Proceedings of Machine Learning Research 103:295–303, 2019 ISIPTA 2019

On Valid Uncertainty Quantification About a Model

Ryan Martin RGMARTI3@NCSU.EDU
Department of Statistics, North Carolina State University, USA

Abstract

Inference on parameters within a given model is familiar, as is ranking different models for the purpose of selection. Less familiar, however, is the quantification of uncertainty about the models themselves. A Bayesian approach provides a posterior distribution for the model but it comes with no validity guarantees, and, therefore, is only suited for ranking and selection. In this paper, I will present an alternative way to view this model uncertainty problem, through the lens of a valid inferential model based on random sets and non-additive beliefs. Specifically, I will show that valid uncertainty quantification about a model is attainable within this framework in general, and highlight the benefits in a classical signal detection problem.

Keywords: Bayesian; inferential model; marginalization; plausibility; random set; variable selection.

1. Introduction

Statistical inference, with a certain model, is already a challenging problem, with a wide range of possible solutions, even several competing philosophies, and has been a source of heated debate for at least a century; see, for example, Mayo (2018). However, it is rare for the statistical model to be certain in real applications, and uncertainty about the model itself creates additional challenges. A major one being that the unknown parameters are often determined by the model so, in the uncertain model case, the data analyst must first select a model before the relevant statistical inference problem is defined. But when the data itself determines, e.g., the hypotheses to be tested, this creates potentially serious biases. One specific example of this bias is discussed in Hong et al. (2018), and Taylor and Tibshirani (2015) describe the general impact of model uncertainty and selection biases. My focus here in this paper is on the model itself, not on any model-specific parameter.

In this uncertain model case, there are a number of existing approaches. In a classical or frequentist framework, the most familiar strategy is to use the data to drive *selection* of a model; and then the quality of a selection strategy is evaluated based on the frequency at which it identifies the "correct" model in repeated sampling. Of course, this approach is analogous to point estimation and, therefore, does not provide any uncertainty quantification. To address this, there are other approaches, most notably, Bayesian (e.g., Clyde and George, 2004), that return as output a probability distribution over the set of candidate models and, hence, provide uncertainty quantification. The critical question, however, is if the uncertainty quantification it provides is valid in any meaningful sense. A basic and uncontroversial requirement is that the probabilities assigned to "wrong" models should be small and those assigned to "not wrong" models should be relatively large. Challenges arise when it comes time to define what "small" and "large" mean, and I will address this in Section 3.

To set the scene for the present developments, consider first the certain model case where inference about the model parameters is the goal. I will define an *inferential model* as a map that takes the data analysts' inputs—including the available data, that certain statistical model, and perhaps other things, such as a prior distribution—to a function whose input is a hypothesis about the parameter and the output represents his/her degrees of belief in the truthfulness of that hypothesis. In other words, an inferential model is just a rule by which the information available to the data analyst gets converted into degrees of belief about the unknowns. This definition is quite general, so perhaps it comes as no surprise that it covers a number of familiar approaches, such as Bayesian, along with some less familiar. More details are given in Section 2; see, also, Martin (2019). When it comes to uncertainty quantification, a desirable property is that the inferential model be calibrated in the sense that the degrees of belief it assigns to false hypotheses tend to be small in a sufficiently precise way. Since magnitudes of degrees of belief will be used by the data analyst to draw inference, such a calibration property will protect him/her from the undesirable risk of making "systematically misleading conclusions" (Reid and Cox, 2015). An interesting and surprising result is that the only inferential models that satisfy (a strong form of) this calibration property, called *validity* in Section 2, are those whose degrees of belief are non-additive. That is, inferential models with additive degrees of belief output, such as Bayesian, cannot satisfy the validity property and, therefore, suffer from what Balch et al. (2017) call *false confidence*. My view is that this calibration property is essential to the logic of statistical inference, so I will focus here on valid inferential models which necessarily return non-additive degrees of belief.

Given that beyond-probability considerations are required for valid statistical inference even in the certain model case, and that the uncertain model case is even more

challenging, one would expect that similar considerations are needed in the uncertain model case. My goal here is to present first a definition of validity in the case where the quantity of interest is the uncertain model, and to construct an inferential model that achieves this property, thereby, a framework for valid uncertainty quantification about a model. Before doing so, I review in Section 2 the construction of a valid inferential model for the unknown parameter in a certain model context, and then, in Section 3, describe the transition from certain to uncertain model under the proposed framework. Section 4 lays out a fairly general approach to separating and isolating the model component from the model-specific parameters and a notion of validity in this context. Details of the inferential model construction are presented in Section 5 in the context of an important Gaussian signal detection example. There I describe the precise formulation, prove a validity theorem, and show how this result can be used to derive a procedure having good frequentist properties; numerical results demonstrate that this inferential model-based procedure is as good or better than traditional methods across a range of settings.

2. Inferential Models

2.1. Definition

Let $Y \in \mathbb{Y}$ be the observable data and denote by $\mathscr{P} = \{\mathsf{P}_{Y|\theta} : \theta \in \Theta\}$ the certain statistical model. Then the goal is to make inference on the unknown θ that determines the distribution of Y based on an observation $Y = y$. In this setting, the *inferential model* is simply a map from the data, certain statistical model, and possibly other things (e.g., a prior distribution) to a belief function defined on the parameter space Θ, i.e.,

$$(y, \mathscr{P}, \ldots) \mapsto \mathsf{bel}_y,$$

where $\mathsf{bel}_y(A) \in [0,1]$ is interpreted as the data analyst's degrees of belief about the hypothesis $A \subseteq \Theta$ based on data $Y = y$, relative to the posited model, etc. This inferential model framework covers familiar modes of inference, such as Bayesian, as well as some less familiar, such as fiducial (e.g., Dempster, 1963; Fisher, 1973; Zabell, 1992), generalized fiducial (e.g., Hannig et al., 2016), confidence distributions (e.g., Schweder and Hjort, 2016), Dempster–Shafer theory (e.g., Kohlas and Monney, 1995; Dempster, 2008; Shafer, 1976), confidence structures (e.g., Balch, 2012), possibility measures (e.g., Dubois and Prade, 1988), and the approach in Martin and Liu (2013, 2016) described in more details below.

2.2. Validity Property

Formally, an inferential model with output bel_y is said to be *valid* if

$$\sup_{\theta \notin A} \mathsf{P}_{Y|\theta}\{\mathsf{bel}_Y(A) > 1-\alpha\} \leq \alpha, \quad \begin{cases} \forall\, \alpha \in (0,1) \\ \forall\, A \subseteq \Theta. \end{cases} \tag{1}$$

In words, false hypotheses—those that do not contain the true θ—being assigned high belief is a rare event relative to the posited model. This prevents false confidence and, thereby, systematically misleading conclusions. It is not difficult to see that an additive belief function cannot satisfy (1), the key being that (1) covers *all hypotheses* A.

If the belief function bel_y is non-additive, then it has a distinct dual, $\mathsf{pl}_y(A) = 1 - \mathsf{bel}_y(A^c)$, called a plausibility function. Since (1) covers all hypotheses, an equivalent condition can be expressed in terms of plausibility:

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

In words, this says that hypotheses which are not false—the ones that contain the true θ—will tend to be assigned relatively high plausibility.

The validity property offers a sort of calibration, so that it is clear what it means for belief in a hypothesis to be "large" or plausibility to be "small." It also immediately leads to decision procedures with guaranteed control on the frequentist error rates. Such properties will show up in what follows, but for the general results, see, e.g., Martin and Liu (2013, Sec. 3.4).

2.3. Construction of Valid IMs

Non-additivity alone is not enough to achieve validity. Discussion of non-additive beliefs having such properties can be found in Balch (2012) and Denœux and Li (2018). But, to my knowledge, what follows is the only general construction of a valid inferential model.

A-step *Define an* association *consistent with* $\mathscr{P}$*. 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*

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u). \tag{3}$$

Then the distribution of $\Theta_y(\mathcal{S})$ *in* (3)*, as a function of* $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$*, for fixed* y*, determines the inferential model output:*

$$\mathsf{bel}_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \subseteq A\}.$$

Under certain weak conditions on the user-specified $\mathcal{S}$, it can be shown that the inferential model constructed above achieves the validity property in (1). For the sake of space, and since these details won't be needed in what follows, I'll refer the reader to Martin and Liu (2013, 2016) for more on the random sets and the validity property.

2.4. Dimension Reduction

In the above discussion, I didn't comment on the dimension of Y, θ, or U. It will often be the case, e.g., in iid-data settings, that the dimension of Y and U will be greater than that of θ and, in such cases, it is advantageous, in terms of efficiency, to reduce the dimension of U before the introduction of a random set to predict its unobserved value. The key observation in Martin and Liu (2015a) is that, in such cases, there are features of U that are observed and, therefore, do not need to be predicted, thus allowing a dimension reduction. Moreover, by conditioning on those observed features, it is possible to sharpen the prediction of unobserved features. Below I briefly summarize this dimension reduction procedure since it plays a role in the developments of Section 4.

Typically, Y and U both are n-dimensional and here I'll assume that θ is d-dimensional, with $d < n$. I'll further assume 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 a known function analogous to the original a. Such a re-expression is possible in "regular" cases where the minimal sufficient statistic is also d-dimensional, among others. I'll assume this regularity here and in Section 4. Then there are two key observations: first, there is no θ in the second expression so the feature $\eta(U)$ is *observed* and does not need to be predicted; and second, the unobservable feature $\tau(U)$ is of lower dimension than U, which simplifies the random set construction and improves efficiency.

For the uncertain model context in Section 4, the re-expression above will play a different role. Specifically, the second expression will be free of the model-specific parameter, but will depend on the model itself. This will make it possible to marginalize over the nuisance parameters and get directly at the uncertain model of interest.

3. The Uncertain Model Problem

My jumping off point is the idea to write the model as part of the "full parameter." That is, if θ denotes the parameter that determines the distribution of observable data Y, then I will re-express θ as (M, θ_M), where $M \in \mathscr{M}$ is a model index and $\theta_M \in \Theta_M$ denotes the model-specific parameter. For example, perhaps Y might have a gamma distribution, a log-normal distribution, or perhaps something else. Then, when $M =$ gamma, θ_M contains a shape and scale parameter whereas, when $M =$ log-normal, θ_M consists of a mean and variance for $\log Y$. Another example is in linear regression, $Y = X\beta + \varepsilon$, where M could represent the set of predictor variables, i.e., columns of X, to be included in the structural part of the linear model, and θ_M could consist of the β coefficients that align with those active variables, along with any other parameters that determine the distribution of ε. The example in Section 5 is of this latter type. In general, in the uncertain model case, the full statistical model can be written as

$$\mathscr{P} = \{\mathsf{P}_{Y|M,\theta_M} : M \in \mathscr{M}, \theta_M \in \Theta_M\}. \tag{4}$$

The model $\mathscr{P}$ is certain and the index M is uncertain. Both $\mathscr{P}$ and M can/will be called "models," but the meaning should be clear from the context. The goal is to convert the observed data $Y = y$, relative to model $\mathscr{P}$, into information about the pair (M, θ_M) and, in particular, about M.

By decomposing the full parameter θ as (M, θ_M), it becomes clear that the uncertain model problem is just a special case where the full parameter can be split into disjoint sets of *interest* and *nuisance* parameters. That is, M is the interest parameter and θ_M is the nuisance parameter. Therefore, the goal is marginal inference on M.

As discussed in Section 1, in the classical/frequentist context, the focus is primarily on model selection, which boils down to point estimation of M. In the inferential model context, there are Bayesian and generalized fiducial strategies available which, of course, produce (additive) posterior probabilities about M. In particular, in the Bayesian setting, the marginal posterior distribution for M is given by

$$\pi_y(M) \propto \pi(M) \int_{\Theta_M} L_y^M(\theta_M)\, g_M(\theta_M)\, d\theta_M, \tag{5}$$

where π is the marginal prior distribution for M, g_M is the conditional prior density for θ_M, given M, L_y^M is the likelihood function under model M, and the proportionality constant is determined by summing the right-hand side above over all the models in $\mathscr{M}$. An example of such a calculation is given in Section 4.2.

Computational difficulties aside, Bayesian marginalization to the model index M is conceptually straightforward, but there are still some challenges. One is that, in the uncertain model case, "non-informative" priors for θ_M, given M, generally cannot be used so, in the typical case where little or no prior information is available, there's a risk of bias from poor prior specification. On top of that, there's a question of whether π_y in (5) provides meaningful or reliable uncertainty quantification about M. That is, will $\pi_Y(M)$

tend to be small and large, in some predictable way, when M is "wrong" and "not wrong," respectively? The existing theory can only establish that, as data become more informative, e.g., as sample size increases, the posterior will put mass 1 on the true M but, again, this only assists in model selection, not in uncertainty quantification. And that additivity of the posterior can cause lack-of-calibration even in the certain model case, gives strong reason to believe that non-additivity is needed in the more difficult uncertain model case to achieve valid uncertainty quantification.

4. Valid Inference About an Uncertain Model

4.1. Definition

Whatever kind of inferential model I'm considering, it will return to me (marginal) data-dependent degrees of belief about hypotheses concerning the uncertain model M. For example, the calculations above described how the Bayesian inferential model assigns its degrees of belief about M. That is, if $A \subseteq \mathscr{M}$ is some hypothesis about M, then, with a slight abuse of notation, one can get $\pi_y(A) = \sum_{M \in A} \pi_y(M)$, and this would the data analyst's degrees of belief in the truthfulness of hypothesis A based on y, with respect to the various inputs, including prior. I'll show a different way to construct (non-additive) degrees of belief about the model in Section 4.3 below.

As the above discussion reveals, there is nothing really special in the uncertain versus certain model context, at least when it comes to interpretation of the inferential model's (marginal) plausibility function output. That is, if a hypothesis about M is not true, then I would expect the plausibility assigned to that hypothesis to tend to be small, otherwise I'd be at risk of making systematically misleading conclusions if I rely on the magnitudes of my plausibility function values. Therefore, nothing really should change when it comes to the properties I'd like my plausibility function to satisfy. Consequently, the validity property can be stated very similar to before.

To emphasize that the present focus is on *marginal* plausibility assigned to hypotheses about the uncertain model M, I'll write mpl_y to denote that marginal plausibility function that depends explicitly on data y and implicitly on other things. Then the corresponding inferential model is valid (for M) if its marginal plausibility function satisfies

$$\sup_{M \in A} \sup_{\theta_M \in \Theta_M} \mathsf{P}_{Y|M,\theta_M}\{\mathsf{mpl}_Y(A) \leq \alpha\} \leq \alpha. \tag{6}$$

As before, in addition to providing a scale on which the plausibility function values can be interpreted, i.e., so one knows what "small" and "large" means, decision procedures with provable control on frequentist error rates can be easily derived; see Section 5.

A question is if it is possible to construct an inferential model that achieves the validity property (6). In the next section, I demonstrate that the Bayesian inferential model, as described in Section 3, *doesn't* achieve validity.

4.2. Illustration

Let $Y = (Y_1, Y_2)$, with $Y_i \sim \mathsf{N}(\theta_i, 1)$, $i = 1, 2$, independent. There are four "models" in this case, one for each zero and non-zero combination for (θ_1, θ_2); that is,

$$\mathscr{M} = \big\{\varnothing, \{1\}, \{2\}, \{1,2\}\big\},$$

where, e.g., $M = \{1\}$ means that $\theta_1 \neq 0$ and $\theta_2 = 0$. For my Bayesian formulation, I will take a uniform prior for M that assigns weight 0.25 to each of the four entries in $\mathscr{M}$. For the model-specific parameters, I take a $\mathsf{N}(0, v)$ prior for each non-zero θ_i in the model, where the to-be-specified variance, $v > 0$, controls the degree of prior uncertainty.

Here I will consider two model hypotheses, namely,

$$A = \varnothing \quad \text{and} \quad A = \big\{\varnothing, \{1\}\big\}.$$

For uncertainty quantification about the model, given $Y = y$, I will evaluate the posterior probabilities at each A above. For those two hypotheses, respectively, the posterior probability is given by

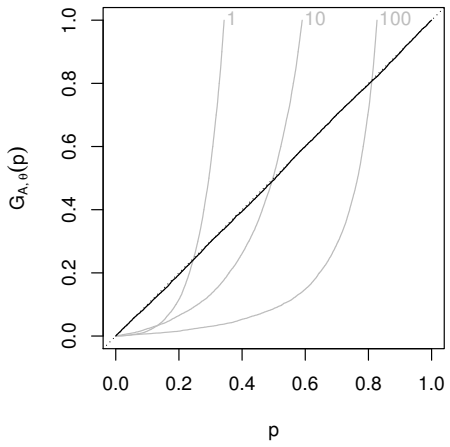
$$\begin{aligned} \pi_y(A) &\propto \mathsf{N}(y_1 \mid 0, 1)\mathsf{N}(y_2 \mid 0, 1) \\ \pi_y(A) &\propto \mathsf{N}(y_1 \mid 0, 1)\mathsf{N}(y_2 \mid 0, 1+v) \\ &\qquad + \mathsf{N}(y_1 \mid 0, 1)\mathsf{N}(y_2 \mid 0, 1), \end{aligned}$$

where the normalizing constant is the sum of all four marginal likelihoods. These are easy to compute, the question is how to interpret them. For example, the posterior probability assigned to the true model ought to be relatively large, but how large is large?

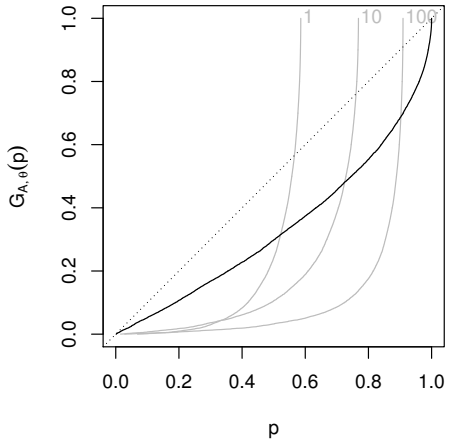
Figures 1–2 shows the distribution function

$$G_{A,\theta}(p) = \mathsf{P}_{Y|\theta}\{\pi_Y(A) \leq p\}, \quad p \in [0,1], \tag{7}$$

for the two hypotheses A of interest, for different values of the true $\theta = (\theta_1, \theta_2)$, and for three different values of v. The key observation here is that the scale on which the posterior probability should be interpreted—i.e., the scale that determines the "small" and "large" probabilities—depends on a number of things, including v, θ, and A. The reader likely is not surprised by this dependence, but contrast this observation with the way we tend to interpret probabilities. Indeed, the magnitudes of $G_{A,\theta}(0.3)$ reveal that a probability 0.3 can be interpreted as small, large, or something in between, depending on $(v, \theta, A, \ldots)$. This complicated dependence makes interpretation of the raw posterior probabilities difficult, which is perhaps why the tendency is to focus on the relative magnitudes, i.e., which posterior probabilities are largest? But the relative magnitudes are

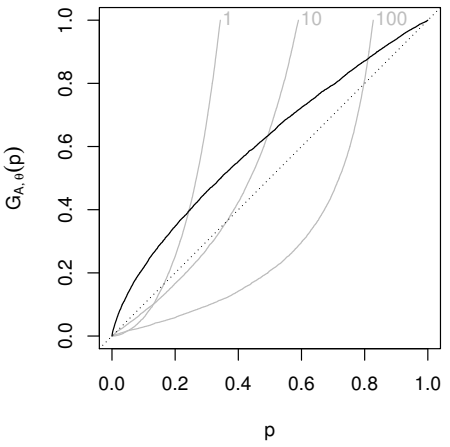


(a) $(\theta_1,\theta_2)=(0,0), A=\varnothing$

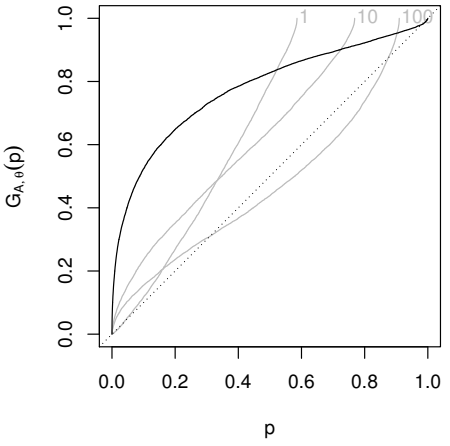


(b) $(\theta_1,\theta_2)=(0,1), A=\{\varnothing,\{1,\}\}$

Figure 1: Plots of the distribution function $G_{A,\theta}$ in (7) for the two hypotheses A and different configurations of the true (θ_1,θ_2) for which A is *true*. Gray lines correspond to the Bayesian posterior with prior variance $v=1,10,100$; black line corresponds to the solution described in Section 5.



(a) $(\theta_1,\theta_2)=(0,1), A=\varnothing$



(b) $(\theta_1,\theta_2)=(2,1), A=\{\varnothing,\{1,\}\}$

Figure 2: Same as in Figure 1 but where θ is such that the hypotheses A are *not true*.

useful only for selection purposes. If the goal is uncertainty quantification about the model, then the raw magnitudes are important and, therefore, a scale that doesn't depend on $(v,\theta,A,\ldots)$, such as $\mathsf{Unif}(0,1)$, is needed.

4.3. Construction

The choice to express the uncertain model case via a "full parameter" (M,θ_M) has the following effect on the association described in Section 2.3,

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

with P_U not depending on either M or θ_M. If I treat M as fixed, then there again is a mis-match between the dimension of U and that of θ_M, so I would apply the dimension reduction strategy outlined in Section 2.4. That is, I would identify maps $y \mapsto (T_M(y), H_M(y))$ and $u \mapsto (\tau_M(u), \eta_M(u))$ such that the above association can be rewritten in the form

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

But M is not fixed, so I don't actually *observe* $\eta_M(U)$ and, therefore, this re-expression of the baseline association can't be used as as in Section 2.4 to reduce the dimension. It is still useful, however, from the point of view of marginalization. That is, θ_M only appears in the first equation and, as Martin and Liu (2015b) argue, the fact that θ_M can be anything implies that the first equation actually carries no direct information about M and, hence, can be ignored. Applying that reasoning here in the uncertain model

context, I obtain a marginal association for M by ignoring the first equality in the above display and keeping only

$$H_M(Y) = \eta_M(U). \tag{8}$$

This marginalization step might or might not result in some dimension reduction, it depends on the example. The key point is that the nuisance parameter θ_M is absent in (8), creating an opportunity to get directly at M.

The intuition behind this argument is familiar even if the specifics are not. Details will be given in Section 5 below, but for now let me explain what's going on in the context of a specific example. Consider a regression problem where M signals which of the p predictor variables (columns in the $n \times p$ design matrix X) are to be included in the expression for the mean response. Then the model-specific parameter is $\theta_M = (\beta_M, \sigma^2)$, the subset of slope parameters and the error variance. One way to represent the data in this case is in terms of (a) the sufficient statistics $T_M(Y)$ for θ_M, given M, and (b) the residuals $H_M(Y)$ of the model M fit. If M is certain, then focus turns to the sufficient statistics and the residuals can be ignored; but if M is uncertain, then the residuals are needed to address questions about the model fit, etc. So, all that I'm really suggesting above is to implement the Stat 101 logic of using the residuals to assess the quality and appropriateness of a posited model.

To put this proposal into action, starting from the marginal association (8), I can introduce the same random set $\mathcal{S}$ for predicting the unobserved value of U; how this looks can vary across applications, and I'll give details in Section 5. Using the observed data, y, and the equation in (8), this random set can be mapped to $\mathscr{M}$ just like in the C-step above, i.e.,

$$\mathscr{M}_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \mathscr{M}_y(u),$$

where $\mathscr{M}_y(u) = \{M \in \mathscr{M} : H_M(y) = \eta_M(u)\}$. This leads to a data-dependent (marginal) belief/plausibility function on $\mathscr{M}$, give by

$$\mathsf{mpl}_y(A) = 1 - \mathsf{P}_{\mathcal{S}}\{\mathscr{M}_y(\mathcal{S}) \subseteq A^c\}, \quad A \subseteq \mathscr{M}, \tag{9}$$

which can be used to quantify uncertainty about M. Properties of this marginal inferential model in the general case are still being developed, but results are available in certain special cases, like in Section 5 below.

5. Gaussian Signal Detection

5.1. Setup and Construction

An important canonical example in statistics, signal processing, and elsewhere is the normal means model, where $Y \sim \mathsf{N}_n(\theta, I_n)$. What makes this example interesting/challenging is that the n-vector θ of unknown means contains some *exact zeros*. That is, those Y_i's with $\theta_i = 0$ correspond to "noise" and those with $\theta_i \neq 0$ correspond to "signal," and the goal is to detect those signals. This example has received considerable attention in the literature for the insights it provides on methods for estimation of high-dimensional parameters with a certain kind of sparsity structure; see, e.g., Martin and Walker (2014), Martin (2017), Martin and Ning (2018), and the references therein.

An association to describe these data is straightforward,

$$Y = \theta + U, \quad U \sim \mathsf{P}_U := \mathsf{N}_n(0, I_n), \tag{10}$$

where $\mathsf{N}_n(0, I_n)$ denotes the joint distribution of n independent standard normal random variables. Here $\theta \in \mathbb{R}^n$ is unknown, except that it contains a relatively small number of signals or non-zero values. This signal/noise separation, along with the fact that a primary goal is to learn where the signals are, suggests the introduction of a configuration M that serves as a model index. That is, $\theta = (M, \theta_M)$, where $M \subseteq \{1, 2, \ldots, n\}$ identifies which indices correspond to signals, and θ_M is the $|M|$-vector that contains the specific non-zero signal values.

Splitting the above baseline association into two parts, one free of the nuisance parameter θ_M, is immediate in this case. For a generic n-vector x and $M \subseteq \{1, 2, \ldots, n\}$, I'll write x_M to denote the subvector $(x_i : i \in M)$. Then (10) can be rewritten as

$$Y_M = \theta_M + U_M \quad \text{and} \quad Y_{M^c} = U_{M^c}.$$

Then, clearly, the marginal association (8) for M is

$$Y_{M^c} = U_{M^c}.$$

This 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 resulting from filling in around y_{M^c} with all 0's, and the marginal plausibility function as in (9). In particular, for the class of hypotheses

$$A_M = \{M' \in \mathscr{M} : M' \subseteq M\}, \quad M \in \mathscr{M}, \tag{11}$$

which corresponds to a claim that M contains all the signals, the plausibility function above simplifies:

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

This makes clear the above intuition that model M is plausible if y_{M^c} resembles a vector of iid standard normals.

What's a good choice of the random set $\mathcal{S}$ in this case? For situations like this, where the relevant signal detection questions correspond to zero/non-zero questions about each

individual parameter, treated simultaneously, I will follow the arguments in Martin et al. (2016) and use the random hyper-cube

$$\mathcal{S} = \{u : \|u\|_\infty \leq \|\tilde{U}\|_\infty\}, \quad \tilde{U} \sim \mathsf{P}_U,$$

where $\|x\|_\infty = \max_i |x_i|$ is the ℓ_∞-norm. Note that realizations of $\mathcal{S}$ correspond to realization of $\tilde{U} \sim \mathsf{P}_U$ and, hence, the distribution of $\mathcal{S}$ is determined by P_U. In this case, the marginal plausibility function at A_M in (11) simplifies:

$$\mathsf{mpl}_y(A_M) = 1 - F_n(\|y_{M^c}\|_\infty), \quad M \in \mathscr{M}, \tag{12}$$

where $F_n(z) = \mathsf{P}\{\mathsf{ChiSq}(1) \leq z^2\}^n$ is the distribution function of the maximum modulus of n iid standard normals, which is easy to compute. Of course, this is not the only choice of random set; see Section 5.3. Note, finally, that this IM construction is different from that presented in Liu and Xie (2014).

5.2. Validity Property

For the class $\{A_M : M \in \mathscr{M}\}$ of hypotheses in (11), which I'll show is practically relevant, a validity result for the inferential model constructed above is available.

Theorem 1 *For the inferential model constructed above, with marginal plausibility function in (9), the following validity result holds:*

$$\sup_{\theta_M \in \mathbb{R}^{|M|}} \mathsf{P}_{Y|M,\theta_M}\{\mathsf{mpl}_Y(A_M) \leq \alpha\} \leq \alpha,$$

for all $\alpha \in (0,1)$ and any $M \in \mathscr{M}$.

Proof Fix any $\alpha \in (0,1)$ and note that

$$\mathsf{mpl}_Y(A_M) \leq \alpha \iff \|Y_{M^c}\|_\infty \geq c_\alpha(M),$$

where $c_\alpha(M) = F_n^{-1}(1-\alpha)$. Therefore,

$$\begin{aligned}
\mathsf{P}_{Y|M,\theta_M}\{\mathsf{mpl}_Y(A_M) \leq \alpha\} &= \mathsf{P}_{Y|M,\theta_M}\{\|Y_{M^c}\|_\infty \geq c_\alpha(M)\} \\
&= 1 - \mathsf{P}\{\mathsf{ChiSq}(1) \leq c_\alpha^2(M)\}^{|M^c|} \\
&= 1 - \{(1-\alpha)^{1/n}\}^{|M^c|} \\
&\leq \alpha.
\end{aligned}$$

Since this equality holds for all $\alpha \in (0,1)$ and for all $M \in \mathscr{M}$, the claim follows. ■

Theorem 1 provides a calibration of the marginal plausibility function values for this special class of hypotheses, which aids in their interpretation. That is, judgments about what values of plausibility are small/large can be made (conservatively) by comparing to quantiles of a $\mathsf{Unif}(0,1)$ distribution. This explains why the black curves in Figure 1—where the corresponding A_M-type hypothesis is *true*—are on or to the right of $\mathsf{Unif}(0,1)$. Specifically, equality holds in the last line of the proof when $M = \varnothing$, which is the scenario in Figure 1(a).

5.3. Model Selection Procedure

If selection of a single model is required, then the validity result suggests the following strategy: for a given $\alpha \in (0,1)$, select the smallest M such that $\mathsf{mpl}_y(A_M) > \alpha$, i.e.,

$$\widehat{M}_\alpha(y) = \text{smallest } M \text{ such that } \mathsf{mpl}_y(A_M) > \alpha. \tag{13}$$

Intuitively, this corresponds to selecting the smallest model that is "sufficiently plausible," where the latter interpretation relies on the scale of mpl_y established by Theorem 1. In addition, I claim that this method satisfies a *selection validity* property:

$$\mathsf{P}_{Y|M,\theta_M}\{\widehat{M}_\alpha(Y) \subseteq M\} \geq 1-\alpha, \quad \forall\, M \in \mathscr{M}. \tag{14}$$

Theorem 2 *The rule (13) satisfies (14).*

Proof Clearly, $\widehat{M}_\alpha(y) \subseteq M$ is implied by $\mathsf{mpl}_y(A_M) > \alpha$. The latter event has probability no less than $1-\alpha$ according to Theorem 1, which proves the claim. ■

Note that (14) is analogous to family-wise error rate control in the context of multiple hypothesis testing, as discussed in, e.g., Lehmann and Romano (2005, Chap. 9). Interestingly, validity considerations in the context of uncertainty quantification about a model correspond to family-wise error control rather than, say, false discovery rate control as in Benjamini and Hochberg (1995); see below.

For this specific model selection task, one might be tempted to try a slightly different inferential model construction, one that uses a M-dependent random set for each individual $\mathsf{mpl}_y(A_M)$ calculation. The point is that, for hypothesis A_M, the n-dimensional random hyper-cube is used to predict a $|M^c|$-dimensional auxiliary variable, and the mis-match in dimension means the set is, in some sense, larger than necessary, which, in turn, suggests some potential inefficiency. To overcome this, an idea is to replace the n-dimensional random hyper-cube with a $|M^c|$-dimensional hyper-cube that is specifically tailored for calculating $\mathsf{mpl}_y(A_M)$ for the given M. This choice would still achieve the validity result in Theorem 1, with equality, and similarly in (14). Despite the apparent benefits of this adaptive-dimension approach, there's an important downside, namely, that it can only produce belief/plausibility for hypotheses of the form A_M, not for general hypotheses $A \subseteq \mathscr{M}$. Moreover, the numerical results below suggest that there is actually no substantial gain in efficiency using the adaptive version. For these latter two reasons, I choose not to emphasize this approach here.

To conclude this section, I'll present simulation results that compare the performance of the two proposed inferential model-driven selection methods—based on the fixed- and adaptive-dimension random sets, respectively, for $\alpha = 0.1$—with a few classical methods, namely, the lasso (Tibshirani, 1996), a universal thresholding rule

R	Method	Subset	Equal	Superset	FDR	FNR
2	IM_{fi}	0.880	0.022	0.004	0.082	0.066
	IM_{ad}	0.878	0.024	0.004	0.081	0.066
	Lasso	0.278	0.022	0.408	0.500	0.036
	Thresh	0.730	0.052	0.018	0.164	0.060
	BH	0.832	0.038	0.020	0.095	0.064
4	IM_{fi}	0.214	0.700	0.060	0.034	0.011
	IM_{ad}	0.208	0.698	0.068	0.037	0.010
	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
8	IM_{fi}	0	0.926	0.074	0.025	0
	IM_{ad}	0	0.918	0.082	0.028	0
	Lasso	0	0.096	0.904	0.582	0
	Thresh	0	0.764	0.236	0.083	0
	BH	0	0.738	0.262	0.098	0

Table 1: Simulation results for various model selection procedures, including those based on inferential models constructed with fixed- and adaptive-dimension random sets, with $n = 25$. Explanation of the methods and summaries is given in the text.

R	Method	Subset	Equal	Superset	FDR	FNR
2	IM_{fi}	0.900	0.000	0.000	0.072	0.087
	IM_{ad}	0.898	0.000	0.000	0.073	0.0871
	Lasso	0.130	0.000	0.238	0.574	0.040
	Thresh	0.786	0.000	0.000	0.135	0.081
	BH	0.826	0.002	0.000	0.099	0.083
4	IM_{fi}	0.562	0.356	0.028	0.018	0.019
	IM_{ad}	0.554	0.362	0.030	0.018	0.018
	Lasso	0.000	0.008	0.978	0.607	0.000
	Thresh	0.346	0.450	0.098	0.041	0.012
	BH	0.208	0.396	0.286	0.089	0.009
8	IM_{fi}	0	0.916	0.084	0.014	0
	IM_{ad}	0	0.904	0.096	0.016	0
	Lasso	0	0.006	0.994	0.624	0
	Thresh	0	0.798	0.202	0.036	0
	BH	0	0.580	0.420	0.088	0

Table 2: Same as in Figure 1 but with $n = 50$.

(e.g., Donoho and Johnstone, 1994), and the Benjamini–Hochberg false discovery rate controlling procedure (Benjamini and Hochberg, 1995). For the simulation settings, I consider $n \in \{25, 50\}$ and, in each case, the n-vector θ is filled with 10% signals of constant size $R \in \{2, 4, 8\}$. The comparisons are made in terms of the probability that $\widehat{M}_\alpha$ is a subset, equal to, and a superset of the true M, and the false discovery and non-discovery rate. Tables 1 and 2 summarize the results based on 500 Monte Carlo samples. A key observation is that, by property (14), both inferential model-based approaches satisfy $\text{Subset} + \text{Equal} \approx 1 - \alpha = 0.90$, but none of the other methods do. Naturally, the problem is harder for smaller signal size R and larger dimension n, and this is reflected in the results. But the inferential model-based methods proposed here perform very well compared to the more traditional methods, across the board.

6. Conclusion

This paper presents some first thoughts on the construction of an inferential model for valid uncertainty quantification about an uncertain model. The focus here was on an important albeit relatively simple normal means problem, but the main ideas—and promising numerical results—should generalize to other cases. I'll conclude here with a discussion of some open questions and perspectives.

In my present formulation, writing the full parameter θ as the pair (M, θ_M) was helpful because the problem could be viewed as one of marginal inference where the goal is to eliminate the nuisance parameter θ_M. But there are other situations where the goal is prediction of a new $\tilde{Y}$ or perhaps inference on some parameter that is common across all M, such as an error variance in regression. An interesting question is if a version of the Bayesian "model averaging" can be formulated in this different context. Techniques similar to those in Martin and Lingham (2016) are expected to be useful, but I've yet to attempt working out the details.

The modern versions of the signal detection problem described in Section 5 are those where the dimension n is very large but the signal is assumed to *sparse* in the sense that most of the θ_i's are zero. A now fairly standard approach is to develop a sparsity-encouraging prior distribution and carry about a Bayesian analysis. However, one often is lacking genuine prior information about the signals and, in such high-dimensional problems, the need to fill in the gap with an artificial prior can create problems. A new perspective was recently presented in Cella and Martin (2019), where they attempt to incorporate incomplete prior information into an inferential model formulation while maintaining validity. Such an approach would, at least in principle, fit in nicely here since an assumption of sparsity is effectively just an incomplete prior for the model index M. This is an important problem and the focus of ongoing work.

References

Michael S. Balch, Ryan Martin, and Scott Ferson. Satellite conjunction analysis and the false confidence theorem. `arXiv:1706.08565`, 2017.

Michael Scott Balch. Mathematical foundations for a theory of confidence structures. *Internat. J. Approx. Reason.*, 53(7):1003–1019, 2012.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to mul-

tiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995. ISSN 0035-9246.

Leonardo Cella and Ryan Martin. Incorporating expert opinion in an inferential model while maintaining validity. Unpublished manuscript, 2019.

Merlise Clyde and Edward I. George. Model uncertainty. *Statist. Sci.*, 19(1):81–94, 2004.

A. P. Dempster. Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, 34:884–891, 1963.

A. P. Dempster. The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, 48(2):365–377, 2008.

Thierry Denœux and Shoumei Li. Frequency-calibrated belief functions: review and new insights. *Internat. J. Approx. Reason.*, 92:232–254, 2018.

David L. Donoho and Iain M. Johnstone. Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, 99(2):277–303, 1994.

Didier Dubois and Henri Prade. *Possibility Theory*. Plenum Press, New York, 1988.

Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner Press, New York, 3rd edition, 1973.

Jan Hannig, Hari Iyer, Randy C. S. Lai, and Thomas C. M. Lee. Generalized fiducial inference: a review and new results. *J. Amer. Statist. Assoc.*, 111(515):1346–1361, 2016.

L. Hong, T. A. Kuffner, and R. Martin. On overfitting and post-selection uncertainty assessments. *Biometrika*, 105 (1):221–224, 2018.

Jürg Kohlas and Paul-André Monney. *A Mathematical Theory of Hints*, volume 425 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 1995.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

Chuanhai Liu and Jun Xie. Probabilistic inference for multiple testing. *Internat. J. Approx. Reason.*, 55(2): 654–665, 2014. ISSN 0888-613X.

R. Martin. Invited comment on the article by van der Pas, Szabó, and van der Vaart. *Bayesian Anal.*, 12(4):1254–1258, 2017.

Ryan Martin. False confidence, non-additive beliefs, and valid statistical inference. `arXiv:1607.05051`, 2019.

Ryan Martin and Rama T. Lingham. Prior-free probabilistic prediction of future observations. *Technometrics*, 58(2): 225–235, 2016.

Ryan Martin and Chuanhai Liu. Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313, 2013.

Ryan Martin and Chuanhai Liu. Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77 (1):195–217, 2015a.

Ryan Martin and Chuanhai Liu. Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110(512):1621–1631, 2015b.

Ryan Martin and Chuanhai Liu. *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2016.

Ryan Martin and Bo Ning. Empirical priors and coverage of posterior credible sets in a sparse normal mean model. `arXiv:1812.02150`, 2018.

Ryan Martin and Stephen G. Walker. Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.*, 8(2):2188–2206, 2014.

Ryan Martin, Huiping Xu, Zuoyi Zhang, and Chuanhai Liu. Valid uncertainty quantification about the model in linear regression. Unpublished manuscript, `arXiv:1412.5139`, 2016.

Deborah G. Mayo. *Statistical Inference as Severe Testing*. Cambridge University Press, Cambridge, 2018.

Nancy Reid and David R. Cox. On some principles of statistical inference. *Int. Stat. Rev.*, 83(2):293–308, 2015.

Tore Schweder and Nils Lid Hjort. *Confidence, Likelihood, Probability*, volume 41 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2016.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proc. Natl. Acad. Sci.*, 112 (25):7629–7634, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

S. L. Zabell. R. A. Fisher and the fiducial argument. *Statist. Sci.*, 7(3):369–387, 1992.