Methods to Compute Prediction Intervals: A Review and New Results

Qinglong Tian, Daniel J. Nordman, William Q. Meeker

Department of Statistics, Iowa State University, Ames, IA 50011

September 28, 2021

Abstract

The purpose of this paper is to review both classic and modern methods for constructing prediction intervals. We focus, primarily, on model-based non-Bayesian methods for the prediction of a scalar random variable, but we also include Bayesian methods with objective prior distributions. Our review of non-Bayesian methods follows two lines: general methods based on (approximate) pivotal quantities and methods based on non-Bayesian predictive distributions. The connection between these two types of methods is described for distributions in the (log-)location-scale family. We also discuss extending the general prediction methods to data with complicated dependence structures as well as some non-parametric prediction methods (e.g., conformal prediction).

MSC 2020 Subject Classifications: primary 62F25; secondary 62F15, 62F40, 62N01.

Keywords: bootstrap; calibration; coverage probability; prediction interval; predictive distribution

1 Introduction

1.1 Prediction History and Notation

While most statistics textbooks and courses emphasize the explanatory and descriptive roles of statistics, the topic of statistical prediction often receives less attention, despite its practical importance, as noted in recent commentaries (cf. Shmueli 2010 and Harville 2014). The goal of this paper is to review important prediction interval methods and show some interesting connections. We start with some history of statistical prediction.

1.1.1 Bayesian Prediction

Bayesian prediction is accomplished via the Bayesian predictive distribution, which is a conditional distribution of future random variables given the observed data. When the future random variable Y is conditionally independent of the sample $\boldsymbol{X}_n$ given $\boldsymbol{\theta}$, the Bayesian predictive distribution, in the form of probability density function (pdf), is computed as

$$p(y|\boldsymbol{x}_n) = \int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x}_n)d\boldsymbol{\theta}, \tag{1}$$

where $p(y|\boldsymbol{\theta})$ is the pdf of Y conditional on $\boldsymbol{\theta}$ and $p(\boldsymbol{\theta}|\boldsymbol{x}_n)$ is the posterior distribution of $\boldsymbol{\theta}$ given the realized value $\boldsymbol{x}_n$ of the random data $\boldsymbol{X}_n$. Similar to a credible interval, a $100(1-\alpha)\%$ Bayesian prediction interval can be obtained from the $\alpha/2$ and $1-\alpha/2$ quantiles of the Bayesian predictive distribution.

Bayesian prediction can be traced back to Laplace's 1774 Memoir which contains a derivation of the Bayesian predictive distribution for a binomial random variable (cf. Stigler 1986). The early work of de Finetti (e.g., de Finetti 1937) is a cornerstone of Bayesian statistics wherein the importance of the Bayesian predictive distribution under de Finetti's subjective viewpoint of probability is emphasized. Geisser (1993) states that "The completely observabilistic view was brought particularly to the attention of British and American statisticians with the translation into English of the book of probability authored by de Finetti (1974)," where the "completely observabilistic view" refers to the principle of assigning probability only to observable events. Fortini and Petrone (2014) provide a review of the Bayesian predictive distribution, noting that "In the *predictive approach* (also referred to as the *generative model*), the uncertainty is directly described through the predictive distributions."

During the 1960s and 1970s, more work was done to apply Bayesian prediction methods. For example, Guttman and Tiao (1964) use a Bayesian predictive distribution to solve "best population problems," while Thatcher (1964) revisits the binomial prediction problem considered by Laplace and compares Bayesian and non-Bayesian prediction methods. Aitchison and Dunsmore (1975) and Geisser (1993) are books that describe methods for Bayesian prediction.

Although Bayesian statistical methods are not constrained to specific models and distributions, they would wait for the rediscovery of MCMC sampling methods in the late 1980s to take advantage of that generality beyond special distributions that have conjugate prior distributions, such as distributions in the natural exponential family (e.g., the normal, binomial, and Poisson distributions). Also, the large increase in computing power over past decades and computer software like Stan have facilitated the application of Bayesian methods to more complicated prediction problems.

1.1.2 Non-Bayesian Prediction

Similar to the confidence interval for parameters and functions of parameters, prediction intervals provide a natural approach for quantifying prediction uncertainty. Let $\boldsymbol{X}_n$ be the sample and Y be the future random variable. An *exact* $100(1-\alpha)\%$ prediction interval method for Y, denoted by $\text{PI}(\boldsymbol{X}_n)$, satisfies $\Pr_{\boldsymbol{\theta}}[Y \in \text{PI}(\boldsymbol{X}_n)] = 1-\alpha$, where $\boldsymbol{\theta}$ contains the parameters that index the joint distribution of $(\boldsymbol{X}_n, Y)$, and $1-\alpha$ is called the nominal confidence level because not all prediction interval methods are *exact*. If $\Pr_{\boldsymbol{\theta}}[Y \in \text{PI}(\boldsymbol{X}_n)] \geq 1-\alpha$ the procedure is said to be *conservative*. If $\Pr_{\boldsymbol{\theta}}[Y \in \text{PI}(\boldsymbol{X}_n)] \to 1-\alpha$ as $n \to \infty$, we say the method is *asymptotically correct*. It is worth noting that the terminology is different in some literature (especially in nonparametric literature), where the term "exact" is used to indicate being conservative.

In one of the earliest works on non-Bayesian prediction, Fisher (1935) uses a fiducial method to construct a prediction interval for a new observation and a future sample mean from a normal distribution, given a previous sample from the same distribution. Around the same time, Baker (1935) considers predicting a future sample mean and implicitly provides the correct frequentist interpretation for Fisher's interval. In a paper describing sampling sizes for setting tolerance limits, Wilks (1941) also gives Fisher's formula and refers to it as limits that will "...include on the average a proportion a of the universe between them...," which would later be called a β-expectation tolerance interval (equivalent to a prediction interval).

The first use of the term "prediction interval" seems to have come somewhat later. Using a frequentist approach, Proschan (1953) derives the same interval as Fisher and writes "such an interval might be called more appropriately a prediction interval, since the term 'confidence interval' generally refers to population parameters." Thatcher (1964) investigates binomial distribution prediction but used "confidence limit for the prediction" to refer to the prediction interval. As documented in Patel (1989), starting in the late 1960s, numerous papers began appearing in engineering and applied statistics journals presenting methods for many specific prediction problems and using the term "prediction interval."

In the following decades, statisticians began to develop general methods to construct prediction intervals. These include methods based on pivotal quantities, fiducial distributions, and non-Bayesian predictive distributions. Recent developments with these approaches often involve resampling and other simulation-based approximations. These general methods will be described and illustrated in the rest of this paper.

1.2 Overview

In this paper, we focus on constructing prediction intervals for problems where the predictand Y is a scalar and a parametric model (with unknown parameters) is used to describe the distributions of Y and the data $\boldsymbol{X}_n$. We mainly consider the cases where $\boldsymbol{X}_n$ and Y are generated through a random sampling process or a slight variant of it. We describe general non-Bayesian methods for prediction that have been proposed in this setting. We view prediction and prediction interval methods from a frequentist perspective where methods are evaluated, primarily, on the basis of coverage probability, relative to a specified nominal confidence level. Although we assess prediction methods using frequentist criteria, our review includes Bayesian methods with non-informative (or objective) prior distributions. In fact, Bayesian methods with non-informative prior distributions provide an important means of defining prediction methods for both complicated and simple statistical models. Such Bayesian-based methods have been shown to have good frequentist properties (i.e., coverage probability close to nominal; e.g., Hulting and Harville 1991, Harville and Carriquiry 1992, and Section 5 of this paper).

The remainder of this paper is organized as follows. Section 2 describes methods to construct a prediction interval based on pivot-type relations. Section 3 discusses the concept of a (non-Bayesian) predictive distribution as an alternative but equivalent approach to prediction intervals, and Section 4 outlines several methods to construct predictive distributions. Section 5 applies some prediction methods to the (log-)location-scale distribution family and provides new results on connections among various prediction methods. Section 6 describes how to apply general prediction methods to other continuous distributions and provides two illustrative examples. Section 7 describes several general methods that can be applied to construct prediction intervals for discrete distributions. While Sections 2–7 primarily focus on independent data or data with a simple dependence structure, Section 8 discusses extensions involving prediction methods when Y and $\boldsymbol{X}_n$ have a complicated dependence structure. Section 9 discusses several model-free prediction methods. Section 10 provides some concluding remarks.

2 Prediction Interval Methods

2.1 Pivotal Methods

Cox (1975) describes the pivotal prediction method, where the main idea is to find a scalar quantity $q(\boldsymbol{X}_n, Y)$ that does not depend on any parameters. Then the $1-\alpha$ prediction set of Y is given by

$$\{y : q(\boldsymbol{x}_n, y) \leq q_{n,1-\alpha}\}\,, \tag{2}$$

where $q_{n,1-\alpha}$ is the $1-\alpha$ quantile of $q(\boldsymbol{X}_n, Y)$. If $q(\boldsymbol{x}_n, y)$ is a monotone function of y, the prediction region (2) becomes a one-sided prediction bound. When $q(\boldsymbol{X}_n, Y)$ is continuous, the pivotal prediction method is exact because $\Pr_{\boldsymbol{\theta}}\left[q(\boldsymbol{X}_n, Y) \leq q_{n,1-\alpha}\right] = 1-\alpha$ for any $\alpha \in (0,1)$. The rest of this section describes two special types of pivotal methods.

2.1.1 Inverting a Hypothesis Test

Cox (1975) suggests a prediction method based on inverting a hypothesis test. Suppose $\boldsymbol{X}_n \sim f(x; \boldsymbol{\theta})$ and that $Y \sim f(y; \boldsymbol{\theta}^{\dagger})$ is independent of $\boldsymbol{X}_n$. Let w_α be a size α critical region for a similarity test $\boldsymbol{\theta} = \boldsymbol{\theta}^{\dagger}$; that is, $\Pr_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\dagger}}\left[(\boldsymbol{X}_n, Y) \in w_\alpha\right] = \alpha$, where $\Pr_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\dagger}}(\cdot)$ denotes any probability function that belongs to the subset $\{\Pr_{(\boldsymbol{\theta},\boldsymbol{\theta}^{\dagger})} : \boldsymbol{\theta} = \boldsymbol{\theta}^{\dagger}\}$. Then a $1-\alpha$ prediction region is defined as $\{y : (\boldsymbol{x}_n, y) \notin w_\alpha\}$.

2.1.2 Using the Probability Integral Transform

If $T(\boldsymbol{X}_n)$ is a scalar statistic with a continuous cumulative distribution function (cdf) $F(\cdot; \boldsymbol{\theta})$, then $F[T(\boldsymbol{X}_n); \theta]$ has a Uniform$(0,1)$ distribution where $F(\cdot; \theta)$ is the cdf of $T(\boldsymbol{X}_n)$. The pivotal cdf method uses the probability integral transform to compute confidence intervals for a scalar parameter (e.g., Casella and Berger 2002, Chapter 9). When $F(\cdot; \theta)$ is a monotone function of θ, a $1-\alpha$ equal-sided confidence interval is given by $\{\theta : \alpha/2 \leq F[T(\boldsymbol{X}_n; \theta)] \leq 1-\alpha/2\}$.

The pivotal cdf method can be extended to define prediction interval methods. Let $T(\boldsymbol{X}_n)$ be a statistic from data $\boldsymbol{X}_n$ and $R(\boldsymbol{X}_n, Y)$ be a statistic from both the data and the predictand. When the conditional cdf of $T(\boldsymbol{X}_n)$ given $R(\boldsymbol{X}_n, Y)$, say $G_{T|R}\left[t | R(\boldsymbol{x}_n, y)\right]$, does not depend on any parameters and is continuous, then $G_{T|R}\left[T(\boldsymbol{X}_n) | R(\boldsymbol{X}_n, Y)\right]$ has a Uniform$(0,1)$ distribution. If $G_{T|R}\left[t | R(\boldsymbol{x}_n, y)\right]$ is a non-increasing function of y, then $1-\alpha$ lower and upper prediction bounds are defined as

$$\underset{\sim}{Y}_{1-\alpha} = \inf\{y : G_{T|R}\left[T(\boldsymbol{x}_n) | R(\boldsymbol{x}_n, y)\right] < 1-\alpha\}, \quad \widetilde{Y}_{1-\alpha} = \sup\{y : G_{T|R}\left[T(\boldsymbol{x}_n) | R(\boldsymbol{x}_n, y)\right] > \alpha\}. \tag{3}$$

Because $G_{T|R}\left[T(\boldsymbol{X}_n) | R(\boldsymbol{X}_n, Y)\right] \sim$ Uniform$(0,1)$, we have $\Pr(Y \geq \underset{\sim}{Y}_{1-\alpha}) = \Pr(Y \leq \widetilde{Y}_{1-\alpha}) = 1-\alpha$. When $G_{T|R}\left[t | R(\boldsymbol{x}_n; y)\right]$ is non-decreasing in y, then

$$\underset{\sim}{Y}_{1-\alpha} = \inf\{y : G_{T|R}\left[T(\boldsymbol{x}_n) | R(\boldsymbol{x}_n, y)\right] > \alpha\}, \quad \widetilde{Y}_{1-\alpha} = \sup\{y : G_{T|R}\left[T(\boldsymbol{x}_n) | R(\boldsymbol{x}_n, y)\right] < 1-\alpha\}.$$

Similar, but conservative, prediction methods can also be formulated with discrete cdfs. Section 7.1.1 describes such applications to discrete distributions with a scalar parameter.

2.2 Approximate Pivotal Methods

Suppose $q(\boldsymbol{X}_n, Y)$ (with cdf $Q_n(\cdot;\boldsymbol{\theta})$) is a quantity that converges in distribution to a pivotal quantity (with cdf $Q(\cdot)$). In the absence of a pivotal quantity, the approximate pivotal quantity $q(\boldsymbol{X}_n, Y)$ can also be used to construct a prediction interval.

Because $G(Y|\boldsymbol{X}_n;\boldsymbol{\theta})$ is Uniform$(0,1)$ distributed when Y given $\boldsymbol{X}_n$ is continuous with conditional cdf $G(\cdot|\boldsymbol{X}_n;\boldsymbol{\theta})$, an approximate pivotal quantity that is available in most cases is $U_n \equiv G(Y|\boldsymbol{X}_n;\widehat{\boldsymbol{\theta}}_n)$, which converges in distribution to Uniform$(0,1)$ if $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$, usually the maximum likelihood (ML) estimator. Letting $u_{n,1-\alpha}$ be the $1-\alpha$ quantile of U_n, we have $1-\alpha = \Pr_{\boldsymbol{\theta}}(U_n \le u_{n,1-\alpha}) = \Pr_{\boldsymbol{\theta}}[Y \le G^{-1}(u_{n,1-\alpha}|\boldsymbol{X}_n;\widehat{\boldsymbol{\theta}}_n)]$, where $G^{-1}(\cdot|\boldsymbol{X}_n;\boldsymbol{\theta})$ is the quantile function of Y given $\boldsymbol{X}_n$. Because $u_{n,1-\alpha}$ often depends on the unknown parameter $\boldsymbol{\theta}$, an estimate of $u_{n,1-\alpha}$ can be used instead. The rest of this section describes three ways to estimate $u_{n,1-\alpha}$.

2.2.1 The Plug-in Method

The plug-in method, also known as the naive or estimative method, is to use $1-\alpha$ (i.e., the $1-\alpha$ quantile of the Uniform$(0,1)$ distribution) to replace $u_{n,1-\alpha}$. The plug-in $1-\alpha$ upper prediction bound is defined as $\{y : y \le y_{1-\alpha}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{X}_n)\}$, where $y_{1-\alpha}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{X}_n) \equiv \inf\{y : G(y|\boldsymbol{X}_n;\widehat{\boldsymbol{\theta}}_n) \ge 1-\alpha\}$. The $1-\alpha$ lower bound can be defined as $\{y : y \ge y_{\alpha}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{X}_n)\}$. Operationally, the plug-in method replaces the unknown parameters $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}_n$ in the quantile $y_{1-\alpha}(\boldsymbol{\theta}, \boldsymbol{X}_n) \equiv \inf\{y : G(y|\boldsymbol{X}_n;\boldsymbol{\theta}) \ge 1-\alpha\}$. The coverage probability of the plug-in method is typically different from the nominal confidence level because the sampling error in $\widehat{\boldsymbol{\theta}}_n$ is ignored. Under certain regularity conditions, the error of the coverage probability of the plug-in method is of order $O(1/n)$ (cf. Cox 1975, Beran 1990, Hall et al. 1999).

2.2.2 Calibration-Bootstrap Method

To reduce the plug-in coverage error, Beran (1990) proposes the calibration-bootstrap method. Instead of using $1-\alpha$ to estimate $u_{n,1-\alpha}$, a bootstrap re-creation of this quantile is used. The cdf of $U_n = G(Y|\boldsymbol{X}_n;\widehat{\boldsymbol{\theta}}_n)$ is denoted by $H_n(\cdot;\boldsymbol{\theta})$, and $H_n^{-1}(1-\alpha;\widehat{\boldsymbol{\theta}}_n)$ is used to estimate $u_{n,1-\alpha}$, where $H_n^{-1}(\cdot;\boldsymbol{\theta})$ is the quantile function of U_n. Then, the $1-\alpha$ upper prediction bound using the calibration-bootstrap method is given as $y_{H_n^{-1}(1-\alpha;\widehat{\boldsymbol{\theta}}_n)}(\widehat{\boldsymbol{\theta}}_n) = \inf\{y : G(y;\widehat{\boldsymbol{\theta}}_n) \ge H_n^{-1}(1-\alpha;\widehat{\boldsymbol{\theta}}_n)\}$. When a closed-form expression for $H_n(\cdot;\cdot)$ is not available, the bootstrap method can be used to approximate $H_n^{-1}(1-\alpha;\widehat{\boldsymbol{\theta}}_n)$. The bootstrap procedure is as follows,

1. Generate a bootstrap sample $\boldsymbol{x}_n^*$ from the cdf $F(\cdot;\widehat{\boldsymbol{\theta}}_n)$.
2. Compute a bootstrap estimate $\widehat{\boldsymbol{\theta}}_n^*$ of $\boldsymbol{\theta}$ using the bootstrap sample $\boldsymbol{x}_n^*$.
3. Generate y^*, which is the bootstrap version of Y, from the cdf $G(\cdot|\boldsymbol{x}_n^*;\widehat{\boldsymbol{\theta}}_n)$.
4. Compute $u^* = G(y^*|\boldsymbol{x}_n^*;\widehat{\boldsymbol{\theta}}_n^*)$.

5. Repeat the above steps B times to obtain a collection $\{u_1^*, \ldots, u_B^*\}$ and define $\widetilde{u}_{1-\alpha}$ as the $1-\alpha$ sample quantile of these values.
6. The $1-\alpha$ upper calibration prediction bound is $G^{-1}(\widetilde{u}_{1-\alpha}|\boldsymbol{x}_n; \widehat{\boldsymbol{\theta}}_n)$.

Beran (1990) proves that, under regularity conditions, the error of coverage probability of the calibration-bootstrap method is of order $O(1/n^2)$, which is faster than the plug-in method rate $O(1/n)$.

2.2.3 Calibration Using an Asymptotic Expansion

Another method to improve on the plug-in method is to use asymptotic expansion (cf., Cox 1975, Barndorff-Nielsen and Cox 1996, Vidoni 1998). To simplify the presentation, we illustrate this method under the assumption that $\boldsymbol{X}_n$ are independent of Y. Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of $\boldsymbol{\theta}$ that satisfies

$$\mathrm{E}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{\theta} + a(\boldsymbol{\theta})/n + o\left(\frac{1}{n}\right), \qquad \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n) = b(\boldsymbol{\theta})/n + o\left(\frac{1}{n}\right).$$

Because $\boldsymbol{X}_n$ and Y are independent, we denote the $1-\alpha$ quantile of Y by $y_{1-\alpha}(\boldsymbol{\theta}) = \inf\{y : G(y; \boldsymbol{\theta}) \geq 1-\alpha\}$, where $G(\cdot; \boldsymbol{\theta})$ is the cdf of Y; we further define $\kappa_{\alpha,\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n) \equiv \mathrm{Pr}_{\boldsymbol{\theta}}\{Y \leq y_{1-\alpha}[\widehat{\boldsymbol{\theta}}_n(\boldsymbol{X}_n)]\}$. Under smoothness conditions, $\kappa_{\alpha,\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n)$ can be approximated using a Taylor-series expansion around $\boldsymbol{\theta}$ so that, upon taking expectations, the coverage probability of the plug-in prediction bound can be expressed as

$$\begin{aligned}
\mathrm{Pr}_{\boldsymbol{\theta}}\left[Y \leq y_{1-\alpha}(\widehat{\boldsymbol{\theta}}_n)\right] =& \mathrm{E}_{\boldsymbol{\theta}}\left\{\kappa_{\alpha,\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n)\right\} \\
=& \mathrm{E}_{\boldsymbol{\theta}}\left\{\kappa_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta}) + (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\kappa'_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^2 \kappa''_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta})\right\} + o\left(\frac{1}{n}\right) \qquad (4) \\
=& 1 - \alpha + c(\boldsymbol{\theta})/n + o\left(\frac{1}{n}\right),
\end{aligned}$$

for $c(\boldsymbol{\theta}) \equiv a(\boldsymbol{\theta})\kappa'_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta}) + b(\boldsymbol{\theta})\kappa''_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta})/2$ depending on the bias and variance of $\widehat{\boldsymbol{\theta}}_n$ as well as the first derivative $\kappa'_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta})$ and second derivatives $\kappa''_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta})$ of $\kappa_{\alpha,\boldsymbol{\theta}}(\boldsymbol{\theta})$. Letting $\alpha_c = \alpha + c(\boldsymbol{\theta})/n$, then by replacing $1-\alpha$ with $1-\alpha_c$ in (4) we have

$$\begin{aligned}
\mathrm{Pr}_{\boldsymbol{\theta}}\left[Y \leq y_{1-\alpha_c}(\widehat{\boldsymbol{\theta}}_n)\right] =& \mathrm{E}_{\boldsymbol{\theta}}\left\{1 - \alpha_c + (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\kappa'_{\alpha_c}(\boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^2 \kappa''_{\alpha_c}(\boldsymbol{\theta})\right\} + o\left(\frac{1}{n}\right) \\
=& \mathrm{E}_{\boldsymbol{\theta}}\left\{1 - \alpha_c + (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\kappa'_{\alpha}(\boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^2 \kappa''_{\alpha}(\boldsymbol{\theta})\right\} + O\left(\frac{1}{n^2}\right) + o\left(\frac{1}{n}\right) \\
=& 1 - \alpha_c + c(\boldsymbol{\theta})/n + o\left(\frac{1}{n}\right) = 1 - \alpha + o\left(\frac{1}{n}\right),
\end{aligned}$$

by expanding $\kappa'_{\alpha_c}(\boldsymbol{\theta})$ and $\kappa''_{\alpha_c}(\boldsymbol{\theta})$ around α. In other words, the error rate of $1-\alpha$ plug-in prediction bounds can be improved from $O(1/n)$ to $o(1/n)$ by using an adjusted quantile

$y_{1-\alpha_c}(\widehat{\boldsymbol{\theta}}_n)$, rather than the $1-\alpha$ quantile $y_{1-\alpha}(\widehat{\boldsymbol{\theta}}_n)$ directly, from plug-in cdf $G(\cdot;\widehat{\boldsymbol{\theta}}_n)$ of Y. A similar expansion can be obtained by replacing α_c with an estimator $\widehat{\alpha}_c = \alpha + c(\widehat{\boldsymbol{\theta}}_n)/n$. The drawback of this method is usually a closed form for $c(\cdot)$ in (4) is not available.

2.2.4 Additional Comments

In some special cases, U_n does not converge to Uniform$(0,1)$ and has a limiting distribution function that depends on an unknown $\boldsymbol{\theta}$ (examples include Tian et al. 2020). The plug-in method then fails because using a Uniform$(0,1)$ to calibrate the distribution of U_n is no longer valid, even asymptotically. Nevertheless, we can still use the calibration-bootstrap method to construct asymptotically correct prediction intervals based on the non-pivotal quantity U_n. In fact, we can use a broader notion of predictive root $q(\boldsymbol{X}_n, Y)$ (cf. Beran 1990) to include both (approximate) pivotal and non-pivotal cases. Although using a non-pivotal predictive root $q(\boldsymbol{X}_n, Y)$ usually leads to an asymptotically correct prediction interval, it does not have the benefit of being exact.

3 The Predictive Distribution Concept

3.1 Bayesian and Non-Bayesian Predictive Distributions

The concept of a predictive distribution (free of unknown parameters) originated in Bayesian statistics, but efforts have been made to extend the predictive distribution idea to the non-Bayesian world. Non-Bayesian predictive distributions have been implemented using terms including "predictive distribution," "predictive density," "predictive likelihood," and "prediction function." Although terminology varies, they have the same goal to "express the relative credibility of the possible outcomes of a future experiment, in the light of a performed experiment" (Mathiasen 1979).

Although they may have similar forms, the Bayesian predictive distribution, however, is fundamentally different from these non-Bayesian predictive distributions. The Bayesian predictive distribution always represents a type of conditional distribution of Y given $\boldsymbol{X}_n = \boldsymbol{x}_n$, but this is not the case for the non-Bayesian predictive distribution. When discussing non-Bayesian predictive distributions (in the form of a cdf), Lawless and Fredette (2005) say "In a loose sense these are conditional distributions $\tilde{F}_p(y|x)$ that provide probability statement about the future random variable Y, given $X = x$. However, their properties are generally considered by treating them as estimators of the distribution function of Y given $X = x$," where "these" means non-Bayesian predictive distributions in the form of a cdf.

Because our focus is on non-Bayesian methods and for the sake of simplicity, *we refer to the non-Bayesian predictive distributions as "predictive distributions"* in the rest of the paper. When discussing the Bayesian method, we still use the term "Bayesian predictive distribution." The predictive distribution comes in three forms: the predictive cdf, pdf, and likelihood. These three forms are closely related. One can obtain a predictive cdf by integrating the corresponding predictive pdf and obtain a predictive pdf by normalizing the corresponding predictive likelihood. We use $F_p(y;\boldsymbol{x}_n)$, $f_p(y;\boldsymbol{x}_n)$, and $L_p(y;\boldsymbol{x}_n)$, respectively, to denote the predictive cdf, pdf, and likelihood. The notational purpose of using a semicolon instead of a vertical bar is because the predictive distribution is not the conditional distribution of Y given $\boldsymbol{X}_n$.

3.2 Prediction Interval Methods and Predictive Distributions

For any prediction interval method, we can construct a predictive distribution for Y by treating the endpoint of the $1-\alpha$ upper prediction bound as the $1-\alpha$ quantile of such distribution. Specifically, there is a corresponding predictive distribution for any given prediction interval method.

For the pivotal and approximate pivotal calibration methods described in Section 2, Lawless and Fredette (2005) give the formulas for finding the associated predictive cdf. If a pivotal quantity $q(\boldsymbol{X}_n, Y)$ exists and has cdf $Q_n(\cdot)$, with the further assumption that $q(\boldsymbol{x}_n, y)$ is a monotone function of y, then the corresponding predictive cdf for the pivotal method based on $q(\boldsymbol{x}_n, y)$ is given by $F_p(y;\boldsymbol{x}_n) = Q_n[q(\boldsymbol{x}_n, y)]$. Similarly, if $q(\boldsymbol{X}_n, Y)$ is an approximate pivotal quantity with cdf $Q_n(\cdot;\boldsymbol{\theta})$, the corresponding predictive cdf is $F_p(y;\boldsymbol{x}_n) = \widetilde{Q}_n[q(\boldsymbol{x}_n, y)]$, where $\widetilde{Q}_n(\cdot)$ is an estimate of $Q_n(\cdot;\boldsymbol{\theta})$ (e.g., $\widetilde{Q}_n(\cdot) = \lim_{n\to\infty} Q_n(\cdot;\boldsymbol{\theta})$).

For the calibration-bootstrap method, $H_n(\cdot;\widehat{\boldsymbol{\theta}}_n)$ is used to approximate $H_n(\cdot;\boldsymbol{\theta})$, where the latter is the cdf of $U_n = G(Y|\boldsymbol{X}_n;\widehat{\boldsymbol{\theta}}_n)$; thus the associated predictive cdf is

$$F_p(y;\boldsymbol{x}_n) = H[G(y|\boldsymbol{x}_n;\widehat{\boldsymbol{\theta}}_n);\widehat{\boldsymbol{\theta}}_n]. \tag{5}$$

When an explicit form of $H(\cdot;\cdot)$ is not available, Fonseca et al. (2012) propose a formula to compute (5) using bootstrap

$$\begin{aligned} F_p(y;\boldsymbol{x}_n) &= \mathrm{E}_{\widehat{\boldsymbol{\theta}}_n}\left(G\left\{G^{-1}\left[G(y|\boldsymbol{x}_n;\widehat{\boldsymbol{\theta}}_n)|X_n^*,\widehat{\boldsymbol{\theta}}_n^*\right]|X_n^*,\widehat{\boldsymbol{\theta}}_n\right\}\right) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} G\left\{G^{-1}\left[G(y|\boldsymbol{x}_n;\widehat{\boldsymbol{\theta}}_n)|\boldsymbol{x}_{n,b}^*,\widehat{\boldsymbol{\theta}}_{n,b}^*\right]|\boldsymbol{x}_{n,b}^*,\widehat{\boldsymbol{\theta}}_n\right\}, \end{aligned} \tag{6}$$

where $\mathrm{E}_{\widehat{\boldsymbol{\theta}}_n}$ is the expectation with respect to the bootstrap sample $\boldsymbol{X}_n^*$ and the corresponding bootstrap estimate $\widehat{\boldsymbol{\theta}}_n^*$; the second expression in (6) represents a Monte Carlo approximation based on the bootstrap estimates $\widehat{\boldsymbol{\theta}}_{n,b}^*$ from independently generated bootstrap samples $b = 1,\ldots,B$ for some B. However, for values of y where $G(y|\boldsymbol{x}_n;\widehat{\boldsymbol{\theta}}_n)$ is close to one, the approximation formula in (6) will fail due to limited precision of floating-point computations.

4 Predictive Distribution Methods

4.1 An Overview

In Section 3.2, we describe finding the associated predictive distribution for prediction interval methods. Conversely, given a predictive distribution, we can obtain the corresponding prediction intervals using the quantiles of that predictive distribution. So, the development of a predictive distribution can be useful for formulating a prediction method.

Bjørnstad (1990) summarizes three types of predictive likelihood methods (equivalently, predictive distribution methods): maximization-based, conditioning-based, and integration-based. In addition to the methods discussed in Bjørnstad (1990), Barndorff-Nielsen and Cox (1996) propose a predictive density that generally yields prediction intervals that have a coverage probability that is close to the nominal confidence level. Komaki (1996) considers constructing predictive distributions from the viewpoint of optimizing the Kullback-Leibler divergence between the true distribution of Y and the predictive distribution of Y. But this idea of constructing non-Bayesian predictive distribution is not without difficulty as Hall et al. (1999) point out that many of the predictive distribution methods "do not reduce coverage error by an order of magnitude, relative to the 'naive' or 'estimative' approach to prediction." They further use bootstrap calibration to improve the coverage. In our review, we focus on the integration-based methods, where more research has been done since the review paper by Bjørnstad (1990).

4.2 Integration-Based Predictive Distributions

The construction of an integration-based predictive distribution is similar to that of the Bayesian predictive distribution in (1). The idea is to assign a data-based distribution to the non-random parameter $\boldsymbol{\theta}$ and use this distribution to marginalize out the parameters in the distribution function $G(y|\boldsymbol{x}_n;\boldsymbol{\theta})$ of Y. The resulting predictive cdf has the form

$$F_p(y;\boldsymbol{x}_n) = \int G(y|\boldsymbol{x}_n;\boldsymbol{\theta})p(\boldsymbol{\theta};\boldsymbol{x}_n)d\boldsymbol{\theta}, \tag{7}$$

where $p(\boldsymbol{\theta};\boldsymbol{x}_n)$ is a data-based pdf assigned to $\boldsymbol{\theta}$. More generally, we do not strictly require a pdf $p(\boldsymbol{\theta};\boldsymbol{x}_n)$ for purposes of defining an integral of $G(y|\boldsymbol{x}_n;\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ in (7). Technically, any data-based distribution over the parameter space can be used to integrate $G(y|\boldsymbol{x}_n;\boldsymbol{\theta})$ (although not all of them have a practical meaning) and, in practice, $F_p(y;\boldsymbol{x}_n)$ is often evaluated through a Monte Carlo approximation as

$$F_p(y; \boldsymbol{x}_n) \approx \frac{1}{B} \sum_{b=1}^{B} G(y|\boldsymbol{x}_n; \boldsymbol{\theta}^{(b)})$$

using a set of independent draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(B)}$ from the chosen distribution over the parameter space that is determined from the data $\boldsymbol{x}_n$. In the rest of this section, we discuss three types of integration-based predictive distributions.

4.2.1 Using a Bootstrap Distribution

Harris (1989) proposes a bootstrap predictive distribution obtained by integrating (7) using a bootstrap distribution of $\widehat{\boldsymbol{\theta}}_n$ in the role of $p(\boldsymbol{\theta}; \boldsymbol{x}_n)$ and shows that the proposed predictive distribution is asymptotically superior to the plug-in method in terms of average Kullback-Leibler divergence for the natural exponential family. Although bootstrap samples are used, the coverage probability of this method can be shown, under assumptions similar to the plug-in method, to exhibit a coverage error of order $O(1/n)$, which is the same error rate as the plug-in method in (4); formal details are given in Section A of the supplementary material. We call this method the "direct-bootstrap" method because the bootstrap draws are used directly to compute the predictive distribution. In Section 5.1, we introduce the generalized pivotal quantity (GPQ) bootstrap method, a variant of this method.

4.2.2 Using a Fiducial Distribution

Fiducial inference was first introduced by R. A. Fisher and applies concepts of transferring randomness from the data to the parameters to produce a fiducial distribution on the parameter space. The resulting fiducial distribution is similar to a Bayesian posterior but does not require a prior distribution. We use an illustrative example to demonstrate the fiducial idea. Suppose $X \sim \text{Norm}(\mu, 1)$, then a structural equation for linking the data to the parameter is given by $X = \mu + Z$ where $Z \sim \text{Norm}(0, 1)$. For a realized $X = x$, this equation is solved for μ as $\mu = x - Z$ and thus the fiducial distribution for μ is $\text{Norm}(x, 1)$. The construction of a fiducial distribution may not be unique. More details about fiducial and generalized fiducial inference can be found in Hannig et al. (2016). When the data $\boldsymbol{X}_n$ and Y are independent (i.e., $G(y|\boldsymbol{x}_n; \boldsymbol{\theta}) = G(y; \boldsymbol{\theta})$), the fiducial predictive cdf has the same form as (7), where $p(\boldsymbol{\theta}; \boldsymbol{x}_n)$ is the fiducial distribution of $\boldsymbol{\theta}$. A detailed discussion of the fiducial prediction, including the case that $\boldsymbol{X}_n$ and Y are dependent, can be found in Wang et al. (2012).

4.2.3 Using a Confidence Distribution

Shen et al. (2018) propose a prediction framework based on the notion of confidence distribution (CD) and prove that the corresponding prediction interval is asymptotically correct for a scalar parameter. The idea is to replace $p(\boldsymbol{\theta}; \boldsymbol{x}_n)$ in (7) with a real-valued confidence distribution. But as stated in Xie and Singh (2013), the definition of confidence distribution for a parameter vector with more than one element remains an open question, and the theoretical properties of CD-based predictive distributions in this more general setting require further development.

5 New Results for Location-Scale Distributions

This section presents some particular results for predicting an independent future random variable from a (log-)location-scale distribution given data from the same distribution. These families of distribution include the most widely used probability distributions, such as the normal, lognormal, logistic, loglogistic, Weibull, Fréchet, and some extreme value distributions. Consider a sample $\boldsymbol{X}_n$ consisting of n iid observations from a member of the location-scale distribution family with cdf $F(x;\mu,\sigma) = \Phi\left[(x-\mu)/\sigma\right]$ depending on parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ and where $\Phi(\cdot)$ is a given continuous cdf with no unknown parameters. The corresponding pdf is then $f(x;\mu,\sigma) = \sigma^{-1}\phi[(y-\mu)/\sigma]$, where $\phi(z) = d\Phi(z)/dz$. The predictand Y is an independent random variable from the same distribution. Suppose that the data $\boldsymbol{X}_n$ can be observed under three different situations: complete $\boldsymbol{X}_n$, time (Type-I) censored $\boldsymbol{X}_n^{\mathrm{I}}$, or failure (Type-II) censored $\boldsymbol{X}_n^{\mathrm{II}}$. For time-to-event data, Type-I censoring means that observation stops at a fixed censoring time, while Type-II censoring means that observation stops once a predetermined number of events have occurred.

5.1 The Calibration-Bootstrap Method and its Predictive Distribution

This section shows that (i) the calibration-bootstrap method (cf. Section 2.2.2) is equivalent to a predictive distribution based on integrating out the parameters with the distribution of the GPQ and (ii) the calibration-bootstrap method is also shown to be equivalent to a pivotal method (cf. Section 2.1) for complete or Type-II censored data, thus having exact coverage probability.

By applying (6) to the location-scale distribution, the predictive cdf of the calibration-bootstrap method is

$$F_p(y;\boldsymbol{x}_n) = \mathrm{E}_{\widehat{\boldsymbol{\theta}}_n}\left(F\{F^{-1}\left[F(z;\widehat{\mu},\widehat{\sigma});\widehat{\mu}^*,\widehat{\sigma}^*\right];\widehat{\mu},\widehat{\sigma}\}\right) = \mathrm{E}_{\widehat{\boldsymbol{\theta}}_n}\Phi\left\{\frac{y-[\widehat{\mu}+\frac{\widehat{\sigma}}{\widehat{\sigma}^*}(\widehat{\mu}-\widehat{\mu}^*)]}{\widehat{\sigma}\frac{\widehat{\sigma}}{\widehat{\sigma}^*}}\right\}; \qquad (8)$$

here $(\widehat{\mu},\ \widehat{\sigma})$ are the ML estimators of (μ,σ) and $\mathrm{E}_{\widehat{\boldsymbol{\theta}}_n}$ denotes expectation with respect to the bootstrap distribution of $(\widehat{\mu}^*,\widehat{\sigma}^*)$, which is a version of $(\widehat{\mu},\widehat{\sigma})$ found from (parametric) bootstrap samples.

We define two new quantities $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ using $(\widehat{\mu}, \widehat{\sigma})$ and $(\widehat{\mu}^{*}, \widehat{\sigma}^{*})$ as,

$$\widehat{\mu}^{**} \equiv \widehat{\mu} + \widehat{\sigma}\frac{\widehat{\mu} - \widehat{\mu}^{*}}{\widehat{\sigma}^{*}}, \quad \widehat{\sigma}^{**} \equiv \widehat{\sigma}\frac{\widehat{\sigma}}{\widehat{\sigma}^{*}}. \tag{9}$$

Then (8) can be written in the form of (7), where the parameters (μ, σ) in the cdf $F(y; \mu, \sigma) = \Phi\left[(y - \mu)/\sigma\right]$ of the predictand Y are integrated out with respect to the joint distribution of $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ as

$$F_p(y; \boldsymbol{x}_n) = \mathrm{E}_{\widehat{\boldsymbol{\theta}}_n} \Phi\left(\frac{y - \widehat{\mu}^{**}}{\widehat{\sigma}^{**}}\right) = \int \Phi\left(\frac{y - \widehat{\mu}^{**}}{\widehat{\sigma}^{**}}\right) \mathrm{Pr}_{\widehat{\boldsymbol{\theta}}_n}(d\widehat{\mu}^{**}, d\widehat{\sigma}^{**}) \approx \frac{1}{B}\sum_{b=1}^{B} \Phi\left(\frac{y - \widehat{\mu}_b^{**}}{\widehat{\sigma}_b^{**}}\right), \tag{10}$$

where $(\widehat{\mu}_b^{**}, \widehat{\sigma}_b^{**})$ are realized values of $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ over independently generated bootstrap samples $b = 1, \ldots, B$. This equivalence shows that the calibration-bootstrap method coincides with a predictive distribution constructed via an integration method.

Next, we introduce the definition of GPQ and illustrate the connection between the calibration-bootstrap method and GPQs. Here we use the definition given in Hannig et al. (2006). Let $\mathbb{S} \in \mathbb{R}^k$ denote a random vector and $\mathbb{S}^*$ is an independent copy of $\mathbb{S}$. The distribution of $\mathbb{S}$ is indexed by $\boldsymbol{\theta}$. Suppose we would like to estimate a function of $\boldsymbol{\theta}$ (possibly a vector) $\xi \equiv \pi(\boldsymbol{\theta})$. A GPQ for ξ, denoted by $\mathcal{R}_\xi$, is a function $(\mathbb{S}, \mathbb{S}^*, \boldsymbol{\theta})$ with the following properties

1. The distribution of $\mathcal{R}_\xi$, conditional on $\mathbb{S} = \boldsymbol{s}$, is free of ξ.
2. For every allowable $\boldsymbol{s} \in \mathbb{R}^k$, $\mathcal{R}_\xi$ depends on $\boldsymbol{\theta}$ only through ξ.

We can use GPQs, for example, to construct confidence intervals for parameters of interest.

Interestingly, for complete data $\boldsymbol{X}_n$ or Type-II censored data $\boldsymbol{X}_n^{\mathrm{II}}$, the pair $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ defined in (9) has the same distribution as the GPQ (μ^{**}, σ^{**}), defined as

$$\mu^{**} = \widehat{\mu} + \left(\frac{\mu - \widehat{\mu}^{\mathbb{S}}}{\widehat{\sigma}^{\mathbb{S}}}\right)\widehat{\sigma}, \quad \sigma^{**} = \left(\frac{\sigma}{\widehat{\sigma}^{\mathbb{S}}}\right)\widehat{\sigma}, \tag{11}$$

where $\mathbb{S}$ denotes an independent copy of the sample $\boldsymbol{X}_n$ (or $\boldsymbol{X}_n^{\mathrm{II}}$), and $(\widehat{\mu}, \widehat{\sigma})$ and $(\widehat{\mu}^{\mathbb{S}}, \widehat{\sigma}^{\mathbb{S}})$ denote the ML estimators of (μ, σ) computed from $\boldsymbol{X}_n$ (or $\boldsymbol{X}_n^{\mathrm{II}}$) and $\mathbb{S}$, respectively. The pair (μ^{**}, σ^{**}) is called the GPQ of (μ, σ) for location-scale distribution (cf. Krishnamoorthy and Mathew 2009, Page 17). Because (11) are also fiducial quantities, (8) is a fiducial predictive cdf (details are given in Section B of the supplementary material).

Because the pair $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ is available for any (log-)location-scale distribution and because this pair is operationally computed from the bootstrap samples $(\widehat{\mu}^{*}, \widehat{\sigma}^{*})$ (i.e., compare (9) to (11)), the prediction method in (10) is called the "GPQ-bootstrap" method in contrast to the "direct-bootstrap" method where the bootstrap pair $(\widehat{\mu}^{*}, \widehat{\sigma}^{*})$ is used directly. Note that under Type-I or random censoring, $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ are no longer GPQs; however, we can still use the prediction method in (10) the resulting prediction intervals are still asymptotically correct.

Note also that the calibration-bootstrap samples are used to approximate the quantity

$$U = G(Y; \widehat{\mu}, \widehat{\sigma}) = \Phi\left(\frac{Y - \widehat{\mu}}{\widehat{\sigma}}\right) = \Phi\left[\frac{(Y - \mu)/\sigma - (\widehat{\mu} - \mu)/\sigma}{\widehat{\sigma}/\sigma}\right],$$

which is a pivotal quantity under complete or Type-II censored data and its bootstrap re-creation also has the same distribution for such data. This implies that, for complete or Type-II censored data, the calibration-bootstrap method has exact coverage probability and so does the GPQ-bootstrap method (i.e., due to producing the same prediction intervals from matching predictive distributions in (8) and (10)). We provide illustrative numerical examples in Sections E.1–4 of the supplementary materials.

5.2 Properties of the Bayesian Predictive Distribution

For location-scale distributions and complete or Type-II censored data, the exact probability matching prior is $\pi(\mu, \sigma) = \sigma^{-1}$ (and this is also known as the modified Jeffreys prior), which implies that using this prior leads to credible intervals that have exact frequentist coverage for either μ or σ (cf. Peers (1965), Lawless (1972), DiCiccio et al. (2017)) and certain functions of these parameters (e.g., quantiles and tail probabilities). The purpose of this section is to show that (i) the prediction interval procedure based on the Bayesian predictive distribution using the prior $\pi(\mu, \sigma) = \sigma^{-1}$ is exact and (ii) the Bayesian predictive distribution using the prior $\pi(\mu, \sigma) = \sigma^{-1}$ is equivalent to a predictive distribution based on the generalized fiducial distribution (GFD) derived from the user-friendly formula in Section 2 of Hannig et al. (2016). The latter GFD for (μ, σ) has a density that is proportional to

$$r(\mu, \sigma; \boldsymbol{x}_n) \propto \frac{J(\boldsymbol{x}_n; \mu, \sigma)}{\sigma^n} \prod_{i=1}^{n} \phi\left(\frac{x_i - \mu}{\sigma}\right),$$

by Theorem 1 of Hannig et al. (2016), where the function $J(\boldsymbol{x}_n; \mu, \sigma)$ is

$$J(\boldsymbol{x}_n; \mu, \sigma) = \sum_{1 \le i < j \le n} \left| \det\left(\begin{bmatrix} 1 & 1 \\ \frac{x_i - \mu}{\sigma} & \frac{x_j - \mu}{\sigma} \end{bmatrix}\right) \right| = \frac{1}{\sigma} \sum_{1 \le i < j \le n} |x_i - x_j|.$$

Theorem 1. *Under a complete sample $\boldsymbol{X}_n$ or a Type-II censored sample $\boldsymbol{X}_n^{II}$ from a location-scale distribution with location parameter μ and scale parameter σ, suppose that the ML estimators are $\widehat{\mu}$ and $\widehat{\sigma}$, Y is an independent random variable from the same distribution as*

$\boldsymbol{X}_n$, and that the quantity (U_1, U_2) is defined as $((\widehat{\mu}-\mu)/\widehat{\sigma}, \widehat{\sigma}/\sigma)$. Then,

1. *The joint posterior distribution of (U_1, U_2) using prior $\pi(\mu,\sigma) \propto \sigma^{-1}$ is the same as the frequentist conditional distribution of (U_1, U_2) conditioned on ancillary statistic $A = ((X_1-\widehat{\mu})/\widehat{\sigma}, \ldots, (X_{n-2}-\widehat{\mu})/\widehat{\sigma})$.*
2. *The $1-\alpha$ Bayesian upper prediction bound, which is defined as*

$$\widetilde{Y}_{1-\alpha}^{Bayes} \equiv \inf\left\{ y : \int_{(\mu,\sigma)\in\boldsymbol{\Theta}} F(y|\mu,\sigma)p(\mu,\sigma|\boldsymbol{X}_n)d\mu d\sigma \geq 1-\alpha \right\}, \tag{12}$$

   has exact coverage probability, i.e., $\Pr\left(Y \leq \widetilde{Y}_{1-\alpha}^{Bayes}\right) = 1-\alpha$, where $p(\mu,\sigma|\boldsymbol{X}_n = \boldsymbol{x}_n)$ is the joint posterior distribution using prior $\pi(\mu,\sigma) = \sigma^{-1}$.

3. *The GFD for (μ,σ) is the same as the Bayesian posterior distribution for (μ,σ) using the prior $\pi(\mu,\sigma)=\sigma^{-1}$, and application of this GFD in (7) produces a predictive distribution $\int_{(\mu,\sigma)} F(y;\mu,\sigma)p(\mu,\sigma|\boldsymbol{X}_n)d\mu d\sigma$ and bounds $\widetilde{Y}_{1-\alpha}^{Bayes}$ that match the corresponding Bayesian analogs in (12).*

The proof of Theorem 1 is provided in Section C of the supplementary material. Although the term "fiducial" is used both here and in Section 5.1 (in the context of $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ there), the GFD for (μ,σ) in Point 3 of Theorem 1 is generally different from (but close to) the distribution of the GPQ pair $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ from (9). This is because the GPQs $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ are based on the unconditional distribution of $(U_1, U_2) = ((\widehat{\mu}-\mu)/\widehat{\sigma}, \widehat{\sigma}/\sigma)$ while GFD is determined by the conditional distribution of (U_1, U_2) given the ancillary statistics $\boldsymbol{A} = (A_1, \ldots, A_{n-2})$ (or $(A_1, \ldots, A_{r-2})$ for Type-II censoring); the latter follows from Points 1 and 3 of Theorem 1. The one exception is for the normal distribution, where the distribution of the GPQ pair $(\widehat{\mu}^{**}, \widehat{\sigma}^{**})$ will match the GFD for (μ,σ) (for which Basu's theorem gives that (U_1, U_2) is independent of $\boldsymbol{A}$).

6 Other Continuous Distributions

This section describes and illustrates prediction methods for two continuous distributions that are not in the (log-)location-scale family.

6.1 The Gamma Distribution

The data $\boldsymbol{X}_n$ and the predictand Y are independent samples from a gamma distribution with pdf $f(x;\alpha,\lambda) = \lambda^{\alpha}x^{\alpha-1}\exp(-\lambda x)/\Gamma(\alpha)$. A small-scale simulation study was done to compare: (i) the plug-in method; (ii) the calibration-bootstrap method; (iii) the direct-bootstrap method; and (iv) the fiducial predictive distribution (cf. Section 4.2). Because the gamma distribution does not belong to the (log-)location-scale family, the GPQ-bootstrap is not applicable. To implement methods (i), (ii), and (iii), the ML estimates were computed using the **egamma** function in R package **EnvStats**. For method (iv), two ways of constructing the fiducial distribution were used.

The first is an approximate method proposed by Chen and Ye (2017). From a gamma sample $\boldsymbol{X}_n$, define a scaled chi-square random variable $W(\alpha)$ as,

$$W(\alpha) \equiv 2n\alpha \log\left(\frac{\bar{X}_n}{\prod_{i=1}^{n} X_i^{1/n}}\right) \sim c\chi^2_v,$$

where c and v can be calculated as $v = 2\mathrm{E}^2(W(\alpha))/\mathrm{Var}(W(\alpha))$ and $c = \mathrm{E}(W(\alpha))/v$. Here the expectation and variance of $W(\alpha)$ are

$$\mathrm{E}_{\boldsymbol{\theta}}(W(\alpha)) = 2n\alpha\mathrm{E}(S_1), \quad \mathrm{Var}_{\boldsymbol{\theta}}(W(\alpha)) = 4n^2\alpha^2\mathrm{Var}_{\boldsymbol{\theta}}(S_1),$$

where $\mathrm{E}(S_1) = -\log n + \psi(\alpha n) - \psi(\alpha)$, $\mathrm{Var}(S_1) = -\psi_1(\alpha n) + \psi_1(\alpha)/n$, $\psi(\cdot)$ is the digamma function, and $\psi_1(\cdot)$ is the trigamma function. Chen and Ye (2017) suggested using a consistent estimator $\widehat{\alpha}$ of α to compute $\widehat{c}$ and $\widehat{v}$. Then the approximate marginal fiducial distribution of α is defined by the distribution of a quantity α_b where

$$\alpha_b \sim \frac{\widehat{c}\chi^2_{\widehat{v}}}{2n\log\left(\frac{\bar{x}_n}{\prod_{i=1}^{n} x_i^{1/n}}\right)}. \tag{13}$$

Given a fiducial draw α_b sampled as (13), the fiducial draw for λ, denoted by λ_b, can be sampled as $\chi^2_{2n\alpha_b}/(2\sum_{i=1}^{n} x_i)$ using the fact that $2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2n\alpha}$. Then the fiducial pairs (α_b, λ_b), $b = 1, \ldots, B$ can be used to compute the fiducial predictive distribution for Y via (7).

The second approach is from Wang et al. (2012) based on the user-friendly formula in Hannig et al. (2016). The fiducial distribution of (α, λ) is given by a density proportional to

$$\begin{aligned} r(\alpha, \lambda; \boldsymbol{x}_n) \propto\ & \frac{\lambda^{n\alpha-1}\exp[-\lambda\sum_{i=1}^{n} x_i + (\alpha-1)\sum_{i=1}^{n}\log x_i]}{\alpha\Gamma(\alpha)^n} \times \\ & \sum_{1\le i<j\le n} x_i x_j \left| \frac{\Gamma(\alpha+1)\frac{\partial}{\partial\alpha}V_{\alpha,1}(\lambda x_i)}{(\lambda x_i)^\alpha \exp(-\lambda x_i)} - \frac{\Gamma(\alpha+1)\frac{\partial}{\partial\alpha}V_{\alpha,1}(\lambda x_j)}{(\lambda x_j)^\alpha \exp(-\lambda x_j)} \right|, \end{aligned} \tag{14}$$

where $V_{\alpha,1}(\cdot)$ is the cdf of a Gamma$(\alpha, 1)$. Wang et al. (2012) used an importance sampling algorithm to generate fiducial draws of (α, λ) from (14).

We used simulation to compare these methods mentioned above, and results are given in

Section E.5 of the supplementary material. The calibration-bootstrap method has the best performance and the estimated coverage probability is close to the nominal confidence level, even when n is small. Two fiducial methods also have good coverage probabilities but not as good as the calibration-bootstrap method when n is small. The direct-bootstrap method has poor coverage probability and does not improve on the plug-in method. As described in Section 4.2.1, it can be shown that prediction bounds from direct bootstrap often share a close correspondence to plug-in prediction bounds. General theory, along with numerical illustrations for the gamma case, appear in Section A of the supplementary materials. Consequently, the plug-in and direct-bootstrap methods perform very similarly to each other, but not as well as the calibration-bootstrap approach.

6.2 The Inverse Gaussian Distribution

The sample $\boldsymbol{X}_n$ and predictand Y are independent samples from an inverse Gaussian distribution with pdf

$$f(x;\mu,\lambda)=\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right].$$

As in Section 6.1, a small scale simulation study was done to compare several methods: (i) plug-in; (ii) calibration-bootstrap; (iii) direct-bootstrap; (iv) fiducial predictive distribution methods.

For methods (i),(ii), and (iii), the ML estimators are $\widehat{\mu}=\bar{X}_n$ and $\widehat{\lambda}=n/\sum_{i=1}^{n}(X_i^{-1}-\bar{X}_n^{-1})$. For method (iv), Nájera and O'Reilly (2017) proposed a method to sample from the fiducial distribution for (μ,λ). Because $\sum_{i=1}^{n}(X_i^{-1}-\bar{X}_n^{-1})\sim\chi^2_{n-1}/\lambda$, the marginal fiducial distribution of λ is given as

$$\lambda\sim\frac{\chi^2_{n-1}}{\sum_{i=1}^{n}(x_i^{-1}-\bar{x}_n^{-1})}.$$

Then given λ_b, which is sampled as above, a fiducial draw μ_b for μ can be obtained using the following steps:

1. Generate u_b from Uniform$(0,1)$.
2. Compute the quantile $q_{+\infty,\lambda_b}(u_b)\equiv$ **qinvgauss**$(u_b,\mu=+\infty,\lambda=\lambda_b)$.
3. If $\bar{x}_n/\lambda_b\geq q_{+\infty,\lambda_b}(u_b)$, $\mu_b=+\infty$. If $\bar{x}_n/\lambda_b<q_{+\infty,\lambda_b}(u_b)$, μ_b is obtained by solving the equation **pinvgauss**$(u_b,\mu_b,n)=\bar{x}_n/\lambda_b$.

The inverse Gaussian quantile function **qinvgauss** and cdf function **pinvgauss** are available in R package **statmod**.

The inverse Gaussian simulation (given in Section E.6 of the supplementary material) gives results that are similar to those for the gamma simulation. Use of the direct-bootstrap method does not improve on the plug-in method in terms of coverage probability. The fiducial method has good coverage probability but, as shown in the examples, sampling from a given fiducial distribution is often non-trivial. Both theoretical results (cf. Section 2.2.2) and simulations have shown that the calibration-bootstrap method has the best coverage for several continuous distributions and it is also easy to implement.

7 Prediction Methods for Discrete Distributions

Previous sections in this paper considered prediction from a continuous distribution, and those methods can be applied to a wide variety of continuous distributions (e.g., see Section E in the supplementary materials). This section focuses on prediction methods for discrete distributions. Following the two prediction principles discussed in the previous sections (i.e., (approximate) pivotal methods and (Bayesian and non-Bayesian) predictive distribution methods), this section first discusses some general methods and then implements these methods for the binomial and Poisson distributions. Additionally, we give some cautionary remarks on using the plug-in method for discrete distributions.

7.1 Some General Methods

7.1.1 The Pivotal Conditional Cdf Method

Let $\boldsymbol{X}_n$ be the data, Y be the predictand, and $T(\boldsymbol{X}_n)$ be a statistic whose conditional distribution given a function of $\boldsymbol{X}_n$ and Y, say $R(\boldsymbol{X}_n, Y)$, is a discrete function that does not depend on any unknown parameters. The pivotal conditional cdf method described in Section 2.1.2 cannot be used directly because $G_{T|R}[T(\boldsymbol{X}_n)|R(\boldsymbol{X}_n, Y)]$ is no longer $\text{Uniform}(0,1)$ distributed. Nevertheless, because $G_{T|R}[T(\boldsymbol{X}_n)|R(\boldsymbol{X}_n, Y)]$ is stochastically ordered with respect to the $\text{Uniform}(0,1)$ distribution (see Section D of the supplementary material), the pivotal conditional cdf method can be extended to discrete distributions with slight modifications as long as $G_{T|R}[T(\boldsymbol{x}_n)|R(\boldsymbol{x}_n, y)]$ is a monotone function of y. Without loss of generality, suppose that $G_{T|R}[T(\boldsymbol{x}_n)|R(\boldsymbol{x}_n, y)]$ is a non-increasing function of y, the $1-\alpha$ lower and upper prediction bounds are defined as

$$\underset{\sim}{Y}_{1-\alpha} = \inf\left\{y : 1 - G_{T|R}\left[T(\boldsymbol{x}_n) - 1|R(T(\boldsymbol{x}_n), y)\right] > \alpha\right\}, \quad \widetilde{Y}_{1-\alpha} = \sup\left\{y : G_{T|R}\left[T(\boldsymbol{x}_n)|R(T(\boldsymbol{x}_n), y)\right] > \alpha\right\}. \tag{15}$$

We call (15) the conservative method because the prediction bounds are guaranteed to have a coverage probability that is greater or equal to the nominal confidence level (details are given in Section D of the supplementary material).

There are other constructions of the pivotal (conditional) cdf method. Suppose the conditional distribution of Y given $R(\boldsymbol{X}_n, Y)$ does not depend on any parameters and the conditional cdf is $G[y|R(\boldsymbol{x}_n, y)]$. Faulkenberry (1973) proposes the conditional method by defining the $1-\alpha$ lower and upper bounds as

$$\underset{\sim}{Y}_{1-\alpha} = \sup\{y : G_{Y|R}[y-1|R(x,y)] \leq \alpha\}, \quad \widetilde{Y}_{1-\alpha} = \inf\{y : G_{Y|R}[y|R(x,y)] \geq 1-\alpha\}. \quad (16)$$

As noted by Dunsmore (1976), however, the prediction bounds in (16) may not exist in some situations (i.e., the set may be empty).

7.1.2 Approximate Pivotal Methods

Similar to the idea in Section 2.2, we can construct prediction intervals using approximate pivotal quantities. Suppose $q(\boldsymbol{X}_n, Y, \boldsymbol{\theta}) \xrightarrow{d} U$, where U does not depend on any parameters. Then, if $\boldsymbol{\theta}$ is known, a $1-\alpha$ prediction interval (or bound) can be defined by $\{y : q(\boldsymbol{x}_n, y, \boldsymbol{\theta}) \leq u_{n,1-\alpha}\}$, where $u_{n,1-\alpha}$ is the $1-\alpha$ quantile of U. When $\boldsymbol{\theta}$ is unknown one can replace $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}_n$, such as $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{X}_n)$, which is the ML estimator of $\boldsymbol{\theta}$ from the data $\boldsymbol{X}_n$. Another choice is to use $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{X}_n, Y)$, which is the estimator from both the data and the predictand $(\boldsymbol{X}_n, Y)$. After replacing $\boldsymbol{\theta}$ with an estimator $\widehat{\boldsymbol{\theta}}_n$, we can construct a prediction interval for Y by solving $\{y : q(\boldsymbol{x}_n, y, \widehat{\boldsymbol{\theta}}_n) \leq u_{n,1-\alpha}\}$, for integer y, as illustrated in Sections 7.2 and 7.3.

7.1.3 Methods Based on Integration

Using an objective prior (e.g., a Jeffreys prior), a Bayesian predictive distribution can be used to construct prediction intervals, which may have good frequentist coverage probability, as illustrated in the rest of this section. Similarly, the fiducial method also works when an (approximate) fiducial distribution is available. The conditioning-based predictive likelihood (cf. Bjørnstad (1990)) can also be used, as illustrated in Sections 7.2 and 7.3.

7.2 The Binomial Distribution

Let $X \sim \text{Binom}(n, p)$ and $Y \sim \text{Binom}(m, p)$, where $p \in (0,1)$ is unknown and n, m are given positive integers. The goal is to construct prediction bounds for Y based on the observed value of $X = x$. When $X = 0$ or $X = n$, the ML estimate is $\widehat{p} = 0$ or $\widehat{p} = 1$ so that prediction methods based on ML estimators (including plug-in, calibration-bootstrap, and direct-bootstrap methods) cannot be used directly; this is because estimated distributions used for prediction are degenerate for the extreme values of X. Several prediction methods for the binomial case are described below. A numerical study was done to compare some of the methods and the results are given in Section E of the supplementary material.

7.2.1 The Conservative Method

Thatcher (1964) notes that a prediction interval can be obtained by using the conditional cdf of X given $X+Y$ and proposes this method, which is an implementation of the method described in Section 7.1.1. Suppose there are $n+m$ balls and $R = X+Y$ are red balls. Then X is the number of red balls out of n balls, which has a hypergeometric distribution $\mathrm{Hyper}(X+Y, n, n+m)$ with cdf $\mathrm{phyper}(\cdot; X+Y, n, n+m)$. After observing $X = x$, the $1-\alpha$ lower and upper prediction bounds using the conservative method are

$$\underset{\sim}{Y}_{1-\alpha} = \inf\left\{y : 1 - \mathrm{phyper}(x-1; x+y, n, n+m) > \alpha\right\}, \quad \widetilde{Y}_{1-\alpha} = \sup\left\{y : \mathrm{phyper}(x; x+y, n, n+m) > \alpha\right\}.$$

7.2.2 Methods based on Approximate Pivots

The methods discussed in this section are implementations of the general method described in Section 7.1.2. By the Central Limit Theorem (CLT), both X and Y have normal limits (as $m, n \to \infty$) in that

$$Z_X = \frac{X/n - p}{\sqrt{(1-p)p/n}} \xrightarrow{d} \mathrm{Norm}(0,1), \quad Z_Y = \frac{Y/m - p}{\sqrt{(1-p)p/m}} \xrightarrow{d} \mathrm{Norm}(0,1).$$

Because X and Y are independent and by standardizing $\sqrt{n}Z_Y - \sqrt{m}Z_X$ (as approximately $\mathrm{Norm}(0, n+m)$), it also holds (as $m, n \to \infty$) that

$$\frac{\sqrt{n}Z_Y - \sqrt{m}Z_X}{\sqrt{\mathrm{Var}(\sqrt{n}Z_Y - \sqrt{m}Z_X)}} = \frac{Y - mX/n}{\sqrt{(n+m)(m/n)p(1-p)}} \xrightarrow{d} \mathrm{Norm}(0,1). \tag{17}$$

As mentioned in Section 7.1.2, we need to replace the parameter (p above) with an estimator to construct prediction intervals.

Nelson (1982) proposes replacing p in (17) with $\widehat{p}_x = X/n$. But numerical studies in Wang (2008) and Krishnamoorthy and Peng (2011) show that this method has poor coverage probability, even for large samples. Instead of using $\widehat{p}_x$, Krishnamoorthy and Peng (2011) propose replacing p with $\widehat{p}_{xy} = (X+Y)/(n+m)$ (along with a continuity adjustment: if $x = 0$, use $x = 0.5$; if $x = n$, use $x = n - 0.5$). Inspired by the Wilson score confidence interval (cf. Wilson (1927)), Wang (2010) proposes another method where $\widehat{p} = (X + Y + z^2_{1-\alpha}/2)/(n + m + z^2_{1-\alpha})$ is used to replace p in (17).

7.2.3 Methods Based on Integration

For the $\mathrm{Binom}(n,p)$, the Jeffreys prior is $\pi_J(p) \propto p^{-1/2}(1-p)^{-1/2}$ (i.e., $\pi_J(p) \sim \mathrm{Beta}(0.5, 0.5)$). The $1-\alpha$ lower and upper Jeffreys (Bayesian) prediction bounds are

$$\underset{\sim}{Y}_{1-\alpha} = \mathrm{qbetabinom}(\alpha; m, x{+}0.5, n{-}x{+}0.5), \quad \widetilde{Y}_{1-\alpha} = \mathrm{qbetabinom}(1{-}\alpha; m, x{+}0.5, n{-}x{+}0.5),$$

where $\mathrm{qbetabinom}(p; n, a, b)$ is the p quantile of the beta-binomial distribution with a sample-size parameter n and shape parameters a and b. Compared with the method proposed by Krishnamoorthy and Peng (2011), this Jeffreys prediction method is slightly more conservative (cf. Meeker et al. 2017, Chapter 6).

A fiducial quantity for parameter p based on observation $X = x$ has the form

$$\mathcal{R}_p = U_{(x)} + D(U_{(x+1)} - U_{(x)}),$$

where $U_{(x)}$ is the xth smallest value out of n independent Uniform$(0, 1)$ random variables ($U_{(0)} = 0$, $U_{(n+1)} = 1$) and $D \sim$ Uniform$(0, 1)$. Integrating out the parameter p in the Binom(m, p) cdf using the density function of $\mathcal{R}_p$, say $r(p|X = x)$, gives the fiducial predictive distribution for Y via (7). The prediction bounds are defined using the appropriate quantiles of the fiducial predictive distribution.

7.2.4 The Hinkley Predictive Likelihood

Hinkley (1979) proposes a conditioning-based predictive likelihood, which is based on the fact that the conditional distribution of X (or Y) given $X + Y$ has a hypergeometric distribution. However, unlike the pivotal conditional cdf method (cf. Section 7.1.1), this method is invariant to whether the conditional cdf of X or Y is chosen because the predictive likelihood of Y is

$$L_p(y; x) = \frac{\binom{n}{x}\binom{m}{y}}{\binom{n+m}{x+y}}, \quad y \in \{0, \ldots, m\},$$

where x and y are interchangeable. The predictive cdf is obtained by normalizing the predictive likelihood as $F_p(y; x) = \sum_{j=0}^{y} L_p(j; x)/\sum_{i=0}^{m} L_p(i; x)$. The $1 - \alpha$ lower and upper prediction bounds are defined as

$$\underset{\sim}{Y}_{1-\alpha} = \sup\{y : F_p(y - 1; x) \leq \alpha\}, \quad \widetilde{Y}_{1-\alpha} = \inf\{y : F_p(y; x) \geq 1 - \alpha\}.$$

Although both the conservative method in (15) and the Hinkley predictive likelihood method are based on the conditional distribution of X given $X + Y$, they lead to different prediction intervals.

7.3 The Poisson Distribution

Suppose $X \sim$ Poi$(n\lambda)$ and $Y \sim$ Poi$(m\lambda)$, where $\lambda > 0$ is unknown, n and m are known positive real values, and Y is independent of X. The goal is to construct a prediction interval for Y based on the observation $X = x$. Methods similar to those used in Section 7.2 are used for Poisson prediction.

7.3.1 The Conservative Method

Because the conditional distribution of X given $X + Y$ is Binom$(x + y, n/(n + m))$, we can use the general method described in Section 7.1.1. The $1 - \alpha$ lower and upper prediction bounds using the conservative method are

$$\begin{aligned}
\underset{\sim}{Y}_{1-\alpha} &= \inf\left\{y : 1 - \text{pbinom}(x-1; x+y, n/(n+m)) > \alpha\right\},\\
\widetilde{Y}_{1-\alpha} &= \sup\left\{y : \text{pbinom}(x; x+y, n/(n+m)) > \alpha\right\}.
\end{aligned}$$

7.3.2 Methods based on Approximate Pivots

This section implements the methods proposed in Section 7.1.2. By the CLT, both X and Y have normal limits (as $m, n \to \infty$) given by

$$Z_X = \frac{X - n\lambda}{\sqrt{n\lambda}} \xrightarrow{d} \text{Norm}(0,1), \quad Z_Y = \frac{Y - m\lambda}{\sqrt{m\lambda}} \xrightarrow{d} \text{Norm}(0,1).$$

Because X and Y are independent, $\sqrt{n}Z_Y - \sqrt{m}Z_X$ has approximately a normal distribution with mean 0 and variance $(n+m)$. Thus, it holds (as $m, n \to \infty$) that

$$\frac{\sqrt{n}Z_Y - \sqrt{m}Z_X}{\sqrt{\text{Var}(\sqrt{n}Z_Y - \sqrt{m}Z_X)}} = \frac{Y - mX/n}{\sqrt{(m + m^2/n)\lambda}} \xrightarrow{d} \text{Norm}(0,1).$$

Nelson (1982) replaces the unknown λ with $\widehat{\lambda}_x = X/n$ and Krishnamoorthy and Peng (2011) replace λ with $\widehat{\lambda}_{xy} = (X+Y)/(n+m)$ (along with a continuity adjustment: if $x = 0$, use $x = 0.5$). Krishnamoorthy and Peng show that their method has better coverage probability properties than Nelson's method.

7.3.3 Methods Based on Integration

The Jeffreys prior for the Poisson rate parameter is $\pi_J(\lambda) = \sqrt{1/\lambda}$, $\lambda > 0$, and the corresponding posterior distribution is gamma(x+1/2, n) with density $p(\lambda|x) \propto \lambda^{x-1/2}\exp(-n\lambda)$, for $\lambda > 0$. Using this posterior, the Bayesian predictive density for the Poisson distribution is

$$p(y|x) = \frac{\Gamma(y + x + 1/2)}{\Gamma(x + 1/2)\Gamma(y+1)} \left(\frac{n}{n+m}\right)^{x+1/2} \left(1 - \frac{n}{n+m}\right)^{y}, \tag{18}$$

which is a negative-binomial distribution $\text{NB}(x + 0.5, n/(n+m))$. The Jeffreys Bayesian prediction method tends to be more conservative than the method proposed by Krishnamoorthy and Peng (2011), especially when the ratio m/n is small (i.e., the expected value of $n\lambda$ of the data X greatly exceeds that λm of the predictand Y), but is less conservative than the conservative method (cf. Meeker et al. 2017, Chapter 7).

An approximate fiducial quantity for λ given observation $X = x$ has a distribution of a scaled chi-square variable $\chi^2_{2x+1}/2n$ (cf. Dempster 2008, Krishnamoorthy and Lee 2010). Using this approximate fiducial distribution in place of the (gamma) posterior $p(\lambda|x)$ in (18) leads to a fiducial predictive distribution for a Poisson predictand.

7.3.4 The Hinkley Predictive Likelihood

Using the fact that the conditional distribution of X given $X+Y$ has a binomial distribution, the Hinkley predictive likelihood is as follows

$$L_p(y;x) = \frac{f(X=x, Y=y)}{f(X+Y=x+y)} = \frac{(x+y)!}{x!y!}\left(\frac{m}{n+m}\right)^y\left(\frac{n}{n+m}\right)^x, y \geq 0.$$

Often, the predictive distribution obtained by normalizing the predictive likelihood has no closed form, but in this example the predictive pmf for Y induced by $L_p(y;x)$ has a negative-binomial mass function

$$f_p(y;x) = f_p(y;x+1, n/(n+m)) = \binom{x+y}{x}[m/(n+m)]^x[n/(n+m)]^y, \quad y = 0, 1, 2, \ldots$$

Thus, the prediction bounds can be obtained by using the appropriate quantiles of the negative-binomial distribution.

7.4 Cautionary Comments about the Plug-in Method

The plug-in method generally works only when the following holds

$$\sup_{y\in\mathbb{R}} |G(y|\boldsymbol{X}_n;\boldsymbol{\theta}_0) - G(y|\boldsymbol{X}_n;\widehat{\boldsymbol{\theta}}_n)| \xrightarrow{p} 0 \tag{19}$$

as $n \to \infty$. Here $G(\cdot|\boldsymbol{x}_n;\boldsymbol{\theta}_0)$ is the conditional distribution function of Y given $\boldsymbol{X}_n$. The convergence (19) implies that the "plug-in" version of the cdf approaches the true cdf as $n \to \infty$. This convergence, however, does not always hold.

Tian et al. (2020) discuss a particular type of within-sample prediction problem, where a Type-I censored time-to-event dataset is given to predict the number of future events during a time period after the censoring time. They show that (19) does not hold in this within-sample prediction problem. Thus, for this situation, the plug-in method is *not* asymptotically correct, which means, no matter how large the sample size n is, the true coverage probability is generally different from the nominal confidence level. Note that, for the within-sample prediction, the sample size of the data is n while the scalar predictand is a summary statistic of $n-r$ Bernoulli random variables, where r is the number of events observed in the data. The plug-in method is invalid for a case where the predictand sample size $n-r$ is potentially large compared to the data sample size n. Alternatively, Tian et al. (2020) propose three bootstrap-based methods that are asymptotically correct.

Relatedly, for the discrete new-sample prediction problems in Sections 7.2 and 7.3, both the data X and the predictand Y may be generally viewed as counts that summarize two samples: one sample of size n for X and another sample of size m for Y. We emphasize that the plug-in prediction method requires cautious consideration of the relative sizes of n and m. Importantly, the plug-in method will similarly fail for the binomial and Poisson prediction cases of Sections 7.2 and 7.3 more broadly, unless m is appropriately small relatively to n (that is, asymptotically, success for the plug-in method requires $m/n \to 0$ implying that the data sample size n dominates the predictand sample size m). Without this condition, the $\text{Binom}(m,p)$ or $\text{Poi}(m\lambda)$ distribution of a predictand Y cannot be consistently approximated by the plug-in prediction method (i.e., by substituting X/n for p or λ). For this reason, plug-in prediction is not included in Section 7.2 or 7.3. The other prediction methods in these sections, however, are valid without restrictions on the relative sizes of m, n. That is, for any ratio m/n, these other prediction methods are asymptotically correct.

8 Overview of Prediction Methods for Dependent Data

In this section, "dependent data" refers to data with a complicated dependence structure. Common examples include time series, spatial data, and random networks, where the strength of dependence among observations often depends on proximity. Other examples include mixed-effects models, longitudinal data, and small area estimation where correlation exists among observations sharing random effects or repeated measures. In such examples, distributional models for predictands often share non-trivial connections to data models through dependence conditions. See Clarke and Clarke (2018) for descriptions of other prediction applications with dependent data.

While the literature for predictions with independent data is more extensive, similar prediction methods exist for dependent data along the lines of the plug-in, calibration-bootstrap, and (integration-based) predictive distribution methods. This section discusses prediction interval methods for dependent data.

Similar to previous sections, the prediction problems considered here are based on a parametric model for the data and the predictand, although the dependence among and between these quantities can create complications. Providing a challenge with model formulation and prediction under dependence, both the data structure and the nature of possible dependence in samples can vary greatly.

A further challenge in prediction with dependent data is that prediction strategies developed for independent observations may fail if directly applied to dependent data, so that caution may be required. As a simple example, the plug-in method provides consistent prediction bounds for many problems with independent data but fails for the within-sample prediction problem described in Section 7.4 where, despite arising from a random sample, the predictand and the observed data are dependent. Additionally, several prediction methods from previous sections involve bootstrap sampling (under independence), where data simulation and the generation of bootstrap samples is more complicated when data are dependent.

Before describing prediction methods, we mention that there exists a variety of ways to generate bootstrap samples under dependence, particularly for time series. For the latter, common formulations of bootstrap include model-residual-based bootstraps (e.g., AR(p) models, cf. Pan and Politis 2016), transformation-based bootstraps aiming to weaken dependence (e.g., Kreiss et al. 2011, Jentsch and Politis 2015), and block-based bootstraps that use data blocks to reconstruct time series (Gregory et al. 2018). These bootstrap methods differ in their mechanics as well as in the amount of time series structure presumed by the bootstrap. For reviews of bootstrap methods with time series and other dependent data, see Politis (2003), Lahiri (2006), and Kreiss and Lahiri (2012).

For parametric-model based predictions from dependent data, Sections 8.1, 8.2, and 8.3 respectively describe plug-in, calibration-bootstrap, and integration-based predictive methods. The bootstrap, when employed, is parametric. These procedures have been largely studied and justified in the context of Gaussian process models, where Gaussian assumptions also facilitate generation of bootstrap samples needed for the calibration-bootstrap. More development is needed to extend these approaches to predictions with non-Gaussian dependent data, with some possibilities suggested in Section 8.4.

8.1 The Plug-in Method

Beran (1990) and Hall et al. (1999) consider the plug-in prediction method for some specially structured dependent data models with independent additive errors (e.g., regression models, and the AR(1) model). To set an h-step ahead prediction interval, given a realization of time series data from an ARMA process, for example, Brockwell and Davis (2016, Chapter 3) suggest using a normality assumption along with an approximation for the best linear predictor found by replacing unknown parameters with consistent estimates. Similarly, by assuming a stationary Gaussian process, the plug-in prediction interval has been suggested in spatial applications based on using a normal approximation with kriging predictors, where unknown parameters are replaced with consistent estimates (cf. Cressie 2015, Chapter 3). With such a Gaussian process, the coverage probability of the plug-in method typically has an error of $O(1/n)$ (cf. Sjöstedt-de Luna and Young 2003, Vidoni 2004). However, it is not generally clear when the plug-in method is asymptotically correct for dependent data, particularly for more complicated and potentially non-Gaussian dependence structures.

8.2 The Calibration Method

Using the calibration-bootstrap method (described in Section 2.2.2), Sjöstedt-de Luna and Young (2003) improve plug-in kriging prediction intervals for a stationary Gaussian process while

De Oliveira and Kone (2015a) establish similar findings for predicting spatial averages from these processes, and Hall and Maiti (2006) use calibration-bootstrap for small area prediction. Similar to the method described in Section 2.2.3, Giummole and Vidoni (2010) calibrate the plug-in method using an asymptotic expansion for a general class of Gaussian models, including time-series, Gaussian state-space models, and Gaussian Markov random fields. Under regularity conditions with certain dependent Gaussian processes, the calibration methods reduce the error of the coverage probability to $o(1/n)$ compared to $O(1/n)$ for the plug-in method (cf. Sjöstedt-de Luna and Young 2003, Giummole and Vidoni 2010).

8.3 Bayesian and Fiducial Predictive Distributions

The fiducial method can be potentially extended to dependent data, but requires the development of an appropriate fiducial distribution under dependence, as described in Wang et al. (2012). The Bayesian method has been studied extensively for dependent data for prediction. For example, West and Harrison (1997) discuss Bayesian prediction in dynamic linear models for time series; recent work includes Aktekin et al. (2018), McAlinn and West (2019), and Berry and West (2020). Handcock and Stein (1993) propose a best linear unbiased prediction procedure within a Bayesian framework for Gaussian random fields and De Oliveira et al. (1997) present prediction methods for some types of non-Gaussian random fields. See Hulting and Harville (1991), Harville and Carriquiry (1992), and Christensen and Waagepetersen (2002) for related results in the context of linear mixed models or generalized mixed models.

8.4 Extensions to Non-Gaussian Dependent Data

As described in Sections 8.1-8.2, many of the existing formal treatments of model-based predictions for dependent data have largely focused on types of Gaussian processes. This simply indicates that the prediction methods based on the plug-in or bootstrap methods rely heavily on tractable forms for the distribution of the dependent data. One option for model-based predictions with non-Gaussian data is to use a Gaussian model in conjunction with a suitable data transformation; for example, De Oliveira and Rui (2009) develop plug-in and bootstrap calibration for log-Gaussian fields. Beyond normal data cases, we mention another general model class for developing predictions could potentially involve Markov random field (MRF) structures. This approach for modeling dependent data involves specifying a full conditional distribution for each observation on the basis of an underlying MRF (cf. Besag (1974)). Model formulation in a conditional, component-wise fashion provides an alternative to direct specification of a full joint distribution for the data. Additionally, such conditional distributions often depend functionally on small subsets of "neighboring" observations, which is a property that may be useful for extending the plug-in and calibration-bootstrap prediction methods to

MRF models. The supplement describes more details about this prediction problem, along with a numerical illustration (see Section F of the supplementary materials). For implementing parametric bootstrap without assumptions of Gaussianity, an attractive feature of MRF models is that data may be simulated rapidly from specified full conditional distributions through Gibbs sampling (Kaplan et al. (2020)). Note that MRF models have applications to both continuous and discrete dependent data (cf. Cressie 2015, Kaiser and Cressie 2000, and Casleton et al. 2017).

Regarding the calibration-bootstrap of Section 8.2, the bootstrap for dependent data can be difficult to establish through the prescription in Section 2.2, which requires an analytic form for the conditional distribution of a predictand Y given the data $\boldsymbol{X}_n$. By ignoring this distribution, an alternative strategy for predictions based on bootstrap is to approximate the distribution of a prediction error $|Y - \hat{Y}|$, where $\hat{Y}$ denotes a statistic based on data $\boldsymbol{X}_n$. See, for example, Politis (2013) and Pan and Politis (2016) for illustrations with time series, and also De Oliveira and Kone (2015b) for similar bootstrap predictions with spatial data.

9 Nonparametric Prediction Methods

Up to this point, we have considered prediction problems based on parametric models. Given a sufficient amount of data, however, nonparametric methods are available to construct prediction intervals. In this section, we discuss two types of nonparametric prediction methods.

9.1 Prediction Intervals Based on Order Statistics

Let $X_1, \ldots, X_n$ be a random sample from some continuous distribution function $F(x)$, then $(X_{(r)}, X_{(s)})$ is a $100\,[(s-r)/(n+1)]\,\%$ prediction interval for an independent future random variable Y from the same distribution, where $1 \leq r < s \leq n$. The coverage probability of this prediction interval method is exact and it does not depend on the form of the underlying continuous distribution (i.e., the method is distribution-free). For a desired nominal coverage probability that cannot be obtained in the form of $[(s-r)/(n+1)]$, Beran and Hall (1993) suggest interpolation to construct nonparametric prediction intervals to approximate the desired coverage probability. Also proposing methods based on order statistics, Fligner and Wolfe (1976) consider constructing distribution-free prediction intervals that contain at least k of m future random variables and prediction intervals for any particular order statistic of a future sample. Meeker et al. (2017, Chapter 5 and Appendix G) describe computational details and illustrate the use of these distribution-free methods. Frey (2013) proposes a shortest nonparametric prediction interval, as opposed to methods that have approximately equal one-sided coverage probabilities.

9.2 Conformal Prediction

Conformal prediction has been gaining popularity recently because it applies to many prediction problems in the area of supervised learning, including regression and classification problems. Conformal prediction has often been presented in the form of an algorithm (e.g., Vovk et al. 2005, Shafer and Vovk 2008). Here we describe the conformal prediction through sampling distributions in order to connect with the pivotal cdf methods described in Section 2.1.2.

Suppose the data sample $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$ and the predictand $Y \equiv X_{n+1}$ are i.i.d. (or a weaker exchangeable assumption). Conformal prediction intervals are based on a choice of distance statistic $d(\boldsymbol{X}_n, Y)$ (or nonconformity measure). One example is $d(\boldsymbol{X}_n, Y) = |\bar{X}_n - Y|$, which is the distance between a new observation Y and the data sample mean $\bar{X}_n$. Denote the cdf of $d(\boldsymbol{X}_n, Y)$ by $G(t) = \Pr[d(\boldsymbol{X}_n, Y) \leq t], t \in \mathbb{R}$, and write the left limit of the cdf as $G(t-) = \Pr[d(\boldsymbol{X}_n, Y) < t] = \lim_{x \uparrow t} G(t), t \in \mathbb{R}$. The conformal prediction uses the probability integral transform in combination with the quantile $1-\alpha$ of a Uniform$(0,1)$ (i.e., $1-\alpha$) to calibrate prediction regions (similar to the approach used in Section 2.1.2), after an initial step of estimating the cdf G with an empirical distribution $\widehat{G}_y$.

Let $\boldsymbol{X}_{n+1} \equiv \{X_1, \ldots, X_n, X_{n+1}\}$. Then a $1-\alpha$ conformal prediction region for Y is given as

$$\left\{ y \in \mathbb{R} : \widehat{G}_y \left[d\left(\boldsymbol{X}_n, y\right) - \right] < 1 - \alpha \right\}, \tag{20}$$

where $\widehat{G}_y$ is given by

$$\widehat{G}_Y(t) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathrm{I}\left(d\left(\boldsymbol{X}_{n+1} \backslash \left\{X_i\right\}, X_i\right) \leq t\right), \quad t \in \mathbb{R}$$

which is the empirical cdf of distances $d(\boldsymbol{X}_{n+1} \backslash \{X_i\}, X_i), i = 1, \ldots, n+1$, found by separating point X_i from $\boldsymbol{X}_{n+1}$. Note that $\widehat{G}_Y(\cdot)$ depends on the unobserved value $X_{n+1} = Y$, so that $\widehat{G}_y(\cdot) \equiv \widehat{G}_{Y=y}(\cdot)$ is computed provisionally using a potential value of $Y = y$ in (20). Furthermore, in (20), the estimated cdf $\widehat{G}_y(t)$ at $t = d(\boldsymbol{X}_n, y)$ is technically replaced with a left limit $\widehat{G}_y(d(\boldsymbol{X}_n, y)-)$ as a type of adjustment to $\widehat{G}_y$ (i.e., as the latter always jumps by $1/(n+1)$ at the argument $t = d(\boldsymbol{X}_n, y)$). For perspective, in other prediction problems based on cdf transforms, a switch from cdfs to left limits of cdfs is known to be helpful for correcting issues of discreteness in data, with the effect of ensuring coverage probabilities are conservative. Jump points of $\widehat{G}_y(t)$ can also be randomized, e.g., replace $\widehat{G}_y(t-)$ with $\widehat{G}_y(t-) + U[\widehat{G}_y(t) - \widehat{G}_y(t)]$ in (20) for a Uniform$(0,1)$ draw U, which then blends the two prediction regions given by using either $\widehat{G}_y(t)$ or $\widehat{G}_y(t-)$ alone. The conformal prediction method is conservative, but if a randomization scheme is used (whereby the realized prediction interval would depend on the outcome of a random draw,) the method can be made to be exact (cf. Vovk et al. 2005).

Conformal prediction can be used in the supervised learning setting (i.e., predicting the response given the features as well as labeled training data). Some recent work includes applying conformal prediction to regression (cf. Lei et al. 2018), quantile regression (cf. Romano et al. 2019), and lasso regression (cf. Lei 2019).

10 Discussion

This paper discusses two major types of methods to construct frequentist prediction intervals. One is based on an (approximate) pivotal quantity and the other is based on a predictive distribution (or likelihood). The extensions of these prediction methods to dependent data are briefly discussed. Here is a summary of our important conclusions.

- Exact prediction methods are available for (log-)location-scale distributions under complete and Type II censoring and good approximations are available for Type I censoring.
- For (log-)location-scale distributions, there are several equivalent methods for computing *exact* intervals. The GPQ predictive distribution method (GPQ-bootstrap) has strong appeal due to its ease of implementation.
- For other continuous distributions, the direct-bootstrap method performs no better than the naive plug-in approach and should be avoided (i.e., due to the increased computational costs versus no performance gain over the plug-in method). The calibration-bootstrap method, however, has good coverage probability properties, even with moderate to small sample sizes. Another potentially useful method is to use a (generalized) fiducial predictive distribution.
- For discrete distributions, we discussed and illustrated the use of three general methods: pivotal cdf (i.e., the conservative method), approximate normal statistics (e.g., based on a Wald-like or a score-like statistic), and integration methods (e.g., a Bayesian method with an objective prior).
- When the prediction problems involve dependent data, the development of prediction intervals, particularly based on parametric bootstrap, requires more investigation for non-Gaussian dependent data.

This paper focuses on prediction intervals and coverage probability while in most of the statistical learning (also known as machine learning) literature, the focus is on algorithms for point prediction, which are evaluated with metrics like mean squared error using cross-validation. However, even with contemporary (nonparametric) prediction algorithms, such as neural networks, boosting, support vector machines, and random forests (cf. Clarke and Clarke 2018), there is increasing interest in developing prediction intervals.

In addition to the conformal prediction, for example, prediction intervals based on random forests (cf. Zhang et al. 2020) may be formulated by estimating a distribution of prediction errors (via left out or "out-of-bag" observations), similar to the approach described in Section 9.2. Ultimately, the development of prediction interval procedures from statistical learning algorithms relates to bridging prediction and estimation, as outlined in the recent overview of Efron (2020). As our focus in this paper has been on prediction intervals, we have not covered the important area of multivariate prediction. Recently, some work has been done on multivariate prediction, especially for Gaussian sequence models (cf. George et al. 2012 and Mukherjee and Johnstone 2015).

Acknowledgements

We would like to thank Professor Jan Hannig for providing a copy of his MATLAB code for computing the Gamma fiducial distribution and Professor Luis A. Escobar for helpful guidance on the use of the pivotal methods for discrete distributions. We would also like to thank the editor, associate editor, and referees for their constructive and helpful comments. The research was partially supported by NSF DMS-2015390.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis.* Cambridge University Press.

Aktekin, T., Polson, N., and Soyer, R. (2018). Sequential Bayesian analysis of multivariate count data. *Bayesian Analysis*, 13:385–409.

Baker, G. A. (1935). The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample. *The Annals of Mathematical Statistics*, 6:197–201.

Barndorff-Nielsen, O. E. and Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli*, 2:319–340.

Beran, R. (1990). Calibrating prediction regions. *Journal of the American Statistical Association*, 85:715–723.

Beran, R. and Hall, P. (1993). Interpolated nonparametric prediction intervals and confidence intervals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55:643–652.

Berry, L. R. and West, M. (2020). Bayesian forecasting of many count-valued time series. *Journal of Business & Economic Statistics*, 38:872–887.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36:192–225.

Bjørnstad, J. F. (1990). Predictive likelihood: a review. *Statistical Science*, 5:262–265.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer.

Casella, G. and Berger, R. L. (2002). *Statistical Inference, Second Edition*. Duxbury.

Casleton, E., Nordman, D., and Kaiser, M. (2017). A local structure model for network analysis. *Statistics and Its Interface*, 10:355–367.

Chen, P. and Ye, Z.-S. (2017). Approximate statistical limits for a gamma distribution. *Journal of Quality Technology*, 49:64–77.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58:280–286.

Clarke, B. S. and Clarke, J. L. (2018). *Predictive Statistics: Analysis and Inference beyond Models*. Cambridge University Press.

Cox, D. R. (1975). Prediction intervals and empirical Bayes confidence intervals. *Journal of Applied Probability*, 12(S1):47–55.

Cressie, N. (2015). *Statistics for Spatial Data, Revised Edition*. John Wiley & Sons.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

de Finetti, B. (1974). *Theory of Probability: A Critical Introductory Treatment*. Wiley, 1 edition.

De Oliveira, V., Kedem, B., and Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, 92:1422–1433.

De Oliveira, V. and Kone, B. (2015a). Prediction intervals for integrals of Gaussian random fields. *Computational Statistics & Data Analysis*, 83:37–51.

De Oliveira, V. and Kone, B. (2015b). Prediction Intervals for Integrals of Some Types of Non-Gaussian Random Fields: A Semiparametric Bootstrap Approach. *JSM Proceedings, Statistics and the Environment Section*, pages 2588–2597.

De Oliveira, V. and Rui, C. (2009). On shortest prediction intervals in log-Gaussian random fields. *Computational Statistics & Data Analysis*, 53:4345–4357.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48:365–377.

DiCiccio, T. J., Kuffner, T. A., and Young, G. A. (2017). A simple analysis of the exact probability matching prior in the location-scale model. *The American Statistician*, 71:302–304.

Dunsmore, I. R. (1976). A note on Faulkenberry's method of obtaining prediction intervals. *Journal of the American Statistical Association*, 71:193–194.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115:636–655.

Faulkenberry, G. D. (1973). A method of obtaining prediction intervals. *Journal of the American Statistical Association*, 68:433–435.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398.

Fligner, M. A. and Wolfe, D. A. (1976). Some applications of sample analogues to the probability integral transformation and a coverage property. *The American Statistician*, 30:78–85.

Fonseca, G., Giummolè, F., and Vidoni, P. (2012). Calibrating predictive distributions. *Journal of Statistical Computation and Simulation*, 84:373–383.

Fortini, S. and Petrone, S. (2014). Predictive distribution (de Finetti's view). *Wiley StatsRef: Statistics Reference Online*, pages 1–9.

Frey, J. (2013). Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference*, 143:1039–1048.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall/CRC.

George, E. I., Liang, F., Xu, X., et al. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statistical Science*, 27:82–94.

Giummole, F. and Vidoni, P. (2010). Improved prediction limits for a general class of Gaussian models. *Journal of Time Series Analysis*, 31:483–493.

Gregory, K. B., Lahiri, S. N., and Nordman, D. J. (2018). A smooth block bootstrap for quantile regression with time series. *Annals of Statistics*, 46:1138–1166.

Guttman, I. and Tiao, G. C. (1964). A Bayesian approach to some best population problems. *The Annals of Mathematical Statistics*, 35:825–835.

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:221–238.

Hall, P., Peng, L., and Tajvidi, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika*, 86:871–880.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35(4):403–410.

Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111:1346–1361.

Hannig, J., Iyer, H., and Patterson, P. (2006). Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101:254–269.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, 76:675–684.

Harville, D. A. (2014). The need for more emphasis on prediction: a "nondenominational" model-based approach. *The American Statistician*, 68:71–83.

Harville, D. A. and Carriquiry, A. L. (1992). Classical and Bayesian prediction as applied to an unbalanced mixed linear model. *Biometrics*, 48:987–1003.

Hinkley, D. (1979). Predictive likelihood. *Annals of Statistics*, 7:718–728.

Hulting, F. L. and Harville, D. A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*, 86:557–568.

Jentsch, C. and Politis, D. N. (2015). Covariance matrix estimation and linear process bootstrap for multivariate time series of possibly increasing dimension. *Annals of Statistics*, 43:1117–1140.

Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis*, 73:199–220.

Kaplan, A., Kaiser, M. S., Lahiri, S. N., and Nordman, D. J. (2020). Simulating Markov random fields with a conclique-based Gibbs sampler. *Journal of Computational and Graphical Statistics*, 29:286–296.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, 83:299–313.

Kreiss, J.-P. and Lahiri, S. N. (2012). Bootstrap methods for time series. In *Handbook of Statistics*, volume 30, pages 3–26. Elsevier.

Kreiss, J.-P., Paparoditis, E., and Politis, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. *Annals of Statistics*, 39:2103–2130.

Krishnamoorthy, K. and Lee, M. (2010). Inference for functions of parameters in discrete distributions based on fiducial approach: Binomial and Poisson cases. *Journal of Statistical Planning and Inference*, 140:1182–1192.

Krishnamoorthy, K. and Mathew, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation.* Wiley.

Krishnamoorthy, K. and Peng, J. (2011). Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference*, 141:1709–1718.

Lahiri, S. N. (2006). Bootstrap methods: A review. In Fan, J. and Koul, H. L., editors, *Frontiers in Statistics*, pages 231–255. World Scientific.

Lawless, J. F. (1972). Conditional confidence interval procedures for the location and scale parameters of the Cauchy and logistic distributions. *Biometrika*, 59:377–386.

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

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106:749–764.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.

Mathiasen, P. E. (1979). Prediction functions. *Scandinavian Journal of Statistics*, 6:1–21.

McAlinn, K. and West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210:155–169.

Meeker, W. Q., Hahn, G. J., and Escobar, L. A. (2017). *Statistical intervals: A Guide for Practitioners and Researchers, Second Edition.* Wiley.

Mukherjee, G. and Johnstone, I. M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *Annals of Statistics*, 43:937.

Nájera, E. and O'Reilly, F. (2017). On fiducial generators. *Communications in Statistics-Theory and Methods*, 46:2232–2248.

Nelson, W. (1982). *Applied Life Data Analysis.* Wiley.

Pan, L. and Politis, D. N. (2016). Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference*, 177:1–27.

Patel, J. (1989). Prediction intervals—a review. *Communications in Statistics-Theory and Methods*, 18:2393–2465.

Peers, H. (1965). On confidence points and Bayesian probability points in the case of several parameters. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27:9–16.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–230.

Politis, D. N. (2013). Model-free model-fitting and predictive distributions. *Test*, 22:183–221.

Proschan, F. (1953). Confidence and tolerance intervals for the normal distribution. *Journal of the American Statistical Association*, 48:550–564.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.

Shen, J., Liu, R. Y., and Xie, M.-G. (2018). Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25:289–310.

Sjöstedt-de Luna, S. and Young, A. (2003). The bootstrap and kriging prediction intervals. *Scandinavian Journal of Statistics*, 30:175–192.

Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–363.

Thatcher, A. R. (1964). Relationships between Bayesian and confidence limits for predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26:176–192.

Tian, Q., Meng, F., Nordman, D., and Meeker, W. (2020). Predicting the number of future events. *Journal of the American Statistical Association*. DOI:10.1080/01621459.2020.1850461.

Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika*, 85:949–953.

Vidoni, P. (2004). Improved prediction intervals for stochastic process models. *Journal of Time Series Analysis*, 25:137–154.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Wang, C., Hannig, J., and Iyer, H. K. (2012). Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142:1980–1990.

Wang, H. (2008). Coverage probability of prediction intervals for discrete random variables. *Computational Statistics & Data Analysis*, 53:17–26.

Wang, H. (2010). Closed form prediction intervals applied for disease counts. *The American Statistician*, 64:250–256.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12:91–96.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212.

Xie, M.-G. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81:3–39.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2020). Random forest prediction intervals. *The American Statistician*, 74:392–406.