Model Selection as Point Estimation
by Eduardo Gutierrez-Pena

Discussion by Ed George
Wharton, University of Pennsylvania

In Honor of Luis Perrichi
OBayes 2022, Santa Cruz, CA
September 8, 2022

Model Selection as Point Estimation
by Eduardo Gutierrez-Pena

Discussion by Ed George
Wharton, University of Pennsylvania

In Honor of Luis Perrichi
OBayes 2022, Santa Cruz, CA
September 8, 2022

Happy Birthday Luis!

Model Selection as Point Estimation
by Eduardo Gutierrez-Pena

Discussion by Ed George
Wharton, University of Pennsylvania

In Honor of Luis Perrichi
OBayes 2022, Santa Cruz, CA
September 8, 2022

M-Closed versus M-Completed

- Different frameworks for performing and evaluating inference
- Which is most common? most realistic?
- Model selection is coherently treated from a Bayesian point of view in the M-Closed framework.
- But can it be coherently treated in the M-Completed framework?
- As Eduardo shows us, the answer is yes!!!

Parametric Inference in the M-Closed Framework

- Problem Considered: Select an estimator of the true predictive density $f(x)$ based on iid data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ from $f(x)$.

- Suppose it can be assumed that $f(x)$ belongs to a simple parametric model class

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

- Some candidates for $\widehat{f}(x)$ discussed by Eduardo:
  - $f(x|\hat{\theta}_{ML})$ using no prior
  - $f(x|\phi^0, \boldsymbol{x}) = \int f(x|\theta)\, p(\theta|\phi^0, \boldsymbol{x})\, \mathrm{d}\theta$ using $p(\theta|\phi^0)$
  - $f(x|\hat{\theta}_{MAP})$ using $p(\theta|\phi^0)$
  - $\frac{1}{B}\sum_{b=1}^{B} f(x|\tilde{\theta}^{(b)})$ using a WLB sample but no prior

Parametric Inference in the M-Closed Framework

- We are here immediately faced with some Plug-in versus Bayes choices
- Such choices were treated as the Estimative vs Predictive controversy in the early 1970's
- This was largely settled for KL risk from a decision theory point of view by Aitkinson in 1975:
  - $\widehat{f}(x) = f(x|\phi^0, \boldsymbol{x})$ is Bayes and hence best under $p(\theta|\phi^0)$
  - $\widehat{f}(x) = f(x|\phi^0, \boldsymbol{x})$ under the uniform prior dominates the plug-in $\widehat{f}(x) = f(x|\hat{\theta}_{ML})$ when all $f \in \mathcal{F}$ are Normal
- Interestingly the WLB predictive $\widehat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f(x|\tilde{\theta}^{(b)})$ converges to $\widehat{f}(x) = f(x|\phi^0, \boldsymbol{x})$ when $p(\theta|\phi^0)$ is the uniform prior.

Parametric Inference in the M-Closed Framework

- Hierarchical elaboration of $\mathcal{F}$ yields larger classes

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

with $f^*(x|\phi) = \int f(x|\theta)\, p(\theta|\phi)\, \mathrm{d}\theta$ for hyperparameter ϕ.

- Candidates here for $\widehat{f}(x)$ discussed by Eduardo:
  - $f^*(x|\hat{\phi}_{EB}, \boldsymbol{x})$ using no hyperprior
  - $f^*(x|\lambda^0, \boldsymbol{x}) = \int f^*(x|\phi, \boldsymbol{x})\, p^*(\phi|\lambda^0, \boldsymbol{x})\, \mathrm{d}\phi$ using prior $p^*(\phi|\lambda^0)$
  - $f^*(x|\hat{\phi}_{MAP}, \boldsymbol{x})$ using $p^*(\phi|\lambda^0)$
  - $\frac{1}{B}\sum_{b=1}^{B} f^*(x|\tilde{\phi}^{(b)}, \boldsymbol{x})$ using a WLB sample but no hyperprior

- It is still clear that posterior predictive estimates are best under their corresponding priors (from complete class theorems).

- This WLB predictive may well be better than the EB plug-in.

Parametric Inference in the M-Closed Framework

- Further model elaboration of $\mathcal{F}^*$ yields
$$\mathcal{F}^{**} = \{f^{**}(x|\lambda) : \lambda \in \Lambda\}$$
with $f^{**}(x|\lambda) = \int f(x|\theta_\lambda, \lambda)\, p(\theta_\lambda|\lambda)\, \mathrm{d}\theta_\lambda$ for model λ

- Candidates here for $\widehat{f}(x)$ discussed by Eduardo:
  - $f^{**}(x|\hat{\lambda}_{BF}, \boldsymbol{x})$ where $\hat{\lambda}_{BF} = \arg\max_{\Lambda} f^{**}(\boldsymbol{x}|\lambda)$
  - $f^{**}(x|\omega^0, \boldsymbol{x}) = \int f^{**}(x|\lambda, \boldsymbol{x})\, p^{**}(\lambda|\omega^0, \boldsymbol{x})\, \mathrm{d}\lambda$ using $p^{**}(\lambda|\omega^0)$
  - $f^{**}(x|\boldsymbol{\omega}^0, \boldsymbol{x}) = \sum_{\lambda=1}^{m} \omega_\lambda^0(\boldsymbol{x})\, f^{**}(x|\lambda, \boldsymbol{x})$ using $p^{**}(\lambda|\omega^0) = \omega_\lambda$
  - $f^{**}(x|\hat{\lambda}_{PO}, \boldsymbol{x})$ where $\hat{\lambda}_{PO} = \arg\max_{\Lambda} p^{**}(\lambda|\omega^0, \boldsymbol{x})$
  - $\frac{1}{B}\sum_{b=1}^{B} f^{*}(x|\tilde{\phi}^{(b)}, \boldsymbol{x})$ using a WLB sample but no model space prior

- Interestingly, WLB provides automatic model weight estimates when the set of models under consideration is discrete. These provide avenues for model averaging and model selection.

Parametric Inference in the M-Closed Framework

- Lastly, the reduction of $\mathcal{F}^{**}$ to discrete model mixtures is considered
$$\mathcal{F}^{***} = \{f^{***}(x|\boldsymbol{\omega}) : \boldsymbol{\omega} \in \Omega\}$$
with $f^{***}(x|\boldsymbol{\omega}) = \sum_{\lambda=1}^{m} \omega_\lambda \, f^{**}(x|\lambda)$

- Candidates here for $\widehat{f}(x)$ discussed by Eduardo:
  - $f^{***}(x|\hat{\boldsymbol{\omega}}_E, \boldsymbol{x})$ where $\hat{\boldsymbol{\omega}}_E = \arg\max_{\Omega} f^{***}(\boldsymbol{x}|\boldsymbol{\omega})$
  - $f^{***}(x|\hat{\boldsymbol{\omega}}, , \boldsymbol{x}) = \sum_{\lambda=1}^{m} \hat{\omega}_\lambda \, f^{**}(x|\lambda, \boldsymbol{x})$ where $\hat{\boldsymbol{\omega}} = \arg\max_{\Omega} p^{***}(\boldsymbol{\omega}|\boldsymbol{\alpha}^0, \boldsymbol{x})$
  - $f^{***}(x|\boldsymbol{\alpha}^0, \boldsymbol{x}) = \sum_{\lambda=1}^{m} E[\omega_\lambda|\boldsymbol{\alpha}^0, \boldsymbol{x}] \, f^{**}(x|\lambda, \boldsymbol{x})$

- Interesting variations of these predictive estimates obtain with different prior choices for ω.

Parametric Inference in the M-Completed Framework

- Turning now to the M-Completed framework, suppose of interest is a class of parametric predictive distributions

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

  such as any of the classes of predictive estimates constructed for $\mathcal{F}, \mathcal{F}^*, \mathcal{F}^{**}, \mathcal{F}^{***}$.

- From the M-Completed perspective, a prior distribution cannot be used to describe the uncertainty surrounding model selection from $\mathbf{F}_K$.

- Indeed, honest acknowledgement of uncertainly here requires a prior that puts probability 1 on

$$\mathbf{F} = \{F : F \text{ is a probability distribution on } \mathcal{X}\},$$

  such as a Dirichlet process prior $F \sim \mathcal{DP}(a_0 F_0)$.

- Fully respecting this limitation for M-Completed contexts, Eduardo and coauthors have proposed a coherent approach that maximizes expected log utility wrt F for selection of $f_\kappa \in \mathbf{F}_K$.

Parametric Inference in the M-Completed Framework

- Their maximum expected utility selection approach for the predictive density problem proceeds as follows.

- The posterior mean of $F \sim \mathcal{DP}(a_0 F_0)$, with $a_0 = 0$ for objectiveness, is simply the empirical cdf of $\boldsymbol{x}$, denoted $\widehat{F}(\cdot)$.

- The posterior expected utility of any $f_\kappa \in \mathbf{F}_K$ is then

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

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

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

- Notice that this $\widehat{\kappa}$ minimizes the KL distance from $f_\kappa \in \mathbf{F}_K$ to $\widehat{F}$.

Parametric Inference in the M-Completed Framework

- Illustrating the potential of this approach, Eduardo considers the class of model averaged predictive estimates of the form

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

where each $f_j(x \mid \boldsymbol{x}) = \int f_j(x \mid \theta_j)\, \pi_j(\theta_j \mid \boldsymbol{x}) \, \mathrm{d}\theta_j$ is obtained with a posterior reference prior.

- From this class, the optimally weighted predictive model is simply obtained by the $\widehat{\omega}$ which maximizes

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

- Note that this strategy is providing an automatic and coherent approach to selection of the averaging weights.

Parametric Inference in the M-Completed Framework

- Overall, it seems clear that the ultimate effectiveness of this M-Complete strategy rests on the quality of the predictive densities that comprise the parametric class $\mathbf{F}_K$ of interest.

- Thus, Eduardo is exactly right to emphasize the importance of the construction of the classes of surrogate predictive densities to be considered.

- In this regard, his development of the wide varieties of predictive estimates for the classes $\mathcal{F}, \mathcal{F}^*, \mathcal{F}^{**}, \mathcal{F}^{***}$, is a generous master class in how we might proceed forward with these constructions.

Congratulations
Eduardo!

Happy Birthday Luis!