Journal of Statistical Planning and Inference 195 (2018) 126–140

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Prediction with confidence—A general framework for predictive inference

Jieli Shen, Regina Y. Liu, Min-ge Xie *

Rutgers University, United States

ARTICLE INFO

Article history:
Available online 14 October 2017

Keywords:
Confidence distribution
Distributional inference
Frequentist coverage
Prediction
Predictive distribution

ABSTRACT

This paper proposes a general framework for prediction in which a prediction is presented in the form of a distribution function, called *predictive distribution function*. This predictive distribution function is well suited for the notion of *confidence* subscribed in the frequentist interpretation, and it can provide meaningful answers for questions related to prediction. A general approach under this framework is formulated and illustrated by using the so-called confidence distributions (CDs). This CD-based prediction approach inherits many desirable properties of CD, including its capacity for serving as a common platform for connecting and unifying the existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. The theory underlying the CD-based predictive distribution is developed and some related efficiency and optimality issues are addressed. Moreover, a simple yet broadly applicable Monte Carlo algorithm is proposed for the implementation of the proposed approach. This concrete algorithm together with the proposed definition and associated theoretical development produce a comprehensive statistical inference framework for prediction. Finally, the approach is applied to simulation studies, and a real project on predicting the incoming volume of application submissions to a government agency. The latter shows the applicability of the proposed approach to dependence data settings.

1. Introduction

Consider the task of predicting the future value of a univariate random variable Y^*, based on a given sample of size n, $\mathbf{Y}_n \equiv \{Y_1, Y_2, \ldots, Y_n\}$. Assume that the vector of the given sample data is from a distribution $G_\theta(\cdot)$ with parameter θ, denoted by $\mathbf{Y}_n \sim G_\theta$, and that the new data point to be predicted is from a distribution $F_\theta(\cdot)$ with the same parameter θ, denoted by $Y^* \sim F_\theta$. Here, $\theta \in \mathbb{R}^p$ is a $p \times 1$ vector of parameter unless specified otherwise. Since G_θ and F_θ share the same θ, information contained in the observed data $\mathbf{Y}_n$ can be channeled through an estimate of θ to assist the prediction of Y^*. To simplify our presentation, we assume that Y^* and $\mathbf{Y}_n$ are independent, except in Section 6 with an example that Y^* and $\mathbf{Y}_n$ are allowed to be dependent. Throughout the paper, the realization of Y^* and $\mathbf{Y}_n$ is denoted by y^* and $\mathbf{y}_n = \{y_1, \ldots, y_n\}$, respectively. Also, when they exist, the corresponding density functions of F_θ and G_θ are denoted by f_θ and g_θ, respectively.

There is a rich literature on predictive inference. Lawless and Fredette (2005) provided an excellent overview on the topic and categorized statistical methods for prediction into two main approaches—frequentist and Bayesian. (I) In frequentist approach, prediction intervals of the specific form $(L_1(\mathbf{Y}_n), L_2(\mathbf{Y}_n))$ are considered, so that the coverage probability

$$\text{CP} \equiv \mathbb{P}_{\mathbb{J}} \left\{ L_1(\mathbf{Y}_n) \le Y^* \le L_2(\mathbf{Y}_n) \right\} \tag{1}$$

* Corresponding author.
E-mail address: mxie@stat.rutgers.edu (Min-ge Xie).

https://doi.org/10.1016/j.jspi.2017.09.012
0378-3758/

can be specified, exactly or asymptotically. Here, $\mathbb{P}_{\mathbb{J}}$ refers to the joint probability for both random variables Y^* and $\mathbf{Y}_n$. Relevant references include Aitchison and Dunsmore (1980), Cox (1975), Beran (1990), Barndor-Nielsen and Cox (1996) and Escobar and Meeker (1999), among others. (II) In Bayesian inference, Bayesian predictive distributions of the form

$$Q_B(y^*; \mathbf{y}_n) = \int_{\theta\in\Theta} F_\theta(y^*)p(\theta|\mathbf{y}_n)d\theta \tag{2}$$

is used. Here, Θ is the parameter space of θ and $p(\theta|\mathbf{y}_n)$ is the posterior density given $\mathbf{Y}_n = \mathbf{y}_n$. Bayesian prediction intervals $(L_1(\mathbf{y}_n), L_2(\mathbf{y}_n))$ can then be obtained from (2). Relevant references include Aitchison (1975), Aitchison and Dunsmore (1980), Geisser (1993), Smith (1998) and others.

The classical frequentist approaches in (I) have the advantage of having a precise and well defined frequentist probabilistic interpretation, analogous to that of "confidence intervals". But those prediction intervals use only two endpoints of the intervals to describe Y^*, and thus are not as informative or flexible as the entire predictive distributions produced by the Bayesian methods in (II) (as well as the approach to be discussed in this paper). This comparative observation is similar to that in comparing inference outcomes from confidence intervals versus confidence distributions (cf. Cox, 2013; Xie, 2013). Specifically, as stated in Cox (1958, 2013), one often has a sense that "when 95% confidence limits of a normal mean are found then, even if the parameter is outside the calculated range, it will not be too far outside". This sense cannot be captured by the definition of a 95% confidence interval, but can be clearly displayed by a confidence distribution. Similar case can be made for using a full-fledged distribution function to describe the prediction outcome, as to convey fuller the prediction outcome and also be sufficiently flexible to admit all forms of prediction outcomes, e.g., point estimates or prediction intervals of all levels, etc.

The Bayesian approach in (II) does use a distribution function to describe the prediction of Y^*, and enjoy the aforementioned "flexibility". But the Bayesian outcomes depend on the additional assumptions of priors. Lawless and Fredette (2005) pointed out that "objective Bayesian methods do not have clear probability interpretations in finite samples", and "subjective Bayesian predictions have a clear personal probability interpretation but it is not generally clear how this should be applied to non-personal predictions or decisions". In addition, many statistical models are developed under non-Bayesian framework and Bayesian predictive distribution methods are not a natural fit for the developments in such practices.

To overcome the above shortcomings of the Bayesian approach Lawless and Fredette (2005) studied frequentist predictive distribution functions in a special setting equipped with pivotal quantities, and referred to this as *the pivotal method*. They further proved the superiority of the predictive distributions obtained from the pivotal method, as having a smaller average Kullback–Leibler distance to the true distribution $f_\theta(y^*)$, over those from the simple plug-in approach by using $f_{\hat\theta}(y^*)$ to derive prediction intervals for all θ. Here, $\hat\theta \equiv \hat\theta(\mathbf{y}_n)$ is the maximum likelihood estimate or any efficient estimate of θ based on the observed data. A related development is the fiducial predictive distributions studied by Wang et al. (2012), who provided a set of conditions under which the fiducial predictive distributions can be used to construct prediction intervals. The fiducial prediction intervals coincide with the exact pivotal-based intervals when available, and otherwise possess correct frequentist coverage asymptotically.

Following the concept of predictive distribution in Lawless and Fredette (2005), we propose in this paper a rigorous definition of a predictive distribution function and develop a general approach for constructing a predictive distribution of Y^* using a confidence distribution (CD) of the unknown parameter θ. The resulting predictive distribution can account for both the variability from the future random variable Y^* and that from estimating the unknown parameter θ using the sample $\mathbf{Y}_n$. It takes the same form as the Bayesian and fiducial predictive distribution functions and thus also enjoys the flexibility of being predictive distribution functions. More importantly, it is anchored on the idea *to always provide prediction intervals with clear frequentist probability interpretations*. This approach was also considered in Schweder and Hjort (2016) under the name of predictive confidence distribution, which also had a comparison with the Bayesian predictive distribution. In this paper, we establish theoretical properties for the CD-based predictive distribution, including the frequentist coverage probabilities of the prediction intervals, and related efficiency and optimality properties. Moreover, we also establish the connections of this approach to other existing prediction approaches. In particular, we show that, under the formulation of the CD-based predictive approach, the frequentist predictive distribution functions derived from the pivotal method in Lawless and Fredette (2005), the fiducial predictive distributions from Wang et al. (2012), and even the Bayesian predictive distribution all amount to the same equivalent expression. This clearly shows that the CD-based approach can provide a unifying platform linking the existing frequentist, Bayesian and fiducial predictive distribution functions.

The rest of this paper is organized as follows: Section 2 defines predictive distribution functions and formulates a CD-based predictive approach. Section 3 examines the theoretical properties of the CD-based predictive distribution function and shows its connections to the Bayesian and fiducial predictive functions, and the frequentist predictive distribution function studied in Lawless and Fredette (2005). This section also presents several properties concerning the efficiency and optimality. Section 4 contains a simple yet broadly applicable Monte Carlo algorithm for carrying out the CD-based approach. Section 5 demonstrates the effectiveness of the proposed CD-based approach using a simulation study under linear and nonlinear regression models. Section 6 presents a real project on predicting the future volume of application submissions to a government agency, showing that the proposed approach applies even to settings with dependent observations. Section 7 provides further comments and discussions.

2. Predictive distribution function and its general formulation based on confidence distribution

Let $\mathcal{Y}^*$ be the sample space of Y^* and $\mathcal{Y}^n$ the sample space of $\mathbf{Y}_n$. Recall that $\mathbf{Y}_n \equiv \{Y_1, Y_2, \ldots, Y_n\} \sim G_\theta$; $Y^* \sim F_\theta$, and $\theta \in \mathbb{R}^p$ is the unknown parameter with parameter space Θ. Denote by θ_0 the true parameter value of θ. We define a predictive distribution function for Y^* based on the sample data $\mathbf{Y}_n$ as follows:

Definition 1. A function $Q(\cdot;\cdot)$ on $\mathcal{Y}^* \times \mathcal{Y}^n \longrightarrow (0, 1)$ is called A PREDICTIVE DISTRIBUTION FUNCTION FOR A NEW OBSERVATION Y^* if it satisfies the two requirements below:
(R1) For each given $\mathbf{Y}_n = \mathbf{y}_n \in \mathcal{Y}^n$, $Q_{\mathbf{y}_n}(\cdot) = Q(\cdot; \mathbf{y}_n)$ is a cumulative distribution function on $\mathcal{Y}^*$;
(R2) $Q(Y^*; \mathbf{Y}_n)$, as a function of both random sample Y^* and $\mathbf{Y}_n$, satisfies the following equation:

$$\mathbb{P}_{\mathbb{J}}(Q(Y^*; \mathbf{Y}_n) \leqslant \alpha) = \alpha, \quad \text{for any } 0 < \alpha < 1, \tag{3}$$

where $\mathbb{P}_{\mathbb{J}}(\cdot)$ is the joint probability measure w.r.t. Y^* and $\mathbf{Y}_n$. Also, the function $Q(\cdot;\cdot)$ is called an asymptotic (or approximate) predictive distribution, if the statement in (3) holds asymptotically.

Requirement (R1) in Definition 1 implies that, in principle, any sample-dependent distribution function on the space of the future random variable Y^* can be used to predict Y^* (i.e., to describe the performance of Y^*). To draw meaningful prediction inference, the additional Requirement (R2) is imposed to ensure that the statements of our prediction have the desired frequentist interpretations. In particular, Requirement (R2) ensures that the coverage probability (CP for short) defined in (1) equals α, $0 < \alpha < 1$, for $L_1(\mathbf{Y}_n) = Q^{-1}_{\mathbf{Y}_n}(\alpha/2)$ and $L_2(\mathbf{Y}_n) = Q^{-1}_{\mathbf{Y}_n}(1-\alpha/2)$.

Note that Definition 1 of prediction functions bears striking resemblance to the definition of confidence distributions (CDs), except that the parameter θ and the corresponding parameter space Θ in CDs are now replaced, respectively, by the "future observation" Y^* and its sample space $\mathcal{Y}^*$. More precisely, a sample-dependent function defined on the parameter space Θ is called a CD for θ if it satisfies the following two requirements: ($R1^c$) For each given sample, it is a distribution function on the parameter space Θ; ($R2^c$) It can provide confidence intervals (regions) of all levels for θ; cf. Xie and Singh (2013), Schweder and Hjort (2016) and references therein. See also Singh et al. (2001), Schweder and Hjort (2002) and Singh et al. (2005) for a formal definition of CD. In general, a CD is a distribution estimate, instead of the usual point or interval estimate, of the parameter of interest.

The statement of Definition 1 is an abstract definition without concrete procedures for constructing predictive distribution functions. We exploit the similarities between the concepts of CDs and predictive distributions, in terms of their capability of summarizing information and quantifying uncertainty, to devise a precise formulation based on CD for constructing predictive distribution functions.

As stated in Cox (2013), a CD is to provide "a simple and interpretable summary of what can reasonably be learned from data (and an assumed model)". It quantifies both the information and uncertainty about the parameter θ from the observed data, and thus should naturally be the first and key ingredient for constructing a predictive distribution function for Y^*. This link of CDs to the construction of predictive distributions will later be seen as desirable in many practices. More specifically, for a given CD for θ derived from the data $\mathbf{y}_n$, denoted by $H_n(\cdot) = H(\cdot; \mathbf{y}_n)$, we can apply the formula below to obtain a predictive distribution function:

$$Q(y^*; \mathbf{y}_n) = \int_{\theta \in \boldsymbol{\Theta}} F_\theta(y^*) dH(\theta; \mathbf{y}_n). \tag{4}$$

In Schweder and Hjort (2016) the same formula was also suggested along with some examples. Strictly speaking, $Q(y^*; \mathbf{y}_n)$ obtained by using (4) may not always satisfy Requirement (R2), but our theoretical results in Section 3 show that (R2) holds exactly under some additional conditions on $F_\theta(y^*)$ and $H(\theta; \mathbf{y}_n)$ and asymptotically under mild conditions.

We now use two simple examples to illustrate the construction formula (4). The first example assumes i.i.d. sequence from the same distribution, but the second allows Y^* and Y_i's to have different distribution to show the flexibility and generality of the proposed approach. These two examples will serve as working examples for illustrating key steps in our development throughout the paper.

Example 1 (*Normal Distribution with Known Variance*)**.** Let $Y_1, \ldots, Y_n$ and Y^* be independent copies from $N(\theta, \sigma^2)$ with a known σ^2. A CD for θ based on the sample $\mathbf{y}_n$ is $N(\bar{y}, \sigma^2/n)$, where $\bar{y} = \sum_{i=1}^n y_i/n$ is the sample mean. This yields $F_\theta(y^*) = \Phi((y^* - \theta)/\sigma)$ and $H(\theta; \mathbf{y}_n) = \Phi((\theta - \bar{y})/(\sigma/\sqrt{n}))$. Thus, by (4), it follows immediately

$$Q(y^*; \mathbf{y}_n) = \int_{-\infty}^{\infty} \Phi\left(\frac{y^* - \theta}{\sigma}\right) d\Phi\left(\frac{\theta - \bar{y}}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{y^* - \bar{y}}{\sigma\sqrt{1 + 1/n}}\right). \tag{5}$$

Since $Q(Y^*; \mathbf{Y}_n) = \Phi((Y^* - \bar{Y})/(\sigma/\sqrt{1 + 1/n})) \sim \text{Uniform}(0, 1)$, the requirements in Definition 1 are satisfied. Note that this $Q(y^*; \mathbf{y}_n)$ is exactly the well-known predictive distribution $N(\bar{y}, \sigma^2(1 + 1/n))$ as well as the Bayesian predictive distribution with a flat prior for θ.

Example 2 (*Exponential Distribution*)**.** Let $Y_1, \ldots, Y_n$ be independent copies from an exponential distribution with scale $\alpha\theta$ where $\alpha > 0$ is a known acceleration parameter (as in an accelerated life testing). Then, the joint density function of

$\mathbf{Y}_n \equiv \{Y_1, Y_2, \ldots, Y_n\}$ is $g_\theta(\mathbf{y}_n) = (\alpha\theta)^{-n} e^{-n\bar{y}/(\alpha\theta)}$, where $\bar{y} = \sum_{i=1}^n y_i/n$ is the sample mean. Let Y^* follow an exponential distribution with scale θ, i.e., with the density function $f_\theta(y^*) = \theta^{-1} e^{-y^*/\theta}$. A CD for θ based on the sample $\mathbf{y}_n$ is $H_n(\theta) = H(\theta; \mathbf{y}_n) = 1 - \Gamma_{n,1}(n\bar{y}/(\alpha\theta))$, where $\Gamma_{n,1}(\cdot)$ is the cumulative distribution function of Gamma$(n, 1)$ distribution. With $F_\theta(y^*) = 1 - e^{-y^*/\theta}$, for $y^* > 0$, it follows from (4) with straightforward calculations that

$$Q(y^*; \mathbf{y}_n) = \int_0^\infty F_\theta(y^*) dH(\theta; \mathbf{y}_n) = 1 - \left(1 + \frac{\alpha y^*}{n\bar{y}}\right)^{-n}. \tag{6}$$

Clearly, the two requirements in Definition 1 hold, since $\alpha Y^*/\bar{Y}$ follows an F-distribution and $Q(Y^*; \mathbf{Y}_n) = \mathcal{F}_{2,2n}(Y^*/\bar{Y}) \sim$ Uniform$(0, 1)$. Here, $\mathcal{F}_{2,2n}(t) = 1 - (1 + t/n)^{-n}$ is the cumulative distribution function of the F-distribution with degrees of freedom $(2, 2n)$. Note that this same predictive distribution can also be obtained using the Bayesian approach with the Jeffreys' prior $\pi(\theta) \propto 1/\theta$.

Note that there are many ways to derive a CD, say from, for instance, normalized likelihood, fiducial distribution, Bayesian posterior distribution, bootstrap distribution, p-value function, among others; cf. Xie and Singh (2013) and references therein. The same paper also stated, "any approach, regardless of being frequentist, fiducial or Bayesian, can potentially be unified under the concept of confidence distributions, as long as it can be used to build confidence intervals of all levels, exactly or asymptotically". This useful property that CD can provide a unified framework to encompass inference procedures from different paradigms is readily inherited by the framework of predictive distribution functions. This makes formula (4) broadly applicable in many general settings. This point will be elaborated further.

3. Theoretical properties

In this section, we investigate theoretical properties of the predictive distribution $Q(y^*; \mathbf{y}_n)$ constructed using formula (4). For the ease of presentation, we focus on the case of scalar θ with $p = 1$ in this section. We will provide comments on extensions to the case of a multivariate θ with $p > 1$ at the end of the section.

The mean, the median and the mode of a CD $H_n(\cdot) = H(\cdot; \mathbf{y}_n)$ have been shown in Singh et al. (2007) to be consistent estimators of the unknown parameter θ under Condition (A) below:

(A) For any δ, $0 < \delta < 1/2$, $L_n(\delta) = H_n^{-1}(1 - \delta) - H_n^{-1}(\delta) \to 0$, in probability, as the sample size $n \to \infty$.

Later, Xie et al., (2011) proved that this is equivalent to Condition (A') below:

(A') For any fixed $\epsilon > 0$, $H_n(\theta_0 - \epsilon) \to 0$ and $H_n(\theta_0 + \epsilon) \to 1$, in probability, as $n \to \infty$,

where θ_0 is the true value of θ. These two conditions can be interpreted as: as the sample size n increases, the probability mass of the CD $H_n(\theta)$ becomes more concentrated around θ_0.

We establish the following theorem to show that, if $H_n(\theta)$ satisfies condition (A) or (A'), then $Q(y^*; \mathbf{y}_n)$ in (4) is an asymptotic predictive distribution function for Y^*. Thus, $Q(y^*; \mathbf{y}_n)$ based on $H_n(\theta)$ has valid frequentist interpretations asymptotically. A proof of the theorem is given in Appendix.

Theorem 1. *Assume that the CD $H_n(\cdot)$ used for constructing the predictive function in (4) satisfies Condition (A), and also that $F_\theta(\cdot)$ is continuous in θ in a neighborhood of θ_0:*

$$\sup_t \left|F_\theta(t) - F_{\theta_0}(t)\right| \le C\,|\theta - \theta_0|\,, \tag{7}$$

for some constant $C > 0$. Then,

$$Q(Y^*; \mathbf{Y}_n) = U + o_p(1), \tag{8}$$

where $U \sim$ Uniform$(0, 1)$.

Theorem 1 ensures an asymptotic coverage in (3) for a broad range of settings, though in some cases such as in Examples 1 and 2, $Q(Y^*; \mathbf{Y}_n)$ follows exactly Uniform$(0, 1)$ independent of the sample size. Next, we provide a set of sufficient conditions, under which the predictive distribution $Q(Y^*; \mathbf{Y}_n)$ always has exact coverage probability. Specifically, consider a condition on the distribution function $F_{\theta_0}(y^*)$:

(I) Suppose that there exists a monotonic mapping $s_1 : \mathcal{Y}^* \times \Theta \to \mathcal{Y}^*$ and a monotonic mapping $s_2 : \Theta \times \Theta \to \Theta$ such that $F_{\theta_0}(y^*)$ is invariant to the transformations s_1 and s_2 in the sense that, for any $\theta \in \Theta$,

$$F_{\theta_0}(y^*) = F_{s_2(\theta_0, \theta)}(s_1(y^*, \theta)). \tag{9}$$

Condition (I) is satisfied in both Examples 1 and 2. For instance, in Example 1, with $s_1(y^*, \theta) = y^* - \theta$, $s_2(\theta_0, \theta) = \theta_0 - \theta$ and $\mathcal{Y}^* \equiv \Theta \equiv (-\infty, \infty)$, we can verify (9), since $F_{\theta_0}(y^*) = \Phi((y^* - \theta_0)/\sigma) = \Phi(\{(y^* - \theta) - (\theta_0 - \theta)\}/\sigma) = F_{\theta_0 - \theta}(y^* - \theta)$ for any $\theta \in (-\infty, \infty)$. Similarly, in Example 2, with $s_1(y^*, \theta_0) = y^*/\theta_0$, $s_2(\theta_0, \theta) = \theta_0/\theta$ and $\mathcal{Y}^* \equiv \Theta \equiv (0, \infty)$, we immediately have (9), since $F_{\theta_0}(y^*) = 1 - e^{-y^*/\theta_0} = 1 - e^{-(y^*/\theta)/(\theta_0/\theta)} = F_{\theta_0/\theta}(y^*/\theta)$ for any $\theta \in (0, \infty)$.

Without loss of generality and to simplify our presentation, we assume from now on that $s_2(\theta_0, \theta)$ is increasing in θ_0 and decreasing in θ. Denote by $S_{\theta_0}(t) = \mathbb{P}_{\theta_0}\{s_1(Y^*, \theta_0) \le t\}$ and $R_{\theta_0}(t) = \mathbb{P}_{\theta_0}\{s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0) \le t\}$, where $\hat{\theta}(\mathbf{Y}_n)$ is the maximum likelihood estimate or some other efficient estimate of θ_0 derived from the observed data. It follows immediately that $R_{\theta_0}(s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0)) \sim$ Uniform(0, 1). If $s_1(Y^*, \theta_0)$ and $s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0)$ are pivotal quantities, then $S_{\theta_0}(t)$ and $R_{\theta_0}(t)$ are independent of θ_0, and thus can be written as $S(t)$ and $R(t)$. In this case, a CD for θ_0 can be obtained by

$$H_R(\theta; \hat{\theta}(\mathbf{y}_n)) = 1 - R(s_2(\hat{\theta}(\mathbf{y}_n), \theta)).$$

Following (4), the corresponding predictive distribution is

$$Q_R(y^*; \mathbf{y}_n) = \int_{\theta \in \Theta} F_\theta(y^*) dH_R(\theta; \hat{\theta}(\mathbf{y}_n)). \tag{10}$$

The following theorem states that the function $Q_R(\cdot; \cdot)$ expressed in (10) is an exact predictive distribution function. This theorem covers a class of cases including Examples 1 and 2. The proof of the theorem is also given in Appendix.

Theorem 2. *Assume that Condition (I) holds, and that $s_1(Y^*, \theta_0)$ and $s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0)$ are pivotal quantities. Then, $Q_R(Y^*; \mathbf{Y}_n) \sim$ Uniform(0, 1).*

The proposed CD-based prediction framework has broad implications. In particular, we present two corollaries which indicate that the CD-based prediction framework can be applied broadly to encompass several existing Bayesian, fiducial and frequentist prediction procedures.

First, note that fiducial and posterior distributions are sample-dependent distribution functions on the parameter space. If their corresponding fiducial or credible intervals have valid frequentist probability coverages (which is a goal in many developments on the topics of fiducial and (objective) Bayes), they satisfy the definition of CDs; cf. Xie and Singh (2013). In this context, Bayesian predictive distributions defined in (2) and the fiducial predictive distributions defined in Wang et al. (2012) are in fact the same as (or treated as special cases of) the general formulation of the predictive distributions in (4). Thus, an immediate result from Theorems 1 and 2 is that the predictive intervals obtained from these fiducial and Bayesian predictive distributions have valid frequentist coverage. This observation is summarized as a corollary below.

Corollary 1. *If a Bayesian posterior or a fiducial distribution of θ can be justified as a CD, then (a) its corresponding predictive distribution also has the valid frequentist probability coverage, and (b) it is also a predictive distribution function with valid frequentist probability coverages, as defined in Definition 1.*

Note that the predictive distribution by the pivotal method of Lawless and Fredette (2005) can also be linked to the general formulation (4), even though it is quite different in appearance. The pivotal method relies on the random variable $W = F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*)$, which is required to be a pivotal quantity so that its cumulative distribution function $K(t) = \mathbb{P}_{\mathbb{J}}(W \le t)$ is parameter-free. By defining our predictive distribution function as

$$Q_{\text{piv}}(y^*; \mathbf{y}_n) \equiv K(F_{\hat{\theta}(\mathbf{y}_n)}(y^*)), \tag{11}$$

we obtain the predictive distribution function proposed in Lawless and Fredette (2005). Clearly, $Q_{\text{piv}}(y^*; \mathbf{y}_n)$ satisfies the requirements in Definition 1. The next corollary states that $Q_{\text{piv}}(y^*; \mathbf{y}_n)$ can actually be expressed in the general formula (4). A proof of Corollary 2 can be found in Appendix.

Corollary 2. *Under the condition of Theorem 2, $Q_{piv}(y^*; \mathbf{y}_n)$ defined in (11) can be expressed as*

$$Q_{piv}(y^*; \mathbf{y}_n) = \int_{\theta \in \Theta} F_\theta(y^*) dH_R(\theta; \hat{\theta}(\mathbf{y}_n)),$$

where $H_R(\theta; \hat{\theta}(\mathbf{y}_n))$ is the CD obtained based on $\hat{\theta}(\mathbf{y}_n)$.

Altogether, Corollaries 1 and 2 suggest that the general formulation of predictive distributions in (4) through CDs provides a common link or a unifying platform for most, if not all, existing frequentist, fiducial and Bayesian predictive distributions.

We next discuss two optimality results regarding choices of different predictive distribution functions. As in the CD development where there may exist different CDs for the same estimation problem, there may also exist different predictive distribution functions for the same prediction problem. The discussions on the optimality issues surrounding CDs in Xie and Singh (2013) and Schweder and Hjort (2016) indicate that a better CD (under a certain criterion) typically leads to a better (under the same or a similar criterion) of the corresponding point estimator or test, and vice versa (cf. Xie and Singh, 2013). A natural question here would be whether a better CD will also lead to a better predictive distribution function.

Assume that we are to predict the random quantity Y^* and also that $Y^* \sim F_{\theta_0}(\cdot)$. Let $U^* = F_{\theta_0}(Y^*)$. Then $U^* \sim$ Uniform(0, 1) and is free of θ_0. Assume that $F_{\theta_0}(\cdot)$ is invertible, then $Y^* = F^{-1}_{\theta_0}(U^*)$. Suppose that we use predictive

distribution function, $Q(\cdot; \mathbf{Y}_n)$, to make inference about Y^*. Denoting by $Q_{\mathbf{Y}_n}^{-1}$ the inverse function of $Q_{\mathbf{Y}_n}(\cdot) \equiv Q(\cdot; \mathbf{Y}_n)$, we can write $Y_Q^* = Q_{\mathbf{Y}_n}^{-1}(U^*) \sim Q(\cdot; \mathbf{Y}_n)$ and then view Y_Q^* as a random copy of the predicted Y^* derived from the predictive distribution function $Q(\cdot; \mathbf{Y}_n)$.

We define the *predictive mean squared error* (PMSE) for quantifying the expected squared deviation between Y_Q^* and Y^*:

$$\mathrm{PMSE}(Q) = \mathbb{E}_{\mathbb{J}}(Y_Q^* - Y^*)^2. \tag{12}$$

In essence, PMSE(Q) quantifies the expected squared deviation between the quantiles of the distributions $Q(\cdot; \mathbf{Y}_n)$ and $F_{\theta_0}(\cdot)$.

Suppose that there are two different predictive distribution functions $Q_1(\cdot; \mathbf{Y}_n)$ and $Q_2(\cdot; \mathbf{Y}_n)$ obtained through the general form (4) using two different CDs, $H_1(\cdot) \equiv H_1(\cdot; \mathbf{Y}_n)$ and $H_2(\cdot) \equiv H_2(\cdot; \mathbf{Y}_n)$, respectively. We study how the underlying properties of CDs $H_1(\cdot)$ and $H_2(\cdot)$ affect the comparison of PMSE(Q_1) and PMSE(Q_2).

Following Singh et al. (2007) and Xie and Singh (2013), $H_1(\cdot)$ is considered *more efficient* than $H_2(\cdot)$ at $\theta = \theta_0$, if for all $u \in (0, 1)$,

$$(H_1^{-1}(u) - \theta_0)^+ \overset{\text{sto}}{\leq} (H_2^{-1}(u) - \theta_0)^+ \quad \text{and} \quad (H_1^{-1}(u) - \theta_0)^- \overset{\text{sto}}{\leq} (H_2^{-1}(u) - \theta_0)^-. \tag{13}$$

Here $\overset{\text{sto}}{\leq}$ is the stochastic order. Also shown in Singh et al. (2007) is: If the *CD random variable* associated with $H_i(\cdot)$ is denoted by $\theta_{\mathrm{CD},i}$, i.e., $\theta_{\mathrm{CD},i} \sim H_i(\cdot)$ for $i = 1, 2$, then

$$(\theta_{\mathrm{CD},1} - \theta_0)^+ \overset{\text{sto}}{\leq} (\theta_{\mathrm{CD},2} - \theta_0)^+ \quad \text{and} \quad (\theta_{\mathrm{CD},1} - \theta_0)^- \overset{\text{sto}}{\leq} (\theta_{\mathrm{CD},2} - \theta_0)^-. \tag{14}$$

The inequality (14) is interpreted that the CD $H_1(\cdot)$ is more "concentrated" around the true parameter θ_0 than $H_2(\cdot)$; cf., Singh et al. (2007). The theorem below shows that a more efficient CD yields a better prediction. A proof of the theorem is provided in Appendix.

Theorem 3. *Let $F_\theta^{-1}(u)$ be nondecreasing in θ for any given $u \in (0, 1)$. If the CD $H_1(\theta)$ is more efficient than another CD $H_2(\theta)$ for θ for the same true value θ_0, then*

$$\mathit{PMSE}(Q_1) \leq \mathit{PMSE}(Q_2), \tag{15}$$

where $Q_i(Y^*; \mathbf{Y}_n)$ is the predictive distribution induced by $H_i(\theta)$ for $i = 1, 2$.

An immediate result in terms of stochastic ordering similar to that of (14) is

$$(Y_{Q_1}^* - Y^*)^+ \overset{\text{sto}}{\leq} (Y_{Q_2}^* - Y^*)^+ \quad \text{and} \quad (Y_{Q_1}^* - Y^*)^- \overset{\text{sto}}{\leq} (Y_{Q_2}^* - Y^*)^-, \tag{16}$$

provided that $F_\theta^{-1}(u)$ is nondecreasing in θ. Similarly, (16) can be interpreted as that the predictive distribution function $Q_1(\cdot; \mathbf{Y}_n)$ is more "concentrated" around the "actual" Y^* than $Q_2(\cdot; \mathbf{Y}_n)$.

If there is a family of uniformly most powerful unbiased (UMPU) tests for testing $K_0 : \theta \leq c$ versus $K_1 : \theta > c$, for every $c \in \Theta$, Theorem 2.2 of Singh et al. (2007) states that the CD corresponding to the p-value function of the UMPU tests is the most efficient. Combining this observation with Theorem 3, we immediately have:

Corollary 3. *Under the setting of Theorem 3 and assume that a CD is derived from a p-value function of a UMPU test, then the corresponding predictive distribution function obtained by using (4) has the smallest PMSE.*

We now use Examples 1 and 2 to elucidate the implications of Theorem 3. Under the setting of Example 1, Singh et al. (2007) showed that $H_1(\theta) = \Phi((\theta - \bar{Y})/(\sigma/\sqrt{n}))$ is the most efficient CD for θ_0. We also consider a CD derived from the sample median, say M. Since $\sqrt{n}(M - \theta_0) \to N(0, \pi\sigma^2/2)$ in distribution, as $n \to \infty$, $H_2(\theta) = \Phi((\theta - M)/(\sigma/\sqrt{2n/\pi}))$ is an asymptotic CD for θ_0. Although H_2 may be more robust, it is known to be less efficient than H_1. Applying (4), the predictive distribution functions based on H_1 and H_2 can be obtained. They are $Q_1(Y^*; \mathbf{Y}_n) = \Phi((Y^* - \bar{Y})/(\sigma/\sqrt{1 + 1/n}))$ and $Q_2(Y^*; \mathbf{Y}_n) = \Phi((Y^* - M)/(\sigma/\sqrt{1 + \pi/2n}))$, respectively. Since $F_\theta^{-1}(u) = \Phi^{-1}(u) + \theta$ is increasing in θ for any $u \in (0, 1)$, Theorem 3 implies that the PMSE of Q_1 is smaller than that of Q_2. Indeed, simple algebra gives $\mathrm{PMSE}(Q_1) = \frac{2}{n}\sigma^2$ and $\mathrm{PMSE}(Q_1) \approx \frac{\pi}{n}\sigma^2$ for a large n. The same argument also holds for Example 2 with the CDs $H_1(\theta) = 1 - \Gamma_{n,1}(n\bar{Y}_n/(\alpha\theta))$ and $H_2(\theta) = 1 - \Gamma_{1,1}(n'\bar{Y}_{n'}/(\alpha\theta))$ with their corresponding predictive distributions $Q_1(Y^*; \mathbf{Y}_n) = \mathcal{F}_{2,2n}(\alpha Y^*/\bar{Y}_n)$ and $Q_2(Y^*; \mathbf{Y}_n) = \mathcal{F}_{2,2n'}(\alpha Y^*/\bar{Y}_{n'})$. Here we assume that $\mathbf{Y}_{n'}$ is a subset of $\mathbf{Y}_n$ with $n > n' > 4$. Clearly, $F_\theta^{-1}(u) = -\theta \log(1 - u)$ is increasing in θ for any $u \in (0, 1)$. Therefore, Theorem 3 implies $\mathrm{PMSE}(Q_1) < \mathrm{PMSE}(Q_2)$.

Finally, we discuss the plug-in predictive distribution $F_{\hat{\theta}}(y^*)$ which has often been used as an approximation to the true distribution $F_{\theta_0}(y^*)$. Although the plug-in predictive distribution has valid asymptotic coverage probability similar to that of (8), it fails to account for the uncertainty in the estimation of θ and typically cannot achieve exact coverage probability in comparison with the result of Theorem 2. In fact, Lawless and Fredette (2005) showed that when the pivot method applies, the predictive distribution $Q_{\mathrm{piv}}(y^*; \mathbf{y}_n)$ in (11) is always better than the plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, as measured by the average Kullback–Leibler distance to the true distribution $F_{\theta_0}(y^*)$; cf. Theorem 1 of Lawless and Fredette (2005).

The next theorem reports a slightly more general result. Let $H_R(\theta; \hat{\theta}(\mathbf{y}_n))$ be a CD for θ obtained based on $\hat{\theta} = \hat{\theta}(\mathbf{y}_n)$. To simplify the notations, we let $Q_{\hat{\theta}}(t) = Q(t; \hat{\theta})$, $q_{\hat{\theta}}(t) = \frac{d}{dt}Q_{\hat{\theta}}(t)$ and $f_{\hat{\theta}}(t) = \frac{d}{dt}F_{\hat{\theta}}(t)$. The theorem below shows that the

predictive distribution function $Q_{\hat{\theta}}(y^*) = Q(y^*; \hat{\theta}(\mathbf{y}_n))$ obtained using $H_R(\theta; \hat{\theta}(\mathbf{y}_n))$ is better than the naive plug-in predictive distribution function $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, as measured by the average Kullback–Leibler distance to the true distribution $F_{\theta_0}(y^*)$. A proof is provided in Appendix.

Theorem 4. *Assume that*

$$\mathbb{E}_{\mathbb{J}}\left\{\frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} \le 1. \tag{17}$$

Then,

$$\bar{D}_{KL}(f_{\theta_0}|q_{\hat{\theta}}) \le \bar{D}_{KL}(f_{\theta_0}|f_{\hat{\theta}}),$$

where $\bar{D}_{KL}(f_{\theta_0}|g_{\hat{\theta}}) = \mathbb{E}_{\mathbb{J}}\left\{\log\frac{f_{\theta_0}(Y^*)}{g_{\hat{\theta}}(Y^*)}\right\}$ *is the average Kullback–Leibler distance between* f_{θ_0} *and any density function of the form* $g_{\hat{\theta}}$.

In the pivot example in Lawless and Fredette (2005), $Q_{\text{piv}}(y^*; \mathbf{y}_n) = K(F_{\hat{\theta}(\mathbf{y}_n)}(y^*))$. So, $q_{\hat{\theta}}(t) = \frac{\partial}{\partial t}Q_{\text{piv}}(t; \mathbf{y}_n) = k(F_{\hat{\theta}(\mathbf{y}_n)}(t))f_{\hat{\theta}(\mathbf{y}_n)}(t)$, where $k(s) = \frac{\partial}{\partial s}K(s)$ is the density function corresponding to the cumulative distribution function $K(\cdot)$. It follows from direct calculation that $\mathbb{E}_{\mathbb{J}}\left\{\frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} = \mathbb{E}_{\mathbb{J}}\{1/k(F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*))\} = \mathbb{E}_U\{1/k(U)\} = 1$, where the last two equations are obtained by variable transformation and the observation that $U = F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*) \sim k(\cdot)$. Thus, (17) holds and Theorem 4 covers the result of Lawless and Fredette (2005) as a special case.

We close this section by addressing the potential extensions of the above theoretic developments to the multivariate $\boldsymbol{\theta} \in \mathbb{R}^p$ setting with $p > 1$. Although, on the outset, we note that Definition 1 of the predictive distribution, the general formulation (4) in Section 2 and even the algorithm to be proposed in Section 4 can be applied directly in the multivariate setting, there remains a technical difficulty in defining a general multivariate CD in very general cases and thus a rigorous presentation of all theoretical results in Section 3 for the general multivariate $\boldsymbol{\theta}$ setting is still being sought. In principle, the concept of a multivariate CD is straightforward (i.e., a sample-dependent distribution function on the multivariate parameter space that can produce confidence regions of all levels), however a precise definition with explicit mathematical formulation to cover very general cases thus far remains elusive. But partial progress can still be made, since under asymptotic settings or wherever the usual likelihood inference or bootstrap theory applies, multivariate CDs can be applied with ease. For instance, under the general setup in likelihood inference, the multivariate normal distribution $N(\hat{\boldsymbol{\theta}}, \hat{\Sigma})$ serves as a first-order asymptotic CD for $\boldsymbol{\theta}$ where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and $\hat{\Sigma}$ is the inverse of the observed Fisher's information using the entire n observations; cf. Yang et al. (2014) and Liu et al. (2015). In addition, if we limit ourselves to center-outwards confidence regions (instead of all Borel sets) in the parameter space, concepts such as the c-CDs derived from the notion of data depth (cf. Liu et al. 1999) considered in Singh et al. (2007) and the confidence curve considered in Schweder (2007) and Schweder and Hjort (2016) offer coherent notions of multivariate CDs in the exact sense. In these cases, we still can generalize most of the theoretical developments to the multivariate setting. This fact has been used in some of our examples, e.g., in Section 6. See also Schweder and Hjort (2016) for related discussions.

4. A computing algorithm

To implement the approach formulated in (4), we propose a Monte Carlo algorithm for computing predictive distributions and prediction intervals. This algorithm is simple yet applicable to a wide range of problems. Specifically, given $\mathbf{Y}_n = \mathbf{y}_n$, a CD $H_n(\cdot) = H(\cdot; \mathbf{y}_n)$ is a distribution function on the parameter space Θ. Conditional on $\mathbf{Y}_n = \mathbf{y}_n$, we can simulate a CD-random variable θ_{CD} by $\theta_{\text{CD}}|\mathbf{y}_n \sim H_n(\cdot)$. The precise algorithm is as follows:

[**Monte Carlo Algorithm**] Obtain a simulated copy of y_S^* from $Q(\cdot; \mathbf{y}_n)$ by: first simulate a CD-random variable $\theta_{\text{CD}}|\mathbf{y}_n \sim H_n(\cdot)$, and then simulate a y_S^* from $y_S^*|\theta_{\text{CD}} \sim f_{\theta_{\text{CD}}}(\cdot)$. Repeat this procedure a large number of times, say N times, to obtain N copies of simulated y_S^*. The histogram of these N copies of y_S^* is then used to approximate a predictive distribution of Y^* and hence its prediction intervals of all levels.

This algorithm applies to any CD $H_n(\cdot) = H(\cdot, \mathbf{y}_n)$ for $\theta \in \mathbb{R}^p$. Note that, any approach, regardless of being frequentist, fiducial or Bayesian, can be used to construct CDs, as long as the produced CDs can be used to build confidence intervals of all levels, exact or asymptotically; cf. Xie and Singh (2013) and references therein. So this algorithm is quite general and can be applied broadly.

As a special case, this algorithm can be carried out using a bootstrap method, noting that a bootstrap distribution is known to be also a CD (see e.g., Efron, 1998; Xie and Singh, 2013). In particular, we can simply simulate a future observation y^* by $y^*|\theta_{\text{boot}} \sim f_{\theta_{\text{boot}}}(\cdot)$, where θ_{boot} is the bootstrap estimate of the parameter θ. Obviously, this simulation method makes the proposed prediction approach very useful in practice, as it is simple and general.

Clearly, the prediction intervals and predictive distributions obtained by using the proposed algorithm above have valid frequentist interpretations, following Theorems 1 and 2 (and their extensions to multivariate θ as discussed at the end of Section 3),

5. Simulation

In this section, we use two numerical examples to demonstrate the proposed approach and computing algorithm for constructing predictive distribution functions, and then examine their frequentist properties. The first example is a simple linear regression model with zero-intercept, for which the well-known exact predictive distribution function can be obtained explicitly. We report and compare the numerical results from this explicit predictive distribution function and those from the computing algorithm described in Section 4. The second example is a nonlinear regression, for which an exact CD function for the underlying parameter does not exist and neither does the corresponding predictive distributions. Nonetheless, we are able to apply the computing algorithm in Section 4 with several different asymptotic CD functions to perform predictions and study their numerical performance.

Simulation Example I: Consider a simple linear regression model with zero-intercept:

$$y_i = \theta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Let $\hat{\theta} = \sum_{i=1}^n y_i x_i / \sum_{i=1}^n x_i^2$ be the ordinary least squares estimate of θ and $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \hat{\theta}x_i)^2$. For a new independent observation Y^* associated with covariate x^*, the well-known predictive distribution is given by

$$Q_t(y^*; \mathbf{y}_n) = T_{n-1}\left(\frac{y^* - \hat{\theta}x^*}{\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}}\right), \tag{18}$$

where $T_{n-1}(\cdot)$ is the cumulative distribution function of t-distribution with degrees of freedom $n-1$. It is easy to verify that $Q_t(y^*; \mathbf{y}_n)$ satisfies Definition 1. This is the same as the predictive distribution considered in Schweder and Hjort (2016). If σ is known, straightforward calculation can yield

$$H(\theta) = \Phi\left(\frac{\theta - \hat{\theta}}{\sigma/\sqrt{\sum_{i=1}^n x_i^2}}\right)$$

as a CD for θ. The corresponding predictive distribution for Y^* using formula (4) is then

$$Q(y^*; \mathbf{y}_n) = \int_{-\infty}^{\infty} \Phi\left(\frac{y^* - \xi x^*}{\sigma}\right) d\Phi\left(\frac{\xi - \hat{\theta}}{\sigma/\sqrt{\sum_{i=1}^n x_i^2}}\right) = \Phi\left(\frac{y^* - \hat{\theta}x^*}{\sigma\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}}\right),$$

and hence

$$Q_a(y^*; y_n) = \Phi\left(\frac{y^* - \hat{\theta}x^*}{\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}}\right) \tag{19}$$

is an asymptotic predictive distribution by replacing σ with its estimate $\hat{\sigma}$.

Alternatively, we can also construct the predictive distribution using bootstrap distribution of θ, since bootstrap distribution is an asymptotic CD (as demonstrated earlier) with which the bootstrap estimator is the corresponding CD-random variable. Specially, we (1) compute the ordinary least squares estimate $\hat{\theta}$, (2) obtain bootstrap samples from the residuals $e_i = y_i - \hat{\theta}x_i$, denoted by $e_{i,\text{boot}}$, and (3) compute the bootstrap least squares estimate $\hat{\theta}_{\text{boot}}$ using the new samples $\{(y_{i,\text{boot}}, x_i)\}_{i=1}^n$, where $y_{i,\text{boot}} \equiv \hat{\theta}x_i + e_{i,\text{boot}}$. Finally in step (4), a sample from the predictive distribution of Y^*_{boot}, say $Q_{\text{boot}}(\cdot; \mathbf{y}_n)$, can be obtained empirically by first generating $\epsilon^* \sim N(0, \hat{\sigma}^2)$ and then computing $y^*_{\text{boot}} = \hat{\theta}_{\text{boot}}x^* + \epsilon^*$. Repeat these four steps for a large number of times to get sufficient many copies of y^*_{boot}. These copies of y^*_{boot} are then used to construct a predictive distribution function as well as prediction intervals.

We compare the empirical coverage probabilities of the prediction intervals from the four different predictive distributions: (1) the naive plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*) = \Phi((y^* - \hat{\theta}x^*)/\hat{\sigma})$, (2) the exact predictive distribution $Q_t(y^*; \mathbf{y}_n)$ in (18) and (3) the asymptotic predictive distributions $Q_a(y^*; \mathbf{y}_n)$ in (19) and (4) $Q_{\text{boot}}(y^*; \mathbf{y}_n)$ described in the previous paragraph. The prediction intervals are obtained by taking the upper and lower $\alpha/2$ quantiles of the corresponding predictive distributions. Comparisons are made with different choices of α and sample sizes in order to provide a general picture of performance comparison. The numerical settings are as follows: $\theta = 1, \sigma = 1, x_i \sim U[-2, 2]$ are fixed once they have been generated, and $x^* = 2$. For $Q_{\text{boot}}(y^*; \mathbf{y}_n)$, 1000 bootstrap samples are utilized. Three sample sizes are considered: $n = 10, 100, 1000$. For each sample size, the analysis is repeated 5000 times with $y_1, \ldots, y_n, y^*$ being simulated anew accordingly.

Table 1 contains the empirical coverage probabilities and median widths of the prediction intervals. Note that the widths of the prediction intervals from $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, $Q_t(y^*; \mathbf{y}_n)$ and $Q_a(y^*; \mathbf{y}_n)$ can be assessed without simulation. They are, respectively, $2z_{\alpha/2}\hat{\sigma}$, $2t_{n-1,\alpha/2}\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}$ and $2z_{\alpha/2}\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}$. Here $t_{n-1,\alpha/2}$ and $z_{\alpha/2}$ are the $(1-\alpha/2)$th percentiles of t distribution with degrees of freedom $n-1$ and the standard normal distribution, respectively. From Table 1, at all nominal levels, the prediction intervals from $Q_t(y^*; \mathbf{y}_n)$ have the correct frequentist coverage probability since it is an exact predictive

Table 1
Comparison of predictive distributions in Simulation Example I: 80%, 90% and 95% prediction intervals.

n	$1-\alpha$	$F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*)$		$Q_t(Y^*;\mathbf{Y}_n)$		$Q_a(Y^*;\mathbf{Y}_n)$		$Q_{\text{boot}}(Y^*;\mathbf{Y}_n)$	
		Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
10	0.80	0.695	2.482	0.793	3.096	0.762	2.869	0.746	2.786
10	0.90	0.803	3.185	0.895	4.103	0.860	3.682	0.845	3.579
10	0.95	0.871	3.796	0.950	5.064	0.916	4.387	0.906	4.258
100	0.80	0.795	2.555	0.805	2.619	0.803	2.601	0.798	2.595
100	0.90	0.893	3.279	0.903	3.370	0.899	3.339	0.897	3.326
100	0.95	0.941	3.907	0.947	4.028	0.944	3.978	0.943	3.958
1000	0.80	0.801	2.562	0.802	2.568	0.801	2.566	0.799	2.562
1000	0.90	0.901	3.288	0.902	3.296	0.901	3.293	0.900	3.284
1000	0.95	0.948	3.918	0.948	3.929	0.948	3.924	0.942	3.903

distribution. For small sample size (such as $n = 10$), the empirical coverage of the prediction intervals from $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$ is far below its nominal level. This is because those prediction intervals do not take into account the uncertainty stemming from the estimation of the unknown parameter and such uncertainty is large relative (or at least comparable) to the noise level (σ^2) for small n. The empirical coverages of the prediction intervals from $Q_a(y^*;\mathbf{y}_n)$ and $Q_{\text{boot}}(y^*;\mathbf{y}_n)$ improve significantly upon those from $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, though are still below the nominal level for small n. This is due to the fact that in $Q_a(y^*;\mathbf{y}_n)$ the estimated $\hat{\sigma}$ is used to approximate the actual σ and that for $Q_{\text{boot}}(y^*;\mathbf{y}_n)$ the bootstrap distribution only works well for at least moderate sample size n. For moderate or large sample size, such as $n = 100$ or $n = 1000$, the coverage probabilities of all the four types of prediction intervals approximate well their nominal levels.

Simulation Example II: Consider a nonlinear regression in the form of

$$y_i = h(x_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where

$$h(x_i, \boldsymbol{\theta}) = \frac{\theta_1 x_i}{\theta_2 + x_i},$$

and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. The parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^t$ can be estimated by nonlinear least squares and solved iteratively using Gauss–Newton algorithm. Denote by $\hat{\boldsymbol{\theta}}$ the nonlinear least squares estimate of θ and let $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - h(x_i, \hat{\boldsymbol{\theta}}))^2$. Although there exists no explicit expression of the exact sampling distribution of $\hat{\boldsymbol{\theta}}$, it can be approximated by $N\big(\boldsymbol{\theta}, \sigma^2\{A(\mathbf{x}, \boldsymbol{\theta})^t A(\mathbf{x}, \boldsymbol{\theta})\}^{-1}\big)$ where $A(\mathbf{x}, \boldsymbol{\theta})$ is the $n \times 2$ matrix of the partial derivatives with the ith row $\left(\frac{\partial}{\partial\theta_1}h(x_i, \boldsymbol{\theta}), \frac{\partial}{\partial\theta_2}h(x_i, \boldsymbol{\theta})\right) = \left(\frac{x_i}{\theta_2 + x_i}, -\frac{\theta_1 x_i}{(\theta_2 + x_i)^2}\right)$. Therefore, the cumulative distribution function of $N\big(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2(A(\mathbf{x}, \hat{\boldsymbol{\theta}})^t A(\mathbf{x}, \hat{\boldsymbol{\theta}}))^{-1}\big)$ can be used as an asymptotic CD function for $\boldsymbol{\theta}$. In this formula, the unknown values are replaced by their estimates.

For a new independent observation Y^* associated with covariate x^*, we can construct asymptotic predictive distribution by using the above asymptotic CD and taking advantage of the approximation

$$h(x^*, \boldsymbol{\theta}) \approx h(x^*, \hat{\boldsymbol{\theta}}) + a(x^*, \hat{\boldsymbol{\theta}})^t(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

where $a(x^*, \boldsymbol{\theta}) = \frac{\partial h(x^*, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Applying formula (4) and some simple algebra, we can obtain the asymptotic predictive distribution

$$Q_a(y^*; y_n) = \Phi\left(\frac{y^* - h(x^*, \hat{\boldsymbol{\theta}})}{\hat{\sigma}\sqrt{1 + a(x^*, \hat{\boldsymbol{\theta}})^t\{A(\mathbf{x}, \hat{\boldsymbol{\theta}})^t A(\mathbf{x}, \hat{\boldsymbol{\theta}})\}^{-1} a(x^*, \hat{\boldsymbol{\theta}})}}\right).$$

Alternatively, we can also see the bootstrap-based predictive distribution, denoted by $Q_{\text{boot}}(y^*;\mathbf{y}_n)$, using the construction almost in the same way as in Simulation Example I.

We proceed to compare the empirical coverage probabilities of the prediction intervals from the three different predictive distributions: (1) the naive plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*) = \Phi((y^* - h(x^*, \hat{\boldsymbol{\theta}}))/\hat{\sigma})$, (2) the asymptotic predictive distributions $Q_a(y^*;\mathbf{y}_n)$, and (3) $Q_{\text{boot}}(y^*;\mathbf{y}_n)$. Once again, comparisons are made at $\alpha = 0.8, 0.9, 0.95$ and $n = 10, 100, 1000$ with 5000 repetitions for each sample size. The numerical settings are: $\theta_1 = 15, \theta_2 = 5, \sigma = 1, x_i \overset{\text{i.i.d.}}{\sim} U[0, 30]$ are fixed once generated, $x^* = 40$. For the bootstrap-based approach, 1000 bootstrap samples are generated. Similar to Table 1, Table 2 lists the empirical coverage probabilities and median widths of the prediction intervals. In the case of small sample size ($n = 10$), the empirical coverage probabilities of the prediction intervals from all the three approaches are below the nominal level since they are all approximate methods. However, both the CD-based predictive distributions, either $Q_a(y^*;\mathbf{y}_n)$ derived from the multivariate normal CD or $Q_{\text{boot}}(y^*;\mathbf{y}_n)$ derived from the bootstrap CD, have outperformed the plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$ in terms of empirical coverage. This is because the CD-based methods have incorporated the uncertainty in the parameter estimation. Again, for moderate or large sample size, such as $n = 100$ or $n = 1000$, the coverage probabilities of all the three prediction intervals are close to the corresponding nominal levels.

Table 2
Comparison of predictive distributions in Simulation Example II: 80%, 90% and 95% prediction intervals.

n	$1-\alpha$	$F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*)$		$Q_a(Y^*;\mathbf{Y}_n)$		$Q_{\text{boot}}(Y^*;\mathbf{Y}_n)$	
		Coverage	Width	Coverage	Width	Coverage	Width
10	0.80	0.698	2.438	0.764	2.817	0.751	2.742
10	0.90	0.809	3.129	0.862	3.615	0.854	3.512
10	0.95	0.872	3.728	0.913	4.308	0.903	4.177
100	0.80	0.782	2.552	0.791	2.610	0.791	2.603
100	0.90	0.888	3.275	0.896	3.350	0.894	3.335
100	0.95	0.941	3.902	0.946	3.992	0.943	3.973
1000	0.80	0.794	2.564	0.795	2.569	0.793	2.564
1000	0.90	0.895	3.291	0.896	3.298	0.896	3.285
1000	0.95	0.949	3.922	0.949	3.929	0.947	3.907

6. Real applications

In this section, we provide a real data example, in which the predictive inference developed is applied to data from a complex time series. We can envision that the development of predictive distributions be applied and generalized to other complex situations such as survival data analysis, multiple regressions and any other fields and applications that involve forecasting and prediction.

Before we start our real data example, we need to extend the general formula (4) discussed in Section 2 to cover the case that Y^* and $\mathbf{Y}_n$ are dependent; for instance, a time series data in which $\mathbf{Y}_n$ are sample observations up to the given data and Y^* is a future response at the time series. Specifically, we propose to consider the conditional distribution of Y^* given $\mathbf{Y}_n$ and modify the general formula (4) to be

$$Q_c(y^*;\mathbf{y}_n)=\int_{\theta\in\Theta} F_\theta(y^*|\mathbf{y}_n)dH(\theta;\mathbf{y}_n). \tag{20}$$

In fact, formula (4) can now be viewed as a special case of (20) when $F_\theta(y^*|\mathbf{y}_n)\equiv F_\theta(y^*)$. Many of the theoretical results developed in Section 3 can be extended straightforwardly. For example, if we modify (7) to be $\sup_t\left|F_\theta(t|\mathbf{y}_n)-F_{\theta_0}(t|\mathbf{y}_n)\right|\le C\,|\theta-\theta_0|$ for some positive constant C, then the result of Theorem 1 applies to $Q_c(y^*;\mathbf{y}_n)$ for the dependent case. This means that the predictive distribution function $Q_c(y^*;\mathbf{y}_n)$ for the dependent case also has valid frequentist interpretations, under a set of very mild conditions.

The real data example is from a research project partially sponsored by the US Department Homeland Security (DHS) through its academic research center DHS University Center of Excellence for Command, Control, and Interoperability (CCICADA) based at Rutgers University. This data example specifically focuses on the analysis of the monthly volume of applications for a certain type of government benefit (the name of the governmental program is masked per a confidentiality agreement).

The main objective of the project is to seek more effective statistical methods that can substantially improve upon the current benchmark model used by the agency in gaining accuracy of forecast. This gain can allow the agency to optimize the human resource allocation and minimize the cost of management.

The data set contains 167 months of application volume. The logarithm transformation of the 167 observed volumes is shown in Fig. 1. We denote the transformed series by $\{y_t\}_{t=1}^{167}$.

It was noted in Chang (2015) the known outliers at $t=105, 106, 107$ due to policy changes in the application process. Thus, we filter out these outliers with three indicator variables $\mathbb{1}_t^{(105)}$, $\mathbb{1}_t^{(106)}$, and $\mathbb{1}_t^{(107)}$, where $\mathbb{1}_t^{(k)}=1$ if $t=k$ and $\mathbb{1}_t^{(k)}=0$ otherwise. Also, the series in seasonal nature exhibits a cyclical pattern with periodicity of 12 that is modeled with seasonal terms. In addition, there is a strong linear relationship between y_t and another type of benefit application x_t. Taking all this information into account and building upon the work by Chang (2015), we propose the following seasonal ARMA model with exogenous variables,

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(y_t-\beta_1 x_t-\beta_2\mathbb{1}_t^{(105)}-\beta_3\mathbb{1}_t^{(106)}-\beta_4\mathbb{1}_t^{(107)})=(1+\Theta_1 B^{12})\varepsilon_t. \tag{21}$$

Here, $\{\varepsilon_t\}$ is a white noise series with variance σ_ε^2, B is the backshift operator such that $B^s y_t=y_{t-s}$ for an integer $s>0$. Also, denote by $\boldsymbol{\theta}=(\phi_1,\Phi_1,\Theta_1,\beta_1,\beta_2,\beta_3,\beta_4)$ the associated coefficients.

Table 3 summarizes the coefficient estimates and their standard errors from model (21). It is easy to see that all the coefficients are significant at the 95% significance level. Fig. 2 shows the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the residuals from model (21). With no significant autocorrelation and partial autocorrelation, we conclude that model (21) is adequate in capturing the patterns of $\{y_t\}_{t=1}^{167}$.

Our ultimate goal is to make prediction on future application volumes given the past observations and construct the corresponding prediction interval and predictive distributions. More specifically, we need to predict a sequence of y_{167+h} for $h=1,2,\ldots$, based on past observations up to time $t=167$. On the other hand, since we do not know the values of the future observations after $t=167$, we cannot really tell how well these predictions are. To this end, we demonstrate the effectiveness of our proposed method by formulating our predictions as of length $h>0$ steps away, on a rolling basis

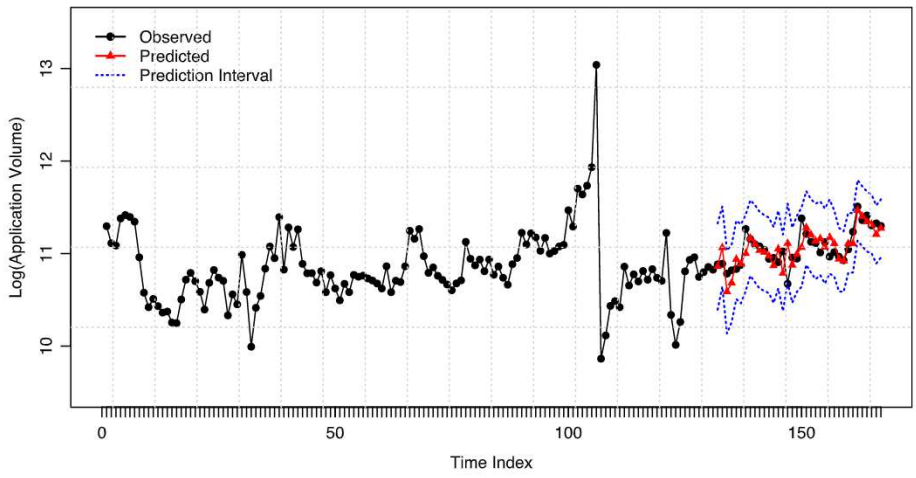

Fig. 1. Time series plot of monthly application volumes for a government benefit and 95% one-step ahead prediction intervals, rolling from $t = 141$ with a rolling window size $d = 120$. The red triangle points show the predicted values and the blue dotted lines show the upper and lower limits of the corresponding 95% prediction intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

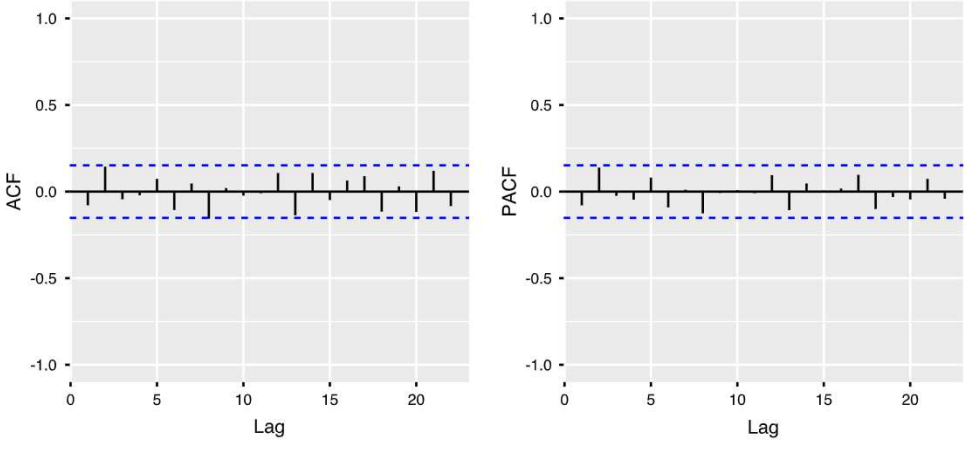

Fig. 2. Sample ACF and PACF plots of the residuals from model (21).

Table 3
Coefficient estimates and their standard errors of model (21).

	ϕ_1	Φ_1	Θ_1	β_1	β_2	β_3	β_4
Coefficient	0.784	0.998	−0.966	0.975	0.633	2.160	−0.696
Std. Error	0.048	0.011	0.111	0.014	0.175	0.200	0.175

with a rolling window of width d, e.g., $d = 120$ corresponding to the data of the past ten years. That is, at time t, we predict y_{t+l} based on the most recent d observations, compare the prediction with the actual value, and then increase t by one and repeat the procedure until $t = 167 - h$. It is well-known that the coverage of the prediction intervals by the so-called plug-in method (described in Section 3) is typically below the nominal level because they fail to consider the uncertainty in parameter estimation, among others. Using our approach, however, it is possible to capture this type of uncertainty, and thus show substantial improvement.

The process to derive simulated predictive distribution of y_{t+h}, given $\{y_{t-d+1}, \ldots, y_t\}$ for any prediction length h, is outlined in four steps as follows:

1. Estimate model (21) using maximum likelihood method and $\{y_{t-d+1}, \ldots, y_t\}$. Denote by $\hat{\boldsymbol{\theta}} = (\hat{\phi}_1, \hat{\Phi}_1, \hat{\Theta}_1, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ the estimated coefficients, and $\hat{\Sigma}$ the covariance matrix of $\hat{\boldsymbol{\theta}}$, and by $\hat{\sigma}_\varepsilon^2$ the estimated variance of the noise term. Let $\hat{y}_s$ be the fitted values of y_s and $e_s = y_s - \hat{y}_s$, for $s \le t$.

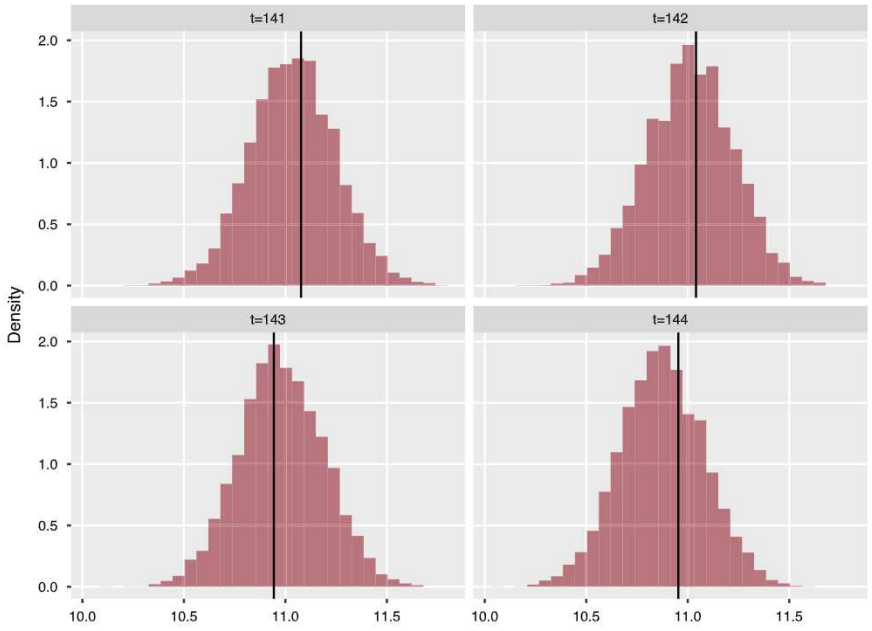

Fig. 3. One-step ahead predictive distribution of Y_t for $t = 141, 142, 143$ and 144.

2. As demonstrated in Section 3, the multivariate normal distribution $N(\hat{\boldsymbol{\theta}}, \hat{\Sigma})$ serves as a first-order asymptotic CD for $\boldsymbol{\theta} = (\phi_1, \Phi_1, \Theta_1, \beta_1, \beta_2, \beta_3, \beta_4)$ for a reasonable d. Thus, we can simulate

$$\hat{\boldsymbol{\theta}}_{\mathrm{CD}} = (\phi_{\mathrm{CD},1}, \Phi_{\mathrm{CD},1}, \Theta_{\mathrm{CD},1}, \beta_{\mathrm{CD},1}, \beta_{\mathrm{CD},2}, \beta_{\mathrm{CD},3}, \beta_{\mathrm{CD},4}) \sim N(\hat{\boldsymbol{\theta}}, \hat{\Sigma}).$$

   We also draw $\varepsilon^*_{t+1}, \ldots, \varepsilon^*_{t+h} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon)$, with the unknown σ^2_ε replaced by $\hat{\sigma}^2_\varepsilon$ under a reasonable d.
3. Recursively solve for y^*_{t+h} through $(1-\phi_{\mathrm{CD},1}B)(1-\Phi_{\mathrm{CD},1}B^{12})(y^*_{t+h} - \beta_{\mathrm{CD},1}x_{t+h} - \beta_{\mathrm{CD},2}\mathbb{1}^{(105)}_{t+h} - \beta_{\mathrm{CD},3}\mathbb{1}^{(106)}_{t+h} - \beta_{\mathrm{CD},4}\mathbb{1}^{(107)}_{t+h}) = (1 + \Theta_{\mathrm{CD},1}B^{12})\varepsilon^*_{t+h}$, where $y^*_s = y_s$ and $\varepsilon^*_s = e_s$ for $s \le t$.
4. Repeat Steps 1 to 3 for, say, $N = 5000$ times and get N copies of prediction value of y^*_{t+h}. These copies of y^*_{t+h} can be used to form a predictive distribution and prediction intervals for y_{t+h}.

Following the algorithm above, we can now make one-step ahead prediction, i.e., $h = 1$, for our data set, rolling from $t = 131$ to 166 (representing three years) with window width $d = 120$. The blue dotted lines in Fig. 1 show the upper and lower limits of the 95% prediction intervals.

We also plot in Fig. 3 the predictive predictions at, for example, $t = 141, 142, 143$ and 144, respectively, with the black lines indicating the actual values of y_t. The predictive distributions provided in our prediction contain a wealth of information and can facilitate the quantification of uncertainty in prediction. Take $t = 141$ for example, we are able to gain insight into issues such as: (1) What is the prediction interval at 90% confidence level? (The 90% prediction interval is [10.7,11.4].) (2) What confident levels are associated with the statements that the untransformed application volume will be greater than 40,000, 50,000 or 60,000? (The confidence is 98.3%, 84.4% and 54.0% respectively.) (3) What is the lowest predicted application volume of original scale at 90% confidence level? (It is with 90% confidence level that the application volume will exceed 47,332.) These are all important questions concerning government officials in their planning of allocating manpower for handling applications.

7. Further comments

In this paper, we develop a comprehensive statistical inference framework for prediction by: (1) providing (in Definition 1) a formal definition of predictive distribution functions, (2) presenting a general approach based on CDs for constructing such predictive distribution functions, and finally, (3) proposing a Monte Carlo algorithm for implementing the CD-based approach to obtain predictive distribution functions and make inference about the predictions. We also establish the supporting theories for the proposed approach, and discuss its optimality issues as well as its connections to other existing prediction approaches, including Bayesian, fiducial and the frequentist pivotal-based predictive distribution proposed in Lawless and Fredette (2005) and also the CD-based method by Schweder and Hjort (2016). The proposed approach is shown to have several desirable features. Particularly notable is its ability to afford a valid frequentist interpretation and yield prediction intervals of all levels with valid frequentist probability coverage.

This framework is very general and the proposed CD-based formulation is broadly applicable, as the CD concept covers a broad range of examples, including: fiducial distribution, bootstrap distributions, likelihood functions (after normalization), p-value functions, and Bayesian posterior distributions. Regardless of different statistical paradigms, these examples can all be used as CDs as long as they provide valid frequentist probability coverage. This entails that the proposed predictive distribution has the desirable property to be flexible and all encompassing. Case in point is that the Bayesian posterior distribution is often a CD, either asymptotically under the Bernstein–von Mises type theorems or exact using probability matching priors. Noting that $Q(y^*; \mathbf{y}_n)$ has the same form of the Bayesian predictive distribution in (2), the Bayesian predictive distribution can be simply viewed as a special case of our CD-based predictive distribution. Similar arguments apply to the fiducial predictive distributions defined in Wang et al. (2012). All these observations show that the general formulation of $Q(y^*; \mathbf{y}_n)$ through CDs provides an ideal platform to unify most of, if not all, the existing frequentist, fiducial and Bayesian predictive distributions.

There are ample discussions in literature on the great generality and utility of CD as an inference tool. Given that CD has succeeded in providing solutions to problems surrounding difficult complex settings such as making inference from combining heterogeneous studies (e.g., Liu et al., 2015; Claggett et al., 2014; Yang et al., 2014) or studies that fail to produce well-defined point or interval estimates (e.g., Liu et al., 2014), it would seem natural to expect that our proposed CD-based approach can be applied to make inference in predictions for such complex problem settings as well. This should be worth studying further.

Finally, there are also some publications in the literature that treat "predictive distributions" as estimators of $F_\theta(y^*|\mathbf{y}_n)$, the distribution function of Y^* given $\mathbf{Y}_n = \mathbf{y}_n$, see, e.g., Aitchison (1975), Murray (1977), Ng (1980), Lejeune and Faulkenberry (1982), Harris (1989), and Vidoni (1998). But, as pointed out by Lawless and Fredette (2005), although an estimator of $F_\theta(y^*|\mathbf{y}_n)$, say $\tilde{F}(y^*|\mathbf{y}_n)$, provides probability statements about the future random variable Y^*, given $\mathbf{Y}_n = \mathbf{y}_n$, the probability statements for Y^* do not have a frequentist interpretation in terms of repeated sampling. For example, even if $a^* = L(\mathbf{y}_n)$ is chosen so that $\tilde{F}(a^*|\mathbf{y}_n) = 0.95$, it is not true in general that $\mathbb{P}_{\mathbb{J}}\{Y^* < L(\mathbf{Y}_n)\} = 0.95$; see Lawless and Fredette (2005) for further elaboration. Furthermore, there are developments of "predictive likelihood function" (see, e.g., Bjornstad, 1990 and references therein), which rely on a so-called *likelihood principle for prediction* (Berger et al., 1988). The general idea here is to eliminate the "nuisance" parameter θ in the joint likelihood function $L(\theta|y^*, \mathbf{y}_n)$ by using different techniques to obtain a new "likelihood" $L(y^*|\mathbf{y}_n)$ which is free of θ, and then use it to make predictive inference. Depending on the techniques use, different versions of predictive likelihood functions can be obtained, and their performance naturally varies. Some may meet the frequentist probability coverage criterion discussed in this paper, but many may not (cf. Bjornstad, 1990). Finally, even though in some special cases the method of the predictive likelihood function coincides with the predictive distribution function developed in this paper, this method does not stress the need of providing a predictive distribution function that has suitable frequentist probabilistic interpretations.

Acknowledgments

The research is supported in part by the research grants NSF-DMS: 1513483, 1107012, 0915139. The authors thank Fred Roberts, Director of the DHS University Center of Excellence, for Command, Control, and Interoperability (CCICADA), as well as an (anonymous) federal agency for the data and research problem in Section 6. The authors also thank Nils Hjort, Tore Schweder and an anonymous reviewer for their many constrictive comments on this paper.

Appendix

Proof of Theorem 1. By Condition (A) and (7), we have, for any $\epsilon > 0$,

$$\begin{aligned}\left|\int_{\theta\in\mathbf{\Theta}} \left\{F_\theta(Y^*) - F_{\theta_0}(Y^*)\right\} dH(\theta; \mathbf{Y}_n)\right| &\le \int_{\theta\in\mathbf{\Theta}} \left|F_\theta(Y^*) - F_{\theta_0}(Y^*)\right| dH(\theta; \mathbf{Y}_n)\\ &= \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} \left|F_\theta(Y^*) - F_{\theta_0}(Y^*)\right| dH(\theta; \mathbf{Y}_n) + 2H(\theta_0-\epsilon) + 2(1 - H(\theta_0+\epsilon))\\ &\le C\epsilon \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} dH(\theta; \mathbf{Y}_n) + o_p(1) \le C\epsilon + o_p(1).\end{aligned}$$

It follows that

$$\int_{\theta\in\mathbf{\Theta}} \left\{F_\theta(Y^*) - F_{\theta_0}(Y^*)\right\} dH(\theta; \mathbf{Y}_n) = o_p(1).$$

Thus, we have

$$\begin{aligned}Q(Y^*; \mathbf{Y}_n) &= \int_{\theta\in\mathbf{\Theta}} F_\theta(Y^*) dH(\theta; \mathbf{Y}_n) = F_{\theta_0}(Y^*) + \int_{\theta\in\mathbf{\Theta}} \left\{F_\theta(Y^*) - F_{\theta_0}(Y^*)\right\} dH(\theta; \mathbf{Y}_n)\\ &= U + o_p(1). \quad \square\end{aligned}$$

Proof of Theorem 2. First, we note that

$$F_\theta(y^*) = \mathbb{P}_\theta(Y^* \le y^*) = \mathbb{P}_\theta(s_1(Y^*, \theta) \le s_1(y^*, \theta)) = S(s_1(y^*, \theta)).$$

Therefore we have

$$\int_{\theta\in\mathbf{\Theta}} F_\theta(y^*)dH_R(\theta; \hat\theta(\mathbf{y}_n)) = \int_{\theta\in\mathbf{\Theta}} S(s_1(y^*, \theta))dH_R(\theta; \hat\theta(\mathbf{y}_n)). \tag{22}$$

Let (W, V) be a transformation from $(Y^*, \hat\theta(\mathbf{Y}_n))$ such that

$$\begin{cases} W = F_{s_2(\hat\theta(\mathbf{Y}_n),\theta_0)}(s_1(Y^*, \theta_0)) \\ V = s_2(\hat\theta(\mathbf{Y}_n), \theta_0), \end{cases}$$

and let $w = F_{s_2(\hat\theta(\mathbf{y}_n),\theta_0)}(s_1(y^*, \theta_0))$ be a realization of W. By the invariance condition $w = F_{s_2(\hat\theta(\mathbf{y}_n),\theta)}(s_1(y^*, \theta))$. Plugging this into the right hand side of (22) gives

$$\int_{\theta\in\mathbf{\Theta}} F_\theta(y^*)dH_R(\theta; \hat\theta(\mathbf{y}_n)) = \int_{\theta\in\mathbf{\Theta}} S(F^{-1}_{s_2(\hat\theta(\mathbf{Y}_n),\theta)}(w))dH_R(\theta; \hat\theta(\mathbf{y}_n)). \tag{23}$$

On the other hand, the cumulative distribution function of W is

$$\mathbb{P}(W \le w) = \int_v \mathbb{P}(W \le w|V = v)dR(v) = \int_v \mathbb{P}(F_v(s_1(Y^*, \theta_0)) \le w)dR(v) \tag{24}$$

$$= \int_v \mathbb{P}(s_1(Y^*, \theta_0) \le F_v^{-1}(w))dR(v) = \int_v S(F_v^{-1}(w))dR(v) \tag{25}$$

$$= \int_{\theta\in\mathbf{\Theta}} S(F^{-1}_{s_2(\hat\theta(\mathbf{Y}_n),\theta)}(w))dH_R(\theta; \hat\theta(\mathbf{y}_n)). \tag{26}$$

Here, the second equation of (24) is true because given $V = v$, $F_V(s_1(Y^*, \theta_0))$ is independent of V; and (26) is true following the transfer of randomness from v to θ through $v = s_2(\hat\theta(\mathbf{y}_n), \theta)$.

By (23) and (26), we have that $Q_R(y^*; \mathbf{y}_n) = \int_{\theta\in\mathbf{\Theta}} F_\theta(y^*)dH_R(\theta; \hat\theta(\mathbf{y}_n))$ is equivalent to $F_W(w) = \mathbb{P}(W \le w)$, where $F_W(\cdot)$ is the cumulative distribution function of W. Therefore, $Q_R(Y^*; \mathbf{Y}_n) = F_W(W)$ and it is uniformly distributed on $(0, 1)$. □

Proof of Corollary 2. By the invariance condition $F_{\hat\theta}(y^*) = F_{s_2(\hat\theta(\mathbf{y}_n),\theta_0)}(s_1(y^*, \theta_0))$. It immediately follows from the proof of Theorem 2 that $K(F_{\hat\theta(\mathbf{y}_n)}(y^*))$, or $K(F_{s_2(\hat\theta(\mathbf{y}_n),\theta_0)}(s_1(y^*, \theta_0)))$, can be expressed as $\int_{\theta\in\mathbf{\Theta}} F_\theta(y^*)dH_R(\theta; \hat\theta(\mathbf{y}_n))$. □

Proof of Theorem 3. Since $F_\theta^{-1}(u)$ is non-decreasing in θ, $(F_\theta^{-1}(u) - F_{\theta_0}^{-1}(u))^2$ as a function of θ is non-increasing for $\theta \le \theta_0$ and non-decreasing for $\theta \ge \theta_0$. Thus by (14) we have

$$(F^{-1}_{\theta_{\mathrm{CD},1}}(u) - F_{\theta_0}^{-1}(u))^2\mathbb{1}(\theta_{\mathrm{CD},1} \ge \theta_0) \overset{\mathrm{sto}}{\le} (F^{-1}_{\theta_{\mathrm{CD},2}}(u) - F_{\theta_0}^{-1}(u))^2\mathbb{1}(\theta_{\mathrm{CD},2} \ge \theta_0)$$

and

$$(F^{-1}_{\theta_{\mathrm{CD},1}}(u) - F_{\theta_0}^{-1}(u))^2\mathbb{1}(\theta_{\mathrm{CD},1} < \theta_0) \overset{\mathrm{sto}}{\le} (F^{-1}_{\theta_{\mathrm{CD},2}}(u) - F_{\theta_0}^{-1}(u))^2\mathbb{1}(\theta_{\mathrm{CD},2} < \theta_0).$$

The above inequalities lead to

$$\mathbb{E}(F^{-1}_{\theta_{\mathrm{CD},1}}(u) - F_{\theta_0}^{-1}(u))^2 \le \mathbb{E}(F^{-1}_{\theta_{\mathrm{CD},2}}(u) - F_{\theta_0}^{-1}(u))^2,$$

for any $u \in (0, 1)$, which further implies

$$\mathbb{E}(F^{-1}_{\theta_{\mathrm{CD},1}}(U) - F_{\theta_0}^{-1}(U))^2 \le \mathbb{E}(F^{-1}_{\theta_{\mathrm{CD},2}}(U) - F_{\theta_0}^{-1}(U))^2, \tag{27}$$

where $U \sim$ Uniform(0, 1), and thus (15) by the relation between Y_Q^* and Y^*. □

Proof of Inequality (16). Since $F_\theta^{-1}(u)$ is nondecreasing in θ for any given $u \in (0, 1)$, $\{F_\theta^{-1}(u) - F_{\theta_0}^{-1}(u) > \varepsilon\}$ has non-zero probability only if $\theta > \theta_0$, for any $\varepsilon \ge 0$. Therefore from (14) we have $(F^{-1}_{\theta_{\mathrm{CD},1}}(u) - F_{\theta_0}^{-1}(u))^+ \overset{\mathrm{sto}}{\le} (F^{-1}_{\theta_{\mathrm{CD},2}}(u) - F_{\theta_0}^{-1}(u))^+$. Similarly, $(F^{-1}_{\theta_{\mathrm{CD},1}}(u) - F_{\theta_0}^{-1}(u))^- \overset{\mathrm{sto}}{\le} (F^{-1}_{\theta_{\mathrm{CD},2}}(u) - F_{\theta_0}^{-1}(u))^-$. Since the two inequalities are true for any $u \in (0, 1)$, it immediately follows that $(Y^*_{Q_1} - Y^*)^+ \overset{\mathrm{sto}}{\le} (Y^*_{Q_2} - Y^*)^+$ and $(Y^*_{Q_1} - Y^*)^- \overset{\mathrm{sto}}{\le} (Y^*_{Q_2} - Y^*)^-$ by substituting u with $U \sim$ Uniform(0, 1). □

Proof of Theorem 4. The average Kullback–Leibler distance of any density function in the form of $g_{\hat\theta}(\cdot)$ to $f_{\theta_0}(\cdot)$ can be expressed as

$$\bar D_{\mathrm{KL}}(f_{\theta_0}|g_{\hat\theta}) = \mathbb{E}_{\mathbb{J}}\left\{\log\frac{f_{\theta_0}(Y^*)}{g_{\hat\theta}(Y^*)}\right\},$$

where the subscript means the expectation is taken jointly over $\mathcal{Y}^* \times \mathcal{Y}^n$ at the true parameter value θ_0, and thus (17) implies

$$\bar{D}_{\mathrm{KL}}(f_{\theta_0}|q_{\hat{\theta}}) - \bar{D}_{\mathrm{KL}}(f_{\theta_0}|f_{\hat{\theta}}) = \mathbb{E}_{\mathbb{J}}\left\{\log\frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} \le \log \mathbb{E}_{\mathbb{J}}\left\{\frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} \le 0. \quad \square$$

References

Aitchison, J., 1975. Goodness of prediction fit. Biometrika 62, 547–554.
Aitchison, J., Dunsmore, I.R., 1980. Statistical Prediction Analysis. CUP Archive.
Barndor-Nielsen, O.E., Cox, D.R., 1996. Prediction and asymptotics. Bernoulli 2, 319–340.
Beran, R., 1990. Calibrating prediction regions. J. Amer. Statist. Assoc. 85, 715–723.
Berger, J.O., Wolpert, R.L., Bayarri, M.J., DeGroot, M.H., Hill, B.M., Lane, D.A., LeCam, L., 1988. The Likelihood Principle. In: Lecture Notes-Monograph Series, vol. 6, pp. 1–199.
Bjornstad, J.F., 1990. Predictive likelihood: a review (with discussion). Statist. Sci. 5, 242–265.
Chang, K., 2015. Topics in Compositional, Seasonal and Spatial-Temporal Time Series. Rutgers University, (Ph. D. thesis).
Claggett, B., Xie, M., Tian, L., 2014. Meta analysis with fixed, unknown, study-specific parameters. J. Amer. Statist. Assoc. 109, 1667–1671.
Cox, D.R., 1958. Some problems connected with statistical inference. Ann. Math. Statist. 29, 357–372.
Cox, D.R., 1975. Prediction intervals and empirical Bayes condence intervals. In: Gani, J. (Ed.), Perspectives Probability and Statistics. Academic Press, London.
Cox, D.R., 2013. Discussion of "confidence distribution, the frequentist distribution estimator of a parameter: a review". Internat. Statist. Rev. 81, 40–41.
Efron, B., 1998. R. A. Fisher in the 21st century. Statist. Sci. 13, 95–122.
Escobar, L.A., Meeker, W.Q., 1999. Statistical prediction based on censored life data. Technometrics 41, 113–124.
Geisser, S., 1993. Predictive Inference: An Introduction. Chapman & Hall, New York.
Harris, I.R., 1989. Predictive fit for natural exponential families. Biometrika 76, 675–684.
Lawless, F., Fredette, M., 2005. Frequentist prediction intervals and predictive distributions. Biometrika 92, 529–542.
Lejeune, M., Faulkenberry, G.D., 1982. A simple predictive density function. J. Amer. Statist. Assoc. 77, 654–657.
Liu, D., Liu, R., Xie, M., 2014. Exact meta-analysis approach for discrete data and its application to 2 × 2 tables with rare events. J. Amer. Statist. Assoc. 109, 1450–1465.
Liu, D., Liu, R.Y., Xie, M., 2015. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. J. Amer. Statist. Assoc. 110, 326–340.
Liu, R.Y., Parelicus, J., Singh, M., 1999. Multivariate analysis by data depth: de- scriptive statistics, graphics and inference. Ann. Statist. 27, 783–858.
Murray, G.D., 1977. A note on the estimation of probability density functions. Biometrika 64, 150–152.
Ng, V.M., 1980. On the estimation of parametric density functions. Biometrika 67, 505–506.
Schweder, T., 2007. Confidence nets for curves. In: Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum. pp. 593–609.
Schweder, T., Hjort, N.L., 2002. Confidence and likelihood. Scand. J. Stat. 29, 309–332.
Schweder, T., Hjort, N., 2016. Confidence, Likelihood and Probability. Cambridge University Press, Cambridge, U.K..
Singh, K., Xie, M., Strawderman, W.E., 2001. Confidence distributions - concept, theory and applications. Tech. rep., Deptartment of Statistics and Biostatistics, Rutgers University.
Singh, K., Xie, M., Strawderman, W.E., 2005. Combining information from independent sources through confidence distributions. Ann. Statist. 33, 159–183.
Singh, K., Xie, M., Strawderman, W.E., 2007. Confidence distribution (CD) - distribution estimator of a parameter. IMS Lecture-Notes Monogr. Ser. 54, 132–150.
Smith, R.L., 1998. Bayesian and frequentist approaches to parametric predictive inference. In: Bayesian Statistics 6. pp. 589–612.
Vidoni, P., 1998. A note on modified estimative prediction limits and distributions. Biometrika 85, 949–953.
Wang, J.C.M., Hannig, J., Iyer, H.K., 2012. Fiducial prediction intervals. J. Statist. Plann. Inference 142, 1980–1990.
Xie, M., 2013. Rejoinder of "confidence distribution, the frequentist distribution estimator of a parameter: a review". Internat. Statist. Rev. 81, 68–77.
Xie, M., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: a review. Internat. Statist. Rev. 81, 3–39.
Xie, M., Singh, K., Strawderman, W.E., 2011. Confidence distributions and a unifying framework for meta-analysis. J. Amer. Statist. Assoc. 106, 320–333.
Yang, G., Liu, D., Liu, R.Y., Xie, M., Hoaglin, D., 2014. A confidence distribution approach for an efficient network meta-analysis. Stat. Methodol. 20, 105–125.