Model-free generalized fiducial inference

Jonathan Williams

North Carolina State University

Centre for Advanced Study, Norwegian Academy of Science and Letters

https://jonathanpw.github.io

3 May 2023

Goals for the talk

→ Introduce the idea of the conformal prediction (CP) algorithm

→ Describe why this idea is so important

→ Explain how the CP framework lacks versatility

→ Propose a resolution relying on:

- → Generalized fiducial (GF) inference
- → Imprecise probability calculus

Conformal prediction algorithm

CP is a relatively general-purpose approach to uncertainty quantification for prediction problems in machine intelligence

Desirable properties:

$\rightarrow$ Finite sample control of type 1 error rates for predictions

$\rightarrow$ Can be built on top of virtually any machine learning algorithm

$\rightarrow$ Requires only weak assumptions on data generating mechanisms

Conformal prediction algorithm

Input: Nonconformity measure $\Psi : \wr\mathbb{R}^n\int \times \mathbb{R} \to \mathbb{R}$, measurable;
Exchangeable examples $y_1, \ldots, y_n$, and an arbitrary y;
Significance level $\alpha \in (0, 1)$

Output: Logical value;
1 indicates that $y_1, \ldots, y_n, y$ are exchangeable;
0 else

Denote $y_{n+1} := y$;
for $i \in \{1, \ldots, n+1\}$ **do**
  Compute $t_i(y_i) = \Psi(y_{-i}^{n+1}, y_i)$;
end
Set $p_{n+1} := \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{t_i(y_i) \geq t_{n+1}(y_{n+1})\}$;
return $1\{p_{n+1} > \alpha\}$;

Conformal prediction algorithm

Finite sample control of type 1 errors:

Let $\{\Gamma_n^\alpha : \alpha \in (0,1)\}$ be a family of CP sets for Y_{n+1} constructed from observed data $Y_1, \ldots, Y_n$

The set Γ_n^α is comprised of the values y_{n+1} such that $p_{n+1} > \alpha$

Theorem

If $Y_1, \ldots, Y_n, Y_{n+1} \sim P$ are exchangeable, then the CP sets are valid *in the sense that for all (α, n, P),*

$$P\left(\Gamma_n^\alpha \ni Y_{n+1}\right) \geq 1 - \alpha$$

Remark: It suffices to assume only that $t_1(Y_1), \ldots, t_n(Y_n), t_{n+1}(Y_{n+1})$ are exchangeable

Why does *validity* matter?

Validity matters because accountability and reliability in uncertainty quantification matters — in the same way that:

$\rightarrow$ Financial reporting standards exist to facilitate security valuation of insurance companies

$\rightarrow$ Building codes and standards exist to ensure the integrity of engineering and construction practices

★ There is no generally accepted standard of accountability of stated uncertainties in all of data science

Why does *validity* matter?

At the American Society of Clinical Oncology conference in Chicago last June:

A new liquid biopsy can help identify the need for adjuvant therapy in stage II colon cancer

$\rightarrow$ thereby avoiding post-operative chemotherapy,

$\rightarrow$ which for bowel cancer can cause peripheral neuropathy

Why does *validity* matter?

Suppose the results of this biopsy is 95% confidence reported . . .

How is this confidence defined?

→ Is it defined as the reported error on a test set?

→ Is it a Bayesian posterior probability?

→ Is it some sort of averaging over a collection of predictions?

★ All are widely accepted notions of *confidence*

★ Varying (if any) guarantees for how the algorithm might perform on future data

Conformal prediction algorithm

The CP algorithm provides valid, general purpose uncertainty quantification, but lacks versatility:

$\rightarrow$ Does not prescribe how to quantify the degree to which a data set provides evidence in support of (or against) an arbitrary event from a general class of events.

e.g., within the Bayesian paradigm, the degree to which a data set provides evidence in support of (or against) an event is quantified by the posterior probability of the event, for any *measurable* event.

Conformal predictions from generalized fiducial inference

Approach:

$\rightarrow$ Construct CP sets from the GF statistical framework

$\rightarrow$ Motivated by a rank-based data generating *association*

$\rightarrow$ Apply imprecise probability tools

$\rightarrow$ e.g., belief/plausibility functions or lower/upper probabilities

$\rightarrow$ Approximate imprecise GF distribution by a precise distribution

Generalized fiducial inference

Assume a data generating model for some Y:

$$Y = G(U, \theta),$$

where

$\rightarrow$ G is a deterministic function

$\rightarrow$ θ is an unknown population parameter(s) of interest

$\rightarrow$ U has a completely known and fully specified distribution

Generalized fiducial inference

Definition (Hannig et al., 2016)

Given an observed data set $y_1, \ldots, y_n$ generated independently from $Y = G(U, \theta)$, a GF distribution on a parameter space Θ is defined as the weak limit,

$$\lim_{\epsilon \to 0} \left\{ \operatorname*{argmin}_{\vartheta \in \Theta} \sum_{i=1}^{n} \|y_i - G(U_i, \vartheta)\|^2 \;\middle|\; \min_{\vartheta \in \Theta} \sum_{i=1}^{n} \|y_i - G(U_i, \vartheta)\|^2 \leq \epsilon \right\}$$

For discrete-valued data:

$\rightarrow$ The limit $\epsilon \to 0$ reduces to setting $\epsilon = 0$

$\rightarrow$ Leads to an imprecise probability distribution over Θ

Generalized fiducial inference

e.g., for $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$,

$$Y_i = \underbrace{1\{U_i < \theta\}}_{=G(U_i,\theta)},$$

where $U_1, \ldots, U_n \overset{\text{iid}}{\sim} \text{uniform}(0,1)$

Leads to the (imprecise) GF distribution for θ:

$$\left\{ (U^\star_{(\sum_1^n y_i)}, U^\star_{(1+\sum_1^n y_i)}] \;:\; \text{where } U^\star_1, \ldots, U^\star_n \overset{\text{iid}}{\sim} \text{uniform}(0,1) \right\} \subseteq \Theta,$$

$\rightarrow$ $U^\star_{(k)}$ denotes the k-th order statistic

Conformal predictions from generalized fiducial inference

Suppose $Y_1, \ldots, Y_{n+1}$ are exchangeable and continuous

A *model-free* data generating *association* for Y_{n+1}:

$$\mathrm{rank}(t_{n+1}(Y_{n+1})) = V \sim \mathrm{uniform}\{1, \ldots, n+1\},$$

where

$\rightarrow$ $t_i(Y_i) := \Psi(Y_{-i}^{n+1}, Y_i)$ is a nonconformity score

$\rightarrow$ $\mathrm{rank}(t_{n+1}(Y_{n+1}))$ denotes position in ascending order

Conformal predictions from generalized fiducial inference

Using the rank-based data association,

The (imprecise) GF distribution of the to-be-predicted value y_{n+1} is a distribution over the random sets:

$$A_n(V^\star) := \underset{y}{\operatorname{argmin}}\Big\{\big|\mathrm{rank}(t_{n+1}(y)) - V^\star\big|\Big\} = \Big\{y \ : \ \mathrm{rank}(t_{n+1}(y)) = V^\star\Big\}$$

where $V^\star \sim \mathrm{uniform}\{1, \ldots, n+1\}$

Conformal predictions from generalized fiducial inference

Illustration of the imprecise GF distribution of y_{n+1}

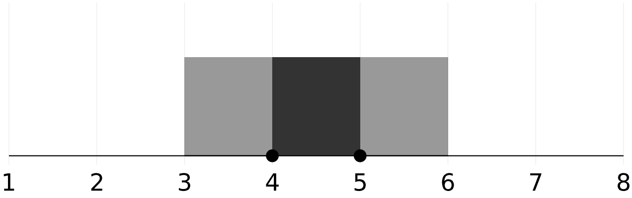

Figure: Hypothetical observed univariate data with $y_1 = 4$, $y_2 = 5$, and $n = 2$. With nonconformity measure $\Psi(y_{-i}^{n+1}, y_i) := |\text{mean}(y_{-i}^{n+1}) - y_i|$

$\rightarrow A_n(1) =$ black region

$\rightarrow A_n(2) =$ grey region

$\rightarrow A_n(3) =$ white region

Conformal predictions from generalized fiducial inference

With respect to the discrete uniform measure μ,

$$\mu\big(A_n(V^\star) \ni y_{n+1}\big) = \mu\bigg(V^\star = \mathrm{rank}(t_{n+1}(y_{n+1}))\bigg) = \frac{1}{n+1}$$

i.e., $A_n(1), \ldots, A_n(n+1)$ are all equally likely to contain y_{n+1}

How to construct a prediction set with at least $1-\alpha$ level confidence?

$\rightarrow$ Accumulate k of the prediction sets such that

$$\frac{k}{n+1} \geq 1 - \alpha$$

Conformal predictions from generalized fiducial inference

Accordingly, for $k \in \{1, \ldots, n+1\}$,

$$\Omega_n(k) := \bigcup_{1 \le v \le k} A_n(v) = \left\{ y \ : \ \text{rank}(t_{n+1}(y)) \le k \right\}$$

is a $\frac{k}{n+1}$ level prediction set

$\rightarrow f_n(y) := \mu\big(\Omega_n(V^\star) \ni y\big)$ is a conformal transducer

Conformal predictions from generalized fiducial inference

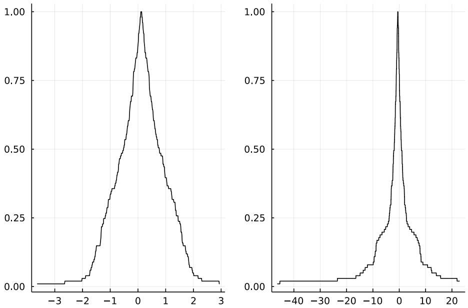

Figure: $f_n(y) = \mu\big(\Omega_n(V^\star) \ni y\big)$; the left and right plots are based on simulated samples of $n = 100$ realizations from the standard Gaussian and standard Cauchy distribution, respectively

$\Upsilon_n^\alpha := \{y : f_n(y) > \alpha\}$ is a CP set, i.e., *valid* in the sense that

$$P(\Upsilon_n^\alpha \not\ni Y_{n+1}) = P\Big(f_n(Y_{n+1}) \leq \alpha\Big) \leq \alpha$$

Model-free generalized fiducial inference

To summarize:

$\to$ The sets $A_n(1), \ldots, A_n(n+1)$ are the atoms of the random set GF predictive distribution

$\to$ Each set has $\frac{1}{n+1}$ GF probability

$\to$ These sets can be arranged to construct any CP set

$\to$ Further, for any assertion B – *not necessarily a CP set*:

$$\text{(lower probability)} \qquad \underline{\Pi}_n(B) := \mu\{A_n(V^\star) \subseteq B \mid A_n(V^\star) \neq \emptyset\}$$
$$\text{(upper probability)} \qquad \overline{\Pi}_n(B) := 1 - \underline{\Pi}_n(B^c)$$

Model-free generalized fiducial inference

What if the lower and upper probabilities are difficult to compute?

$\rightarrow$ Construct a precise approximation for model-free GF inference

$\rightarrow$ Uniform sampling over $A_n(1), \ldots, A_n(n+1)$ seems to be the sensible thing to do

$\rightarrow$ Leads to the *precise* model-free GF (MFGF) distribution with density:

$$\begin{aligned}
\pi_y(y_{n+1}) &= \sum_{v^\star=1}^{n+1} \pi_{y,v}(y_{n+1}, v^\star) \\
&= \sum_{v^\star=1}^{n+1} \frac{1}{\mu\{A_n(v^\star)\}} \cdot \frac{1}{n+1} \cdot 1\{y_{n+1} \in A_n(v^\star)\},
\end{aligned}$$

Model-free generalized fiducial inference

Algorithm 1: Computing the MFGF predictive distribution.

Input: Prediction regions $A_n(1), \ldots, A_n(n+1)$ and a desired sample size N.

Output: A sample from the MFGF distribution of Y_{n+1}.

Initialize an N-dimensional vector $\widetilde{y}$;
for $j \in \{1, \ldots, N\}$ **do**
    Sample $v^{\star} \sim \text{uniform}\{1, \ldots, n+1\}$;
    Sample $y_{n+1} \sim \text{uniform}\big(A_n(v^{\star})\big)$;
    Set $\widetilde{y}_j := y_{n+1}$;
end
return $\widetilde{y}$;

Model-free generalized fiducial inference

Theorem

If $Y_1, \ldots, Y_n, Y_{n+1} \overset{iid}{\sim} P$ is a collection of continuous random variables, then for any $\epsilon > 0$, $\alpha \in (0,1)$, and $v \in \{1, \ldots, n\}$,

$$P\left(n^{\alpha}\left|\pi_y\{A_n(v)\} - P\{A_n(v)\}\right| > \epsilon\right) \leq e^{-n^{1-\alpha}\epsilon}.$$

$\rightarrow$ MFGF distribution converges to the true distribution of Y_{n+1}.

Model-free generalized fiducial inference

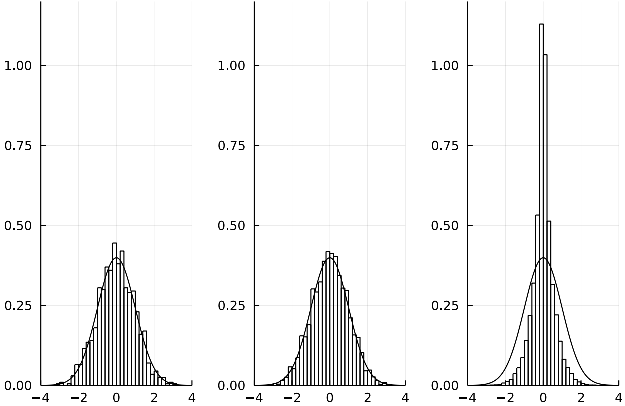

Figure: The middle and right panels display histograms of samples of size 100,000 drawn from the MFGF distribution and CP-induced distribution, respectively, based on $n = 1,000$ data points drawn from a Gaussian distribution

Model-free generalized fiducial inference

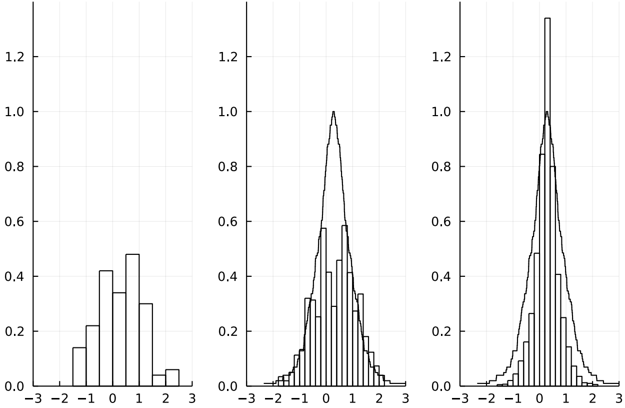

Figure: The middle and right panels display histograms of samples of size 10,000 drawn from the MFGF distribution and CP-induced distribution, respectively, based on $n = 100$ data points drawn from the standard Gaussian distribution

Model-free generalized fiducial inference

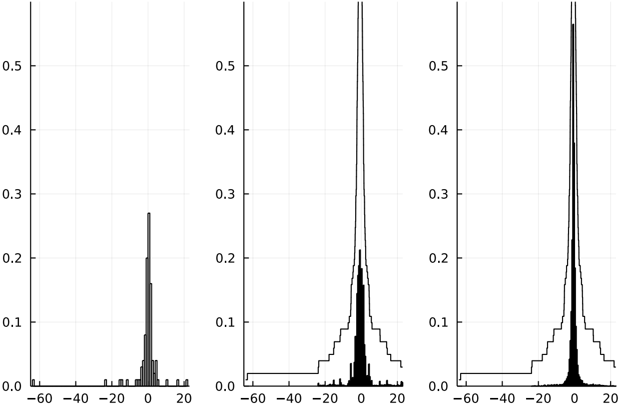

Figure: The middle and right panels display histograms of samples of size 10,000 drawn from the MFGF distribution and CP-induced distribution, respectively, based on $n = 100$ data points drawn from the standard Cauchy distribution

Model-free generalized fiducial inference

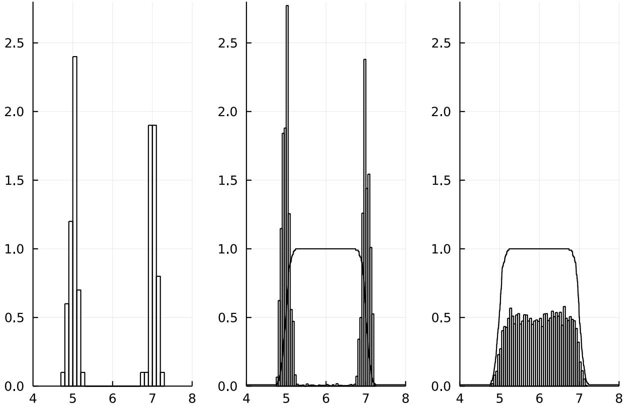

Figure: The middle and right panels display histograms of samples of size 10,000 drawn from the MFGF distribution and CP-induced distribution, respectively, based on $n = 100$ data points drawn from a mixture of two Gaussian distributions

Link to preprint:

Coming soon

My personal academic website:

`https://jonathanpw.github.io`

The end