Valid uncertainty quantification about a model

Ryan Martin
Department of Statistics, North Carolina State University
www4.stat.ncsu.edu/~rmartin/

NC STATE

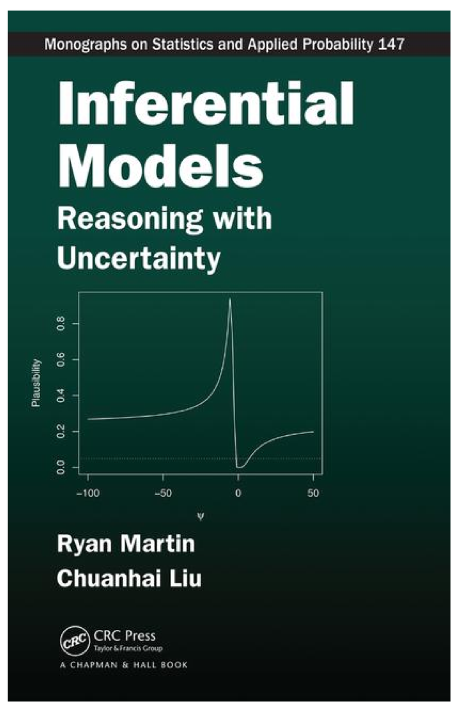
1 Introduction

We are familiar with uncertainty quantification about parameters for a given model, but we are much less familiar with uncertainty quantification about the model itself. Relevant questions include: how to attach reliable measures of uncertainty to the candidate models?, what form should these measures take?, and what does it mean to be "reliable" in this context?

First thought: construct a Bayesian posterior distribution for the model. But there are challenges: first, the prior choice of prior distribution matters in these cases; second, the posterior comes with no reliability guarantees. To achieve a certain kind of reliability—what I call *validity*—additive probabilities won't do, the beliefs must be non-additive (arXiv:1607.05051). So then the question is: *how to construct valid non-additive beliefs about models?*

2 What's in the paper?

- Inference under fixed model
  - basic inferential model (IM) construction
  - validity property
  - dimension reduction
- The uncertain model problem
- Valid uncertainty quantification about a model
- Bayesian lack of validity
- IM-based model assessment
  - a few general concepts
  - focus on Gaussian signal detection
  - validity result
  - selection rule properties



3 Inferential model basics

3.1 IM construction

Start with data, Y, and a statistical model $\mathsf{P}_{Y|\theta}$ depending on a parameter $\theta \in \Theta$. For now, the model is taken as *given*. The general IM construction is as follows.

A-step. *Define an* association *consistent with the statistical model. That is, introduce* $a : \Theta \times \mathbb{U} \to \mathbb{Y}$ *such that data* $Y \sim \mathsf{P}_{Y|\theta}$ *can be simulated by the algorithm* $Y = a(\theta, U)$, $U \sim \mathsf{P}_U$, *where* $U \in \mathbb{U}$ *is an auxiliary variable and its distribution,* P_U *does not depend on any unknown parameters. Then define the set-valued map*

$$\Theta_y(u) = \{\vartheta : y = a(\vartheta, u)\}, \quad u \in \mathbb{U}.$$

P-step. *Introduce a suitable random set* $\mathcal{S}$, *with distribution* $\mathsf{P}_{\mathcal{S}}$, *taking values in* $2^{\mathbb{U}}$, *designed to predict the unobserved value of the auxiliary variable* U.

C-step. Combine Θ_y *and* $\mathcal{S}$ *to get a new random set*

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u). \tag{1}$$

Then the distribution of $\Theta_y(\mathcal{S})$ *in* (1), *as a function of* $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$, *for fixed* y, *determines the IM output:*

$$\mathsf{bel}_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \subseteq A\} \quad \textit{and} \quad \mathsf{pl}_y(A) = 1 - \mathsf{bel}_y(A^c), \quad A \subseteq \Theta.$$

3.2 Validity property

Under mild conditions on the user-specified random set $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$, the IM is *valid* in the sense that

$$\sup_{\theta \in A} \mathsf{P}_{Y|\theta}\{\mathsf{pl}_Y(A) \leq \alpha\} \leq \alpha, \quad \forall\, \alpha \in (0,1), \quad \forall\, A \subseteq \Theta. \tag{2}$$

In words, true hypotheses—those that contain the true θ—being assigned low plausibility is a rare event relative to the posited model. Among other things, this implies that hypothesis tests and confidence regions based on the plausibility function have frequentist error rate control.

3.3 Dimension reduction

Often $\dim(U) > \dim(\theta)$, so it's advantageous to reduce the dimension of U before introducing the random set. Suppose that there exists a pair of one-to-one mappings $y \mapsto (T(y), H(y))$ and $u \mapsto (\tau(u), \eta(u))$ such that the original association, $Y = a(\theta, U)$, can be re-expressed as

$$T(Y) = b(\theta, \tau(U)) \quad \text{and} \quad H(Y) = \eta(U),$$

where b is a known function analogous to the original a. Two key observations:

- there is no θ in the second expression, so $\eta(U)$ is *observed* and does not need to be predicted;
- the unobservable feature $\tau(U)$ is of lower dimension than U, which simplifies the random set construction and improves efficiency via conditioning.

4 Uncertain model problem

Let $\mathscr{M}$ be a model index set and write $\{\mathsf{P}_{Y|(M,\theta_M)} : M \in \mathscr{M}, \theta_M \in \Theta_M\}$. This boils down to working with an "expanded" parameter, namely, (M, θ_M). This suggests treating the uncertain model problem as one where θ_M as a nuisance parameter to be marginalized out.

5 Marginalizing out the model-specific parameters

When model is uncertain, the association depends on $\theta = (M, \theta_M)$ and takes the form

$$Y = a_M(\theta_M, U), \quad U \sim \mathsf{P}_U.$$

Those maps discussed previously implicitly depend on the assumed model, so if M is uncertain then we actually have

$$T_M(Y) = b_M(\theta_M, \tau_M(U)) \quad \text{and} \quad H_M(Y) = \eta_M(U).$$

Note that the second equation depends on M but not θ_M; under certain conditions, described in the general IM marginalization theory, the first equation can be ignored, leaving

$$H_M(Y) = \eta_M(U)$$

as the relevant *marginal association* for inference on M. Think of this as a generalization of the concept taught in basic linear regression: use the residuals to assess the quality of the model itself. Then an IM for the model M can be constructed exactly as before, and a validity property, like (2), should emerge automatically from the general theory.

6 Gaussian signal detection

Consider the classical normal means problem, $Y = \theta + U$, where $U \sim \mathsf{P}_U = \mathsf{N}_n(0, I_n)$. A number of means are exactly zero (noise) and others are non-zero (signal). The goal is to identify the signals, i.e., what configuration $M \subseteq \{1, 2, \ldots, n\}$ of indices corresponds to non-zero θ's?

Re-express the full parameter vector θ as

$$\theta = (M, \theta_M) := (\text{non-zero indices}, \text{non-zero values}).$$

Splitting the baseline association into two parts and marginalizing θ_M is immediate in this case:

$$\left.\begin{array}{l} Y_M = \theta_M + U_M \\ Y_{M^c} = U_{M^c} \end{array}\right\} \xrightarrow{\text{marginalize}} Y_{M^c} = U_{M^c}.$$

The right-most expression carries some nice intuition: model M is plausible if the observed y_{M^c} resembles a vector of iid standard normals. This can be made precise by introducing a random set, $\mathcal{S}$, to predict the unobserved value of U. That is, if $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$ is a random set on the U-space, then

$$\mathscr{M}_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \{M : y_{M^c} = u_{M^c}\} = \{M : \mathcal{S}_{M^c} \ni (0_M, y_{M^c})\},$$

where $(0_M, y_{M^c})$ is the n-vector with 0's filling in around y_{M^c}. For the class of hypotheses

$$A_M = \{M' \in \mathscr{M} : M' \subseteq M\}, \quad M \in \mathscr{M}, \tag{3}$$

which corresponds to a claim that M contains all the signals, the marginal plausibility function is

$$\mathsf{mpl}_y(A_M) = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \ni (0_M, y_{M^c})\}, \quad M \in \mathscr{M}.$$

If $\mathcal{S}$ is the n-hypercube, centered at the origin, with random half-width, i.e.,

$$\mathcal{S} = \{u \in \mathbb{R}^n : \|u\|_\infty \leq \|U\|_\infty\}, \quad U \sim \mathsf{P}_U,$$

then the above marginal plausibility function at A_M equals

$$\mathsf{mpl}_y(A_M) = 1 - F_n(\|y_{M^c}\|_\infty), \quad M \in \mathscr{M}, \tag{4}$$

where $F_n(z) = \mathsf{P}\{\mathsf{ChiSq}(1) \leq z^2\}^n$ is easy to compute.

Theorem. *The IM with* $\mathsf{mpl}_y(A_M)$ *given by* (4) *satisfies the following validity property:*

$$\sup_{\theta_M \in \mathbb{R}^{|M|}} \mathsf{P}_{Y|M,\theta_M}\{\mathsf{mpl}_Y(A_M) \leq \alpha\} \leq \alpha, \quad \forall\, \alpha \in (0,1), \quad \forall\, M \in \mathscr{M}, \quad A_M \textit{ in } (3).$$

This provides justification for the use of mpl_y for uncertainty quantification, at least along those hypotheses A_M in (3). Moreover, it yields a selection rule with good properties.

Theorem. *Fix* $\alpha \in (0,1)$ *and define a selection rule*

$$\widehat{M}_\alpha(y) = \textit{smallest } M \textit{ such that } \mathsf{mpl}_y(A_M) > \alpha. \tag{5}$$

Then $\mathsf{P}_{Y|M,\theta_M}\{\widehat{M}_\alpha(Y) \subseteq M\} \geq 1 - \alpha$ *for all* $M \in \mathscr{M}$.

The paper gives some numerical examples to show that this selection rule compares favorably to a number of more traditional methods, such as lasso, Benjamini–Hochberg, etc.

Here's a tidbit, with $n = 20$ and $\alpha = 0.10$. Note that the IM procedure (5) satisfies

$$\text{Subset} + \text{Equal} = 0.914 > 0.90 = 1 - \alpha$$

as predicted by the above theorem.

Signal	Method	Subset	Equal	Superset	FDR	FNR
4	IM	0.214	0.700	0.060	0.034	0.011
	Lasso	0.002	0.094	0.890	0.595	0.001
	Thresh	0.102	0.674	0.188	0.088	0.006
	BH	0.126	0.644	0.206	0.092	0.007