Conformal predictive distributions: an approach to nonparametric fiducial prediction

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #30

August 19, 2020

Project web site:
http://alrw.net

Abstract

The subject of this chapter is conformal predictive distributions, which represent the only approach to nonparametric fiducial prediction that is available at this time. It starts from reviewing parts of traditional fiducial inference that are relevant to conformal predictive distributions and places the latter in the more general context of fiducial prediction. Among key desiderata for fiducial prediction procedures are their validity and efficiency; the first is usually attained automatically, and the other one require careful design. We will discuss various formalizations of these desiderata.

The requirement to present results of the process of prediction in the form of predictive distributions imposes severe restrictions on the process and is not always possible or even desirable. At the end of the chapter we discuss the wider area of conformal prediction, whose predictions can be interpreted as families of p-values; we no longer insist on the kind of consistency of the p-values implicit in conformal predictive distributions.

Contents

1 Introduction

The most developed approach to nonparametric fiducial prediction is via conformal predictive distributions, and these are the main subject of this chapter. We start in Section 2 from reviewing parts of traditional fiducial inference that are relevant to conformal predictive distributions. At the end of the section we discuss key desiderata for fiducial prediction, validity and efficiency.

Fiducial prediction in the nonparametric setting was at least implicit in Fisher's work, and has been greatly developed in the work of Dempster, Hill, and Coolen. It will be the topic of Section 3, where we describe what we call the Dempster–Hill procedure and embed it into fiducial prediction by expressing it in terms of a randomized pivot.

What makes (supervised) machine learning a powerful practical approach to various kinds of prediction problems is that it deals with observations that are pairs (x, y), where x is a potentially complicated object (such as a movie, or all available information about a patient) and y is an associated label (such as a movie's sales figures). In nonparametric fiducial prediction, as developed in statistics, the xs are absent and the observations are just the ys. Conformal predictive distributions add the objects x to the Dempster–Hill picture, which only involves the labels y. This is discussed in Section 4.

Conformal predictive distributions can be derived only under serious restrictions. Getting rid of these restrictions extends their application area and logically leads to the more general area of conformal prediction, which is the topic of Section 5. In that section we will also discuss a generalization of conformal prediction to general repetitive structures (including many of the models considered by Fisher in his parametric fiducial inference) and its application to online methods of hypothesis testing.

We will not make notational distinction between random variables and their possible values, which should always be clear from context. The phrase "uniformly distributed on the interval $[0, 1]$" (applied to a random variable) will sometimes be abbreviated to "uniformly distributed" or even "uniform".

2 Parametric fiducial prediction

Fisher's fiducial inference, at least in his publications, can be divided into two parts (not strictly disjoint). One part concerns inference about the parameters of the unknown true distribution generating the data. This part is controversial and is often regarded as Fisher's greatest blunder [12]. Now it has been revived under the name of confidence distributions (see, e.g., [37] and [52]), but it is still common to deemphasize connections with Fisher's ideas, to stay away from controversy. This chapter is about the other part of fiducial inference, namely inference about future observations, or fiducial prediction. This part remains feasible even without parametric assumptions.

Fiducial prediction is much less prominent in Fisher's oeuvre, and it is much less controversial. This section reviews Fisher's work and some other closely related work in this area.

Remark 1. In this chapter we only consider the simple case of predicting one future observation, although Fisher and later authors are sometimes interested in predicting several future observations (which may blur the difference between fiducial prediction and fiducial inference about parameters, since in many interesting cases parameters can be recovered as limiting values of some function of k future observations as $k \to \infty$).

2.1 Fisher's fiducial prediction

Fisher's two main publications that discuss fiducial prediction are his 1935 paper [15] and his 1956 book [19]. One of the examples in his 1935 paper (Section II) is particularly simple, treating the Gaussian IID model (in this chapter the adjective "IID" signals being related to a sequence of independent and identically distributed random variables, but rarely simply stands for "independent and identically distributed"). Consider an IID sequence of observations $y_1, y_2, \dots$, each observation distributed as $y_i \sim N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2. After observing IID $y_1, \dots, y_n$ (our *past data* y^{past}) we can compute

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i, \qquad s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Then

$$t := \sqrt{\frac{n}{n+1}} \frac{y_{n+1} - \bar{y}}{s}, \tag{1}$$

where y_{n+1} is the next observation, has Student's t-distribution with $n-1$ degrees of freedom; such a statistic with an invariant (independent of the parameters) distribution is known as a *pivot*.

Once we have a pivot, we can project its distribution onto the next observation. For example, if $\epsilon \in (0,1)$ is our target probability of error, we may predict that the pivot will be in the interval $(F^t_{n-1}(\epsilon/2), F^t_{n-1}(1-\epsilon/2))$, where F^t_{n-1} is the t-distribution function with $n-1$ degrees of freedom. This gives us the prediction

$$y_{n+1} \in \left(\bar{y} + sF^t_{n-1}(\epsilon/2), \bar{y} + sF^t_{n-1}(1-\epsilon/2)\right) \tag{2}$$

for the next observation. This set prediction is valid in the sense that it is correct with probability $1-\epsilon$.

The general scheme of fiducial prediction in Fisher's work is that we combine the past observations y^{past} and future observation y to obtain a pivot U,

$$U := Q(y^{\text{past}}, y).$$

The distribution of U is independent of the parameters, and without loss of generality we can assume that U is uniformly distributed (on $[0,1]$), at least when it is continuous: if not, replace U by $F_U(U)$, where F_U is U's distribution function.

The existence of, and using, a pivot is a hallmark of fiducial inference. Fisher's approach was invariably *direct*, in that he saw the pivot immediately, without any auxiliary tools. Some kind of normalization was used to get rid of the parameters. For example, in the case of (1), we get rid of the location parameter μ by subtracting $\hat{y}$ from the future observation, and we get rid of the scale parameter σ by dividing the difference by s. In future sections, where we consider nonparametric situations, pivots will be derived indirectly via conditioning on sufficient statistics.

If $y \in \mathbb{R}$ and uniformly distributed $Q(y^{\text{past}}, y)$ is increasing in y, $y \mapsto Q(y^{\text{past}}, y)$ will often be a distribution function, and we may call it the *fiducial (predictive) distribution.* In Fisher's example (1), the pivot

$$t = \sqrt{\frac{n}{n+1}} \frac{y - \bar{y}}{s}$$

becomes the fiducial (predictive) distribution (function)

$$Q(y^{\text{past}}, y) := F^t_{n-1}\left(\sqrt{\frac{n}{n+1}} \frac{y - \bar{y}}{s}\right). \tag{3}$$

Therefore, a fiducial predictive distribution can be defined simply as a uniform pivot, provided the pivot *is* a distribution function (i.e., is increasing from 0 to 1 over $(-\infty, \infty)$). Notice that no explicit "fiducial inversion" is needed in this exposition (unless we want prediction sets).

Fisher imposed some further restrictions in both fiducial prediction and non-predictive fiducial inference, emphasizing both continuity (in many publications) and monotonicity (in 1962 in letters to Barnard and Sprott [4, pp. 44, 218–219], for parametric fiducial inference). In a non-predictive context, he writes in his last letter to Barnard in March (?) 1962 (Bennett [4, p. 44] and Barnard [3]; the question mark is Bennett's):

> A pivotal quantity is a function of parameters and statistics, the distribution of which is independent of all parameters. To be of any use in deducing probability statements about parameters, let me add
>
> (a) it involves only one parameter,
>
> (b) the statistics involved are jointly exhaustive for that parameter,
>
> (c) it varies monotonically with that parameter.

The publications where Fisher prominently mentions *continuity* includes his 1956 book [19] ("the observations should not be discontinuous").

Even if the pivot is increasing in the future observation, there is still a possibility that it will fail to be a distribution function. Namely, there is no guarantee that $Q(y^{\text{past}}, -\infty) = 0$ and $Q(y^{\text{past}}, \infty) = 1$ is satisfied for all y^{past}. For example, $Q(y^{\text{past}}, -\infty) > 0$ means that there is a positive mass at $-\infty$. Using the pivot for computing prediction sets is more flexible.

2.2 Validity and efficiency of fiducial prediction

Probabilistic forecasting has become very popular in many application areas, such as economics and weather forecasting (see, e.g., [21, Subsection 1.1]). The most fundamental property of validity is probabilistic calibration, promoted by, among others, Philip Dawid [9] and Tilmann Gneiting (see, e.g., [21, Definition 3(b)]). In general, a random distribution function Q is *probabilistically calibrated* if the random variable

$$Q(y^{\text{past}}, y-) + \tau \left(Q(y^{\text{past}}, y) - Q(y^{\text{past}}, y-) \right)$$

has a uniform distribution on $[0,1]$ provided τ is a uniform random variable that is independent of the past and future observations y^{past} and y. In Fisher's continuous case probabilistic calibration simply means the uniform distribution of $Q(y^{\text{past}}, y)$ and, therefore, is achieved automatically. In fact this simplified understanding of probabilistic calibration will be sufficient in the whole of this chapter, even when we move on to the Dempster–Hill procedure and conformal predictive distributions.

Probabilistic calibration may be the most fundamental notion of validity, but there are several other requirements of this kind, such as marginal calibration [21, Definition 3(a)]. The strongest requirement of validity would be that $Q(y^{\text{past}}, y)$ have the uniform distribution given the σ-algebra generated by the past observations y^{past}. Such strong validity is attainable only with the full knowledge of the stochastic mechanism generating the data (such as under Bayesian assumptions). But we might hope to achieve the uniformity of the distribution of $Q(y^{\text{past}}, y)$ given a smaller σ-algebra $\mathcal{F}$. An interesting example of such conditional probabilistic calibration is due to Peter McCullagh [33, 34]. Our model is that of linear regression

$$y_i = \beta \cdot x_i + \sigma \xi_i, \tag{4}$$

where x_i are fixed vectors in $\mathbb{R}^p$, $\beta \in \mathbb{R}^p$ is a vector of parameters, $\sigma > 0$ is another parameter, and ξ_i are IID noise random variables with a known distribution P (which does not have to be Gaussian). Let $\mathcal{F}$ be the σ-algebra of events invariant under the transformations $(y_1, y_2, \dots) \mapsto (a \cdot x_1 + by_1, a \cdot x_2 + by_2, \dots)$, $a \in \mathbb{R}^p$ being a vector and $b > 0$ a positive number. Then the fiducial predictive distribution constructed by McCullagh is probabilistically calibrated given $\mathcal{F}$. Intuitively, the σ-algebra $\mathcal{F}$ corresponds to forgetting just $p+1$ numbers, where, remember, $p+1$ is the total number of parameters.

Fisher was motivated by a search for fiducial statements that "may claim unique validity" [35, footnote in Fisher's comment]. He was not satisfied with fiducial statements that were merely true (or valid), he wanted them to be the whole truth.

This was the reason for his introduction of various restrictions, such as the use of exhaustive statistics, as in his letter to Barnard. The insistence on unique validity, leading to serious difficulties and numerous paradoxes, was perhaps the main reason for the rejection of fiducial inference by many statisticians. In various extensions of fiducial inference (such as by Hannig [23]) the requirement of uniqueness has been abandoned. To compare various valid procedures we then need additional desiderata. An appealing statement of the overall goal of probabilistic forecasting is due to Gneiting and his co-authors (see, e.g., [21, Section 1.2]):

> Probabilistic forecasting has the general goal of maximizing the sharpness of predictive distributions, subject to calibration.

The idea is that validity is the requirement of agreement between the predictive distributions and the actual observations (they should tell the truth), whereas the sharpness characterizes the predictive distributions only (measures how concentrated they are). More generally, instead of the sharpness we can also talk about the efficiency of predictive distributions (the truth should be informative, even if it is not the whole truth), without insisting that the efficiency does not depend on the actual observations. This leads to Martin and Liu's [31, Section 3.3] *efficiency principle*:

> Subject to the validity constraint, probabilistic inference should be made as efficient as possible.

3 Dempster–Hill procedure

This section makes a first step towards nonparametric prediction. In the nonparametric setting Fisher's requirement of continuity becomes unnecessary if we are allowed to randomize (at least a little).

3.1 Fisher's nonparametric fiducial inference

There are only hints of nonparametric fiducial prediction in Fisher's work [39], but numerous authors trace their ideas in this area to Fisher: see, e.g., Dempster 1963 [10], Lane and Sudderth 1984 [28], Hill 1992 [26], and Coolen 1998 [8].

However, Fisher definitely introduced nonparametric fiducial inference for parameter values. In [17] he traced the idea back to Student. In the case of two observations y_1 and y_2 from $N(\mu, 1)$, the probability that $\mu < y_{(1)}$ is 1/4, the probability that $\mu \in (y_{(1)}, y_{(2)})$ is 1/2, and the probability that $\mu > y_{(2)}$ is 1/4, where $y_{(1)} := \min(y_1, y_2)$ and $y_{(2)} := \max(y_1, y_2)$ are the order statistics. In that paper Fisher extended this statement to an arbitrary sample size n and to the pth quantile μ_p (dropping the Gaussian assumption). Another extension, stated in [17] and in his 1948 paper [18, Subsection 4.VII], is to simultaneous inference about several quantiles μ_p. In view of Remark 1, this may be regarded as a kind of prediction, namely predicting some features of infinitely many future observations.

3.2 Dempster–Hill procedure

Genuine nonparametric fiducial prediction may be said to have started by Jeffreys in his 1932 paper [27]. He discussed a very special case, predicting a third observation (but it was sufficient for his goal of justifying his noninformative prior for the parameter $1/\sigma$ of the Gaussian family $N(\mu, \sigma^2)$). Assuming (implicitly) IID observations $y_1, y_2, \dots$ from a continuous distribution, what is the probability that $y_3 \in (y_1, y_2)$? According to Jeffreys, the answer is easily seen to be one-third. Indeed, all orders of observations in $\{y_1, y_2, y_3\}$ (assumed different) have the same probability, and the middle observation (not the largest and not the smallest) will be y_3 with probability 1/3.

Fisher [14] did not accept Jeffreys's argument, let alone accept it as fiducial. One reason might be that Jeffreys did not use the fiducial language (this was done by Seidenfeld in 1995 [38]) and instead couched it in improper Bayesian terms, which Fisher did not like. But another important reason may be that Jeffreys's argument was blatantly discontinuous. We may consider the rank of y_3 as pivot, but we can't even say that it is approximately continuous (which will often be the case for the general Dempster–Hill procedure), since there are only three observations in Jeffreys's picture. In his later paper [16, p. 51] (1935) Fisher did introduce the device of randomization to turn "a discontinuous distribution, leading to statements of fiducial inequality, into a continuous distribution, capable of yielding exact fiducial statements", but he never reconsidered his rejection of Jeffreys's argument.

Jeffreys's procedure is a special case of what I will call the *Dempster–Hill procedure*; the latter extends the former to the case of any finite number of observations and any interval between adjacent observations. Dempster [10, (5.7)] derives it by modifying Fisher's fiducial argument into what he calls a direct probability argument. Dempster makes it look as if his method is known, but he refers to Wilks [51, Chapter 11], who describes prediction intervals rather than predictive distributions. Hill [24,25] refers to the Dempster–Hill procedure as A_n; this is the statement that, given the data $y_1, \dots, y_n$, the probability that the next observation y falls in $(y_{(i)}, y_{(i+1)})$ is $1/(n+1)$, for each $i = 0, \dots, n$; by definition, $y_{(0)} := -\infty$, and $y_{(n+1)} := \infty$.

As already mentioned, both Dempster and Hill trace their ideas back to Fisher. In his 1992 paper [26] Hill writes:

> Note that for all three of these authors [Student, Fisher, Dempster] the justification for A_n seems to be purely intuitive. Thus none give anything vaguely representing a "proof" for A_n....

Hill [25] also referred to his procedure as *Bayesian nonparametric predictive inference*. This was abbreviated to *nonparametric predictive inference* (or *NPI*) by Frank Coolen [1, 2], which makes it very wide; in fact, this whole chapter belongs to the area of nonparametric predictive inference.

To embed the Dempster–Hill procedure into fiducial prediction, let us define formally a continuous pivot. As usual, we can achieve continuity by using randomization: if τ is distributed uniformly on $[0, 1]$ and is independent of the

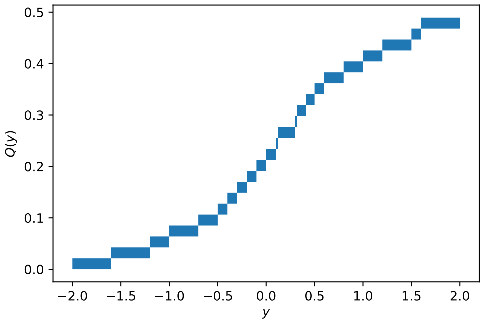

Figure 1: Sample predictive distribution

observations, the *Dempster–Hill pivot*

$$Q(y^{\text{past}}, y, \tau) := \frac{|\{i \mid y_i < y\}| + \tau + \tau\,|\{i \mid y_i = y\}|}{n+1} \tag{5}$$

(i ranging over $\{1, \dots, n\}$, where n is the number of past observations) is distributed uniformly on $[0, 1]$. The last addend in the numerator in (5) takes care of possible ties, but even in the continuous case we need the penultimate addend to achieve the uniformity of the distribution. For large n, the distribution will be approximately uniform even if we ignore τ (say, set $\tau := 0$ or $\tau := 1$).

As function of y, $Q(y^{\text{past}}, y, \tau)$ as defined by (5) can be considered to be a predictive distribution; for a large number n of past observations it is an approximate distribution function. An example of such a predictive distribution is given in Figure 1. The horizontal axis is y, and the vertical axis gives the interval of the values $Q(y^{\text{past}}, y, \tau)$ for all $\tau \in [0, 1]$, where the past observations y^{past} are fixed.

3.3 Validity and efficiency of the Dempster–Hill procedure

As other fiducial procedures, the Dempster–Hill procedure is automatically valid in that the pivot (5) is distributed uniformly on $[0, 1]$. As for its efficiency, we do not really need it in this section for the purpose of comparison, since the Dempster–Hill procedure is uniquely determined (efficiency will become important in the next section, where this procedure will be greatly generalized). Still the efficiency of the Dempster–Hill procedure in the sense of the closeness of the empirical distribution function to the true data-generating distribution function has been extensively studied under the rubric of empirical processes (see, e.g., [41]).

4 Conformal predictive distributions

As mentioned in Section 1, a serious limitation of the Dempster–Hill procedure is that it does not cover typical problems of machine learning, such as regression and classification, where the observations are pairs (x, y) and not just the labels y; the task is to predict the unknown label y of a test object x. The procedure has been criticized on this account; e.g., Genest and Kalbfleisch [20] say in their response to Berliner and Hill [5]: "To be truly useful, however, the methods need extension to regression models with unknown regression parameters".

In most of the rest of this paper our statistical model will be the one standard in machine learning and much of nonparametric statistics: the observations are assumed to be IID (and nothing more is assumed). We will refer to it as the *IID model* (or *hypothesis of randomness*). There are several related ways, treated in Subsections 4.1 and 4.2, to construct a *conformal pivot*, a random variable that is distributed uniformly on $[0, 1]$ (under the IID model) and can be used as a predictive distribution (*conformal predictive distribution*).

While the Dempster–Hill distributions are uniquely determined, there is a huge variety of conformal predictive distributions. For example, outputs of many prediction algorithms in machine learning and statistics can be turned (usually in more than one way) into predictive distributions.

In this section we consider the case of regression, where the labels y are real numbers, $y \in \mathbb{R}$; the objects x are elements of a measurable space (often Euclidean space $\mathbb{R}^p$).

4.1 Full conformal predictive distributions

A *conformity measure* is a measurable function A mapping finite sequences of observations $(z_1, \dots, z_l)$ to *conformity scores* $(\alpha_1, \dots, \alpha_l)$ (sequences of real numbers of the same length) that is equivariant: for any l and any permutation π of $\{1, \dots, l\}$,

$$A(z_1, \dots, z_l) = (\alpha_1, \dots, \alpha_l) \Longrightarrow A\left(z_{\pi(1)}, \dots, z_{\pi(l)}\right) = \left(\alpha_{\pi(1)}, \dots, \alpha_{\pi(l)}\right).$$

The origin of the terminology is that α_i measures how well z_i conforms to the other observations among $z_1, \dots, z_l$. In many cases we can build A on top of some *underlying* algorithm. A simple example is where the conformity scores are defined by

$$\alpha_i := y_i - \hat{y}_i, \tag{6}$$

where $\hat{y}_i$ is the prediction for the label of x_i produced by a prediction algorithm trained on $z_1, \dots, z_l$. Another example, that might better agree with the terminology, is

$$\alpha_i := -\left|y_i - \hat{y}_i\right|. \tag{7}$$

The *conformal pivot* determined by a conformity measure A is

$$Q(z^{\text{past}}, (x, y), \tau) := \frac{\left|\{i \mid \alpha_i^y < \alpha^y\}\right| + \tau + \tau \left|\{i \mid \alpha_i^y = \alpha^y\}\right|}{n+1}, \tag{8}$$

where z^{past} is a sequence $z_1, \ldots, z_n$ of observations of length n (our *training sequence*), $i = 1, \ldots, n$,

$$(\alpha_1^y, \ldots, \alpha_n^y, \alpha^y) := A(z_1, \ldots, z_n, (x, y)), \tag{9}$$

and $\tau \in [0,1]$. For a proof that the distribution of $Q(Z^{\mathrm{past}}, Y)$ is uniform on $[0,1]$, provided τ is distributed uniformly on $[0,1]$ and independent of the observations, see, e.g., [46, Proposition 2.4].

Notice that the direct implementation of (8) involves heavy computations, even if we limit ourselves to a finite grid of values for the label y (which can be done rigorously in some cases [6]). If we have several test objects x for which we would like to have predictive distributions, we need to recompute all the $n+1$ conformity scores (9) for each possible value of y and for each test object x.

We say that (8) is a *conformal predictive system* if:

- $Q(z^{\mathrm{past}}, (x, y), \tau)$ is an increasing function of $y \in \mathbb{R}$ (by definition it is an increasing and linear function of τ);
- for each training sequence z^{past} and each test object x,

$$\lim_{y \to -\infty} Q(z^{\mathrm{past}}, (x, y), 0) = 0 \tag{10}$$

$$\lim_{y \to \infty} Q(z^{\mathrm{past}}, (x, y), 1) = 1. \tag{11}$$

In this case we will say that $Q(z^{\mathrm{past}}, (x, y), \tau)$ as function of y is a *conformal predictive distribution.*

With a careful choice of the conformity measure A the conformal pivot will be an efficiently computable conformal predictive system. A trivial example is where A strips a sequence of observations of the objects:

$$A\left((x_1, y_1), \ldots, (x_l, y_l)\right) := (y_1, \ldots, y_l)$$

(or, alternatively, we have no objects and A is the identity function). In this case we obtain the Dempster–Hill pivot (5).

Our first nontrivial example is the *Least Squares Predictive Machine* (LSPM). This is the case of the conformity measure (6), where $\hat{y}_i$ is the Least Squares (LS) estimate of the label of ith object, which is assumed to be a vector in $\mathbb{R}^p$. The answer to the question of monotonicity of the conformal pivot in y depends on the details of the definition of the LS estimate [50, Section 3]:

- If $\hat{y}_i$ is the LS estimate based on all of $z_1, \ldots, z_l$ (so that α_i is the "full residual"), monotonicity can be violated (albeit only in pathological cases of a high-leverage test object x).
- If $\hat{y}_i$ is the LS estimate based on $z_1, \ldots, z_l$ with z_i removed (so that α_i is the "deleted residual"), monotonicity can be violated (albeit only in pathological cases of high-leverage training objects in z^{past}).

- But in an intermediate situation (of a "studentized residual" α_i), monotonicity always holds.

Therefore, only the studentized LSPM is a conformal predictive system [50, Proposition 5]. Its predictions (conformal predictive distributions) can be computed in time $O(n^2 + kn)$, where n is the length of the training sequence and k is the length of the test sequence (preprocessing takes time $O(n^2)$ and then processing each test object takes time $O(n)$).

Fisher's idea of using exhaustive statistics to obtain uniquely valid predictive distributions does not work at all in the case of conformal predictive systems. Even when constructing a conformity measure from an underlying algorithm, we have plenty of choice, and arguably some underlying algorithms available in machine learning may involve an element of intelligence (such as artificial neural networks). The variety of conformal predictive systems is real, all of them are valid (in the sense of being probabilistically calibrated), and we need some notion of efficiency to distinguish between them. Fisher's whole truth [35, footnote in Fisher's comment] is usually not attainable.

A useful notion of efficiency for a conformal predictive system is how close the predictive distributions that it outputs are to the true predictive distributions. One way to answer this question is to impose strong assumptions on the data-generating distribution. In the case of the LSPM, it is natural to assume the model (4), where ξ_i are IID standard Gaussian random variables, in which case we will refer to it as the *Gauss linear model.* Under this parametric assumption, we can construct nearly optimal, or *oracular*, predictive distributions, and it turns out that, under natural regularity conditions, the conformal predictive distributions output by the LSPM and the oracular predictive distributions approach each other at the usual rate $O(n^{-1/2})$: see [50], Theorems 2–4.

Why are such efficiency results useful? Can't we just use the oracular predictive distributions? This would be risky in situations where we are willing to accept the nonparametric IID model but reluctant to accept the parametric Gauss linear model. We have validity under the IID model, so the conformal predictive distributions will not be misleading (at worst they will be useless). But if we are lucky and the Gauss linear model also holds, we will quickly approach the true predictive distributions.

The LSPM provides an example of the procedure of *conformalization.* We take a point prediction procedure, Least Squares, that is optimal, in some sense, under the Gauss linear model. Then we pass it through the "conformal machine" (8)–(9) to obtain guaranteed validity under the IID model.

A disadvantage of the LSPM is that its underlying prediction algorithm, Least Squares, is linear, and therefore, the LSPM is likely to be efficient only when the true relation between the labels and objects is linear. A method of extending linear methods to nonlinear situations that is standard in machine learning is to apply a feature mapping to the objects, which corresponds to replacing the dot product by another kernel in the object space. The LSPM can be "kernelized" in this way while maintaining its computational efficiency [47]. Pre-processing a training sequence of length n takes, asymptotically, the same time as inverting an $n \times n$ matrix ($O(n^3)$ with the schoolbook method) and, after that, processing a test object takes time $O(n^2)$.

If we are only interested in a weak form of asymptotic efficiency of conformal predictive distributions, their universal consistency, there is no need to adopt narrow statistical models such as the Gauss linear model, and it holds under the IID model for a suitable conformity measure [42, Theorem 28]. Universally consistent conformal predictive systems can be built on top of many classical universally consistent algorithms, such as nearest neighbours, and their construction adapts standard arguments for universal consistency in classification and regression [11, 22, 40].

An advantage of predictive distributions over other, less complete, forms of prediction is that in decision-making problems predictive distributions can be combined with a utility function and the expected utility maximization principle to obtain optimal decisions. Using the predictive distributions produced by a universally consistent predictive system leads, under natural conditions, to decisions whose regret tends to 0 in probability under the IID model [44, Theorem 3].

4.2 Split-conformal and cross conformal predictive distributions

As mentioned earlier, conformal predictive distributions, also called *full* conformal predictive distributions, are difficult to compute apart from a small number of conformity measures that are particularly mathematically tractable. It may also be difficult to check that a full conformal pivot is an increasing function of the label. Both problems simplify drastically if we have enough data.

A *split-conformity measure* is a measurable function A that maps an observation z and a sequence of observations $z_1, \dots, z_l$ into a *conformity score* $A(z; z_1, \dots, z_l) \in \mathbb{R}$; intuitively, the conformity score shows how similar z is to the elements of the sequence $z_1, \dots, z_l$. An example is again given by (6), which in our current notation becomes

$$A((x, y); z_1, \dots, z_l) := y - \hat{y}, \tag{12}$$

where $\hat{y}$ is the prediction for the label of x produced by a prediction algorithm trained on $z_1, \dots, z_l$.

Let us divide the training sequence $z^{\text{past}} = (z_1, \dots, z_n)$ into two parts:

- the *training sequence proper*, $z_1, \dots, z_m$, of length m,
- and the *calibration sequence*, $z_{m+1}, \dots, z_n$, of length $n - m$.

The *split-conformal pivot* for a test object x is

$$Q(z^{\text{past}}, (x, y), \tau) := \frac{|\{i \mid \alpha_i < \alpha\}| + \tau + \tau\,|\{i \mid \alpha_i = \alpha\}|}{n - m + 1}, \tag{13}$$

where i ranges over $m + 1, \dots, n$ and

$$\alpha_i := A(z_i; z_1, \dots, z_m), \quad \alpha := A((x, y); z_1, \dots, z_m). \tag{14}$$

It is easy to construct computationally efficient split-conformal pivots. For example, if the split-conformity measure is (12), there is no need to retrain the underlying prediction algorithm for each calibration and test object.

It is also easy to ensure that the split-conformal pivot $Q(z^{\text{past}}, (x, y), \tau)$ is an increasing function of y: it is enough to require that $A((x, y); \dots)$ is increasing in y. If, in addition, the convex closure of $A((x, \mathbb{R}); \dots)$ does not depend on x, the split-conformal pivot will satisfy (10)–(11), in which case it is called a *split-conformal predictive system* and its output is called a *split-conformal predictive distribution*. For details, see [48, Section 3].

Split-conformal predictive systems are computationally efficient but may lose predictive efficiency as compared with full conformal predictive systems, which use the full training sequence as both training sequence proper and calibration sequence. A natural way out is to divide the training sequence into a number of folds (as in cross-validation), use each fold in turn as calibration sequence, and combine the corresponding split-conformal predictive distributions. The resulting *cross-conformal predictive distributions* [48, Section 4] lose guaranteed validity but are well-calibrated in practice (in the absence of excessive randomization in the underlying algorithm [30, 48]). Under a suitable choice of a split-conformity measure, split-conformal and cross-conformal predictive distributions are universally consistent [48, Theorem 7.2].

One limitation of the conformity measure (6) and the split-conformity measure (12) is that they implicitly assume homoscedasticity. As a result, the predictive distributions for all test objects are, essentially, horizontal translates of each other (even though the shape of the predictive distribution can be very adaptive, as is the case for the kernelized LSPM [47] with a universal kernel). We can generalize, e.g., (12) to

$$A((x, y); z_1, \dots, z_l) := \frac{y - \hat{y}}{s},$$

where $s > 0$ is an estimate, based on $z_1, \dots, z_l$ and x, of the accuracy of the prediction $\hat{y}$ for the label of x. This will still produce essentially the same shape of the predictive distributions, and we will just add a scale parameter.

The split-conformal (and by extension cross-conformal) method allows a much more radical solution. As split-conformity measure we can take

$$A((x, y); z_1, \dots, z_m) := F(y),$$

where F is a predictive distribution function, not required to satisfy any properties of validity under the IID model, for the label of x computed from $z_1, \dots, z_m$ as training sequence. Examples of possible F are the Nadaraya–Watson procedure [36], random forest regression, procedures in the TensorFlow probability module, and various Gaussian processes. See [49] for experimental results. This gives another example of conformalization; the application of the split-conformal machine (13)–(14) ensures the validity under the IID model.

5 Conformal prediction

The requirement of monotonicity of conformal pivots in the label is restrictive and dropping it greatly extends the application area of the conformal method. For example, it then extends immediately to the case of classification, where the label is only allowed to take values in a finite set (which does not even have to be ordered). Instead of probability distributions we only have p-values, (8) or (13) (they are valid p-values since they are distributed uniformly on $[0, 1]$). In order for p-values to become probabilities (to form a distribution function, at least approximate) we need to impose strict discipline, first of all monotonicity, but even mere p-values have many important uses.

The step from conformal predictive distributions to p-values is somewhat similar to the steps from fiducial distributions for parameter values to Neyman's confidence intervals [35] and from fiducial predictive distributions to prediction intervals (e.g., from (3) to (2)). The theory simplifies and generalizes. We lose something but enough survives. Our goal is still to design prediction algorithms that are automatically valid and, under the constraint of validity, are as efficient as possible, in various formal and informal senses.

In this section we will concentrate on full conformal prediction, although many ideas can be extended to split-conformal and cross-conformal prediction. We will refer to the system of p-values (8) (without further requirements such as monotonicity) as *conformal predictor*.

5.1 Validity of conformal prediction

We know that conformal predictors are valid in the sense of (8) being distributed uniformly on $[0, 1]$. This allows us to compute prediction sets with a guaranteed probability of error; e.g., defining the prediction set as the set of all labels leading to a p-value greater than a fixed significance level $\epsilon \in (0, 1)$ ensures the probability of error ϵ. (In order to obtain a bounded prediction interval we should use (7) rather than (6).)

This property of validity can be strengthened in the online prediction protocol. In this protocol, the observations $z_n = (x_n, y_n)$, $n = 1, 2, \dots,$ arrive sequentially, and at the nth step we predict the label y_{n+1} of x_{n+1} given x_{n+1} itself and the previous observations $z^{\text{past}} := (z_1, \dots, z_n)$. It turns out that the p-values $p_n := Q(z^{\text{past}}, (x_{n+1}, y_{n+1}), \tau_n)$ are independent provided the uniform random numbers τ_n are independent between themselves and of the observations [46, Proposition 2.4]. Therefore, by the law of large numbers, the guaranteed probability of error for the prediction set will be reflected in the frequency of errors.

5.2 Efficiency of conformal prediction and training conformal predictors

The question of measuring the efficiency of conformal predictors is difficult, and during the short history of conformal prediction several unsuitable measures have been used. In the case of classification, a natural desideratum for a criterion of efficiency is that the true conditional probability of the label y of an object x should be an ideal conformity measure as evaluated by that criterion. This desideratum is formalized in [45, Section 4] and criteria of efficiency that satisfy it are called *probabilistic*.

An example of a probabilistic criterion of efficiency is evaluating the quality of a conformal predictor by the average p-value computed for a test object x with postulated label y, with x ranging over the test sequence and y ranging over all possible labels. This criterion does not depend on the true labels of the test objects, and so agrees with the notion of sharpness in Gneiting et al.'s version of the efficiency principle. The rationale behind it is that we would like the p-values for labels different from the true one to be as small as possible (the p-value for the true label is not under our control; because of the validity, it is distributed uniformly on $[0,1]$).

The disadvantage of this criterion of efficiency is that the average contains a lot of noise created by the true label. Another criterion of efficiency ("observed fuzziness") is defined in the same way except that (x,y) ranges over all pairs where x is a test object and y is a label different from x's true label. A disadvantage is that this criterion does depend on the true labels of the test objects, but the average now involves less noise; in the ideal case it will be close to zero.

The original goal of conformal prediction was to complement the predictions output by state-of-the-art algorithms of mainstream machine learning by provably valid measures of their accuracy and reliability [46]. Recently, first steps have been made in designing conformal predictors *ab initio* [7]; we can train conformal predictors directly by minimizing the observed fuzziness on a calibration sequence (or in the framework of cross-validation).

5.3 Conformal martingales and online hypothesis testing

An interesting application of the validity of conformal predictors, including the independence of p-values in the online protocol, is to testing the hypothesis of randomness online [43]. Suppose an IID sequence of observations at some point ceases to be IID. How can we detect the change and, for example, raise an alarm as soon as possible after the change occurs? (A special case is where the hypothesis of randomness never holds, which corresponds to testing randomness.)

Existing methods can deal with the case where the pre-change distribution is known. In combination with conformal prediction, we can apply these methods to the uniformly distributed and independent p-values. This gives us procedures for change detection in the nonparametric situation where the pre-change distribution is only known to belong to the IID model.

5.4 Repetitive structures

Conformal prediction, as discussed so far, produces predictions that are valid under the IID model. However, it can be generalized in a straightforward manner to arbitrary *repetitive structures*, as introduced by Per Martin-Löf [32] and further developed by Lauritzen [29].

The IID model is an important repetitive structure, but another example is provided by the Gaussian IID model [46, Section 8.5]. Interestingly, Fisher's fiducial distribution (3) coincides with the conformal predictive distribution for the Gaussian IID model and trivial conformity measure (the identity function). This implies, in particular, that Fisher's fiducial distribution (3) satisfies a stronger property of validity than probabilistic calibration; we can also say that the random variables $Q(y^{\text{past}}, y_n)$ are independent under the Gaussian IID model (as defined at the beginning of Subsection 2.1) in the online prediction protocol. Therefore, the prediction intervals (2) fail to cover the true future observations with probability ϵ independently.

Yet another example of repetitive structure is provided by graphical models [13]. The testing methods described in Subsection 5.3 are applicable to any repetitive structure, including graphical models.

6 Conclusion

The key message of this chapter is that there are ways to extend ideas of fiducial prediction to nonparametric settings, including those useful in regression problems. The notion of validity known as probabilistic calibration is attained automatically, but there are other notions of validity that deserve to be explored in modern approaches to fiducial prediction. Efficiency is never automatic and is an extensive and largely unexplored area of research. This includes developing suitable criteria of efficiency for predictive systems, fiducial and conformal.

Acknowledgments

This chapter is essentially a written version of my talk at the Sixth BFF meeting in 2019 (29 April 2019, Duke University, Durham, NC). Out of the six BFF meetings so far I have attended two (Fourth in 2017 and Sixth in 2019), and both deeply affected this chapter. I am grateful to the organizers of and participants in both meetings. The work on this paper has been partially supported by Amazon, Astra Zeneca, and Stena Line.

References

[1] Thomas Augustin and Frank P. A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124:251–272, 2004.

[2] Thomas Augustin, Gero Walter, and Frank P. A. Coolen. Statistical inference. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, chapter 7, pages 135–189. Wiley, Chichester, 2014.

[3] George A. Barnard. Fisher's contributions to mathematical statistics. *Journal of the Royal Statistical Society A*, 126:162–166, 1963.

[4] J. H. Bennett, editor. *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher.* Clarendon Press, Oxford, 1990.

[5] L. Mark Berliner and Bruce M. Hill. Bayesian nonparametric survival analysis. *Journal of the American Statistical Association*, 83:772–779, 1988.

[6] Wenyu Chen, Kelli-Jean Chun, and Rina Foygel Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7:e173, 2018.

[7] Nicolo Colombo and Vladimir Vovk. Training conformal predictors. *Proceedings of Machine Learning Research*, 128:55–64, 2020. COPA 2020.

[8] Frank P. A. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Statistics and Probability Letters*, 36:349–357, 1998.

[9] A. Philip Dawid. Statistical theory: the prequential approach (with discussion). *Journal of the Royal Statistical Society A*, 147:278–292, 1984.

[10] Arthur P. Dempster. On direct probabilities. *Journal of the Royal Statistical Society B*, 25:100–110, 1963.

[11] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

[12] Bradley Efron. R. A. Fisher in the 21st century. *Statistical Science*, 13:95–122, 1998.

[13] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Hypergraphical conformal predictors. *International Journal on Artificial Intelligence Tools*, 24(6):1560003, 2015. COPA 2013 Special Issue.

[14] Ronald A. Fisher. The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London A*, 139:343–348, 1933.

[15] Ronald A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935.

[16] Ronald A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98:39–54, 1935.

[17] Ronald A. Fisher. "Student". *Annals of Eugenics*, 9:1–9, 1939.

[18] Ronald A. Fisher. Conclusions fiduciaires. *Annales de l'Institut Henry Poincaré*, 10:191–213, 1948.

[19] Ronald A. Fisher. *Statistical Methods and Scientific Inference.* Hafner, New York, third edition, 1973. First edition: 1956.

[20] Christian Genest and Jack Kalbfleisch. Bayesian nonparametric survival analysis: comment. *Journal of the American Statistical Association*, 83:780–781, 1988. Comment on [5].

[21] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

[22] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York, 2002.

[23] Jan Hannig. On generalized fiducial inference. *Statistica Sinica*, 19:491–544, 2009.

[24] Bruce M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63:677–691, 1968.

[25] Bruce M. Hill. De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In Dennis V. Lindley, José M. Bernardo, Morris H. DeGroot, and Adrian F. M. Smith, editors, *Bayesian Statistics 3*, pages 211–241. Oxford University Press, Oxford, 1988.

[26] Bruce M. Hill. Bayesian nonparametric prediction and statistical inference. In Prem K. Goel and N. Sreenivas Iyengar, editors, *Bayesian Analysis in Statistics and Econometrics*, volume 75 of *Lecture Notes in Statistics*, chapter 4, pages 43–94. Springer, New York, 1992.

[27] Harold Jeffreys. On the theory of errors and least squares. *Proceedings of the Royal Society of London A*, 138:48–55, 1932.

[28] David A. Lane and William D. Sudderth. Coherent predictive inference. *Sankhyā A*, 46:166–185, 1984.

[29] Steffen L. Lauritzen. *Extremal Families and Systems of Sufficient Statistics*, volume 49 of *Lecture Notes in Statistics.* Springer, New York, 1988.

[30] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. On the calibration of aggregated conformal predictors. *Proceedings of Machine Learning Research*, 60:154–173, 2017. COPA 2017.

[31] Ryan Martin and Chuanhai Liu. *Inferential Models: Reasoning with Uncertainty*, volume 147 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, FL, 2016.

[32] Per Martin-Löf. Repetitive structures and the relation between canonical and microcanonical distributions in statistics and statistical mechanics. In Ole Barndorff-Nielsen, Preben Blæsild, and Geert Schou, editors, *Proceedings of Conference on Foundational Questions in Statistical Inference*, pages 271–294, Aarhus, 1974.

[33] Peter McCullagh. Fiducial prediction. Manuscript, `http://www.stat.uchicago.edu/~pmcc/reports/fiducial.pdf`, 2004.

[34] Peter McCullagh, Vladimir Vovk, Ilia Nouretdinov, Dmitry Devetyarov, and Alex Gammerman. Conditional prediction intervals for linear regression. In *Proceedings of the Eighth International Conference on Machine Learning and Applications (ICMLA 2009)*, pages 131–138, 2009. Available from `http://www.stat.uchicago.edu/~pmcc/reports/predict.pdf`.

[35] Jerzy Neyman. On the two different aspects of the representative method: the method of stratifiedsampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society*, 97:558–625, 1934. Fisher's comment: 614–619.

[36] Murray Rosenblatt. Conditional probability density and regression estimators. In Paruchuri R. Krishnaiah, editor, *Multivariate Analysis II*, pages 25–31. Academic Press, New York, 1969.

[37] Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge, 2016.

[38] Teddy Seidenfeld. Jeffreys, Fisher, and Keynes: predicting the third observation, given the first two. In Allin F. Cottrell and Michael S. Lawlor, editors, *New Perspectives on Keynes*, pages 39–52. Duke University Press, Durham, NC, 1995.

[39] Teddy Seidenfeld. Personal communication. Fourth BFF meeting, May 2017.

[40] Charles J. Stone. Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5:595–645, 1977.

[41] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[42] Vladimir Vovk. Universally consistent conformal predictive distributions. *Proceedings of Machine Learning Research*, 105:105–122, 2019. COPA 2019.

[43] Vladimir Vovk. Testing randomness online. Technical Report arXiv:1906.09256 [math.PR], arXiv.org e-Print archive, March 2020.

[44] Vladimir Vovk and Claus Bendtsen. Conformal predictive decision making. *Proceedings of Machine Learning Research*, 91:52–62, 2018. COPA 2018.

[45] Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman. Criteria of efficiency for set-valued classification. *Annals of Mathematics and Artificial Intelligence*, 81:21–46, 2017.

[46] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[47] Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Conformal predictive distributions with kernels. In Lev Rozonoer, Boris Mirkin, and Ilya Muchnik, editors, *Braverman's Readings in Machine Learning: Key Ideas from Inception to Current State*, volume 11100, pages 103–121. Springer, Cham, Switzerland, 2018.

[48] Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Computationally efficient versions of conformal predictive distributions. *Neurocomputing*, 397:292–308, 2020. COPA 2018 Special Issue.

[49] Vladimir Vovk, Ivan Petej, Paolo Toccaceli, Alex Gammerman, Ernst Ahlberg, and Lars Carlsson. Conformal calibration. *Proceedings of Machine Learning Research*, 128:84–99, 2020. COPA 2020.

[50] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Minge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474, 2019. COPA 2017 Special Issue.

[51] Samuel S. Wilks. *Mathematical Statistics*. Wiley, New York, 1962.

[52] Minge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81:3–39, 2013.