and hence the Lindley-Jeffreys padadox may kick in.
2. If some models are embedded (nested) into others, e.g. $\mathscr{M}_1 \subseteq \mathscr{M}_2$, possibly the marginal model prior should be $\pi(\mathscr{M}_1) \leq \pi(\mathscr{M}_2)$, in order to be coherent –this is not a panacea.
3. Common parameters between models (such as σ^2 in Example) may be treated as separate parameters. However, it is common to be regarded as the same parameter and be assigned the same prior; this is an additional approximation/simplification that modeler does for computational convenience –less parameters to learn ;-).
    - In the Example, σ^2 is common to all models. If we wanted to be in accordance with this note, we should have specified σ^2 separately for each model $\mathscr{M}_k$, as σ_k^2, but we do not do this now (naughty).

Note 11. The full Bayesian model can be summarized as

$$\begin{cases} y|\theta_k, \mathscr{M}_k \sim F_k(y|\theta) & \text{, data generation fromt he sampling distribution} \\ \theta_k|\mathscr{M}_k \sim \Pi(\theta_k|\mathscr{M}_k) & \text{, within model parameter generation from the conditional prior} \\ \mathscr{M}_k \sim \Pi(\mathscr{M}_k) & \text{, model generation fromt he marginal model prior} \end{cases} \tag{2}$$

hence the joint distribution is such that $\mathrm{d}P(y, \theta_k, \mathscr{M}_k) = \mathrm{d}F_k(y|\theta)\mathrm{d}\Pi(\theta_k|\mathscr{M}_k)\mathrm{d}\Pi(\mathscr{M}_k)$.

Example 12. (Cont...) One can specify the following Bayesian linear regression model

$$y|\beta_k, \sigma^2, \mathscr{M}_k \sim \mathrm{N}\left(\Phi_k\beta_k, I\sigma^2\right) \quad \text{, the sampling distribution} \tag{3}$$

$$\beta_k|\sigma^2, \mathscr{M}_k \sim \mathrm{N}\left(\mu_k, \sigma^2 V_k\right) \quad \text{, conditional prior of the within model parameter} \tag{4}$$

$$\sigma^2|\mathscr{M}_k \sim \mathrm{IG}\left(a, \lambda\right) \quad \text{conditional prior of the within model parameter} \tag{5}$$

$$\mathscr{M}_k \sim \pi(k) = \frac{1}{|\mathcal{K}|} = \frac{1}{2^d} \quad \text{, marginal model prior} \tag{6}$$

Eq. 3 is the sampling distribution $F(y|\beta_k, \sigma^2, \mathscr{M}_k)$ for model $\mathscr{M}_k$. Eq. 4 and 5 are the conditional (or within) model priors $\Pi(\beta_k, \sigma^2|\mathscr{M}_k)$; recall that they are conjugate to $F(y|\beta_k, \sigma^2, \mathscr{M}_k)$ and hence chosen for our computational convenience. Notice that parameter σ^2, which is common to all available models $\{\mathscr{M}_k\}$, is treated as obtaining the same values among all models, and following the same prior. This is against the suggestion in Note 10((3)) but we assume that the computational benefits of reducing the dimensionality of the parametric space dominate the losses from the (possibly) worse approximation of the reality. In Eq. 6, the marginal model prior $\pi(k)$ is chosen to be uniform across models (we have $|\mathcal{K}| = 2^d$ possible combinations or regressors).

Notation 13. To make the notation easier, we will denote $\mathscr{M}_k$ as k and the probability measures as

- $F(y|\theta_k, k) := F_k(y|\theta_k)$ and $f(y|\theta_k, k) := f_k(y|\theta_k)$, as well as
- $\Pi(\theta_k|k) := \Pi(\theta_k|\mathscr{M}_k)$, $\pi(\theta_k|k) := \pi(\theta_k|\mathscr{M}_k)$, $\Pi(k) := \Pi(\mathscr{M}_k)$, and $\Pi(k, \theta_k) := \Pi(\mathscr{M}_k, \theta_k)$, etc...

3 Posterior distributions

According to the Bayesian, the following posteriors can be derived.

Proposition 14. *Given the Bayesian model (2), the joint posterior distribution* $\Pi(k, \theta_k|y)$ *has pdf/pmf*

$$\pi(k, \theta_k|y) = \frac{f(y|\theta_k, k)\pi(\theta_k|k)\pi(k)}{\sum_{k\in\mathcal{K}}\int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k\pi(k)}, \ \ \forall(k, \theta_k) \in \Theta; \tag{7}$$

Proposition 15. *The joint posterior density (7) can be factorized as*

$$\pi(k, \theta_k|y) = \pi(\theta_k|y, k)\pi(k|y)$$

where the first part is the pdf/pmf of the conditional posterior $\Pi(\theta_k|y, k)$ *of* θ_k *given model* $\mathscr{M}_k$ *i.e.*

$$\pi(\theta_k|y, k) = \frac{f(y|\theta_k, k)\pi(\theta_k|k)}{\int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k} = \frac{f(y|\theta_k, k)\pi(\theta_k|k)}{f(y|k)}, \ \forall\theta_k \in \Theta_k\,; \tag{8}$$

and the second term is the pdf/pmf marginal model posterior $\Pi(k|y)$ of model $\mathscr{M}_k$ with pdf/pdf

$$\pi(k|y) = \frac{f(y|k)\pi(k)}{f(y)},\ \forall k \in \mathcal{K}\,.$$

Here, $f(y|k)$ is the conditional marginal likelihood (or the model evidence) of model $\mathscr{M}_k$ with pdf/pdf

$$f(y|k) = \int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k$$

and $f(y)$ is the marginal likelihood with

$$f(y) = \sum_{k\in\mathcal{K}} \left[\int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k\right]\pi(k) = \sum_{k\in\mathcal{K}} f(y|k)\pi(k)$$

Appendix A

Example 16. To calculate the conditional posterior pdf $\pi\left(\beta_k, \sigma^2|y\right)$, we use the Bayes Theorem

$$\begin{aligned}\pi\left(\beta_k, \sigma^2|y, k\right) \propto & f(y|\theta_k, k)\pi(\theta_k|k) \;=\; \mathrm{N}\left(y|\Phi\beta_k, I\sigma^2\right)\mathrm{N}\left(\beta_k|\mu_k, V_k\sigma^2\right)\mathrm{IG}\left(\sigma^2|a, \lambda\right)\\ \propto & \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(\beta_k-\mu_k^*\right)^\top\left(V_k^*\right)^{-1}\left(\mu_k-\beta_k^*\right)\right)}\times\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a+1}\exp\left(-\frac{\lambda}{\sigma^2}\right)}\\ \propto & \mathrm{N}\left(\beta_k|\mu_k^*, V_k^*\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^*, \lambda_k^*\right)\end{aligned}$$

where

$$V_k^* = \left(V_k^{-1}+\Phi_k^\top\Phi_k\right)^{-1};\qquad \mu_k^* = V_k^*\left(V_k^{-1}\mu_k+\Phi_k^\top y\right)\mathscr{M}_k;\quad a^* = \frac{n}{2}+a;\quad \lambda_k^* = \lambda+\frac{1}{2}S_k^*;$$
$$S_k^* = \mu_k^\top V_k^{-1}\mu_k - \left(\mu_k^*\right)^\top\left(V_k^*\right)^{-1}\left(\mu_k^*\right) + y^\top y$$

Hence

$$\implies \begin{cases}\beta_k|y\sigma^2, k & \sim \mathrm{N}\left(\mu_k^*, V_k^*\sigma^2\right)\\ \sigma^2|y, k & \sim \mathrm{IG}\left(a^*, \lambda^*\right)\end{cases},\ \forall k\in\mathcal{K}$$

Appendix B

Example 17. To calculate the model evidence / conditional marginal likelihood for $\mathscr{M}_k$, we integrate

$$f(y|k) = \int \overbrace{\mathrm{N}\left(y|\Phi_k\beta_k, I\sigma^2\right)}^{f(y|\theta_k,k)}\overbrace{\mathrm{N}\left(\beta_k|\mu, V_k\sigma^2\right)\mathrm{IG}\left(\sigma^2|a,\lambda\right)}^{\pi(\theta_k|k)}\mathrm{d}\beta_k\mathrm{d}\sigma^2 = \text{...calc...} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\frac{\Gamma\left(a^*\right)}{\left(\lambda_k^*\right)^{a^*}}$$

Since $\mathcal{K}$ is finite, the marginal likelihood $f(y)$ of y can be computed as

$$f(y) = \sum_{k\in\mathcal{K}} f(y|k)\pi(k) \;=\; \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\lambda^a\frac{\Gamma\left(a^*\right)}{\Gamma(a)}\frac{1}{2^d}\sum_{k\in\mathcal{K}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}}$$

So the marginal model posterior probability $\pi(k|y)$ of $\mathscr{M}_k$ is

$$\pi(k|y) = \frac{f(y|k)\pi(k)}{f(y)} = \quad \left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}}\Bigg/\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_{k'}^*\right)^{a^*}} \tag{9}$$

Proposition 18. *Let $z = (y_{n+1}, ..., y_{n+m})$ be a sequence of future outcomes. The conditional predictive distribution $G(z|y, k)$ of z given observables and model $\mathscr{M}_k$ has pdf/pmf*

$$g(z|y, k) = \int f(z|y, \theta_k, k)d\Pi(\theta_k|y, k) \overset{z \text{ indep. } y|(\theta_k,k)}{=} \int f(z|\theta_k, k)d\Pi(\theta_k|y, k) \tag{10}$$

The marginal predictive distribution $G(z|y)$ of z given the observables y has pdf/pmf

$$g(z|y) = \sum_{k\in\mathcal{K}} \int f(z|y,\theta_k,k) d\Pi(\theta_k|y,k)\pi(k|y) = \sum_{k\in\mathcal{K}} g(z|y,k)\pi(k|y)$$

Example 19. The predictive distribution $G\left(z|y,k\right)$ of future observables $z=(y_{n+1},...,y_{n+m})$ at any new Φ_k^{new} given the observables y and the model $\mathscr{M}_k$; i.e. $z=\Phi_k^{\text{new}}\beta_k+\epsilon$ and $\epsilon\sim \mathrm{N}\left(0,I_m\sigma^2\right)$, has pdf Appendix C

$$\begin{aligned} g(z|y,k) = \int f(z|\theta_k,k)\pi(\theta_k|y,k)\mathrm{d}\theta_k &= \int \mathrm{N}\left(z|\Phi_k^{\text{new}}\beta_k, I\sigma^2\right)\mathrm{N}\left(\beta_k|\mu_k^*, V_k^*\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda_k^*\right)\mathrm{d}\beta_k\mathrm{d}\sigma^2 \\ &= ...\text{calc}... = \mathrm{T}_m\left(z|\Phi^{\text{new}}\mu_k^*, [I+V_k^*]\,\frac{\lambda_k^*}{a^*}, 2a^*\right) \end{aligned}$$

Example 20. The marginal predictive distribution $G\left(z|k\right)$ of future observables $z=(y_{n+1},...,y_{n+m})$ at any new Φ_k^{new} given the observables y has pdf

$$\begin{aligned} g(z|y) &= \sum_{k\in\mathcal{K}} g(z|y,k)\pi(k|y) \\ &= \sum_{k\in\mathcal{K}} \mathrm{T}_m\left(z|\Phi^{\text{new}}\mu_k^*, V_k^*\frac{\lambda_k^*}{a^*}, 2a^*\right)\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}} \Bigg/ \sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_{k'}^*\right)^{a^*}} \end{aligned} \tag{11}$$

4 Inference based on a single Best model

For parametric or predictive inference, the statistician may proceed as follows:

1. select a single 'Best' model $\mathscr{M}_{k^*}$ (optimal in some sense) from a set of available models $\mathcal{M}=\{\mathscr{M}_1,\mathscr{M}_2,...\}$ and then
2. perform Bayesian inference conditional on the selected model $\mathscr{M}_{k^*}$ only and as if it had generated the data; i.e. based on

$$\begin{cases} y|\theta_{k^*},\mathscr{M}_{k^*}\sim F(y|\theta_{k^*},k^*) & \text{, data generation fromt he sampling distribution} \\ \theta_{k^*}|\mathscr{M}_{k^*}\sim \Pi(\theta_{k^*}|k^*) & \text{, within model parameter generation from the conditional prior} \end{cases}$$

4.1 Bayesian model selection: How to select the single 'Best' model $\mathscr{M}_{k^*}$.

Note 21. Model selection/choice is the selection of the "Best" statistical model from a set of competing statistical models $\mathcal{M}=\{\mathscr{M}_k;\ k\in\mathcal{K}\}$. It can be naturally performed as a parametric point estimation problem, statistical decision problem, or hypothesis test (all equivalent). See Criteria 22 and 23.

Criterion 22. *We produce the Bayes estimator (of Bayes rule) $\hat{\delta}$ of parameter $k\in\mathcal{K}$ under a loss function $\ell(\delta,k,\theta_k)$ with decision space $\mathcal{D}=\mathcal{K}$ and posterior distribution $(k,\theta_k)\sim\Pi(k,\theta_k|y)$. I.e find $\hat{\delta}$:*

$$\hat{\delta} = \arg\min_{\forall\delta\in\mathcal{K}} E_\Pi\left(\ell(\delta,k,\theta_k)\,|y\,\right) = \arg\min_{\forall\delta\in\mathcal{K}}\left(\int_{\Theta_k}\ell(\delta,k,\theta_k)d\Pi(k,\theta_k|y)\right)$$

Criterion 23. *Under the $0-1$ loss $\ell(\delta,k,\theta_k)=1\left(\delta\neq k\right)$ the best model corresponds to the MAP $\hat{\delta}$ that is model with the highest marginal model posterior probability $\pi(k|y)$, I.e $\hat{\delta}$* such as*:*

$$\hat{\delta} = \arg\max_{\forall k\in\mathcal{K}}\left\{\pi(k|y)\right\}$$

or equivalently $\hat{\delta}$ such that $\pi(\hat{\delta}|y)\geq\pi(k|y)$, for all $k\in\mathcal{K}$.

Bayes factors can be used for model selection. We re-define the Bayes factor to be suitable to the model selection framework.

Definition 24. Bayes factor $B_{k,j}(y)$ of model $\mathscr{M}_k$ against model $\mathscr{M}_j$ is defined as

$$B_{k,j}(y) = \frac{\pi\left(k|y\right)/\pi\left(k\right)}{\pi\left(j|y\right)/\pi\left(j\right)} = \frac{f(y|k)}{f(y|j)}; \quad \forall k,j \in \mathcal{K}$$

Example 25. (Cont.) The marginal model posterior $\pi(k|y)$ has been calculated in Example 17 in Eq. 9. Under the 0-1 loss $\ell(\delta, k, \theta_k) = 1\left(\delta \neq k\right)$, the best regression model is that with the largest marginal model posterior $\pi(k|y)$ (9).

Consider data-set airquality {datasets} available from R repository, that involves $n = 153$ pairs of observations on the response variable y 'Ozone (ppb) ' and the possible regressors

ϕ_0: Constant value 1

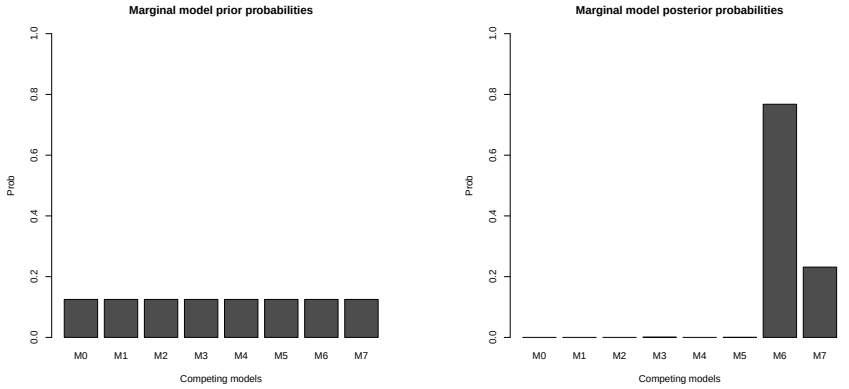
ϕ_1: Solar R (lang)

ϕ_2: Wind (mph)

ϕ_3: Temperature (degrees F)

The regressors have been standardized to adjust their unites in the model. The available models are: $\mathscr{M}_0 = \{1\}$, $\mathscr{M}_1 = \{1,2\}$, $\mathscr{M}_2 = \{1,3\}$, $\mathscr{M}_3 = \{1,4\}$, $\mathscr{M}_4 = \{1,2,3\}$, $\mathscr{M}_5 = \{1,2,4\}$, $\mathscr{M}_6 = \{1,3,4\}$, and $\mathscr{M}_7 = \{1,2,3,4\}$. The prior hyper-parameters are $\mu_k = 0$, $V_k = 100I_{d_k}$, $a_k = 1.0$, and $\lambda_k = 1.0$. The marginal model prior was $\pi(k|y) = 1/8$ for $k = \{0, ..., 7\}$.

In Figure 1, we see the marginal model prior which is uniform, and the marginal model posterior which denotes that the MAP 'best' model is $\mathscr{M}_6$.



(a) Marginal model prior probabilities (b) Marginal model posterior probabilities

Figure 1: Marginal model prior and posterior probabilities produced in Example 25.
`https://github.com/georgios-stats/Bayesian_Statistics/blob/master/HandoutsSupplementary/LinearRegressionModelUncertainty/MarginalModelPosteriroProbability.R`

The Bayes factor is computed as

$$B_{k,j}(y) = \frac{f(y|k)}{f(y|j)} = \left(\frac{|V_j|}{|V_k|}\right)^{\frac{1}{2}} \left(\frac{|V_k^*|}{|V_j^*|}\right)^{\frac{1}{2}} \left(\frac{\lambda_j^*}{\lambda_k^*}\right)^{a^*} \tag{12}$$

4.2 Within model inference

Note 26. Given that $\mathscr{M}_{k^*}$ is the selected "Best" model, interest may lie on learning the parameter $\theta_{k^*} \in \Theta_{k^*}$ (or any function of it), or future outputs $z = (y_{n+1}, ..., y_{n+m})$. Then inference is performed based on the Bayesian model

$$\begin{cases} y|\theta_{k^*}, \mathscr{M}_{k^*} \sim F(y|\theta_{k^*}, k^*) & \text{, data generation fromt he sampling distribution} \\ \\ \theta_{k^*}|\mathscr{M}_{k^*} \sim \Pi(\theta_{k^*}|k^*) & \text{, within model parameter generation from the conditional prior} \end{cases}$$

- Parametric inference about θ_{k^*} (or functions of it) is performed based on the conditional posterior in (8):

$$\theta_{k^*}|y, k^* \sim \Pi(\theta_{k^*}|y, k^*)$$

- Predictive inference about a sequence of future outcomes $z = (y_{n+1}, ..., y_{n+m})$ is performed based on the conditional predictive distribution :

$$z|y, k^* \sim G(z|y, k^*)$$

in (10) with pdf

$$g(z|y, k^*) = \int f(z|y, \theta_{k^*}, k^*) \mathrm{d}\Pi(\theta_{k^*}|y, k^*)$$

5 Bayesian model averaging (BMA)

Note 27. Selecting a single (best) model from a set of available ones and then proceeding as if the selected model had generated may not be optimal. It ignores the uncertainty due to model selection, which leads to over-confident inferences and decisions. BMA offers a coherent mechanism to account model uncertainty.

Definition 28. Model averaging is called the process where a posterior summary is obtained as a weighted mean of the summaries under each model with weights given by the marginal model posterior probability.

Parametric inference There may be certain model parameters or functions of model parameters which are present in all available models and retain a consistent interpretation across all models. Interesting may be to average out all the models.

Note 29. Let partition $\theta_k = (\psi, \phi_k)$, where $\psi \in \Psi$ is common in all models in $\mathcal{M}$. It is possible to obtain the marginal posterior distribution $\Pi(\psi|y)$ of ψ as a mixture of posterior distributions under each model with weights given by the marginal posterior probability of each model. i.e. $\Pi(\psi|y)$ has pdf/pmf

$$\pi(\psi|y) = \sum_{k \in \mathcal{K}} \pi(\psi|y, k)\pi(k|y); \qquad \text{with} \quad \pi(\psi|y, k) = \int \pi(\psi, \phi_k|y, k)\mathrm{d}\phi_k.$$

Note 30. Then any moment (expectation, variance, probability, etc...) of $\psi \in \Psi$ can be expressed in the aforesaid weighted average form. Let $h(\cdot)$ be a function defined on Ψ. Then

$$\mathrm{E}_\Pi\left(h(\psi)|y\right) = \sum_{k \in \mathcal{K}} \int_{\Theta_k} h(\psi)\mathrm{d}\Pi\left(\psi|y, k\right)\pi(k|y) = \sum_{k \in \mathcal{K}} \mathrm{E}_\Pi\left(h(\psi)|y, k\right)\pi(k|y)$$

Example 31. The sampling distribution variance σ^2 is the common parameter across all models, so

$$\pi\left(\sigma^2|y\right) = \sum_{k \in \mathcal{K}} \overbrace{\mathrm{IG}\left(\sigma^2|a^*, \lambda_k^*\right)}^{\pi(\sigma^2|y,k)}\pi(k|y) = \frac{\sum_{k \in \mathcal{K}} \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a^*+1} \exp\left(-\frac{\lambda_k^*}{\sigma^2}\right) \left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}} \Big/ \left(\lambda_k^*\right)^{a^*}}{\sum_{k' \in \mathcal{K}} \left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}} \Big/ \left(\lambda_{k'}^*\right)^{a^*}} 1\left(\sigma^2 > 0\right)$$

and, because $\mathrm{E}_{\sigma^2\sim\mathrm{IG}(a^*,\lambda_k^*)}\left(\sigma^2\right)=\frac{\lambda_k^*}{a^*-1}$ for $a^*>1$, we can get its marginal posterior expectation as

$$\mathrm{E}\left(\sigma^2|y\right)=\sum_{k\in\mathcal{K}}\mathrm{E}_{\Pi}\left(\sigma^2|y,k\right)\pi(k|y)=\frac{\sum_{k\in\mathcal{K}}\frac{\lambda_k^*}{a^*-1}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\Big/\left(\lambda_k^*\right)^{a^*}}{\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\Big/\left(\lambda_{k'}^*\right)^{a^*}}$$

Predictive inference BMA plays an important role in the predictive inference.

Note 32. Let $z=(y_{n+1},...,y_{n+m})$ be a sequence of future outcomes. Then the predictive distribution $\mathrm{d}G(z|y)$ of z given the observables y has pdf/pmf

$$g(z|y)=\sum_{k\in\mathcal{K}}g(z|y,k)\pi(k|y)$$

where

$$g(z|y,k)=\int f(z|y,\theta_k,k)\pi(\theta_k|y,k)\mathrm{d}\theta_k$$

is the predictive pdf of z given model $\mathscr{M}_k$.

Example 33. The BMA predictive distribution for the Normal linear regression model has been calculated in Example 19, Eq. c11.

Note 34. By averaging out all models in this fashion, BMA provides better predictive ability as measured by the logarithmic scoring rule $-\log(\pi(A))$, than using any single model $\mathscr{M}_k$, conditional on $\mathcal{M}$. See the exercise below.

6 Discussion

In model uncertainty framework, we often meet two situations;

$\mathcal{M}$-close perspective: the encompassing Bayesian model includes the real data-generation process $R(\cdot)$. It is argued that this is not a realistic situation. The introduced model uncertainty framework is suitable in this situation.

$\mathcal{M}$-open perspective: the encompassing Bayesian model does not include the real data-generation process $R(\cdot)$. This is a more realistic situation. It is argued that the introduced model uncertainty framework is not suitable in this situation in the sense that $\pi(\mathscr{M}_k)$ does not make sense to exist.

What do we do? This is a rather open problem, and may depends on the application under consideration. Advise; Try to set-up an encompassing Bayesian model rich enough to promise, in some sense, that it acceptably approximate (at least) the real data generating process.

Question 35. *For practice address the Exercises 69, 70, and 71, from the Exercise sheet.*

Exercise 36. Show that for any $k\in\mathcal{K}$

$$-\mathrm{E}\left(\log\left(\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)\right)\right)\leq-\mathrm{E}\left(\log\left(\pi(z|y,k)\right)\right)$$

where the expectations are with respect to $\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)$.

Solution. It is

$$-\mathrm{E}\left(\log\left(\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)\right)\right)\leq-\mathrm{E}\left(\log\left(\pi(z|y,k)\right)\right)\Longleftrightarrow 0\leq\underbrace{-\mathrm{E}\left(\log\left(\frac{\pi(z|y,k)}{\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)}\right)\right)}_{=\mathrm{KL}(\pi(z|y)||\pi(z|y,k))}$$

A Solution to Example 16

Use the following identity

$$(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)=(\beta-\mu^*)^\top(V^*)^{-1}(\beta-\mu^*)+S^*;$$
$$S^*=\mu^\top V^{-1}\mu-(\mu^*)^\top(V^*)^{-1}(\mu^*)+y^\top y;\qquad V^*=\left(V^{-1}+\Phi^\top\Phi\right)^{-1};\qquad \mu^*=V^*\left(V^{-1}\mu+\Phi^\top y\right)$$

Solution. For simplicity, we suppress the indexing $\cdot_k$ denoting the sub-set of the regressors. It is

$$\begin{aligned}
\pi\left(\beta_k,\sigma^2|y,k\right)\propto & f(y|\theta_k,k)\pi(\theta_k|k) \;=\mathrm{N}\left(y|\Phi\beta_k,I\sigma^2\right)\mathrm{N}\left(\beta_k|\mu_k,V_k\sigma^2\right)\mathrm{IG}\left(\sigma^2|a,\lambda\right)\\
&\propto\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^\top(y-\Phi\beta)\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu)^\top V^{-1}(\beta-\mu)\right)\\
&\qquad\times\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^\top(y-\Phi\beta)-\frac{1}{2\sigma^2}(\beta-\mu)^\top V^{-1}(\beta-\mu)\right)\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}\left[(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)\right]\right)\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu^*)^\top(V^*)^{-1}(\mu-\mu^*)-\frac{1}{2\sigma^2}S^*\right)\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu^*)^\top(V^*)^{-1}(\mu-\mu^*)\right)\exp\left(-\frac{1}{\sigma^2}\left(\lambda+\frac{1}{2}S^*\right)\right)\\
&\propto\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu^*)^\top(V_k^*)^{-1}(\mu-\mu^*)\right)}_{\propto\mathrm{N}(\beta_k|\mu^*,V^*\sigma^2)}\times\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\lambda+\frac{1}{2}S^*\right)\right)}_{\propto\mathrm{IG}(\sigma^2|a^*,\lambda^*)}\\
&\propto\mathrm{N}\left(\beta|\mu^*,V^*\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda^*\right)
\end{aligned}$$

where

$$S^*=\mu^\top V^{-1}\mu-(\mu^*)^\top(V^*)^{-1}(\mu^*)+y^\top y$$
$$a^*=\frac{n}{2}+a;\quad \lambda^*=\lambda+\frac{1}{2}S^*;\qquad V^*=\left(V^{-1}+\Phi^\top\Phi\right)^{-1};\qquad \mu^*=V^*\left(V^{-1}\mu+\Phi^\top y\right)$$

The result in Example 16 results by indexing with $\cdot_k$ the quantities that depend on it. I.e.

$$\pi\left(\beta_k,\sigma^2|y,k\right)=\mathrm{N}\left(\beta_k|\mu_k^*,V_k^*\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda_k^*\right)$$
$$S_k^*=\mu_k^\top V_k^{-1}\mu_k-(\mu_k^*)^\top(V_k^*)^{-1}(\mu_k^*)+y^\top y$$
$$a^*=\frac{n}{2}+a;\quad \lambda_k^*=\lambda+\frac{1}{2}S_k^*;\qquad V_k^*=\left(V_k^{-1}+\Phi_k^\top\Phi_k\right)^{-1};\qquad \mu_k^*=V_k^*\left(V_k^{-1}\mu_k+\Phi_k^\top y\right)$$

Hence

$$\implies\begin{cases}\beta_k|y\sigma^2,k & \sim\mathrm{N}\left(\mu_k^*,V_k^*\sigma^2\right)\\ \sigma^2|y,k & \sim\mathrm{IG}\left(a^*,\lambda^*\right)\end{cases},\ \forall k\in\mathcal{K}$$

B Solution to Example 17

Use the following identity

$$(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)=(\beta-\mu^*)^\top(V^*)^{-1}(\beta-\mu^*)+S^*;$$

$$S^*=\mu^\top V^{-1}\mu-(\mu^*)^\top(V^*)^{-1}(\mu^*)+y^\top y;\qquad V^*=\left(V^{-1}+\Phi^\top\Phi\right)^{-1};\qquad \mu^*=V^*\left(V^{-1}\mu+\Phi^\top y\right)$$

Solution. For simplicity, we suppress the indexing $\cdot_k$ denoting the sub-set of the regressors.

The conditional marginal likelihood $f(y|k)$ (or model evidence of $\mathscr{M}_k$) is

$$\begin{aligned}
f(y|k)&=\int_{\Theta_k} f(y|\theta_k,k)\pi(\theta_k|k)\mathrm{d}\theta_k \;=\; \int \mathrm{N}\left(y|\Phi\beta,I\sigma^2\right)\mathrm{N}\left(\beta|\mu,V\sigma^2\right)\mathrm{IG}\left(\sigma^2|a,\lambda\right)\mathrm{d}\beta\mathrm{d}\sigma^2\\
&=\int\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^\top(y-\Phi\beta)\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu)^\top V^{-1}(\beta-\mu)\right)\\
&\qquad\times\frac{\lambda^a}{\Gamma(a)}\exp\left(-\frac{\lambda}{\sigma^2}\right)\mathrm{d}\beta\mathrm{d}\sigma^2\\
&=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\int\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)}{2}+\lambda\right)\right)\\
&=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\int\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}S^*+\lambda\right)\right)\\
&\qquad\times\int\left[\exp\left(-\frac{1}{2}\frac{1}{\sigma^2}(\beta-\beta^*)^\top(V^*)^{-1}(\beta-\beta^*)\right)\mathrm{d}\beta\right]\mathrm{d}\sigma^2\\
&=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\int\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}S^*+\lambda\right)\right)\times\left[(2\pi)^{\frac{d}{2}}\left(\sigma^2\right)^{\frac{d}{2}}|V^*|^{\frac{1}{2}}\right]\mathrm{d}\sigma^2\\
&=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V^*|}{|V|}\right)^{\frac{1}{2}}\frac{k^a}{\Gamma(a)}\frac{\Gamma\left(\frac{n}{2}+a\right)}{\left(\frac{1}{2}S^*+\lambda\right)^{\frac{n}{2}+a}}\;=\;\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V^*|}{|V|}\right)^{\frac{1}{2}}\frac{k^a}{\Gamma(a)}\frac{\Gamma(a^*)}{(\lambda^*)^{a^*}}
\end{aligned}$$

where

$$S^*=\mu^\top V^{-1}\mu-(\mu^*)^\top(V^*)^{-1}(\mu^*)+y^\top y$$

$$a^*=\frac{n}{2}+a;\quad \lambda^*=\lambda+\frac{1}{2}S^*;\qquad V^*=\left(V^{-1}+\Phi^\top\Phi\right)^{-1};\qquad \mu^*=V^*\left(V^{-1}\mu+\Phi^\top y\right)$$

The result in Example 17 results by indexing with $\cdot_k$ the quantities that depend on it. I.e.

$$f(y|k)=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\frac{\Gamma(a^*)}{(\lambda_k^*)^{a^*}}$$

$$S_k^*=\mu_k^\top V_k^{-1}\mu_k-(\mu_k^*)^\top(V_k^*)^{-1}(\mu_k^*)+y^\top y$$

$$a^*=\frac{n}{2}+a;\quad \lambda_k^*=\lambda+\frac{1}{2}S_k^*;\qquad V_k^*=\left(V_k^{-1}+\Phi_k^\top\Phi_k\right)^{-1};\qquad \mu_k^*=V_k^*\left(V_k^{-1}\mu_k+\Phi_k^\top y\right)$$

The marginal likelihood $f(y)$ is

$$f(y)=\sum_{k\in\mathcal{K}}f(y|k)\pi(k)\;=\;\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\lambda^a\frac{\Gamma(a^*)}{\Gamma(a)}\frac{1}{2^d}\sum_{k\in\mathcal{K}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{(\lambda_k^*)^{a^*}}$$

The marginal model posterior probability $\pi(k|y)$ of $\mathscr{M}_k$ is

$$\pi(k|y)=\frac{f(y|k)\pi(k)}{f(y)}=\quad\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{(\lambda_k^*)^{a^*}}\Bigg/\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{(\lambda_{k'}^*)^{a^*}}$$

C Solution to Example 19

Use the following identity:

$$\begin{cases} y|\beta \sim \mathrm{N}_m\left(\Phi\beta + c, \Sigma\right) \\ \\ \beta \sim \mathrm{N}_d\left(\mu, V\right) \end{cases} \implies y \sim \mathrm{N}_m\left(\Phi\mu + c, \Sigma + \Phi V \Phi^\top\right)$$

Use the following property of the Student T distribution

$$\begin{cases} x|\xi \sim \mathrm{N}_m(\mu, \Sigma\xi v) \\ \\ \xi \sim \mathrm{IG}(\frac{v}{2}, \frac{1}{2}) \end{cases} \implies x \sim \mathrm{T}_m(\mu, \Sigma, v)$$

Solution. For simplicity, we suppress the indexing $\cdot_k$ denoting the sub-set of the regressors. The conditional predictive distribution $G(z|y,k)$ given model $\mathscr{M}_k$ and data y has pdf

$$g(z|y,k) = \int f(z|\theta_k,k)\pi(\theta_k|y,k)\mathrm{d}\theta_k = \int \mathrm{N}_m\left(z|\Phi^{\text{new}}\beta, I\sigma^2\right)\mathrm{N}_d\left(\beta|\mu^*, V^*\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda^*\right)\mathrm{d}\beta\mathrm{d}\sigma^2$$

$$= \int \left[\int \mathrm{N}_m\left(z|\Phi^{\text{new}}\beta, I\sigma^2\right)\mathrm{N}_d\left(\beta|\mu^*, V^*\sigma^2\right)\mathrm{d}\beta\right]\mathrm{IG}\left(\sigma^2|a^*,\lambda^*\right)\mathrm{d}\sigma^2$$

$$= \int \mathrm{N}_m\left(z|\Phi^{\text{new}}\mu^*, \sigma^2\left[I + V^*\right]\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda^*\right)\mathrm{d}\sigma^2$$

$$= \int \mathrm{N}_m\left(z|\Phi^{\text{new}}\mu^*, \sigma^2\left[I + V^*\right]\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda^*\right)\mathrm{d}\sigma^2$$

$$= \int \mathrm{N}_m\left(z|\Phi^{\text{new}}\mu^*, \sigma^2\frac{\lambda^*}{\lambda^*}\frac{2a^*}{2a^*}\left[I + V^*\right]\right)\mathrm{IG}\left(\sigma^2|a^*,\lambda^*\right)\mathrm{d}\sigma^2$$

$$\overset{\xi=\frac{\sigma^2}{2\lambda^*}\sim\mathrm{IG}\left(a^*,\frac{1}{2}\right)}{=} \int \mathrm{N}_m\left(z|\Phi^{\text{new}}\mu^*, \xi\lambda^*\frac{2a^*}{a^*}\left[I + V^*\right]\right)\mathrm{IG}\left(\xi|a^*,\frac{1}{2}\right)\mathrm{d}\xi$$

$$= \int \mathrm{N}_m\left(z|\Phi^{\text{new}}\mu^*, \xi 2a^*\frac{\lambda^*}{a^*}\left[I + V^*\right]\right)\mathrm{IG}\left(\xi|\frac{a^*}{2},\frac{1}{2}\right)\mathrm{d}\xi$$

$$= \mathrm{T}_m\left(z|\Phi^{\text{new}}\mu^*, \left[I + V^*\right]\frac{\lambda^*}{a^*}, 2a^*\right)$$

The result in Example 19 results by indexing with $\cdot_k$ the quantities that depend on it. I.e.

$$g(z|y,k) = \mathrm{T}_m\left(z|\Phi^{\text{new}}\mu_k^*, \left[I + V_k^*\right]\frac{\lambda_k^*}{a^*}, 2a^*\right)$$

The marginal predictive distribution $G\left(z|y\right)$ given data y has pdf/pmf

$$g(z|y) = \sum_{k\in\mathcal{K}} g(z|y,k)\pi(k|y)$$

$$= \sum_{k\in\mathcal{K}} \mathrm{T}_m\left(z|\Phi^{\text{new}}\mu_k^*, V_k^*\frac{\lambda_k^*}{a^*}, 2a^*\right)\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}} \Bigg/ \sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_{k'}^*\right)^{a^*}}$$