Prior-free probabilistic prediction of future observations

Ryan Martin

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
`rgmartin@uic.edu`

Rama T. Lingham
Division of Statistics
Northern Illinois University
`rama@math.niu.edu`

June 11, 2018

Abstract

Prediction of future observations is a fundamental problem in statistics. Here we present a general approach based on the recently developed inferential model (IM) framework. We employ an IM-based technique to marginalize out the unknown parameters, yielding prior-free probabilistic prediction of future observables. Verifiable sufficient conditions are given for validity of our IM for prediction, and a variety of examples demonstrate the proposed method's performance. Thanks to its generality and ease of implementation, we expect that our IM-based method for prediction will be a useful tool for practitioners.

Keywords and phrases: Disease count data; environmental data; inferential model; plausibility; prediction interval; system breakdown data; validity.

1 Introduction

The prediction of future observations based on the information available in a given sample is a fundamental problem in statistics. For example, in engineering applications, such as computer networking, one might want to predict, to some degree of certainty, the time at which the current system might fail in order to have resources available to fix it. Despite this being a fundamental problem, the available literature does not seem to give any clear guidelines about how to approach a prediction problem in general. From a frequentist point of view, there are a host of techniques available for constructing prediction intervals in specific examples; see, for example, the book by Hahn and Meeker (1991) and papers by Fertig and Mann (1977), Bhaumik and Gibbons (2004), Krishnamoorthy et al. (2008),

and Wang (2010). Some general approaches to the frequentist prediction problem are presented in Beran (1990) and Lawless and Fredette (2005). From a Bayesian point of view, if a prior distribution is available for the unknown parameter, then the prediction problem is conceptually straightforward. The Bayesian model admits a joint distribution for the observed data and future data, so a conditional distribution for the latter given the former—the Bayesian predictive distribution—is the natural tool. Hamada et al. (2004) presents some applications of Bayesian prediction. The catch is that often there is no clear choice of prior. Default, or non-informative priors can be used but, in that case, it is not clear that the resulting inference will be meaningful in either a personal probabilistic or frequentist sense. Wang et al. (2012) propose a fiducial approach for prediction which, at a high-level, can be viewed as a sort of compromise to the frequentist and Bayesian approaches. They propose a very natural predictive distribution that obtains from the usual fiducial distribution for the parameter (Hannig 2009, 2013). They show that the prediction intervals obtained from the fiducial predictive distribution are asymptotically correct and perform well, in examples, compared to existing prediction intervals.

The fiducial approach is attractive because no prior distributions are required. However, like Bayesian posterior or predictive distributions based on default priors, fiducial distributions may not be calibrated for meaningful probabilistic inference, except possibly in the limit (Liu and Martin 2015). Recently, Martin and Liu (2013) proposed a general framework for prior-free probabilistic inference, called *inferential models* (IMs). This framework has some parallels with fiducial (e.g., Fisher 1959; Hannig 2009, 2013) and Dempster–Shafer theory (e.g., Dempster 2008; Shafer 1976) in that work is carried out in terms of unobservable auxiliary variables. There are also some connections with the frequentist confidence distributions (e.g., Xie and Singh 2013). The key difference between IMs and these other frameworks is the way that the auxiliary variables are handled. The important feature is that IMs provide probabilistic summaries of the information in data concerning the quantity of interest, and that these summaries are provably meaningful, not just in an asymptotic sense.

In this paper, we provide a IM-based solution to the problem of predicting future observations. The critical observation that drives the approach here is that predicting future observations is a marginal inference problem, one in which the full parameter itself is a nuisance parameter to be marginalized out. With this view, in Section 2.3 we apply the general marginalization principles in Martin and Liu (2015b) to eliminate the nuisance parameter, directly providing a marginal IM for the future observations. In Section 2.5 we give general conditions under which the resulting IM for prediction is valid, and discuss under what circumstances these conditions hold, in what cases they can be weakened, and other consequences. The key point is that the plausibility function obtained from a valid IM provides a probabilistic summary of the information in the observed data concerning the future data to be predicted; this function can be plotted to provide some visual summary. Moreover, the validity theorem demonstrates that the predictive plausibility interval, defined in (8), has the nominal frequentist coverage for all finite samples, not just in the limit. Our focus here is on the case of predicting a univariate future observable, but the multivariate case, discussed briefly in Section 4.3, requires some additional considerations. Several practical examples of prediction in the IM context are worked out in Section 3. These examples involve a variety of common models, and prediction problems in quality control, environmental, system breakdown,

and disease count applications are considered. To compare our IM-based solution to other existing methods, we focus on frequentist performance of our prediction intervals. In all the examples we consider, the IM intervals are competitive with the existing methods. R code for these examples is provided at `www.math.uic.edu/~rgmartin`. The take-away message is that the IM approach provides an easily implementable and general method for constructing meaningful prior-free probabilistic summaries of the information in observed data for inference or prediction; that these summaries can be converted to frequentist procedures with fixed-n performance guarantees and comparable efficiencies compared to existing methods is an added bonus.

2   Inferential models for prediction

2.1   Basic notation and terminology

The basic IM framework is introduced in Martin and Liu (2013), and further developments are presented in Martin and Liu (2015a,b). Here we want to briefly introduce the necessary notation and terminology. Suppose that the goal is inference on an unknown parameter θ. Martin and Liu (2013) present a three-step IM construction: association (A), prediction (P), and combination (C) steps. The starting point is identifying an association that links the data Y and the parameter θ to an unobservable auxiliary variable U. Often, a naive association will have auxiliary variables of higher dimension than the parameter, so special conditioning and/or marginalization techniques are needed to satisfactorily reduce the dimension of U. In any case, once this "baseline" association is specified, the A-step of the IM construction is completed. The P-step, unique to the IM approach, introduces a random set for predicting U. Finally, the C-step combines the association with the predictive random set in a natural way, yielding a pair of belief and plausibility functions for probabilistic inference on θ. The aforementioned papers give a number of examples of this approach, along with further explanation and theory.

2.2   Preview for prediction

Before getting into the general details about the prediction problem, we present a relatively simple example as a preview of our proposed IM approach. Consider a homogeneous Poisson process $\{N(t) : t \geq 0\}$ with rate $\theta > 0$. The arrival times $T_0, T_1, T_2, \ldots$ are such that $T_0 \equiv 0$ and the inter-arrival times $T_i - T_{i-1}$, $i \geq 1$, are independent exponential random variables with rate θ.

If the sampling scheme is to wait for the n-th arrival, then the sufficient statistic for θ in this model is $Y = T_n$, the last arrival time. Based on the arguments in Martin and Liu (2015a), the baseline association for θ is

$$Y = (1/\theta)G_n^{-1}(U), \quad U \sim \mathsf{Unif}(0,1),$$

where G_n is the $\mathsf{Gamma}(n, 1)$ distribution function. If inference on θ was the goal, then this would complete the A-step. However, suppose the goal is to predict $\tilde{Y} = T_{n+k}$, the time of the $(n+k)$-th arrival, for some fixed integer $k \geq 1$. Then θ itself is a nuisance parameter, and the quantity of interest is $\tilde{Y}$. From the baseline association above, we

can easily solve for θ in terms of (Y, U), i.e.,

$$\theta(Y, U) = G_n^{-1}(U)/Y.$$

Since $\tilde{Y} = T_{n+k}$, for given $Y = T_n$, equals Y plus an independent gamma random variable with shape k and rate θ, following Martin and Liu (2015b), we have a marginal association for $\tilde{Y}$ given by

$$\tilde{Y} = Y + \frac{1}{\theta(Y, U)} G_k^{-1}(\tilde{U}) = Y\left(1 + \frac{G_k^{-1}(\tilde{U})}{G_n^{-1}(U)}\right).$$

This completes the A-step for prediction. If R denotes the ratio in the far right-hand side above, then R has a generalized gamma ratio distribution (Coelho and Mexia 2007) with density function $f(r) \propto (1 + r)^{-(n+k)}$, $r > 0$. If F is the corresponding distribution function, then we may rewrite the marginal association as

$$\tilde{Y} = Y\{1 + F^{-1}(W)\}, \quad W \sim \mathsf{Unif}(0, 1).$$

Thus, we have successfully marginalized out the unknown parameter, directly associating the quantity to be predicted, $\tilde{Y}$, to the observed data, Y, and an auxiliary variable, W. Then, the general IM principles (Martin and Liu 2013, 2015a,b) can be applied directly. In particular, we apply the P- and C-steps to the association for $\tilde{Y}$, resulting in prior-free probabilistic prediction of the future arrival time. The next two subsections will describe the proposed approach in more detail, and our examples in Section 3 will demonstrate its generality, its quality performance, and its simplicity in applications.

2.3   General setup and the A-step

In the prediction problem, there is observed data Y and future data $\tilde{Y}$ to be predicted; the two are linked together through a common parameter θ. Here we assume that $\tilde{Y}$ is a scalar, though it could be a function of several future observations; see Section 4.3 for discussion on the multivariate prediction problem. Write the sampling model $\mathsf{P}_{Y|\theta}$ for Y in association form:

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

where P_U is known and free of θ. We call this the "baseline" association, and it connects observable data Y and unknown parameter θ to an unobservable auxiliary variable U. Despite its simple form, the baseline association is quite general, i.e., it covers cases outside the structural models in Fraser (1968); see Sections 3.3–3.4. For example, for any iid model with a smooth distribution function F_θ, take the i-th component of $a(\theta, U)$ to be $F_\theta^{-1}(U_i)$ for $U_i \sim \mathsf{Unif}(0, 1)$. Intuitively, any model that can be simulated has a form (1).

As our first step, assume that this baseline association can be re-expressed as

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

for functions (T, H) and (τ, η) such that $y \mapsto (T(y), H(y))$ and $u \mapsto (\tau(u), \eta(u))$ are one-to-one. A key feature of this decomposition is that a solution $\theta = \theta(y, v)$ of the equation $T(y) = b(\theta, v)$ is available for all (y, v). By conditioning on the observed value, $H(Y)$, of $\eta(U)$, this association can then be reduced as follows:

$$T(Y) = b(\theta, V), \quad V \equiv \tau(U) \sim \mathsf{P}_{\tau(U)|\eta(U)=H(Y)}.$$

Martin and Liu (2015a) show that such a decomposition exists in broad generality. For simplicity, we assume here that $\tau(U)$ and $\eta(U)$ are independent, so the conditioning can be dropped, i.e., $\mathsf{P}_{\tau(U)|\eta(U)=H(Y)} \equiv \mathsf{P}_{\tau(U)}$. This assumption holds for many problems, including those in Section 3. Dependence in this context is only a technical complication, not conceptual, so we focus here on the simpler case of independent $\tau(U)$ and $\eta(U)$; the dependent case is discussed further in Section 4.2.

For the observed data Y and the future data $\tilde{Y}$, write a joint association:

$$T(Y) = b(\theta, V) \quad \text{and} \quad \tilde{Y} = \tilde{a}(\theta, \tilde{U}),$$

where $(V, \tilde{U}) \sim \mathsf{P}_{(V,\tilde{U})}$. When Y and $\tilde{Y}$ are independent, V and $\tilde{U}$ are likewise independent, but in time series problems, for example, the auxiliary variables will be correlated. The use of "$\tilde{a}$" for the mapping instead of simply "a" is to cover the case where Y and $\tilde{Y}$ are related through a common parameter θ, but possibly have different distributions. For example, Y might be an iid normal sample, while $\tilde{Y}$ is the maximum of ten future normal samples; similarly, in a regression context, Y and $\tilde{Y}$ might have different values of the predictor variables.

Solving for θ in the first equation and plugging in to the second gives

$$T(Y) = b(\theta, V) \quad \text{and} \quad \tilde{Y} = \tilde{a}(\theta(Y, V), \tilde{U}),$$

Since prediction is a marginal inference problem, where θ itself is the nuisance parameter, it follows from the general theory in Martin and Liu (2015b) that the first equation in the above display can be ignored. This leaves a marginal association for $\tilde{Y}$:

$$\tilde{Y} = \tilde{a}(\theta(Y, V), \tilde{U}). \tag{2}$$

This marginalization has some similarities to the Bayesian and fiducial predictive distributions. That is, the model for $\tilde{Y}$ in (2) is that of a mixture of the distribution of $\theta(Y, V)$, for fixed Y, with the distribution of $\tilde{a}(\theta, \tilde{U})$ for fixed θ. This, of course, is not the "true" distribution of $\tilde{Y}$ given Y; the idea is that the future observable $\tilde{Y}$ is being modeled as a Y-dependent function of $(V, \tilde{U})$. We claim that equation (2) describes a sort of predictive distribution of $\tilde{Y}$ for a given Y, similar to the frequentist predictive distributions in, e.g., Lawless and Fredette (2005). To see this better, let G_Y be the distribution of the right-hand side of (2) as a function of $(V, \tilde{U})$ for fixed Y. Then, in the case this is an absolutely continuous distribution, (2) can be rewritten as

$$\tilde{Y} = G_Y^{-1}(W), \quad W \sim \mathsf{Unif}(0, 1), \tag{3}$$

so G_Y plays the role of a predictive distribution for $\tilde{Y}$. This completes the A-step in the construction of the IM for prediction. That is, (3) is the association that links the observable data Y, the unobservable auxiliary variable $(U, \tilde{U})$, and the future data $\tilde{Y}$.

Though (2) has some connection to Bayesian and fiducial prediction, it differs from a plug-in or parametric bootstrap prediction. The difference is that the quantity $\theta(Y, V)$ plugged in is not fixed. That is, we consider the distribution of $\tilde{a}(\theta(Y, V), \tilde{U})$ as a function of $(V, \tilde{U})$, not the distribution of $\tilde{a}(\hat{\theta}_Y, \tilde{U})$, as a function of $\tilde{U}$, for fixed $\hat{\theta}_Y$.

2.4   P- and C-steps

After the A-step in (3), the P-step requires specification of a suitable predictive random set $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$ for W. A rigorous presentation on the theory of random sets is given in Molchanov (2005), including a general definition. For our purposes here, it suffices to define a random set by first specifying a probability space $(\mathbb{W}, \mathcal{B}_{\mathbb{W}}, \mathsf{P}_W)$ and a map $S : \mathbb{W} \to \mathcal{B}_{\mathbb{W}}$ which is measurable in the sense that $\{w : S(w) \cap K \neq \varnothing\} \in \mathcal{B}_{\mathbb{W}}$ for all compact $K \subseteq \mathbb{W}$. Then $\mathcal{S} = S(W)$, for $W \sim \mathsf{P}_W$ is a random set, and its distribution $\mathsf{P}_{\mathcal{S}}$ is the push-forward measure $\mathsf{P}_W S^{-1}$. Martin and Liu (2013) argue that the choice of predictive random set ought to depend on the assertion A of interest. There are three kinds of assertions about $\tilde{Y}$ that will be of interest here in the prediction problem: two one-sided assertions, and a singleton assertion. Given a predictive random set and an assertion of interest, the C-step proceeds by combining the A- and P-step results. Martin and Liu (2013) give a general explanation, but here this amounts to computing the *plausibility of A*, i.e.,

$$\mathsf{pl}_Y(A) = \mathsf{P}_{\mathcal{S}}\{G_Y^{-1}(\mathcal{S}) \cap A \neq \varnothing\}.$$

Next we discuss, in turn, the P- and C-steps for each of these kinds of assertions.

- *Right-sided.* A right-sided assertion is of the form $A = \{\tilde{Y} > \tilde{y}\}$ for a fixed $\tilde{y}$. For this assertion, by Theorem 4 in Martin and Liu (2013), the optimal predictive random set is one-sided: $\mathcal{S} = [0, W]$ for $W \sim \mathsf{Unif}(0, 1)$. In this case, the C-step gives the plausibility function

$$\mathsf{pl}_Y(A) = \mathsf{P}_{\mathcal{S}}\{G_Y^{-1}(\mathcal{S}) \cap A \neq \varnothing\} = 1 - G_Y(\tilde{y}) \tag{4}$$

  The plausibility function is a non-increasing function of $\tilde{y}$; see Figure 1(a) described in Section 3.2. Hence the prediction region (8) based on the plausibility function in (4) will be an upper prediction bound for $\tilde{Y}$.

- *Left-sided.* A left-sided assertion is of the form $A = \{\tilde{Y} \leq \tilde{y}\}$ for a fixed $\tilde{y}$. Similar to the right-sided case, the optimal predictive random set is $\mathcal{S} = [W, 1]$ for $W \sim \mathsf{Unif}(0, 1)$. Then the C-step gives the plausibility function

$$\mathsf{pl}_Y(A) = \mathsf{P}_{\mathcal{S}}\{G_Y^{-1}(\mathcal{S}) \cap A \neq \varnothing\} = G_Y(\tilde{y}) \tag{5}$$

  The plausibility function is an non-decreasing function of $\tilde{y}$; see Figure 2(a). Hence the prediction region (8) based on the plausibility function in (5) will be a lower prediction bound for $\tilde{Y}$.

- *Singleton.* A singleton assertion is of the form $A = \{\tilde{Y} = \tilde{y}\}$ for a fixed $\tilde{y}$. The optimal predictive random set worked out in Martin and Liu (2013) for this assertion is complicated, but a natural choice that is suitable in most cases (and optimal in some cases) is the "default" predictive random set $\mathcal{S} = \{w : |w - 0.5| \leq |W - 0.5|\}$, for $W \sim \mathsf{Unif}(0, 1)$. Then the C-step gives the plausibility function

$$\mathsf{pl}_Y(A) = \mathsf{P}_{\mathcal{S}}\{G_Y^{-1}(\mathcal{S}) \cap A \neq \varnothing\} = 1 - |2G_Y(\tilde{y}) - 1| \tag{6}$$

  The prediction region (8) based on the plausibility function in (6) will be a two-sided prediction bound for $\tilde{Y}$.

It is important to note that, although the general P- and C-steps may appear rather technical, implementation of the IM approach for prediction requires only that one be able to evaluate, either analytically or numerically, the distribution function G_Y. Section 3 gives several examples and applications to demonstrate that our IM-based plausibility intervals are good general tools for the prediction problem, and that such intervals are often better than what other methods provide.

2.5 Prediction validity

Here we give the main distributional property of the plausibility function for prediction. The key requirement is a mild condition on the predictive random set $\mathcal{S}$. Following Martin (2014), define the contour function $f_\mathcal{S}(w) = \mathsf{P}_\mathcal{S}(\mathcal{S} \ni w)$. Then the predictive random set $\mathcal{S}$ is *valid* if

$$f_\mathcal{S}(W) \geq_{\mathrm{st}} \mathsf{Unif}(0,1) \quad \text{when } W \sim \mathsf{P}_W, \tag{7}$$

where $\geq_{\mathrm{st}}$ means "stochastically no smaller than." Martin and Liu (2013) demonstrate that this is a very mild condition. (Though not required for the theorem, they also recommend to consider only predictive random sets with nested support. Those discussed in the previous section are all nested.) The three assertions A described in Section 2.4 depend on a generic $\tilde{y}$. Here we write $\mathsf{pl}_Y(\tilde{y})$ for the plausibility function for such an assertion; the specific kind of assertion will be clear from the context.

Theorem 1. *For the marginal association* (3) *for* $\tilde{Y}$, *let* $\mathcal{S} \sim \mathsf{P}_\mathcal{S}$ *be a valid predictive random set for* $W \sim \mathsf{Unif}(0,1)$, *i.e.,* (7) *holds, which is non-empty with* $\mathsf{P}_\mathcal{S}$-*probability* 1. *If* $G_Y(\tilde{Y}) \sim \mathsf{Unif}(0,1)$ *for* $(Y, \tilde{Y}) \sim \mathsf{P}_{(Y,\tilde{Y})|\theta}$ *for all* θ, *then*

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

This holds whether $\mathsf{pl}_Y(\tilde{Y})$ *is based on right-sided, left-sided, or singleton assertions.*

Proof. Since $\mathsf{pl}_y(\tilde{y}) = f_\mathcal{S}(G_y(\tilde{y}))$, the result follows from the assumed validity of $\mathcal{S}$ and the assumption that $G_Y(\tilde{Y}) \sim \mathsf{Unif}(0,1)$ as a function of $(Y, \tilde{Y})$. $\qquad\square$

The following sequence of remarks discusses the assumptions, interpretations, and various extensions of Theorem 1. See, also, Section 4.

Remark 1. Martin and Liu (2013) argue that validity gives the plausibility function a scale on which the numerical values can be interpreted. For example, like in the familiar case of p-values, if the plausibility function is small, e.g., $\mathsf{pl}_y(\tilde{y}) < 0.05$, then, for the given $Y = y$, the value $\tilde{y}$ is not a plausible prediction; see, also, Remark 2.

Remark 2. A consequence of Theorem 1 is that the set

$$\{\tilde{y} : \mathsf{pl}_y(\tilde{y}) > \alpha\} \tag{8}$$

is a $100(1-\alpha)\%$ prediction plausibility region, i.e., the probability that $\tilde{Y}$ falls inside the region (8) is at least $1 - \alpha$ under the joint distribution of $(Y, \tilde{Y})$ for any parameter value θ. Then, for the three kinds of assertions, namely, right, left, and singleton, discussed in Section 2.4, one gets $100(1-\alpha)\%$ upper, lower, and two-sided prediction intervals,

respectively. Moreover, the region (8) has the following desirable interpretation: each point $\tilde{y}$ it contains is individually sufficiently plausible. No frequentist, Bayes, or fiducial prediction interval assigns such a meaning to the individual elements it contains.

Remark 3. Suppose that $\mathcal{S}$ is such that $f_{\mathcal{S}}(V) \sim \mathsf{Unif}(0,1)$ for $V \sim \mathsf{Unif}(0,1)$. Then $\mathsf{pl}_Y(\tilde{Y}) \sim \mathsf{Unif}(0,1)$ as a function of $(Y, \tilde{Y}) \sim \mathsf{P}_{(Y,\tilde{Y})|\theta}$ for all θ. These conditions hold in many examples (see Section 3) and they imply that the plausibility region in (8) has exact prediction coverage, $1 - \alpha$, not just conservative.

Remark 4. A natural question is: under what conditions does $G_Y(\tilde{Y}) \sim \mathsf{Unif}(0,1)$ hold? An important example is the case we shall call "separable," where the effect of Y on the right-hand side of (2) can be separated from the auxiliary variables, i.e., (2) can be rewritten as $p(Y, \tilde{Y}) = \varphi(V, \tilde{U})$ for some functions p and φ. In the language of Lawless and Fredette (2005), the quantity $p(Y, \tilde{Y})$ is an exact pivot. Many problems with a group transformation structure (e.g. Eaton 1989) are separable, and are covered by Theorem 1. Some of the examples in Section 3 are of this type, but the numerical results even for the non-separable models (see Sections 3.3–3.4) suggest that the validity result holds broadly. Section 4.1 has more discussion on the non-separable case.

Remark 5. An advantage of the IM's handling of the auxiliary variables, revealed in the previous remarks, is that one has finite-sample control on the prediction coverage. The fiducial approach to prediction, on the other hand, can only guarantee asymptotic control of frequentist prediction coverage (Wang et al. 2012, Theorem 1).

Remark 6. The uniformity condition in Theorem 1 can be relaxed to a stochastic ordering condition, but then the conclusion holds only for certain predictive random sets and certain assertions. For example, suppose $G_Y(\tilde{Y})$ is stochastically no smaller than $\mathsf{Unif}(0,1)$. Then the conclusion of Theorem 1 holds for the one-sided predictive random set $\mathcal{S} = [0, W]$, $W \sim \mathsf{Unif}(0,1)$. In this case, by taking assertions $A = \{\tilde{Y} > \tilde{y}\}$, the lower plausibility bounds obtained via (8) have the nominal frequentist coverage probability as described in Remark 2. Similar conclusions hold if $G_Y(\tilde{Y})$ is stochastically no larger than $\mathsf{Unif}(0,1)$, with obvious changes to the predictive random set and assertion.

3 Examples and applications

3.1 Normal models and a quality control application

Let $Y = (Y_1, \ldots, Y_n)$ be an iid sample from a $\mathsf{N}(\mu, \sigma^2)$ population, where $\theta = (\mu, \sigma)$ is unknown. Our first goal is to predict the next independent observation $\tilde{Y} = Y_{n+1}$. To start, consider the baseline association involving the original data

$$Y_i = \mu + \sigma Z_i, \quad i = 1, \ldots, n,$$

where $Z_1, \ldots, Z_n$ are iid $\mathsf{N}(0,1)$. Based on the arguments in Martin and Liu (2015a), a conditional IM for $\theta = (\mu, \sigma)$ has association

$$\bar{Y} = \mu + \sigma n^{-1/2} U_1 \quad \text{and} \quad S = \sigma U_2,$$

where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ is the sample mean, $S^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the sample variance, $U_1 \sim \mathsf{N}(0,1)$, and $(n-1)U_2^2 \sim \mathsf{ChiSq}(n-1)$, with U_1 and U_2 independent. Then

it is easy to see that

$$\theta(Y, U) = \big(\mu(Y, U), \sigma(Y, U)\big) = \Big(\bar{Y} - \frac{S}{n^{1/2}} \frac{U_1}{U_2}, \frac{S}{U_2}\Big).$$

For the next observation $\tilde{Y} = Y_{n+1}$, the association is just like the baseline association above, i.e., $\tilde{Y} = \mu + \sigma \tilde{U}$, where $\tilde{U}$ is independent of (U_1, U_2). As discussed above, we can insert $\theta(Y, U)$ in place of θ in this association to get a marginal association for $\tilde{Y}$:

$$\tilde{Y} = \bar{Y} - \frac{S}{n^{1/2}} \frac{U_1}{U_2} + \frac{S}{U_2} \tilde{U} = \bar{Y} + S\Big(\frac{1}{n^{1/2}} \frac{U_1}{U_2} - \frac{\tilde{U}}{U_2}\Big). \tag{9}$$

This is clearly one of those separable cases as described in Remark 4. Also,

$$V = \frac{1}{n^{1/2}} \frac{U_1}{U_2} - \frac{\tilde{U}}{U_2}$$

is distributed as $(n^{-1} + 1)^{1/2}\mathsf{t}(n - 1)$, with distribution function F_n. Then the marginal association (9) can be written as $\tilde{Y} = \bar{Y} + SF_n^{-1}(W)$, with $W \sim \mathsf{Unif}(0, 1)$. If we are interested in a two-sided prediction interval, then, as in Section 2.4, we take a singleton assertion $A = \{\tilde{Y} = \tilde{y}\}$ and get the following plausibility function:

$$\mathsf{pl}_Y(\tilde{y}) = 1 - \Big|2F_n\Big(\frac{\tilde{y} - \bar{Y}}{S}\Big) - 1\Big|.$$

Then the corresponding two-sided $100(1 - \alpha)\%$ plausibility interval (8) for $\tilde{Y}$ is

$$\bar{Y} \pm t^{\star}_{n-1, 1-\alpha/2} S(1 + n^{-1})^{1/2},$$

where $t^{\star}_{\nu, p}$ is the $100p$th percentile of the t-distribution with ν degrees of freedom. This is exactly the classical Student-t prediction interval discussed in, e.g., Geisser (1993).

The ideas just discussed extend quite naturally to the case of normal linear regression. The details of the IM calculations would be similar to those presented in Wang et al. (2012) for the fiducial case and, hence, omitted here.

As a more sophisticated example, Odeh (1990) gives a quality control application involving sprinkler systems for fire prevention in a hotel. In this application, based on a sample of $n = 20$ sprinklers, whose activation temperatures are normally distributed, the goal is to give a two-sided prediction interval for the temperature at which at least $k = 36$ of $m = 40$ new sprinklers will activate. In other words, the goal is to predict the temperature at which at least k of the m new sprinklers will activate. The IM methodology can be used for this problem. Let $\tilde{Y}$ be the k-th largest of m future independent normal observations $Y_{n+1}, \ldots, Y_{n+m}$. The corresponding association for $\tilde{Y}$ is

$$\tilde{Y} = \mu + \sigma \tilde{U}, \quad \text{where} \quad \tilde{U} = k\text{-th largest of } U_{n+1}, \ldots, U_{n+m},$$

and $U_{n+1}, \ldots, U_{n+m}$ are iid $\mathsf{N}(0, 1)$. Then the marginal association for $\tilde{Y}$ can be written exactly as in (9) and the problem is still separable. The only difference here is that $V = (n^{-1/2}U_1 - \tilde{U})/U_2$ has a non-standard distribution. As before, write $\tilde{Y} = \bar{Y} + SF_{n,m,k}^{-1}(W)$, where $F_{n,m,k}$ is the distribution function of V, and $W \sim \mathsf{Unif}(0, 1)$. The distribution $F_{n,m,k}$ can be simulated and, therefore, one can easily get a Monte Carlo approximation of the plausibility function (6) for $\tilde{Y}$ and, in turn, a two-sided prediction interval. The IM prediction interval for $\tilde{Y}$ in this normal prediction problem is the same as the fiducial interval in Wang et al. (2012) and the interval in Fertig and Mann (1977).

3.2 Log-normal models and an environmental application

Let $Y = (Y_1, \ldots, Y_n)$ be an iid sample from a log-normal population, with unkown parameter $\theta = (\mu, \sigma)$. Log-normal models are frequently used in environmental statistics (Ott 1995). In this case, $X = (X_1, \ldots, X_n)$, with $X_i = \log(Y_i)$, will be an iid $\mathsf{N}(\mu, \sigma^2)$ sample, and the prediction problem can proceed as in Section 3.1 above. In particular, predicting the next observation $\tilde{Y} = Y_{n+1}$ is straightforward, so we focus here on something more challenging. Consider, as in Bhaumik and Gibbons (2004), the problem of finding the upper prediction limit for the arithmetic mean of m future log-normal observations, i.e., $\tilde{Y} = m^{-1} \sum_{j=1}^{m} Y_{n+j}$. Working on the log-scale, with the X_i's, we can first reduce dimension according to sufficiency and then solve for θ as follows:

$$\theta(Y, U) = \big(\mu(Y, U), \sigma(Y, U)\big) = \Big(\bar{X} - \frac{S}{n^{1/2}} \frac{U_1}{U_2}, \frac{S}{U_2}\Big),$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, $S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, $U_1 \sim \mathsf{N}(0, 1)$, and $(n-1)U_2^2 \sim \mathsf{ChiSq}(n-1)$, with U_1, U_2 being independent. The marginal association for $\tilde{Y}$, the arithmetic mean of m future log-normal observations, is

$$\tilde{Y} = \frac{1}{m} \sum_{j=1}^{m} e^{\log Y_{n+j}} = \frac{1}{m} \sum_{j=1}^{m} \exp\Big\{\Big(\bar{X} - \frac{S}{n^{1/2}} \frac{U_1}{U_2}\Big) + \Big(\frac{S}{U_2} \tilde{U}_{n+j}\Big)\Big\},$$

where $\tilde{U} = (\tilde{U}_{n+1}, \ldots, \tilde{U}_{n+m})$ are iid $\mathsf{N}(0, 1)$, independent of U_1 and U_2. Here we use this association for $\tilde{Y}$ to construct an upper plausibility prediction limit.

The above association is not of the separable form in Remark 4. However, for a given Y, if G_Y is the distribution of the right-hand side in the previous display, then the marginal association for $\tilde{Y}$ can be written in the form $\tilde{Y} = G_Y^{-1}(W)$, for $W \sim \mathsf{Unif}(0, 1)$, just like in (3). This completes the A-step. Since we seek to determine an upper prediction limit, the plausibility function for $\tilde{Y}$ is given by (4); see Section 2.4.

For illustration, we consider an environmental study presented in Bhaumik and Gibbons (2004) concerning lead concentration in soil. It is a "brownfield" investigation in which a now-closed plating facility was being investigated for future industrial use. In April 1996, $m = 5$ soil borings were installed to delineate the extent of lead-impacted soil at the portion of the facility that may have been used for plating. An important environmental question, which Bhaumik and Gibbons (2004) addressed using frequentist prediction methods, is to determine whether the on-site mean lead concentration at this area of the facility exceeded background. To facilitate this determination, $n = 15$ off-site soil samples were collected in areas that were uninfluenced by the activities at the facility. The data are reproduced in Table 1. Using the Shapiro-Wilk normality test, Bhaumik and Gibbons (2004) ascertained that, at the 5% significance level, a log-normal model provides adequate fit to this data. Our main goal in this application is therefore to demonstrate that the on-site concentrations, on average, do not significantly exceed the backgrund. To this end, we will use the IM framework discussed above to produce an upper prediction limit for the arithmetic mean of lead contents, $\tilde{Y}$, of $m = 5$ on-site soil samples based on the $n = 15$ off-site soil samples, Y, and then we will compare it to the arithmetic mean of the data collected on the on-site lead concentration.

The plausibility for $\tilde{Y}$, for right-sided assertions $A = \{\tilde{Y} > \tilde{y}\}$, as a function of $\tilde{y}$, is shown in Figure 1(a). Those $\tilde{y}$ values with plausibility function exceeding 0.05 provide

Off-site	26	63	3	70	16	5	1	57	5	3	24	2	1	48	3
On-site	50	82	95	103	88										

Table 1: Lead (mg/kg) for soil boring samples in off-site and on-site locations.

an upper prediction bound for $\tilde{Y}$ which, in this case, is 136.16 mg/kg. For comparison, Bhaumik and Gibbons (2004) provide the bound 152.26 mg/kg based on their Gram–Charlier approximation, and Kim (2007) provides the bound 139.30 mg/kg based on a Bayesian approach. All three prediction bounds contain the realized arithmetic mean of the on-site data in Table 1, which was 83.6 mg/kg. We therefore conclude that the on-site concentrations do not significantly exceed the backgrund. However, since smaller upper prediction limits are more precise, our IM-based bound is preferred. An additional advantage of our IM-bound is that it, per Remark 2, also has a clearer interpretation than the above Bayesian and frequentist bounds.

To check the prediction performance for settings similar to the soil example, we take 5000 samples of size $n = 15$ from a log-normal distribution with $\mu = 2.173$ and $\sigma^2 = 2.3808$, the maximum likelihood estimates based on the off-site data in Table 1. A Monte Carlo estimate of the distribution function of $G_Y(\tilde{Y})$ is shown in Figure 1(b). Apparently, $G_Y(\tilde{Y})$ is $\mathsf{Unif}(0, 1)$, so the plausibility function for prediction is valid, by Theorem 1.

For further comparison, we performed a simulation study similar to the one presented in Bhaumik and Gibbons (2004). We considered three values for μ (2, 3, 10), six values for σ^2 (0.0625, 0.2, 0.5, 1, 2, 10), five values for n (5, 10, 20, 30, 100), and three values for m (1, 5, 10). For each combination, we evaluated the coverage probability of both the lower and upper 90% prediction intervals. In all cases, the coverage probability equals the nominal level, up to Monte Carlo error; these estimates are based on 10,000 Monte Carlo samples. Unlike the Gram–Charlier approximation method in Bhaumik and Gibbons (2004), our IM-based interval method does not need technical tools for derivation, and achieves the nominal coverage probability even when $\sigma^2 > 3$. Moreover, the other two frequentist approximation methods reported in Bhaumik and Gibbons (2004) do not achieve the nominal coverage probability.

3.3 Gamma models and a system breakdown application

Let $Y = (Y_1, \ldots, Y_n)$ be an iid sample from a gamma distribution with shape parameter $\theta_1 > 0$ and scale parameter $\theta_2 > 0$, both unknown. Gamma models are often used in system reliability applications. Following Martin and Liu (2015a, Sec. 5.3), a conditional association for (θ_1, θ_2) based on sufficient statistics is given by

$$T_1 = \theta_2 \Gamma_{n\theta_1}^{-1}(U_1) \quad \text{and} \quad T_2 = F_{\theta_1}^{-1}(U_2),$$

where $U = (U_1, U_2)$ are iid $\mathsf{Unif}(0, 1)$, $T_1 = \sum_{i=1}^{n} Y_i$, $T_2 = n^{-1} \sum_{i=1}^{n} \log Y_i - \log(n^{-1}T_1)$, Γ_a is the $\mathsf{Gamma}(a, 1)$ distribution function, and F_b is a distribution function without a familiar form. First, suppose the goal is to predict the next (independent) observation $\tilde{Y} = Y_{n+1}$, with the following association:

$$\tilde{Y} = \theta_2 \Gamma_{\theta_1}^{-1}(\tilde{U}), \quad \tilde{U} \sim \mathsf{Unif}(0, 1). \tag{10}$$

(a) Plausibility function of $\tilde{Y}$ (b) Distribution function of $G_Y(\tilde{Y})$
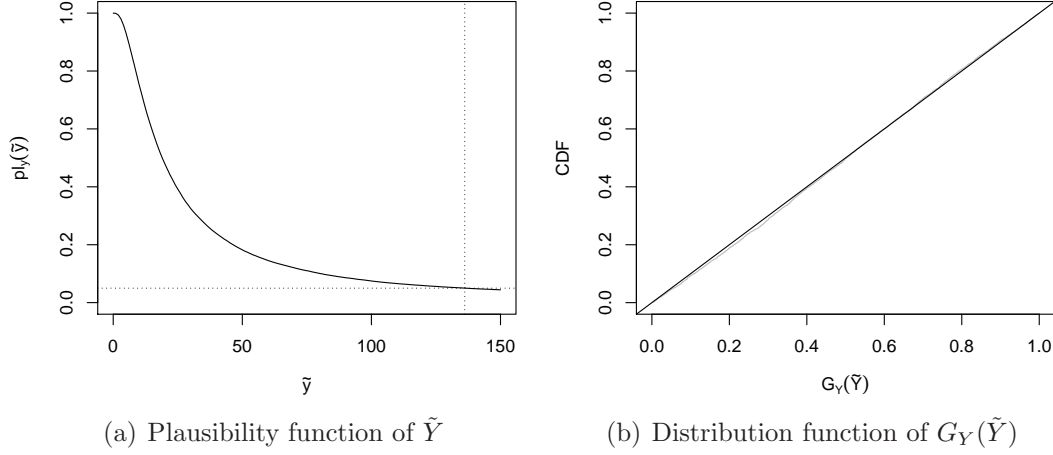
Figure 1: Panel (a): Plausibility function of $\tilde{Y}$ in the log-normal data example. Panel (b): Distribution function of $G_Y(\tilde{Y})$ (gray) compared with that of $\mathsf{Unif}(0,1)$ (black) based on Monte Carlo samples from the log-normal distribution with $\mu = 2.173$ and $\sigma^2 = 2.3808$, the maximum likelihood estimates in the Bhaumik and Gibbons (2004) example.

Specifically, we want to give a lower prediction limit for $\tilde{Y}$. The general strategy is to solve for $\theta = (\theta_1, \theta_2)$ in the conditional association, and then plug this solution in for θ in the association for the new observation. In particular, for a given $U = (U_1, U_2)$, write

$$\theta(Y, U) = \big(\theta_1(Y, U), \theta_2(Y, U)\big) \tag{11}$$

for this solution; it depends on Y only through (T_1, T_2). The solution exists and is unique, though there is no closed-form expression. A proof of this claim, along with some details about computing the solution in (11), are given in the Appendix. Plugging (11) in to the association for $\tilde{Y}$ gives the marginal association

$$\tilde{Y} = \theta_2(Y, U)\Gamma^{-1}_{n\theta_1(Y,U)}(\tilde{U}). \tag{12}$$

This association is not of the separable form in Remark 4. In any case, if G_Y denotes the distribution function of the quantity on the right-hand side of (12), then we can write $\tilde{Y} = G_Y^{-1}(W)$ for $W \sim \mathsf{Unif}(0,1)$. This completes the A-step for IM prediction. Since we are interested in lower prediction limits, we compute the plausibility function in (5).

In some system reliability applications, like in Hamada et al. (2004) and Wang et al. (2012), interest may be in the largest among a collection of m future observations. In that case, we have an association that looks exactly like (12), except that $\tilde{Y} = \max\{Y_{n+1}, \ldots, Y_{n+m}\}$ is a maximum of m future gamma observations and $\tilde{U}$ is the maximum of m independent uniforms, independent of U. This involves the same solution $\theta(Y, U)$ as before, so nothing changes except the distributions being used in the Monte Carlo simulation of the plausibility function.

For illustration, consider the data $Y = (Y_1, \ldots, Y_n)$ on the first breakdown times of $n = 20$ machines given in Hamada et al. (2004). These data are reproduced in Table 2. At the 5% significance level, the Kolmogorov–Smirnov test cannot reject the null hypothesis that these data are gamma, so the goal is to use the IM machinery described above to

12

18	23	29	409	24	74	13	62	46	4
57	19	47	13	19	208	119	209	10	188

Table 2: Machine first breakdown times, in hours.

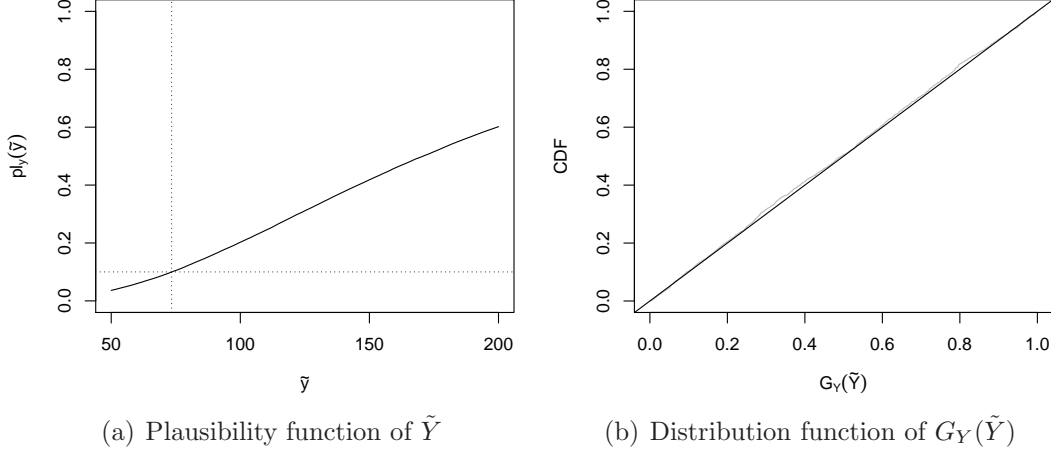(a) Plausibility function of $\tilde{Y}$ (b) Distribution function of $G_Y(\tilde{Y})$

Figure 2: Panel (a): Plausibility function of $\tilde{Y}$ in the gamma data example. Panel (b): Distribution function of $G_Y(\tilde{Y})$ (gray) compared with that of $\mathsf{Unif}(0,1)$ (black) based on Monte Carlo samples from $\mathsf{Gamma}(\hat{\theta}_1, \hat{\theta}_2)$, where $\hat{\theta}_1 = 0.8763$ and $\hat{\theta}_2 = 90.91$ are the maximum likelihood estimates.

produce a lower prediction limit for $\tilde{Y}$, the maximum of $m = 5$ future breakdown times. Proceeding as described above, the plausibility function of $\tilde{Y}$ is given by $\mathsf{pl}_y(\tilde{y}) = G_y(\tilde{y})$, which can be easily evaluated via Monte Carlo. A plot of this plausibility function for $A = \{\tilde{Y} \leq \tilde{y}\}$, as a function of $\tilde{y}$, is given in Figure 2(a). A one-sided 90% plausibility interval is the set of all $\tilde{y}$ values such that $\mathsf{pl}_y(\tilde{y}) > 0.10$, and the lower bound in this case is 73.53 hours. For comparison, our lower bound is bigger, i.e., more precise, than the Bayesian lower bound (71.8 hours) in Hamada et al. (2004) and slightly smaller than the fiducial lower bound (74.36 hours) in Wang et al. (2012). The IM bound, per Remark 2, also has a clearer interpretation than the Bayesian and fiducial bounds. To assess the performance of the method in problems similar to this one, we simulate 2000 data sets based on the maximum likelihood estimates based on the failure time data. A Monte Carlo estimate of the distribution function of $G_Y(\tilde{Y})$ is shown in Figure 2(b). This distribution function is sufficiently close to that of $\mathsf{Unif}(0,1)$, so we can conclude our one-sided IM-based 90% prediction interval has exact coverage.

For further illustration, we consider a simulation experiment, similar to that in Wang et al. (2012), with three values of the sample size n (10, 25, 125), four values of the shape parameter θ_1 (0.5, 1, 5, 10), and two values of m (1, 5); we keep the scale parameter θ_2 fixed at 1. For each combination, we evaluated the coverage probability of both the lower and upper 90% prediction intervals based on 10,000 Monte Carlo samples. In all cases, the coverage probability is within an acceptable range of the target 0.90.

13

3.4 Binomial models and a disease count application

Let $Y \sim \mathsf{Bin}(n, \theta)$ and $\tilde{Y} \sim \mathsf{Bin}(m, \theta)$ be independent binomial random variables, where n and m are known. The goal is to predict $\tilde{Y}$ based on observing Y. There is considerable literature on this fundamental problem: see, e.g., Wang (2010) for frequentist prediction intervals and Tuyl et al. (2009) for Bayesian prediction intervals. The starting point for our IM-based analysis is the following joint association for Y and $\tilde{Y}$,

$$F_{n,\theta}(Y-1) \leq 1 - U < F_{n,\theta}(Y) \quad \text{and} \quad F_{m,\theta}(\tilde{Y}-1) \leq 1 - \tilde{U} < F_{m,\theta}(\tilde{Y}),$$

where U and $\tilde{U}$ are independent uniforms, and $F_{n,\theta}$ is the $\mathsf{Bin}(n, \theta)$ distribution function. To marginalize over θ, we need a known identity linking the binomial and beta distribution functions, i.e., $F_{n,\theta}(y) = 1 - G_{y+1,n-y}(\theta)$, where $G_{a,b}$ is the $\mathsf{Beta}(a, b)$ distribution function. Now rewrite the first expression in the joint association as a θ-interval:

$$G_{Y,n-Y+1}^{-1}(U) \leq \theta < G_{Y+1,n-Y}^{-1}(U). \tag{13}$$

Next, rewrite the $\tilde{Y}$ association as

$$F_{m,\theta}^{-1}(1 - \tilde{U}) < \tilde{Y} < F_{m,\theta}^{-1}(1 - \tilde{U}).$$

Klenke and Mattner (2010) show that $F_{m,\theta}^{-1}(v)$ is an increasing function of θ for all v, so we can "plug in" the Y-dependent interval for θ in to this latter inequality, to get

$$F_{m,\theta_1(Y,U)}^{-1}(1 - \tilde{U}) < \tilde{Y} < F_{m,\theta_2(Y,U)}^{-1}(1 - \tilde{U}), \tag{14}$$

where $\theta_1(Y, U)$ and $\theta_2(Y, U)$ are, respectively, the left and right endpoints of the interval in (13). This completes the A-step. We are interested in two-sided prediction intervals here; see the P- and C-steps for singleton assertions in Section 2.4. Note that this association is an interval, compared to the singletons in the previous examples. This is a consequence of the discreteness of the binomial, not a limitation of the IM approach; but see below. Some minor adjustments to the C-step in Section 2.4 is needed to handle this discreteness. Since we can easily get a Monte Carlo approximation for the distribution of the two endpoints, constructing a plausibility function for $\tilde{Y}$ is no problem.

In medical applications, it may be desirable to obtain accurate prediction of the number of future cases of a disease based on the counts in previous years. Wang (2010, Sec. 5) gives the following example. The total number of newborn babies with permanent hearing loss is $Y = 23$ out of $n = 23061$ normal nursery births over a two-year period. The goal is to predict $\tilde{Y}$, the number of newborns with hearing loss in the following year, based on $m = 12694$ normal births. For a two-sided, IM-based 90% prediction interval for $\tilde{Y}$, we compute the 5th and 95th percentiles of the distribution of the lower and upper endpoints, respectively, in (14). The interval obtained is $(6, 21)$, which contains the true $\tilde{Y} = 20$ and is essentially the same as the intervals in Wang (2010); see Remark 2.

The plausibility function obtained based on the above construction is a bit conservative. One possible adjustment, based on an idea presented by Wang et al. (2012) in the fiducial context, is to eliminate the interval association for $\tilde{Y}$ by first eliminating the interval association (13) for θ in terms of the limits $\theta_1(Y, U)$ and $\theta_2(Y, U)$. The idea is

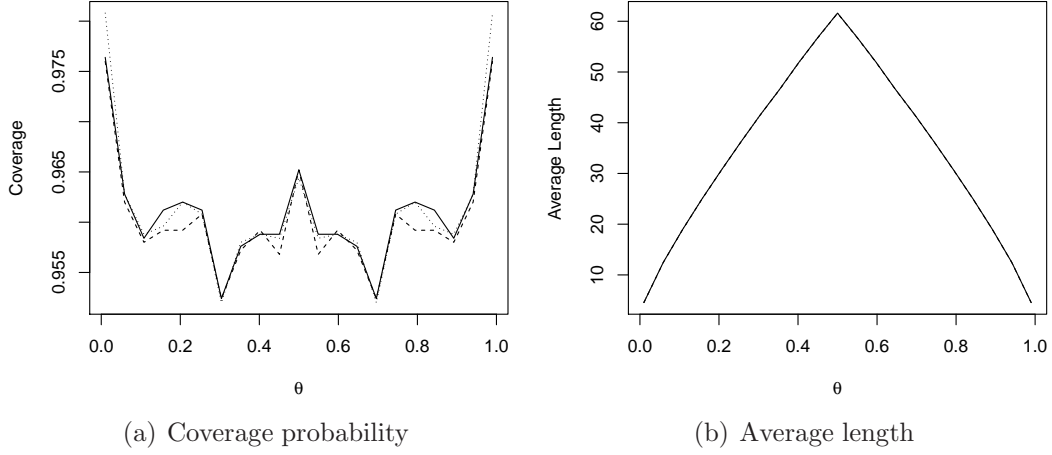(a) Coverage probability (b) Average length

Figure 3: Coverage probability and average length of the modified IM (solid), fiducial (dashed), and Jeffreys prior Bayes (dotted) upper 95% prediction intervals, as functions of θ; the three length curves in Panel (b) are indistinguishable. Here $n = m = 100$ and estimates are based on 2500 simulated data sets.

to sample a value, $\hat{\theta}(Y, U)$, of θ at random from the interval $(\theta_1(Y, U), \theta_2(Y, U))$. This results in a modified association for $\tilde{Y}$:

$$\tilde{Y} = F^{-1}_{m, \hat{\theta}(Y, U)}(1 - \tilde{U}). \tag{15}$$

The intuition is that the uncertainty due to the interval association has been replaced by the uncertainty from sampling. Since the sampled point is "less extreme" than both of the endpoints, this modified association gives a more efficient plausibility function for prediction, which we now demonstrate. Consider binomial samples of size $n = m = 100$ over a range of θ values. Here we compare the coverage probability and average lengths of 95% upper prediction limits based on the modified IM, fiducial, and Jeffreys prior Bayes methods. We simulated 2500 data sets, and each computation of the prediction interval (modified IM, fiducial, and Bayes) used 10,000 Monte Carlo samples. In Figure 3, we see that all three methods have coverage slightly above the nominal level over the entire range of θ; this is to be expected, given the discreteness of the binomial model. The modified IM intervals based on (15) tend to have slightly higher coverage probability than the others, but with no perceptible difference in length.

4   Some further technical details

4.1   Asymptotic validity

Outside the separable class in Remark 4, or in cases where $\tilde{Y}$ is a non-linear function of several future observables, the theory of prediction validity is more challenging. However, our examples in Section 3 demonstrate that the uniformity assumption of Theorem 1 holds at least approximately. Here we give a theoretical argument to explain this phenomenon.

Write $Y^n = (Y_1, \ldots, Y_n)$ for data consisting of n iid components. Suppose that the solution $\theta(Y^n, V)$ converges in probability to θ, as a function of (Y^n, V); this usually

15

is easy to arrange, see the examples in Section 3. If $\tilde{Y}$ has a continuous distribution, then, without loss of generality, we can write $G_{Y^n}(\tilde{Y}) = \tilde{F}_{\theta(Y^n,V)}(\tilde{Y})$, where $\tilde{F}_\theta$ is the true distribution of $\tilde{Y}$. Trivially, we have $(\theta(Y^n,V),\tilde{Y}) \to (\theta,\tilde{Y})$ in distribution so, if $(\theta,\tilde{y}) \mapsto \tilde{F}_\theta(\tilde{y})$ is continuous, then the continuous mapping theorem implies that

$$G_{Y^n}(\tilde{Y}) = \tilde{F}_{\theta(Y^n,V)}(\tilde{Y}) \to \tilde{F}_\theta(\tilde{Y}) \sim \mathsf{Unif}(0,1) \quad \text{in distribution.}$$

Therefore, we can generally be sure that the distribution of $G_{Y^n}(\tilde{Y})$ will be approximately $\mathsf{Unif}(0,1)$ when the sample size n is large. This argument holds even if $\tilde{Y}$ is some scalar function of several future observations.

In addition to providing an asymptotic validity result, the argument above is also relevant to prediction accuracy. That is, we have demonstrated that the IM "predictive distribution" for $\tilde{Y}$, which mixes $\tilde{F}_{\theta(Y^n,V)}$ over the distribution of V, converges to $\tilde{F}_\theta$, the true distribution of $\tilde{Y}$. A precise prediction accuracy result requires computing a measure of the distance/divergence of the IM predictive from the truth. We expect that results comparable to those in Lawless and Fredette (2005) can be derived, but we leave this as a question to be considered in future work.

4.2 Case of dependent $\tau(U)$ and $\eta(U)$

Recall that, in Section 2.3, it was assumed that the original association, $Y = a(\theta, U)$, could be decomposed as $T(Y) = b(\theta, \tau(U))$ and $H(Y) = \eta(U)$, and, furthermore, that $\tau(U)$ and $\eta(U)$ are independent. Of concern here is the case where $\tau(U)$ and $\eta(U)$ are not independent, which would arise, say, in models that are not regular exponential families. The IM framework is equipped to handle this, but the details are more complicated. Here, we describe the three-step construction of an IM for prediction in the case where the model is a location shift of a Student-t distribution.

Let $Y = (Y_1, \ldots, Y_n)^\top$ be iid, with $Y_i = \theta + U_i$ and $U_i \sim \mathsf{t}_d$, i.e., θ is a location parameter and the error has a Student-t distribution with known degrees of freedom d. The case with an additional unknown scale parameter can be handled similarly. Let $T(Y)$ be the maximum likelihood estimator for θ, and let $H(Y) = Y - T(Y)1_n$ be the vector of residuals, where 1_n is a (column) n-vector of unity. Since T is equivariant in this example, we have

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

so that $\tau = T$ and $\eta = H$. However, $\tau(U)$ and $\eta(U)$ are not independent here. As suggested in Martin and Liu (2015a) and in Section 2.3 above, we want to consider the conditional distribution of $\tau(U)$, given that $\eta(U)$ equals the observed value of $H(Y)$. Let H_0 denote the observed value of $H(Y)$. Then Martin and Liu (2015a) give a formula for the conditional distribution of $\tau(U)$, given $\eta(U) = H_0$. Write $\mathsf{P}_{V|H_0}$ for this conditional distribution of $V = \tau(U)$, so that

$$T(Y) = \theta + V, \quad V \sim \mathsf{P}_{V|H_0}.$$

This can be solved for θ, and a marginal association connecting the observed Y, the next $\tilde{Y}$ to be predicted, and the pair $(V, \tilde{U})$ of auxiliary variables is of the form

$$\tilde{Y} = T(Y) + \tilde{U} - V, \quad \text{where} \quad V \sim \mathsf{P}_{V|H_0}, \; \tilde{U} \sim \mathsf{t}_d.$$

Computation with the conditional distributions is more cumbersome, but our claim that the dependent case is conceptually no different than the independent case should now be clear. Moreover, using the law of iterated expectation, it can be shown that the conditioning does not affect validity result in Theorem 1.

4.3 Multivariate prediction

The focus of this paper was on the case of predicting a scalar $\tilde{Y}$, which is possibly a scalar-valued function of several future observables. However, in some cases there could be interest in simultaneous prediction of several future observables. One example is in regression, where interest may be in predicting the response value corresponding to several values of the predictor variables simultaneously. While the general IM framework is well-equipped to handle the multivariate case, our developments here have employed a few scalar-specific steps. Our goal in this section is simply to highlight those scalar-specific steps, defining a roadmap to extend the present developments to the multivariate case.

First, note that, up to the simplified association formula in (3), there is nothing in the developments in Section 2.3 specific to the scalar $\tilde{Y}$ case. There, we introduced a distribution G_Y and a probability integral transform, which is only valid in the scalar case. However, (2) is well-defined for vector $\tilde{Y}$, and would conclude the A-step for multivariate prediction. From here, the P-step proceeds by introducing a predictive random set for the pair $(V, \tilde{U})$. Validity, as usual, would not be a major obstacle, but an efficient choice of predictive random set would be problem specific. For a scalar auxiliary variable, the only reasonable choice of predictive random set is an interval, but in the multivariate case, there are lots of "reasonable" shapes, and the choice among them makes a difference in terms of the corresponding IM's efficiency. Work on the construction of efficient predictive random sets, in general, is ongoing, and results to be obtained will have immediate application to the multivariate prediction problem.

Second, observe that the discussion of one-sided assertions in Section 2.4 is not appropriate in the multivariate setting, where there is no proper ordering. The two most natural assertions would be the singletons and assertions defined via level sets of some scalar-valued function of $\tilde{Y}$, e.g., balls $\{\tilde{y} : \|\tilde{y}\| \leq r\}$ for some fixed radius $r > 0$. The latter case reduces to the scalar prediction problem covered in this paper. For the singleton assertion case, nothing in the present development needs to change, except that the predictive random set, in general, must be specified for the pair $(V, \tilde{U})$ directly. If this predictive random set is valid, then the conclusion of Theorem 1 holds. Again, the challenge is that the shape of the predictive random set is directly related to the shape and efficiency of, say, the IM prediction regions for $\tilde{Y}$. So, more work on constructing good/optimal predictive random sets for multivariate auxiliary variables is needed.

5 Conclusion

In this paper, we have proposed a method for prediction of future observables based on the recently developed IM framework. The key to the IM approach in general is the association of data and parameters with unobservable auxiliary variables, and the use of random sets on the auxiliary variable space to construct belief and plausibility functions

on the parameter space. In the context of prediction of future observables, all the model parameters are nuisance, and an extreme form of the marginalization technique described in Martin and Liu (2015b) is required, which allows us to reduce the dimension of the auxiliary variables, increasing efficiency. We give conditions which guarantee that the IM for prediction is valid, and we argue that this notion of IM validity translates to frequentist coverage guarantees for our plausibility intervals for future observables. A sequence of practical examples demonstrates the quality performance of the proposed method, along with its generality and overall simplicity.

The methodology described here covers both discrete and dependent-data problems. However, these problems present unique challenges. For example, in the binomial example in Section 3.4, our standard IM approach was valid but conservative. A modified and more efficient IM was proposed, and its validity was confirmed numerically, but a theoretical basis for this modification is required. For dependent data problems, marginalization to reduce the dimension of auxiliary variables as described here is possible, but the details would be more challenging. Moreover, as discussed in Section 4.3, additional work is needed to properly extend the developments in the present paper to the multivariate prediction problem. These are all topics for future research.

To conclude, recall the take-away message from Section 1. The IM approach provides a general and easy-to-implement method for constructing valid prior-free probabilistic summaries of the information in observed data relevant for inference or prediction. The fact that these summaries can be converted to frequentist procedures with fixed-n performance guarantees and comparable efficiency compared to existing methods is an added bonus. We expect further developments and applications of IMs in years to come.

Acknowledgments

A Technical details for Section 3.3

A.1 Existence and uniqueness of the solution (11)

Here the issue is existence and uniqueness of the solution $\theta(T, U) = (\theta_1(T, U), \theta_2(T, U))$ in the gamma problem in Section 3.3. The only non-trivial part is the solution θ_1 of equation $F_{\theta_1}(t_2) = u_2$, involving only (t_2, u_2). The challenge is that F_{θ_1} is a non-standard distribution. Glaser (1976), in his notation, considers the random variable

$$U^{\star} = \left\{ \frac{(\prod_{i=1}^{n} Y_i)^{1/n}}{\frac{1}{n} \sum_{i=1}^{n} Y_i} \right\}^n,$$

the n-th power of the ratio of geometric and arithmetic means of an iid $\mathsf{Gamma}(\theta_1, 1)$ sample. Then $T_2 = n^{-1} \log(U^{\star})$, i.e., our T_2 is a monotone increasing function of Glaser's $U^{\star}$. A consequence of Glaser's Corollary 2.2 is that $U^{\star}$ is stochastically strictly increasing

in θ_1, which implies that $F_{\theta_1}(t_2)$ is a decreasing function of θ_1 for all t_2. Therefore, if a solution exists for θ_1 in (11), it must be unique by monotonicity.

Turning to the existence of a solution for θ_1, we need to show that, for any t_2, $F_{\theta_1}(t_2)$ spans all of the interval $(0, 1)$ for u_2 as θ_1 varies. By monotonicity, it suffices to consider the limits $\theta_1 \to \{0, \infty\}$. Jensen (1986) considers the random variable $W = -1/T_2$ and shows, in his Equation (9), that θ_1/W has a limiting distribution as $\theta_1 \to \{0, \infty\}$, which implies the same for $\theta_1 T_2$. It is now clear that $F_{\theta_1}(t_2)$ converges to 1 and 0 as θ_1 converges to 0 and ∞, respectively, for all t_2. Therefore, a solution for θ_1 in (11) exists for all (t_2, u_2) pairs, as was to be shown.

A.2 Computing the solution (11)

Here we consider computing the solution $\theta(Y, U)$ in (11). The only challenging part is solving for θ_1, so we shall focus on this. Suppose t_2 and u_2 are given, and define a function $r(x) = F_x(t_2) - u_2$; the goal is to find the root for r. One can evaluate $r(x)$ by simulating $\mathsf{Gamma}(x, 1)$ variables, giving a Monte Carlo approximation of $F_x(t_2)$, and the root can then be found with any standard method, e.g., bisection. However, this can be fairly expensive computationally. A more efficient alternative approach is available based on large-sample theory. By Theorem 5.2 in Glaser (1976) and the delta theorem, if n is large, then F_x can be well approximated by a normal distribution function with mean $\psi(x) - \log(x)$ and variance $n^{-1}\{\psi'(x) - 1/x\}$, where ψ and ψ' are the digamma and trigamma functions, respectively. With this normal approximation, it is easy to evaluate $r(x)$ and find the root numerically. Though this is based on a large-sample approximation, in our experience, there is no significant loss of accuracy, even for small n.

The normal approximation discussed above is simply a tool to find the solution $\theta(Y, U)$. It also provides some intuition related to the asymptotic argument in Section 5.1. When n is large, the variance in the normal approximation is $O(n^{-1})$, so the distribution function F_x will have a steep slope in the neighborhood of the solution to the equation $t_2 = \psi(x) - \log(x)$ and, therefore, the root for $r(x)$ will be in that same neighborhood, no matter the value of u_2. The solution to the equation $t_2 = \psi(x) - \log(x)$ is the maximum likelihood estimator of θ_1 (e.g., Fraser et al. 1997), which is consistent. Therefore, when n is large, (12) and (10) are essentially the same, so the approximate validity of the corresponding prediction plausibility function is clear.

One last modification that we found to be helpful was to modify that normal approximation discussed above by replacing the normal distribution function with a gamma. That is, find solutions for the mean and variance of the normal approximation as before, but then use a gamma distribution function with mean and variance matching those obtained for the normal. See the R code available at the first author's website.

References

Beran, R. (1990). Calibrating prediction regions. *J. Amer. Statist. Assoc.*, 85(411):715–723.

Bhaumik, D. K. and Gibbons, R. D. (2004). An upper prediction limit for the arithmetic mean of a lognormal random variable. *Technometrics*, 46(2):239–248.

Coelho, C. A. and Mexia, J. T. (2007). On the distribution of the product and ratio of independent generalized gamma-ratio random variables. *Sankhyā*, 69(2):221–255.

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

Eaton, M. L. (1989). *Group Invariance Applications in Statistics.* Institute of Mathematical Statistics, Hayward, CA.

Fertig, K. W. and Mann, N. R. (1977). One-sided prediction intervals for at least p out of m future observations from a normal population. *Technometrics*, 19:167–167.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference.* 2nd ed., revised. Hafner Publishing Company, New York.

Fraser, D. A. S. (1968). *The Structure of Inference.* John Wiley & Sons Inc., New York.

Fraser, D. A. S., Reid, N., and Wong, A. (1997). Simple and accurate inference for the mean of a gamma model. *Canad. J. Statist.*, 25(1):91–99.

Geisser, S. (1993). *Predictive Inference*, volume 55 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, New York.

Glaser, R. E. (1976). The ratio of the geometric mean to the arithmetic mean for a random sample from a gamma distribution. *J. Amer. Statist. Assoc.*, 71(354):480–487.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals:A Guide for Practitioners.* Wiley, New York.

Hamada, M., Johnson, V., Moore, L. M., and Wendelberger, J. (2004). Bayesian prediction intervals and their relationship to tolerance intervals. *Technometrics*, 46(4):452–459.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, 19(2):491–544.

Hannig, J. (2013). Generalized fiducial inference via discretization. *Statist. Sinica*, 23(2):489–514.

Jensen, J. L. (1986). Inference for the mean of a gamma distribution with unknown shape parameter. *Scand. J. Statist.*, 13(2):135–151.

Kim, H.-J. (2007). A Monte Carlo method for estimating prediction limit for the arithmetic mean of lognormal sample. *Comm. Statist. Theory Methods*, 36(9-12):2159–2167.

Klenke, A. and Mattner, L. (2010). Stochastic ordering of classical discrete distributions. *Adv. Appl. Probab.*, 42(2):392–410.

Krishnamoorthy, K., Mathew, T., and Mukherjee, S. (2008). Normal-based methods for a gamma distribution: prediction and tolerance intervals and stress-strength reliability. *Technometrics*, 50(1):69–78.

Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542.

Liu, C. and Martin, R. (2015). Frameworks for prior-free posterior probabilistic inference. *WIREs Comp. Stat.*, to appear.

Martin, R. (2014). Random sets and exact confidence regions. *Sankhyā A*, 76:288–304.

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

Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, to appear; `arXiv:1211.1530`.

Martin, R. and Liu, C. (2015b). Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, to appear; `arXiv:1306.3092`.

Molchanov, I. (2005). *Theory of Random Sets*. Probability and its Applications (New York). Springer-Verlag London Ltd., London.

Odeh, R. E. (1990). Two-sided prediction intervals to contain at least k out of m future observations from a normal distribution. *Technometrics*, 32(2):203–216.

Ott, W. R. (1995). *Environmental Statistics and Data Analysis*. CRC Press, Boca Raton, FL.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Tuyl, F., Gerlach, R., and Mengersen, K. (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Anal.*, 4(1):151–158.

Wang, C. M., Hannig, J., and Iyer, H. K. (2012). Fiducial prediction intervals. *J. Statist. Plann. Inference*, 142(7):1980–1990.

Wang, H. (2010). Closed form prediction intervals applied for disease counts. *Amer. Statist.*, 64(3):250–256.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution of a parameter – a review. *Int. Statist. Rev.*, 81(1):3–39.