Model selection as point estimation

Eduardo Gutiérrez-Peña

Department of Probability and Statistics
National University of Mexico
Mexico

Happy 70th birthday Luis Raul!

First COBAL – Ubatuba, Brazil (2002).

Second COBAL – Los Cabos, Mexico (2005).

Fourth COBAL – Medellín, Colombia (2011).

Outline

1. Introduction
   - Perspectives on model comparison
   - Motivation
2. M-closed view
   - Parametric modelling
   - Model selection
   - Model averaging
3. M-completed view
   - Parametric procedures from a non-parametric perspective
   - Decision theoretical setting
   - Constructing $\mathbf{F}_\Lambda$
4. Concluding remarks

Perspectives on model comparison (Bernardo & Smith, 2000)

M-closed view:

Corresponds to believing that one of the models $\{M_i, i \in I\}$ is "true".

M-completed view:

Corresponds to an individual acting as if $\{M_i, i \in I\}$ simply constitute a range of specified models currently available for comparison, to be evaluated in the light of the individual's separate actual belief model, M_t.

From this perspective, assigning probabilities $\{P(M_i), i \in I\}$ does not make sense.

M-open view:

Also acknowledges that $\{M_i, i \in I\}$ are simply a range of specified models available for comparison, so that assigning probabilities $\{P(M_i), i \in I\}$ does not make sense.

However, in this case, there is no separate overall actual belief specification M_t, perhaps because we lack the time or competence to provide it.

Motivation

This talk presents a general reflection concerning the similarities between point estimation, model selection and model averaging, and how such similarities could be useful to better understand various procedures currently in use and perhaps to propose new ones.

Model selection and model averaging can both be regarded as point estimation problems over suitably defined extended classes of models. We explore this idea in an objective Bayesian context.

Parametric modelling

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$$

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n); \quad \text{independent realizations of } X \sim f(x|\theta^0)$$

Aim: Estimate the true density $f(x)$, which is assumed equal to $f(x|\theta^0)$ for some $\theta^0 \in \Theta$.

Frequentist approach

Maximum likelihood (ML)

$$\hat{\theta}_{ML} = \arg\max_{\Theta} f(\boldsymbol{x}|\theta)$$

Predicting density

$$\widehat{f}(x) = f(x|\hat{\theta}_{ML})$$

Parametric modelling

Bayesian approach

Prior

$$\mathcal{P} = \{p(\theta|\phi) : \phi \in \Phi\}$$

Posterior

$$p(\theta|\phi^0, \boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)\, p(\theta|\phi^0)$$

where ϕ^0 specifies the elicited prior.

Maximum a posteriori (MAP)

$$\hat{\theta}_{MAP} = \arg\max_{\Theta} p(\theta|\phi^0, \boldsymbol{x})$$

Posterior predictive density

$$\begin{aligned}\widehat{f}(x) &= f(x|\phi^0, \boldsymbol{x}) \\ &= \int f(x|\theta, \boldsymbol{x})\, p(\theta|\phi^0, \boldsymbol{x})\, \mathrm{d}\theta \\ &= \int f(x|\theta)\, p(\theta|\phi^0, \boldsymbol{x})\, \mathrm{d}\theta\end{aligned}$$

Parametric modelling

If $f(x|\phi^0, \boldsymbol{x})$ is difficult to compute, we can approximate it by the 'Bayesian predicting density'

$$\widehat{f}(x) \approx f(x|\hat{\theta}_{MAP})$$

or using standard (MC)MC methods.

Another possibility (with a nonparametric, objective Bayes flavour) is the following.

Weighted likelihood bootstrap (WLB); Newton & Raftery (1994)

For $b = 1, 2, \ldots, B$

1. Generate $\boldsymbol{w}^{(b)} = (w_1^{(b)}, w_2^{(b)}, \ldots, w_n^{(b)}) \sim \text{Dir}_n(1, 1, \ldots, 1)$
2. Find $\tilde{\theta}^{(b)} = \underset{\Theta}{\arg\max} \prod_{i=1}^n f(x_i|\theta)^{w_i^{(b)}}$

We can use $\tilde{\theta}^{(1)}, \tilde{\theta}^{(2)}, \ldots, \tilde{\theta}^{(B)}$ to make inferences about θ ('estimate' the distribution of θ) or make predictive inferences, e.g.

$$\widehat{f}(x) \approx \frac{1}{B} \sum_{b=1}^{B} f(x|\tilde{\theta}^{(b)})$$

Just as with ML, and unlike MAP, we do not need to specify ϕ^0.

Hierarchical modelling

$$\mathcal{F}^* = \{f^*(x|\phi) : \phi \in \Phi\}$$

with

$$f^*(x|\phi) = \int f(x|\theta)\, p(\theta|\phi)\, \mathrm{d}\theta$$

This is the prior predictive density from the previous case. For $f^*(x \,|\, \phi)$ to be well defined, we need a proper prior on θ.

'Frequentist' approach

Empirical Bayes (EB)

$$\hat{\phi}_{EB} = \arg\max_{\Phi} f^*(\boldsymbol{x}|\phi)$$

Note: EB is akin to ML; i.e., we are estimating ϕ by maximizing its 'likelihood function' $f^*(\boldsymbol{x} \,|\, \phi)$.

Predicting density

$$\widehat{f}(x) = f^*(x|\hat{\phi}_{EB}, \boldsymbol{x})$$

Hierarchical modelling

Bayesian approach

(Hyper)prior

$$\mathcal{P}^* = \{p^*(\phi|\lambda) : \lambda \in \Lambda\}$$

Posterior

$$p^*(\phi|\lambda^0, \boldsymbol{x}) \propto f^*(\boldsymbol{x}|\phi)\, p^*(\phi|\lambda^0)$$

where λ^0 specifies the elicited hyperprior.

Maximum a posteriori empirical Bayes (MAP-EB)

$$\hat{\phi}_{MAP} = \arg\max_{\Phi} p^*(\phi|\lambda^0, \boldsymbol{x})$$

Posterior predictive density

$$\begin{aligned} \widehat{f}(x) &= f^*(x|\lambda^0, \boldsymbol{x}) \\ &= \int f^*(x|\phi, \boldsymbol{x})\, p^*(\phi|\lambda^0, \boldsymbol{x})\, \mathrm{d}\phi \end{aligned}$$

Hierarchical modelling

Similar to the previous case, if $f^*(x|\lambda^0, \boldsymbol{x})$ is difficult to compute, we can approximate it by the 'Bayesian predicting density'

$$\widehat{f}(x) \approx f^*(x|\hat{\phi}_{MAP}, \boldsymbol{x})$$

or using standard MCMC methods.

Weighted likelihood bootstrap

The WLB becomes more involved in this case due to the dependence of the observations under the joint predictive distribution $f^*(\boldsymbol{x}|\phi)$

For $b = 1, 2, \ldots, B$

1. Generate $\boldsymbol{w}^{(b)} = (w_1^{(b)}, w_2^{(b)}, \ldots, w_n^{(b)}) \sim \mathrm{Dir}_n(1, 1, \ldots, 1)$
2. Find $\tilde{\phi}^{(b)} = \underset{\Phi}{\arg\max} \prod_{i=1}^n f^*(x_i|\phi, x_1, \ldots, x_{i-1})^{w_i^{(b)}}$

Different orderings of the data yield different 'weighted likelihood' functions.

Then we would have

$$\widehat{f}(x) \approx \frac{1}{B}\sum_{b=1}^{B} f^*(x|\tilde{\phi}^{(b)}, \boldsymbol{x})$$

Just as with EB, and unlike MAP-EB, we do not need to specify λ^0.

Model selection

$$\mathcal{F}^{**} = \{f^{**}(x|\lambda) : \lambda \in \Lambda\}$$

with

$$\begin{aligned} f^{**}(x|\lambda) &= \int f_\lambda(x|\theta_\lambda)\, p_\lambda(\theta_\lambda)\, \mathrm{d}\theta_\lambda \\ &= \int f(x|\theta_\lambda, \lambda)\, p(\theta_\lambda|\lambda)\, \mathrm{d}\theta_\lambda \end{aligned}$$

Similar to $\mathcal{F}^*$, but now $\mathcal{F}^{**}$ contains prior predictive densities arising from different parametric families indexed by λ.
For $f^{**}(x|\lambda)$ to be well defined, we need proper priors on θ_λ. The parameters θ_λ can have different dimensions.

'Frequentist' approach

Bayes factors (BF)

$$\hat{\lambda}_{BF} = \arg\max_{\Lambda} f^{**}(\boldsymbol{x}|\lambda)$$

Note 1: BF$(\hat{\lambda}_{BF}, \lambda) = f^{**}(\boldsymbol{x}|\hat{\lambda}_{BF})/f^{**}(\boldsymbol{x}|\lambda) \geq 1$ for all $\lambda \in \Lambda$.
Note 2: BF is related with EB; i.e., we are 'estimating' λ by maximizing its 'likelihood function' $f^{**}(\boldsymbol{x}|\lambda)$.
Note 3: We can in principle use 'priors' on θ_λ derived from intrinsic BF (Berger and Pericchi, 1996), fractional BF (O'Hagan, 1995), posterior BF (Aitkin, 1991), etc.

Predicting density (selected model)

$$\widehat{f}(x) = f^{**}(x|\hat{\lambda}_{BF}, \boldsymbol{x})$$

Model selection

Bayesian approach

Prior (over the class of models)

$$\mathcal{P}^{**} = \{p^{**}(\lambda|\omega) : \omega \in \Omega\}$$

Posterior

$$p^{**}(\lambda|\omega^0, \boldsymbol{x}) \propto f^{**}(\boldsymbol{x}|\lambda)\, p^{**}(\lambda|\omega^0)$$

Posterior odds (PO)

$$\hat{\lambda}_{PO} = \arg\max_{\Lambda} p^{**}(\lambda|\omega^0, \boldsymbol{x})$$

Note 1: PO$(\hat{\lambda}_{PO}, \lambda) = p^{**}(\hat{\lambda}_{PO}|\omega^0, \boldsymbol{x})/p^{**}(\lambda|\omega^0, \boldsymbol{x}) \geq 1$ for all $\lambda \in \Lambda$.
Note 2: PO is related with MAP; i.e., we are 'estimating' λ by maximizing its posterior density $p^{**}(\lambda|\omega^0, \boldsymbol{x})$.

Posterior predictive density (Bayesian model averaging); Draper (1995)

$$\begin{aligned} \widehat{f}(x) &= f^{**}(x|\omega^0, \boldsymbol{x}) \\ &= \int f^{**}(x|\lambda, \boldsymbol{x})\, p^{**}(\lambda|\omega^0, \boldsymbol{x})\, \mathrm{d}\lambda \end{aligned}$$

Discrete model selection

$$\mathcal{F}^{**} = \{f^{**}(x|\lambda) : \lambda \in \Lambda\}$$

with

$$\Lambda = \{1, 2, \ldots, m\}$$

'Frequentist' approach

Bayes factors (BF): $\hat{\lambda}_{BF} = \arg\max_{\Lambda} f^{**}(\boldsymbol{x}|\lambda)$

The 'EB estimator' $\hat{\lambda}_{BF}$ is such that $f^{**}(\boldsymbol{x}|\hat{\lambda}_{BF}) \geq f^{**}(\boldsymbol{x}|\lambda)$ for all $\lambda = 1, 2, \ldots, m$.

Predicting density (selected model)

$$\widehat{f}(x) = f^{**}(x|\hat{\lambda}_{BF}, \boldsymbol{x})$$

Discrete model selection

Bayesian approach

Discrete prior (over the class of models)

$$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m), \qquad \Omega = \left\{ \boldsymbol{\omega} \in \mathbb{R}^m : \omega_j \geq 0, j = 1, 2, \ldots, m; \sum_{j=1}^{m} \omega_j = 1 \right\}$$

$$p^{**}(\lambda | \boldsymbol{\omega}) = \omega_\lambda, \quad \lambda = 1, 2, \ldots, m$$

Posterior

$$\begin{aligned} p^{**}(\lambda | \boldsymbol{\omega}^0, \boldsymbol{x}) &\equiv \omega_\lambda^0(\boldsymbol{x}) \\ &\propto \omega_\lambda^0 \, f^{**}(\boldsymbol{x} | \lambda) \end{aligned}$$

Posterior odds (PO): $\hat{\lambda}_{PO} = \arg\max_{\Lambda} p^{**}(\lambda | \boldsymbol{\omega}^0, \boldsymbol{x})$

Note 1: The 'MAP-EB estimator' $\hat{\lambda}_{PO}$ is such that $p^{**}(\hat{\lambda}_{PO} | \boldsymbol{\omega}^0, \boldsymbol{x}) \geq p^{**}(\lambda | \boldsymbol{\omega}^0, \boldsymbol{x})$ for all $\lambda = 1, 2, \ldots, m$.
Note 2: If $\omega_\lambda = 1/m$ for all λ, then PO $\equiv$ BF.

Posterior predictive density (Bayesian model averaging); Clyde (1999)

$$\begin{aligned} \widehat{f}(x) &= f^{**}(x | \boldsymbol{\omega}^0, \boldsymbol{x}) \\ &= \sum_{\lambda=1}^{m} \omega_\lambda^0(\boldsymbol{x}) \, f^{**}(x | \lambda, \boldsymbol{x}) \end{aligned}$$

Discrete model selection

As in the hierarchical modelling case, if $f^{**}(x|\boldsymbol{\omega}^0, \boldsymbol{x})$ is difficult to compute, we can approximate it by the 'Bayesian predicting density'

$$\widehat{f}(x) \approx f^{**}(x|\hat{\lambda}_{PO}, \boldsymbol{x})$$

or using MCMC methods. This may involve using RJ-MCMC (Green, 1995) or a similar algorithm.

Weighted likelihood bootstrap

Here, the WLB is also more involved due to the dependence of the observations under the joint predictive distribution $f^*(\boldsymbol{x}|\lambda)$

For $b = 1, 2, \ldots, B$

1. Generate $\boldsymbol{w}^{(b)} = (w_1^{(b)}, w_2^{(b)}, \ldots, w_n^{(b)}) \sim \text{Dir}_n(1, 1, \ldots, 1)$
2. Find $\tilde{\lambda}^{(b)} = \underset{\Lambda}{\arg\max} \prod_{i=1}^n f^{**}(x_i|\lambda, x_1, \ldots, x_{i-1})^{w_i^{(b)}}$

Different orderings of the data yield different 'weighted likelihood' functions.

Discrete model selection

Then we would have

$$\begin{aligned} \widehat{f}(x) &\approx \frac{1}{B}\sum_{b=1}^{B} f^{**}(x|\tilde{\lambda}^{(b)}, \boldsymbol{x}) \\ &= \sum_{\lambda=1}^{m} \tilde{\omega}_{\lambda}\, f^{**}(x|\lambda, \boldsymbol{x}) \end{aligned}$$

where

$$\tilde{\omega}_{\lambda} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\lambda}(\tilde{\lambda}^{(b)})$$

Thus we can 'estimate' the distribution of λ using the WLB sample.

Just as with BF, and unlike PO, we do not need to specify $\boldsymbol{\omega}^{0}$.

Note that we could use $\tilde{\omega}_{\lambda}$ to perform model selection: choose the model corresponding to

$$\hat{\lambda}_{WLB} = \arg\max_{\Lambda}\{\tilde{\omega}_{\lambda}\}$$

Discrete model averaging

$$\mathcal{F}^{***} = \{f^{***}(x|\boldsymbol{\omega}) : \boldsymbol{\omega} \in \Omega\}$$

with

$$\begin{aligned} f^{***}(x|\boldsymbol{\omega}) &= \sum_{\lambda=1}^{m} f^{**}(x|\lambda)\, p^{**}(\lambda|\boldsymbol{\omega}) \\ &= \sum_{\lambda=1}^{m} \omega_\lambda\, f^{**}(x|\lambda) \end{aligned}$$

The ‘prior predictive’ densities in $\mathcal{F}^{***}$ are ‘model averages’ (mixtures).

‘Frequentist’ approach

‘Empirical’ model averaging

$$\hat{\boldsymbol{\omega}}_E = \arg\max_{\Omega} f^{***}(\boldsymbol{x}|\boldsymbol{\omega})$$

Predicting density (‘estimated’ model average)

$$\widehat{f}(x) = f^{***}(x|\hat{\boldsymbol{\omega}}_E, \boldsymbol{x})$$

Discrete model averaging

Bayesian approach

Prior (on model weights)

$$\mathcal{P}^{***} = \{p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in A\}$$

Posterior

$$p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}^0, \boldsymbol{x}) \propto f^{***}(\boldsymbol{x}|\boldsymbol{\omega})\, p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}^0)$$

MAP-MA

$$\hat{\boldsymbol{\omega}} = \arg\max_{\Omega} p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}^0, \boldsymbol{x})$$

Bayesian predicting density ('estimated' model average)

$$\widehat{f}(x) = f^{***}(x|\hat{\boldsymbol{\omega}}) = \sum_{\lambda=1}^{m} \hat{\omega}_\lambda\, f^{**}(x|\lambda, \boldsymbol{x})$$

Posterior predictive density ('hierarchical' Bayesian model averaging)

$$\begin{aligned}\widehat{f}(x) &= f^{***}(x|\boldsymbol{\alpha}^0, \boldsymbol{x}) \\ &= \int f^{***}(x|\boldsymbol{\omega}, \boldsymbol{x})\, p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}^0, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\omega}\end{aligned}$$

Posterior computations may involve sophisticated MCMC techniques.

Discrete model averaging

i.e.,

$$\begin{aligned}\widehat{f}(x) &= \int \left\{ \sum_{\lambda=1}^{m} \omega_\lambda \, f^{**}(x|\lambda, \boldsymbol{x}) \right\} p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}^0, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{\omega} \\ &= \sum_{\lambda=1}^{m} E[\omega_\lambda|\boldsymbol{\alpha}^0, \boldsymbol{x}] \, f^{**}(x|\lambda, \boldsymbol{x})\end{aligned}$$

As before, we could use $E[\omega_\lambda|\boldsymbol{\alpha}^0, \boldsymbol{x}]$ or $\hat{\omega}_\lambda$ to perform model selection by choosing the model corresponding to

$$\hat{\lambda}_{HBMA} = \arg\max_{\Lambda} \{E[\omega_\lambda|\boldsymbol{\alpha}^0, \boldsymbol{x}]\}$$

or

$$\hat{\lambda}_{MAP} = \arg\max_{\Lambda} \{\hat{\omega}_\lambda\}$$

respectively.

Discrete model averaging

Priors on $\boldsymbol{\omega}$

1. *Dirichlet*: $p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\omega}|\boldsymbol{\alpha})$

2. *Spike-and-slab (e.g., for large m)*; George & McCulloch (1997):

$$\omega_j = \frac{\delta_j \gamma_j}{\sum_{k=1}^{m} \delta_k \gamma_k}$$

with

$$\gamma_j \sim \mathrm{Ga}(\gamma_j|\alpha_j, 1) \quad \text{and} \quad \delta_j \sim \mathrm{Ber}(q)$$

for $j = 1, 2, \ldots, m$

Parametric procedures from a non-parametric perspective

Traditionally, Bayesian model selection has relied on the use of Bayes factors. However, Bayesian model selection is more than that.

Here we view the traditional parametric procedures discussed above as statistical decision problems where the uncertainty on the unknown true model is modelled non-parametrically.

Each of those parametric procedures can be associated with a specific family of parametric predictive distributions.

In other words, the problem is recast as one of finding a *surrogate* predictive distribution to be used as a simpler alternative to a non-parametric predictive distribution or as an estimate of the unknown true distribution, $f(x)$.

We start by considering a class of predictive distributions entertained by the parametric statistician

$$\mathbf{F}_K = \{f_\kappa(x) : \kappa \in K\},$$

where the forms of $f_\kappa(x)$ and K depend on the specific parametric procedure of interest.

For example, $\mathbf{F}_K$ can be $\mathcal{F}$, $\mathcal{F}^*$, $\mathcal{F}^{**}$ or $\mathcal{F}^{***}$.

Decision theoretical setting

Space of actions: K

Space of unknown states of nature: $\mathbf{F} = \{F : F \text{ is a probability distribution on } \mathcal{X}\}$

Prior distribution: $F \sim \mathcal{DP}(a_0 F_0)$, a Dirichlet process on $\mathbf{F}$

Utility function: based on the logarithmic score

$$U(\kappa, F) = \int \log f_\kappa(x) \; \mathrm{d}F(x)$$

Decision theoretical setting

Given $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, independent realizations of $X \sim F(x)$:

Posterior distribution: $\mathcal{DP}(a_n F_n)$, with $a_n = a_0 + n$ and $F_n(\cdot) = [a_0 F_0(\cdot) + n\widehat{F}(\cdot)]/(a_0 + n)$, where $\widehat{F}(\cdot)$ denotes the empirical distribution function of the sample $\boldsymbol{x}$.

Posterior expected utility: taking $a_0 = 0$,

$$U_n(\kappa) = \int \log f_\kappa(x)\ \mathrm{d}\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \log f_\kappa(x_i)$$

which is maximized by the same $\widehat{\kappa}$ that maximizes

$$\prod_{i=1}^{n} f_\kappa(x_i).$$

Constructing $\mathbf{F}_\Lambda$

We construct the family $\mathbf{F}_K$ of surrogate predictive densities on the basis of a collection of entertained parametric models

$$\mathcal{M} = \{M_j : j = 1, \ldots, m\}$$

with

$$M_j = \{f_j(x|\theta_j),\ p_j(\theta_j) : \theta_j \in \Theta_j\}$$

From this non-parametric, decision theoretical perspective, the densities $p_j(\theta_j)$ are to be regarded simply as convenient building blocks of the parametric predictives $f_\kappa(\cdot)$ and not as actual priors.

Model averaging

Let $\kappa = \boldsymbol{\omega}$ and $K = \Omega$, where

$$\Omega = \left\{ \boldsymbol{\omega} \in \mathbb{R}^m : \omega_j \geq 0, j = 1, 2, \ldots, m; \sum_{j=1}^m \omega_j = 1 \right\}$$

Now set $p_j(\theta_j) = \pi_j(\theta_j|\boldsymbol{x})$, the 'reference posterior' of θ_j.

The corresponding surrogate predictive density is then given by

$$f_\omega(x) = \sum_{j=1}^m \omega_j \, f_j(x|\boldsymbol{x})$$

where

$$f_j(x|\boldsymbol{x}) = \int f_j(x|\theta_j) \, \pi_j(\theta_j|\boldsymbol{x}) \, \mathrm{d}\theta_j$$

The optimal model corresponds to the value $\hat{\boldsymbol{\omega}}$ that maximizes the posterior expected utility $U_n(\boldsymbol{\omega})$ or, equivalently, that maximizes

$$\prod_{i=1}^n f_\omega(x_i) = \prod_{i=1}^n \sum_{j=1}^m \omega_j \, f_j(x_i|\boldsymbol{x})$$

Virtually all the traditional parametric procedures, ranging from point estimation to model averaging, can be accommodated within a slightly more general formulation of this framework by restricting the form of the corresponding κ and K (Gutiérrez-Peña and Walker, 2005).

Model selection

Model selection is a special case of the above where the weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$ degenerate at one of the m models, so that

$$\Omega = \left\{ \boldsymbol{\omega} \in \mathbb{R}^m : \omega_j = \mathbf{1}_j(\lambda);\ j, \lambda = 1, \ldots, m \right\},$$

In this case we can identify κ with λ and let $K = \{1, \ldots, m\}$.

The corresponding surrogate predictive density is then given by

$$f_\lambda(x) = \int f_\lambda(x|\theta_\lambda)\, \pi_\lambda(\theta_\lambda|\boldsymbol{x})\, \mathrm{d}\theta_\lambda\,.$$

The optimal model corresponds to the value of λ that maximizes the posterior expected utility $U_n(\lambda)$ or, equivalently, that maximizes

$$\prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} f(x_i|\lambda, \boldsymbol{x})$$

Using estimated weights for model selection

Once again, we can perform model selection as a by-product of model averaging: choose the model corresponding to

$$\hat{\lambda}_{MA} = \arg\max_{\Lambda}\{\hat{\boldsymbol{\omega}}\}$$

Weighted likelihood bootstrap

If calculating the optimal weights $\hat{\boldsymbol{\omega}}$ is not feasible, we can use the following version of the WLB (Gutiérrez-Peña et al., 2009)

For $b = 1, 2, \ldots, B$

1. Generate $\boldsymbol{w}^{(b)} = (w_1^{(b)}, w_2^{(b)}, \ldots, w_n^{(b)}) \sim \mathrm{Dir}_n(1, 1, \ldots, 1)$
2. Find $\tilde{\lambda}^{(b)} = \arg\max_{\Lambda} \prod_{i=1}^{n} f(x_i|\lambda, \boldsymbol{x})^{w_i^{(b)}}$

Then we would have

$$\widehat{f}(x) \approx \frac{1}{B}\sum_{b=1}^{B} f(x|\tilde{\lambda}^{(b)}, \boldsymbol{x}) = \sum_{\lambda=1}^{m} \tilde{\omega}_{\lambda}\, f(x|\lambda, \boldsymbol{x})$$

with

$$\tilde{\omega}_{\lambda} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\lambda}(\tilde{\lambda}^{(b)})$$

Concluding remarks

In the M-closed view, estimating the 'parameter' that describes the true model leads to various forms of Bayes factors, but the idea can also give rise to other interesting procedures, both for model selection and model averaging.

On the other hand, in both the M-closed and M-completed views, the resulting procedures are akin to finding a 'maximum likelihood estimator'. If necessary, a weighted likelihood bootstrap scheme can be used to determine appropriate weights for model averaging. Such weights can in turn be used as alternatives to Bayes factors for model comparison.

We have at our disposal a wide range of procedures for model selection and model averaging. Feasibility and computational cost usually dictate which one we use in practice. The WLB procedures arising from the M-completed view are simpler than those arising from the M-closed view.

Related ideas have been recently discussed from an M-open perspective by Yao, et al. (2018), who take the idea of *stacking* from the point estimation literature and generalize it to the combination of predictive distributions. They also propose a bootstrapped MA as an approximation for situations where computational cost is an issue.

Thank you very much for your attention!

References

- Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society B*, **53**, 111–142.
- Berger, J.O., Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.
- Bernardo, J.M. & Smith, A.F.M. (2000). *Bayesian Theory*. Chichester: Wiley.
- Clyde, M. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics 6*, pp. 157–185. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.) Oxford: University Press.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* **57**, 45–97.
- George, E.I. & McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-–373.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- Gutiérrez-Peña, E., Rueda, R. & Contreras-Cristán, A., (2009). Objective parametric model selection procedures from a Bayesian nonparametric perspective. *Computational Statistics and Data Analysis* **53**, 4255–4265.
- Gutiérrez-Peña, E. & Walker, S.G. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review* **73**, 309–330.
- Newton, M.A. & Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society B*, **56**, 3–48.
- O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, **57**, 99–138.
- Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis* **13**, 917–944.