Valid inferential models and conformal prediction\(^1\)

Ryan Martin  
North Carolina State University  
Researchers.One

Theoretical and Applied Data Science Seminar  
Iowa State University  
March 18th, 2021

\(^1\)Partial support from NSF. Joint work with Leonardo Cella, the paper is available at https://researchers.one/articles/20.01.00010
Prediction is a fundamental problem in statistics, machine learning, and data science in general.

A single point-prediction usually isn’t enough.

Goal is to quantify uncertainty about the predictions.

First thought: a prediction region.

Prediction regions and their properties are familiar.

Can/should we do more...?
Next step: a *predictive distribution*.

This is standard in the Bayesian context.

Other non-Bayesian predictive distributions:
- “frequentist” (Lawless and Fredette 2005)
- fiducial (Wang et al 2012)
- confidence distributions (Vovk et al 2018)

A common belief is that a full predictive distribution is “more informative” than prediction intervals.

In what sense? Is there theory to support this?
Theory focuses on properties of prediction intervals derived from the predictive distributions. But the original motivation behind a predictive distribution was to do more than get prediction intervals. What about other tasks, e.g., prediction probabilities? Is the predictive distribution “good” for these other tasks? For example, ideally we want to avoid assigning large predictive probabilities to events that don’t happen. How to formulate/achieve this?
Problem setup and “probabilistic predictors”

New notion of prediction validity and consequences

A little background on inferential models (IMs)

Leads to a connection with conformal prediction:
  - probabilistic predictor via conformal + consonance
  - validity follows from IM construction

Covariates, examples, a bit about spatial data

Concluding remarks
Problem setup

- Exchangeable process $Z_1, Z_2, \ldots$ with distribution $P$.
- We observe $Z^n = (Z_1, \ldots, Z_n)$, goal is to predict “$Y_{n+1}$”
- Two cases:
  - U. $Z_i = Y_i$, unsupervised
  - S. $Z_i = (X_i, Y_i)$, supervised
- Start with unsupervised, notation is a bit easier.
- Note: no model assumptions beyond exchangeability.
- By “prediction” I mean:
  - more than a point or interval prediction
  - probabilistic quantification of uncertainty about $Y_{n+1}$
Probabilistic predictors

Definition.

A probabilistic predictor is a $y^n$-dependent pair $(\Pi_{y^n}, \overline{\Pi}_{y^n})$ of lower and upper probabilities on the space $\mathbb{Y}$ of $Y_{n+1}$.

- For data $y^n$ and an assertion $A \subseteq \mathbb{Y}$ about $Y_{n+1}$:
  - $\Pi_{y^n}(A)$ measures how believable “$Y_{n+1} \in A$” is
  - $\overline{\Pi}_{y^n}(A)$ measures how plausible “$Y_{n+1} \in A$” is

- Basic properties:
  - $\Pi_{y^n}(A) \leq \overline{\Pi}_{y^n}(A)$
  - $\overline{\Pi}_{y^n}(A) = 1 - \Pi_{y^n}(A^c)$

- Predictive distributions are special cases: $\Pi_{y^n} = \Pi_{y^n} = \overline{\Pi}_{y^n}$.

- We prefer the simplicity of probability, if it works...
Probabilistic predictors, cont.

Brief pause: examples of lower/upper probabilities.

1. Probabilities are lower/upper probabilities.
2. Lower/upper envelopes:
   - Let \( \mathcal{D} \) be a family of probability measures on \( \mathcal{Y} \).
   - Define upper probability as
     \[
     \overline{\Pi}(A) = \sup_{Q \in \mathcal{D}} Q(A), \quad A \subseteq \mathcal{Y}.
     \]
3. Possibility measures:\(^2\)
   - Take a function \( f : \mathcal{Y} \rightarrow [0, 1] \) such that \( \sup_y f(y) = 1 \).
   - Define upper probability as
     \[
     \overline{\Pi}(A) = \sup_{y \in A} f(y), \quad A \subseteq \mathcal{Y}.
     \]

---

\(^2\)Close connection to distributions of nested random sets.
The probabilistic predictor assigns scores to various assertions, we use these scores to make decisions.

For example, I might be willing to accept \( \Pi_{Y^n}(A) \) dollars in exchange for \( 1\{Y_{n+1} \in A\} \) dollars.

In that case, the following events would be bad for me:

\[
\{ Y^{n+1} : \Pi_{Y^n}(A) \text{ is small and } Y_{n+1} \in A \}, \quad \text{any } A.
\]

My idea: *put a constraint on the probabilistic predictor that ensures these bad events have small \( \mathbb{P} \)-probability.*
Definition.

A probabilistic predictor \( y^n \mapsto (\Pi_y^n, \bar{\Pi}_y^n) \) is valid if

\[
P\{\bar{\Pi}_Y^n(A) \leq \alpha \text{ and } Y_{n+1} \in A\} \leq \alpha \quad \forall (\alpha, A, n).
\]

- Some slight adjustments in \( \alpha \) are needed later...
- Why “for all \( A \)”?
  - Implies an analogous condition for \( \Pi_Y^n \)
  - Betting: once I advertise my \( \Pi_Y^n \)-based prices, the choice of \( A \) isn’t up to me anymore.
- Could restrict the class of \( A \)'s in advance, but doing so too much would defeat the purpose
An equivalent statement of the validity property is

\[ E\left[ 1\{\prod_{Y^n}(A) \leq \alpha \} \ P(Y_{n+1} \in A \mid Y^n) \right] \leq \alpha. \]

Suggests “\(\prod_{Y^n}(\cdot)\) dominates \(P(\cdot \mid Y^n)\)” in some sense.

Next results help clarify this “dominance” property.

Note: I’m still working to understand all this myself.
Consider a collection $\mathcal{Q}$ of distributions $Q$.

For a candidate $Q \in \mathcal{Q}$, we can do prediction using the conditional prob $Q(Y_{n+1} \in A \mid Y^n)$.

Take the upper envelope

$$\overline{\Pi}_{y^n}(A) = \sup_{Q \in \mathcal{Q}} Q(Y_{n+1} \in A \mid y^n).$$

Proposition: everywhere dominance implies validity.

If $P \in \mathcal{Q}$, then the upper envelope is valid. In particular, if $P$ were known, then $\Pi_{y^n}(\cdot) = P(Y_{n+1} \in \cdot \mid y^n)$ is valid.
The condition’s apparent dependence on $\alpha$ is inconvenient.

All that’s required is that “$\prod Y_n(\cdot)$ dominates $P(\cdot \mid Y^n)$ on average” in some sense.

Proposition: “dominance on average” implies validity.

The following dominance property implies validity:

$$E\left\{ \frac{P(Y_{n+1} \in A \mid Y^n)}{\prod Y_n(A)} \right\} \leq 1, \quad \forall A.$$
Take-aways

- The dominance properties above all indicate that the probabilistic predictor probably isn’t a probability.
- Conjecture: the only *precise probability* that’s valid is the true conditional probability.
- Since P is unknown, we have to look at *imprecise probabilities*, i.e., genuinely non-additive probabilistic predictors.
- How?
  - Everywhere dominance is inefficient.
  - I don’t know yet how to use “dominance on average”
- Try a different approach...
To achieve validity, it seems the probabilistic predictor can’t be a probability distribution.

We have some experience with constructing non-additive measures that achieve validity in inference.

I’ll give you a little background about this next.

But rather than go through all the details, I’ll skip to the end and draw connections to conformal prediction.
In the context of inference, there are again many ways to construct "posterior" probability distributions:

- Bayesian
- fiducial
- confidence distributions

But these are generally not valid.\(^3\)

The *false confidence theorem* helps explains this.\(^4\)

Fortunately, we have some experience constructing genuine lower/upper probabilities that are valid.

---

\(^3\) The notion of validity in inference is similar to that for prediction.
\(^5\) M. (2019), https://researchers.one/articles/19.02.00001
This theory relies heavily on nested random sets
Efficient IMs are based on nested random sets.

This suggests a close connection with possibility theory and *consonant* lower/upper probabilities.\(^6\)

Consonance means there exists a function \(\pi_{y^n}\) such that

\[
\sup \tilde{y} \pi_{y^n}(\tilde{y}) = 1
\]

\[
\Pi_{y^n}(A) = \sup \tilde{y} \in A \pi_{y^n}(\tilde{y}).
\]

Simplifies the lower/upper prob construction.

Possibility theory is crucial to frequentist inference.\(^7\)

---

\(^6\) Liu and M. (2020), https://researchers.one/articles/20.08.00004

\(^7\) M. (2021), https://researchers.one/articles/21.01.00002
The original IM construction can be modified. Leads to what we call a *generalized IM*.\(^8\)

Construction naturally leads to a valid, consonant lower/upper probability defined on parameter space.

Cella and M. extend this idea to the prediction setting.

Details are too complicated for a seminar talk...

but the end result has close connections to something more familiar, Vovk et al’s *conformal prediction*.

---

Conformal prediction is one of those very special ideas, simple, elegant, and powerful.

- Relies on exchangeability and the distribution of ranks.
- Provides a universal procedure for making valid predictions.
- Doesn’t obviously fit in the standard statistical toolbox, so can be hard to understand.
- I think the connection to IMs helps.
A non-conformity measure is a function of two arguments:
- a collection/bag of data for making a prediction
- a candidate value of the to-be-predicted thing

Evaluates how closely the to-be-predicted value represents the data in the bag.

Simple example: \( \Psi(y^n, \tilde{y}) = |\text{avg}(y^n) - \tilde{y}| \).

Many other examples, especially in regression-like cases.

Virtually no restrictions on how one can construct a prediction based on data in the bag.
Conformal prediction, cont.

**Algorithm.**

Input: data $y^n$, generic $\tilde{y}$, non-conformity measure $\Psi$.

1. Provisionally set $y_{n+1} = \tilde{y}$.
2. With $y_{n+1}^{n+1}$ being $y^{n+1}$ with $y_i$ removed, define

$$T_i = \Psi(y_{i}^{n+1}, y_i), \quad i = 1, \ldots, n, n+1.$$ 

3. Return

$$\pi_{y^n}(\tilde{y}) = \frac{1}{n+1} \sum_{i=1}^{n+1} 1(T_i \geq T_{n+1}).$$
Conformal prediction, cont.

**Theorem.**
If $P$ is exchangeable and $\Psi$ is invariant bag shuffling, then

$$P\{\pi_{Y^n}(Y_{n+1}) \leq k_n(\alpha)\} \leq \alpha, \quad \forall (\alpha, n, P),$$

where $k_n(\alpha) = (n+1)^{-1}\lfloor(n+1)\alpha\rfloor \approx \alpha$.

**Corollary.**
The set $\mathcal{P}_\alpha(y^n) = \{\tilde{y} : \pi_{y^n}(\tilde{y}) > k_n(\alpha)\}$ satisfies

$$P\{\mathcal{P}_\alpha(Y^n) \ni Y_{n+1}\} \geq 1 - \alpha, \quad \forall (\alpha, n, P).$$

**Why does it work?**

- $T_i = \Psi(Y_{n+1}^i, Y_i)$ are exchangeable
- $\pi_{Y^n}(Y_{n+1})$ is $\propto$ to the rank of $T_{n+1}$ relative to $T_i$’s.
- Ranks of exchangeable random variables are uniform.
For continuous data, \( \sup_{\tilde{y}} \pi_{y^n}(\tilde{y}) = 1 \).

Therefore, we can treat the conformal prediction output as a plausibility contour...

Define a consonant probabilistic predictor

\[
\bar{\Pi}_{y^n}(A) = \sup_{\tilde{y} \in A} \pi_{y^n}(\tilde{y}), \quad \text{any } A.
\]

**Theorem.**

The *conformal + consonance* probabilistic predictor defined above is valid in roughly the same sense as before, i.e.,

\[
P\{\bar{\Pi}_{Y^n}(A) \leq k_n(\alpha) \text{ and } Y_{n+1} \in A\} \leq \alpha.
\]
The choice to convert the conformal prediction output into a consonant lower/upper probability might seem ad hoc.

However:
- this is exactly what we get in Cella and M.
- it’s a consequence of using nested random sets, motivated by efficiency considerations in the general IM theory.

The connection to IMs provides some insight as to
- why conformal prediction works and
- why, for validity, it needs consonance instead of additivity.
Examples

- Plots of two plausibility contours:
  - simple one-dim case
  - two-dim case with non-conformity based on data depth
There’s now an extensive statistics literature on conformal prediction-related things.

Interests include:

- speeding up computations
- “conditional validity”
- beyond exchangeability, e.g., spatial data

I need to say some things about supervised learning...
Conformal with covariates

- Now let $Z_i = (X_i, Y_i)$
  - pairs are exchangeable
  - interest still is in the response $Y$
  - depends on a (possibly high-dim) covariate $X$
- A model for $Y | X$ can be used, but isn't required.
- Like above, we need a non-conformity measure, e.g.,
  \[
  \Psi(Z_{n+1}, Z_i) = |\hat{\mu}_{-i}(X_i) - Y_i|,
  \]
  where $\hat{\mu}_{-i}$ is a fitted mean using only data in the bag
  - could be a high-dim linear model fit with lasso
  - could be something more complex, like a neural net or whatever is the state-of-the-art in deep learning.
Conformal with covariates, cont.

- Same basic algorithm can be applied, returning

\[ (z^n, \tilde{z}) \mapsto \pi_{z^n, x}(\tilde{y}) = \frac{1}{n+1} \sum_{i=1}^{n+1} 1(T_i \geq T_{n+1}). \]

- For continuous responses, the same conformal + consonance procedure can be carried out as before, giving
  - prediction intervals with nominal (marginal) coverage
  - a probabilistic predictor that's (marginally) valid, i.e.,

\[ P\{\prod_{z^n, x_{n+1}}(A) \leq k_n(\alpha) \text{ and } Y_{n+1} \in A\} \leq \alpha. \]

- There's a parallel IM construction...\(^9\)

\(^9\)Cella and M. (2021), see my website.
Regression example

- \( n = 200 \)
- \( X_i \sim \text{Unif}(0, 1) \)
- \( Y_i = \mu(X_i) + 0.1 \, t(5) \)
- \( \mu(x) = \sin^3(2\pi x^3) \).

- \( \Psi \) based on a B-spline \( \hat{\mu} \) with \( \text{df} = 12 \).
- Conformal + consonance gives a plausibility contour for \( Y_{n+1} \) at each candidate \( \tilde{x} \) value.
Classification

- Problem is basically the same as regression
- Key difference is that $Y$ is a discrete label
- This means consonance won’t hold automatically
- In a “closed-world” classification problem, consonance still makes sense, but needs to be enforced manually.
- There are several reasonable ways to do this...
- Here the IM connection provides some guidance:
  - consonance is enforced just by setting the max plausibility contour value to 1
  - achieved via use of “elastic” random sets
  - motivated by efficiency in the IM framework
Spatial data generally aren’t exchangeable.

But local exchangeability is possible:

- Response \( Y = \{ Y(s) : s \in S \} \)
- Given a point \( s^\star \), define the localized process

\[
\tilde{Y}_r(u) = Y(s^\star + ru), \quad \|u\| \leq 1.
\]

- \( Y \) is locally exch. if \( \tilde{Y}_r \overset{d}{\to} \) an exchangeable process, \( r \to 0 \).

Infill asymptotics regime: lots of data near \( s^\star \).

If local exchangeability holds:

- \( \{ Y(s_i) : s_i \text{ near } s^\star \} \) approx exchangeable
- can do basic conformal prediction using only data nearby \( s^\star \)

---

\(^{10}\)Mao, M., & Reich, https://researchers.one/articles/20.06.00006
Conclusion

- Defined valid probabilistic predictors.
- Interesting coherence-related consequences of validity.
- **Claim.** Additive probabilistic predictors can’t be valid.
- Need non-additivity, esp., consonance.
- IM construction leads to valid probabilistic predictors
- Conformal + consonance is a shortcut construction.
- Very flexible, especially in supervised learning.
Open questions

- Proof that additive probabilistic predictors can’t be valid?
- What, if anything, can be said about conditional validity of probabilistic predictors?
  - The following is probably impossible

\[
P\{\prod_{Z^n,x}(A) \leq \alpha \text{ and } Y_{n+1} \in A \mid X_{n+1} = x\} \leq \alpha...
\]

- What about approximately/asymptotically?
- Alternative formulations?

- Model assessment applications? A model fits well if it accurately predicts the data you have.

- .....
The end – thanks!

- Links to papers, talks, etc. can be found on my website:
  
  www4.stat.ncsu.edu/~rmartin/

- Questions? Email me at rgmarti3@ncsu.edu.
- Please check out Researchers.One:
  - articles and open peer review
  - new virtual conference feature
  - more stuff coming soon

- COVID-19 vaccine panel discussion March 25th:
  
  https://researchers.one/conferences/covid19

---

11 https://researchers.one
12 www.twitter.com/ResearchersOne