Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On an inferential model construction using generalized associations

Ryan Martin

Department of Statistics, North Carolina State University, United States

ARTICLE INFO

Article history:
Received 11 May 2016
Received in revised form 25 October 2016
Accepted 28 November 2016
Available online xxxx

Keywords:
Likelihood
Marginalization
Monte Carlo
Plausibility function
Random set
Validity

ABSTRACT

The inferential model (IM) approach, like fiducial and its generalizations, depends on a representation of the data-generating process. Here, a particular variation on the IM construction is considered, one based on generalized associations. The resulting generalized IM is more flexible in that it does not require a complete specification of the data-generating process and is provably valid under mild conditions. Computation and marginalization strategies are discussed, and two applications of this generalized IM approach are presented.

1. Introduction

An advantageous feature of the mainstream approaches to statistical inference is simplicity. On one hand, likelihood-based approaches, including "Frasian" inference (e.g., Reid, 2003; Fraser, 1990, 1991; Barndorff-Nielsen, 1991; Fraser, 2011) and certain forms of Bayesian inference (e.g., Bernardo, 1979; Ghosh, 2011; Berger et al., 2009, 2015), are simple in the sense that the calculations relevant to data analysis are largely (or completely) determined by the posited sampling model. On the other hand, frequentist approaches are also simple because the "do whatever works well" viewpoint is extremely flexible. This is in sharp contrast with fiducial inference (Fisher, 1973; Dawid and Stone, 1982; Barnard, 1995; Taraldsen and Lindqvist, 2013), its generalizations (Hannig, 2009; Hannig et al., 2016), and the recently proposed inferential model (IM) framework (Martin and Liu, 2013, 2015a,c,b), which appear to be not-so-simple in the sense that their construction depends on something more than the data and sampling model. In particular, the fiducial and IM construction begins with a specific representation of the data-generating mechanism, one that determines but is not determined by the sampling model. This data-generating mechanism identifies an auxiliary variable, or pivotal quantity, that controls the random variation in the observable data. A familiar example of this kind is the regression model, $Y = X\beta + \sigma\varepsilon$, where the random "$\varepsilon$" part controls the variation of the response Y around the deterministic "$X\beta$" part. That the fiducial and IM solutions depend on the choice of the data-generating mechanism may be seen as a shortcoming of these approaches.

One approach to deal with the choice of data-generating mechanism is to find one that is "best" in some sense; for example, Pal Majumdar and Hannig (2015) compare different data-generating mechanisms using higher-order asymptotics in the fiducial context. Since defining and identifying the "best" is difficult, I want to take a different approach. In this paper, building on Martin and Liu (2015b, Ch. 11), I want to incorporate the familiar frequentists' flexibility into the IM construction. This allows the user to construct a *generalized IM* without specifying a full data-generating mechanism, simplifying the

E-mail address: rgmarti3@ncsu.edu.

http://dx.doi.org/10.1016/j.jspi.2016.11.006
0378-3758/© 2016 Elsevier B.V. All rights reserved.

construction in several ways. First, just like in the likelihood-based approaches mentioned above, a generalized IM can be constructed based on the sampling model alone, or some function thereof, easing the burden on the user. Second, the generalized IM can be constructed based on a *generalized association* that involves only a one-dimensional auxiliary variable, which simplifies user's task of selecting a good predictive random set. Compare this to the basic IM approach where the user must first specify a data-generating mechanism and carry out some potentially non-trivial dimension-reduction steps (e.g., Martin and Liu, 2015a). Despite making substantial simplifications to the IM construction, it can be shown that this generalization preserves the IM's guaranteed validity property under mild conditions. Therefore, the generalized IM framework is a simple and widely applicable tool for valid, prior-free, probabilistic inference.

This paper's main contribution is the new perspective it brings to some more-or-less familiar ideas, results, and techniques. Specifically, all of the familiar considerations used in constructing statistical procedures fit within the seemingly rigid IM framework, and this has at least two useful consequences. First, working within the IM framework does not require that one abandon all the classical tools and ways of thinking—these can be merged seamlessly into the framework itself. Second, new insights concerning these classical tools can be gained when looking from an IM point of view; see Section 3.3.

The remainder of the paper is organized as follows. After some background on IMs in Section 2, a generalized IM approach is presented in Section 3, with a validity theorem and a special case that is relatively easy to implement, involving only a scalar auxiliary variable. Important practical considerations, namely, computation and marginalization, are discussed in Section 4, and two interesting and challenging applications – inference on the odds ratio in a 2 × 2 tables and inference on the error variance in a mixed-effects model – are presented in Section 5. Concluding remarks are made in Section 6.

2. Background on IMs

Let $Y \in \mathbb{Y}$ be the observable data, and write $\mathsf{P}_{Y|\theta}$ for the sampling model, which depends on an unknown parameter $\theta \in \Theta$. In the basic IM framework, described in Martin and Liu (2013), the starting point – the *A-step* – is to associate Y and θ with an unobservable auxiliary variable $U \in \mathbb{U}$ with known distribution P_U. Formally, write

$$Y = a(\theta, U), \quad U \sim \mathsf{P}_U. \tag{1}$$

Martin and Liu (2015a,c) argue that some dimension-reduction steps should be taken first before an association mapping is defined, so the left-hand side may be something different than the observable data, e.g., a minimal sufficient statistic. This dimension-reduction step is recommended, but it is not necessary to describe these details here. The result of the A-step is a set-valued mapping

$$\Theta_y(u) = \{\theta : y = a(\theta, u)\}, \quad u \in \mathbb{U}, \tag{2}$$

indexed by the observed $Y = y$. The main point is that the association determines the sampling model $\mathsf{P}_{Y|\theta}$ or, alternatively, the ingredients in (1) must be chosen to be consistent with the given sampling model. However, there may be several versions of the association that are consistent with the sampling model, and different versions may produce different inferences. This is not unlike the frequentists' choice of (approximate) pivot for constructing a test, confidence region, etc. In any case, the question of which association (1) to take, for given sampling model $\mathsf{P}_{Y|\theta}$, is an important one.

The second step in the basic IM construction – the *P-step* – is to predict the unobserved value of U in (1), corresponding to the observed $Y = y$, with predictive random set $\mathcal{S}$. The P-step is the defining feature of the IM framework, driving its essential properties and separating it from the approach described in Dempster (2008). The distribution $\mathsf{P}_\mathcal{S}$ of $\mathcal{S}$ is to be chosen by the user, subject to a certain "validity" condition, namely, that, if $f_\mathcal{S}(u) = \mathsf{P}_\mathcal{S}(\mathcal{S} \ni u)$, then

$$f_\mathcal{S}(U) \geq_{\text{st}} \mathsf{Unif}(0,1), \quad \text{as a function of } U \sim \mathsf{P}_U,$$

where "$\geq_{\text{st}}$" means "stochastically no smaller than," i.e.,

$$\mathsf{P}_U\{f_\mathcal{S}(U) \leq \alpha\} \leq \alpha, \quad \forall\, \alpha \in (0,1). \tag{3}$$

Intuitively, the random set $\mathcal{S}$ is meant to be "good" at predicting samples from P_U and (3) makes this precise: the $\mathsf{P}_\mathcal{S}$-probability of the event "$\mathcal{S} \ni u$" is small only for a set of u values with relatively small P_U-probability. Sufficient conditions for (3) are mild, so it is easy to find a valid predictive random set; in fact, most applications of IMs employ a simple "default" predictive random set, see (13).

The third and final step in the basic IM construction – the *C-step* – is to combine the association at the observed data $Y = y$ with the predictive random set $\mathcal{S}$. Specifically, one obtains a random subset of Θ:

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u). \tag{4}$$

The intuition behind this is as follows: if one believes that $\mathcal{S}$ contains the value of U corresponding to the observed $Y = y$ and the true θ, which is justified by (3), then one must also believe, with equal conviction, that $\Theta_y(\mathcal{S})$ contains the true θ. The IM output is the distribution of the random set $\Theta_y(\mathcal{S})$, which I will summarize with a plausibility function. Specifically, if $A \subset \Theta$, then the plausibility function at A is

$$\mathsf{pl}_y(A) = \mathsf{P}_\mathcal{S}\{\Theta_y(\mathcal{S}) \cap A \neq \varnothing\}.$$