Conformal e-prediction in the presence of confounding

Vladimir Vovk and Ruodu Wang

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 #46

First posted March 12, 2026. Last revised March 14, 2026.

Project web site:
http://alrw.net

Abstract

This note extends conformal e-prediction to cover the case where there is observed confounding between the random object X and its label Y. We consider both the case where the observed data is IID and a case where some dependence between observations is permitted.

Contents

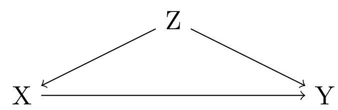

Figure 1: The main causal graph of this note

1 Introduction

Conformal prediction in its basic form is only applicable to IID sequences of observations. In causal inference, including Pearl's [2, Chap. 3] do calculus, we typically observe IID data but then would like to predict results of interventions into stable stochastic mechanisms generating the data. In this note we apply conformal e-prediction to this prediction problem in order to obtain finite-sample guarantees of validity. Formally, however, the specific approach that we take in this note goes beyond conformal prediction and is closer to "randomness prediction" (as defined in, e.g., [6]).

The main question asked in this note is about the simplest setting of causal inference. We are interested in the causal effect of a random variable X on a random variable Y with a confounder Z. (See Figure 1.) Namely, after setting $X := x$, we would like to say something about Y, e.g., to output a prediction region for it. The available data comes from an observational study. We consider two settings: in Sect. 2 we start from the standard IID one, while in Sect. 3 we allow X to be chosen by a non-trivial strategy, as in [4] and [2, Sect. 3.6.1, third extension]. Our mathematical results in Sect. 2 are simple and likely to be known.

For a positive integer N, we set $[N] := \{1, \dots, N\}$.

2 The IID setting

Let P be a positive probability measure on $\mathbf{X} \times \mathbf{Y} \times \mathbf{Z}$ generating random variables (X, Y, Z); for simplicity we assume that $\mathbf{X} \times \mathbf{Y} \times \mathbf{Z}$ is finite (and equipped with the discrete σ-algebra). Let us fix $\tilde{x} \in \mathbf{X}$; we will set X to $\tilde{x}$, $X := \tilde{x}$. For a given value $y \in \mathbf{Y}$, define

$$p_y := \sum_{z \in \mathbf{Z}} P(Z = z) P(Y = y \mid X = \tilde{x}, Z = z). \tag{1}$$

The interpretation of (1) is that it is the probability of $Y = y$ in the mutilated causal model in which the arrow from Z to X in Figure 1 has been removed and X has been set to $\tilde{x}$. We do not mention $\tilde{x}$, which is fixed, in our notation.

Generate an IID random sample (X_n, Y_n, Z_n), $n \in [N]$, of size $N \geq 1$ from

P. For each $y \in \mathbf{Y}$ define an estimate F_y of p_y by

$$F_y := \sum_{z \in \mathbf{Z}} \frac{|\{n \in [N] : Z_n = z\}| + 1}{N+1} \times \frac{|\{n \in [N] : (X_n, Y_n, Z_n) = (\tilde{x}, y, z)\}| + 1}{|\{n \in [N] : (X_n, Z_n) = (\tilde{x}, z)\}| + 1}. \tag{2}$$

Lemma 1. *For each $y \in \mathbf{Y}$, it is true that*

$$\mathbb{E} \frac{p_y}{F_y} \le 1. \tag{3}$$

We will prove Lemma 1 in Appendix B.

Let Y_{N+1} be a random variable, independent of the sample (X_n, Y_n, Z_n), $n \in [N]$, with values in $\mathbf{Y}$ taking each value $y \in \mathbf{Y}$ with probability p_y given by (1). In causal inference we will use the following corollary (see [3] for hypothesis testing using e-variables).

Corollary 2. *For each probability measure Q on $\mathbf{Y}$, the random variable E defined by*

$$E := \frac{Q(\{Y_{N+1}\})}{F_{Y_{N+1}}} \tag{4}$$

is an e-variable (i.e., is nonnegative with expectation at most 1).

Proof. The statement of the corollary follows from

$$\mathbb{E} \frac{Q(\{Y_{N+1}\})}{F_{Y_{N+1}}} = \sum_{y \in \mathbf{Y}} p_y \mathbb{E} \frac{Q(\{y\})}{F_y} = \sum_{y \in \mathbf{Y}} Q(\{y\}) \mathbb{E} \frac{p_y}{F_y} \le \sum_{y \in \mathbf{Y}} Q(\{y\}) = 1,$$

where the first equality follows from the independence of Y_{N+1}. □

Two particularly natural choices for Q in Corollary 2 are the uniform probability measure on $\mathbf{Y}$ and the probability measure concentrated on some $y^* \in \mathbf{Y}$. The former choice treats all labels $y \in \mathbf{Y}$ symmetrically. The latter choice is appropriate when a specific label y^* is much more important for us than the other labels and we would like to exclude it confidently when possible; an example of such a label is "patient's death".

Let us apply Corollary 2 to a simple problem of causal inference. After generating the IID random sample (X_n, Y_n, Z_n), $n \in [N]$, we generate Y_{N+1} according to the probability measure assigning probability p_y to each $y \in \mathbf{Y}$ (conditionally on the random sample, of course). This is exactly the probability distribution on Y in the mutilated causal network obtained by setting $X := \tilde{x}$.

Corollary 2 allows us to output e-prediction regions for Y_{n+1}. Given an alternative Q (say, the uniform probability measure on $\mathbf{Y}$) and a significance level $\alpha > 0$ (typically a large number, such as 10 or 100), we define the corresponding e-prediction region by

$$\Gamma^\alpha := \left\{ y \in \mathbf{Y} : \frac{Q(\{y\})}{F_y} < \alpha \right\}. \tag{5}$$

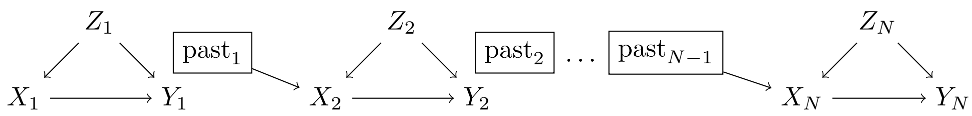

Figure 2: The repeated causal graph

The prediction regions Γ^α are nested and grow as α increases. The main property of validity for this predictor is

$$\int_0^\infty \mathbb{P}\left(Y \notin \Gamma^\alpha\right) \mathrm{d}\alpha \leq 1; \tag{6}$$

in words, the probability of error at significance level $\alpha > 0$ should integrate to at most 1 [5, the end of Appendix B]. Since the probability of error decreases in α, the probability of error at level α does not exceed $1/\alpha$ by Markov's inequality. This is a simpler property of validity; however, the property of validity expressed by (6) is much stronger (and in fact is equivalent to the E defined by (4) being an e-variable).

For a large N and small $|\mathbf{Z}|$, the e-prediction regions (5) are close to being optimal in some sense. Namely, the "oracle" e-prediction regions are

$$\Gamma^\alpha := \left\{ y \in \mathbf{Y} : \frac{Q(\{y\})}{p_y} < \alpha \right\} \tag{7}$$

if Q is a genuine alternative [3, Chap. 3]. Since F_y as defined by (2) is an estimate of p_y, (5) is an approximation to (7) that can be computed from the data.

One special case mentioned earlier is where Q is concentrated on some outcome y^*, which we would like to avoid ("death of the patient"). For a large α, we are justified in predicting $Y_n \neq y^*$ when we observe $F_{y^*} \leq 1/\alpha$. In particular, the probability that we will make an error does not exceed $1/\alpha$ (where the probability is marginal, not conditional on $F_{y^*} \leq 1/\alpha$).

Remark 1. For simplicity, in this note we concentrate on the simplest causal graph given in Figure 1. However, it is easy to extend our approach to any causal graph covered by the popular back-door criterion: see [2, Theorem 3.3.2]. Now Z becomes a set of variables in general, called an adjustment set.

3 No stable stochastic mechanism for X

As pointed out in [4, Sect. 1] (see also [2, Sect. 3.6.1]), the assumption that X_n, $n \in [N]$, are output by a stable stochastic mechanism is unnatural in the context of causal inference, since we are contemplating setting X to some value $\tilde{x}$. In this section we will still assume that Z_n and Y_n are generated by stable stochastic mechanisms but will drop this assumption for X_n.

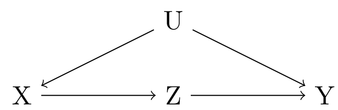

Figure 3: A causal graph illustrating the front-door criterion; the variable U is unobservable

Our assumptions are represented graphically in Figure 2, which essentially consists of a series of causal triangles analogous to the one in Figure 1. The triangles are arranged chronologically from left to right (while the order inside the triangles is not chronological, of course; within each triangle the arrows represent the causal order). The boxes labelled past_n, $n \in [N-1]$, stand for sets of variables. In this and next sections we will consider two possible interpretations of these boxes; in any case, past_n represents some variables in the first n triangles.

The most satisfying analogue of Lemma 1 obtains when each box past_n stands for all the variables X_i and Z_i, $i \in [n]$. We will call this the *Y-oblivious interpretation* of Figure 2. Under this interpretation, each X_{n+1}, $n \in [N-1]$, has incoming arrows from all X_i and Z_i, $i \in [n]$.

Lemma 3. *Lemma 1 continues to hold under the Y-oblivious interpretation of Figure 2.*

The proof of Lemma 3 is given in Appendix B. Corollary 2 also holds in the Y-oblivious setting, and therefore, we still have e-prediction regions (5) satisfying the property of validity (6).

4 A finite-sample version of the front-door criterion

In Remark 1 we noticed that our prediction method is applicable quite generally, not just in the situation of Figure 1. In particular, they are applicable in the context of the back-door criterion and under strong ignorability [2, Sect. 11.3.2]. It is not applicable, however, in the situation of Figure 3, which is covered by Pearl's "front-door criterion". While the back-door criterion is often regarded as more widely applicable than the front-door criterion (e.g., Cox and Wermuth say in their comments on Pearl's discussion paper [1], "Situations where this [the conditions for the front-door criterion] could be assumed with any confidence seem likely to be exceptional"), it is still interesting and popular, and in this section we will develop a prediction method for it.

For simplicity, we will talk about Figure 1, but our argument will be applicable to the front-door criterion in general if we make X, Y, Z, and U sets of variables rather than single variables. Since now we have more variables, the notation that we used in the previous sections becomes too cumbersome, and we will use Pearl's convention of abbreviating $P(X = x)$ to $P(x)$, $P(Y = y)$ to $P(y)$, etc., as explained in [2, Sects. 1.1.4 and 1.1.5]. We will often omit mentioning that $x \in \mathbf{X}$, $y \in \mathbf{Y}$, etc.

Similarly to what we did in Sect. 2, let P be a positive probability measure on finite $\mathbf{X} \times \mathbf{Y} \times \mathbf{Z} \times \mathbf{U}$ that generates random variables (X, Y, Z, U). As before, $\tilde{x} \in \mathbf{X}$ is fixed, and we set $X := \tilde{x}$. For a given value $y \in \mathbf{Y}$,

$$p_y := \sum_{x,z} P(x) P(z \mid \tilde{x}) P(y \mid x, z) \tag{8}$$

is the causal effect of X on Y according to [2, Theorem 3.3.4].

For an IID random sample (X_n, Y_n, Z_n, U_n), $n \in [N]$, and $y \in \mathbf{Y}$, the estimate F_y of p_y is now defined by

$$F_y := \sum_{x,z} \frac{\#x + 1}{N + 1} \, \frac{\#\tilde{x}z + 1}{\#\tilde{x} + 1} \, \frac{\#xyz + 1}{\#xz + 1}, \tag{9}$$

where $\#x := |\{n \in [N] : X_n = x\}|$, $\#xz := |\{n \in [N] : (X_n, Z_n) = (x, z)\}|$, etc. We still have the analogue of Lemma 1.

Lemma 4. *For each $y \in \mathbf{Y}$,* (3) *holds under the definitions* (8) *and* (9)*.*

We will prove Lemma 4 in Appendix B. We still have Corollary 2 and thus e-prediction regions (5) with (6) as property of validity.

5 Conclusion

This note extends conformal e-prediction to a simple setting of causal inference. (And in Appendix B we will see that our causal e-predictor is a combination of $|\mathbf{Z}|$ conformal e-predictors.) In this section we discuss some open questions and directions of further research.

In this note we only discussed causal inference using the back-door (Figure 1 and Remark 1) and front-door (Figure 3 and Sect. 4) criteria. There is no doubt that our method can be extended to other cases in which Pearl's do calculus [2] and its variations allow us to identify the causal effect of X on Y. In the case of the back-door criterion, our method works best for an adjustment set Z that has the smallest product of the sizes of its elements' domains (assuming Z is not unique).

Our proof of Lemma 3 does not work for the *strong interpretation* of Figure 2, where each past_n stands for all the previous variables, namely (X_i, Y_i, Z_i), $i \in [n]$. One possible approach in this situation is to use conformal test martingales [7, Chaps. 8 and 9].

We assumed that each variable takes only finitely many values, but another natural setting is regression, in which Y is allowed to be any real number and we are interested in prediction intervals for it.

We have not discussed the optimality of our finite-sample results. Our simulation studies suggest that (3) will hold even if we use less heavy regularization in (2) (e.g., replacing the entries of " + 1" by " + c" for $c < 1$). What are the admissible constants in (3)? Improving them will lead to an automatic improvement in the e-prediction regions (5). A slack in (3) is also discussed in Sect. A.3.

References

[1] Judea Pearl. Causal diagrams for empirical research (with discussion). *Biometrika*, 82:669–710, 1995.

[2] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, Cambridge University Press, second edition, 2009.

[3] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2):1–390, 2025.

[4] Vladimir Vovk. Another semantics for Pearl's action calculus. In Alex Gammerman, editor, *Computational Learning and Probabilistic Reasoning*, chapter 7, pages 127–146. Wiley, New York, 1996.

[5] Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:article 111674, 2025. Special Issue on Conformal Prediction and Distribution-Free Uncertainty Quantification.

[6] Vladimir Vovk. Universality of conformal prediction under the assumption of randomness. *Proceedings of Machine Learning Research*, 313:1–18, 2026. ALT 2026.

[7] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

A Connections with conformal prediction

This note is motivated by conformal e-prediction, which is applicable to the case where we have only one random variable Y and to the case where we have two random variables X, Y. Both cases will be used in the proofs of Lemmas 1 and 3.

A.1 Simple conformal e-prediction

First suppose that we have only one random variable Y. Let P be a positive probability measure on $\mathbf{Y}$ generating Y. Generate an IID random sample Y_n, $n \in [N+1]$, from P of size $N+1$. For a fixed $y \in \mathbf{Y}$, consider the nonconformity scores

$$\alpha_i := \frac{N+1}{|\{n \in [N+1] : Y_n = y\}|} 1_{\{Y_i = y\}}, \quad i \in [N+1]$$

[5, Sect. 2]. The standard property of validity for conformal e-prediction gives

$$\mathbb{E}\left(\frac{N+1}{|\{n \in [N+1] : Y_n = y\}|} 1_{\{Y_{N+1}=y\}}\right) \le 1,$$

i.e.,

$$\mathbb{E}\frac{N+1}{|\{n \in [N] : Y_n = y\}| + 1} \le \frac{1}{P(Y=y)}. \tag{10}$$

It is instructive to see what Corollary 2 becomes in the case of simple conformal e-prediction. We can rewrite (10) as

$$\mathbb{E}\frac{P(Y=y)}{\hat{P}(Y=y)} \le 1,$$

where

$$\hat{P}(Y=y) := \frac{|\{n \in [N] : Y_n = y\}| + 1}{N+1}$$

(so that $\hat{P}$ is an estimate of P, although it is a "super-probability measure" rather than a probability measure in that $\hat{P}(Y=y)$ sum to more than 1 over $y \in \mathbf{Y}$). The same argument as in the proof of Corollary 2 shows that, for any probability measure Q on $\mathbf{Y}$,

$$\mathbb{E}\frac{Q(\{Y_{N+1}\})}{\hat{P}(\{Y_{N+1}\})} \le 1.$$

In particular, the conformal e-prediction set at significance level $\alpha > 0$ (typically $\alpha \gg 1$) is

$$\left\{y \in \mathbf{Y} : Q(\{y\})/\hat{P}(\{y\}) < \alpha\right\}.$$

A.2 Conditional conformal e-prediction

Now suppose that we have a positive probability measure P on $\mathbf{X}\times\mathbf{Y}$ generating random variables X and Y. We generate an IID random sample (X_n, Y_n), $n \in [N+1]$, from it. For a given $y \in \mathbf{Y}$, we use the nonconformity scores

$$\alpha_i := \frac{|\{n \in [N+1] : X_n = X_i\}|}{|\{n \in [N+1] : (X_n, Y_n) = (X_i, y)\}|} 1_{\{Y_i = y\}}, \quad i \in [N+1],$$

in the object-conditional conformal e-predictor [5, Sect. 4] (0/0 is interpreted as 1 here). Now we have, for a fixed $x \in \mathbf{X}$,

$$\mathbb{E}\left(\frac{|\{n \in [N+1] : X_n = x\}|}{|\{n \in [N+1] : (X_n, Y_n) = (x, y)\}|} 1_{\{(X_{N+1}, Y_{N+1})=(x,y)\}} \;\middle|\; X_1, \dots, X_{N+1}\right) \le 1,$$

i.e.,

$$\mathbb{E}\left(\frac{|\{n\in[N]:X_n=x\}|+1}{|\{n\in[N]:(X_n,Y_n)=(x,y)\}|+1}\middle| X_1,\dots,X_N\right)\le\frac{1}{P(Y=y\mid X=x)}. \tag{11}$$

It is clear that for the validity of conditional conformal e-prediction it suffices to assume that the Y_n are IID conditional on X_n; the distribution of X_n can be arbitrary [7, Sect. 4.6.1].

A.3 Slack in (3)

This section continues discussion started in Sect. 5. The presence of a slack in (3) becomes transparent if we assume that $P(X=\tilde{x})=1$. Then (2) can be rewritten as

$$\begin{aligned}F_y&=\sum_{z\in\mathbf{Z}}\frac{|\{n\in[N]:Z_n=z\}|+1}{N+1}\times\frac{|\{n\in[N]:(Y_n,Z_n)=(y,z)\}|+1}{|\{n\in[N]:Z_n=z\}|+1}\\&=\sum_{z\in\mathbf{Z}}\frac{|\{n\in[N]:(Y_n,Z_n)=(y,z)\}|+1}{N+1}=\frac{|\{n\in[N]:Y_n=y\}|+|\mathbf{Z}|}{N+1},\end{aligned}$$

and so (3) is weaker than what can be obtained with conformal e-prediction, which allows us to have 1 in place of the $|\mathbf{Z}|$. Namely, when $P(X=\tilde{x})=1$, (3) becomes

$$\mathbb{E}\left(P(Y=y)\middle/\frac{|\{n\in[N]:Y_n=y\}|+|\mathbf{Z}|}{N+1}\right)\le 1$$

and so is weaker than (10).

B Some proofs

Proof of Lemma 1

We start from noticing that p_y/F_y is the weighted harmonic mean over $z\in\mathbf{Z}$ of $p_{y,z}/F_{y,z}$ taken with the weights $p_{y,z}/p_y$, where

$$p_{y,z}:=P(Z=z)P(Y=y\mid Z=z,X=\tilde{x})$$

and

$$F_{y,z}:=\frac{|\{n\in[N]:Z_n=z\}|+1}{N+1}\times\frac{|\{n\in[N]:(X_n,Y_n,Z_n)=(\tilde{x},y,z)\}|+1}{|\{n\in[N]:(X_n,Z_n)=(\tilde{x},z)\}|+1}.$$

This follows from the equalities $F_y=\sum_z F_{y,z}$ and $p_y=\sum_z p_{y,z}$, which imply

$$\frac{p_y}{F_y}=\frac{1}{\sum_{z\in\mathbf{Z}}F_{y,z}/p_y}=\frac{1}{\sum_{z\in\mathbf{Z}}\frac{p_{y,z}}{p_y}\frac{F_{y,z}}{p_{y,z}}}.$$

Since a weighted harmonic mean does not exceed the corresponding weighted arithmetic mean, it suffices to prove $\mathbb{E}(p_{y,z}/F_{y,z}) \le 1$ for a fixed $z \in \mathbf{Z}$, i.e.,

$$\mathbb{E}\left(\frac{N+1}{|\{n \in [N] : Z_n = z\}|+1} \times \frac{|\{n \in [N] : (X_n, Z_n) = (\tilde{x}, z)\}|+1}{|\{n \in [N] : (X_n, Y_n, Z_n) = (\tilde{x}, y, z)\}|+1}\right) \le \frac{1}{P(Z=z)P(Y=y \mid X=\tilde{x}, Z=z)}. \quad (12)$$

We can deduce (12) from the property of validity for conformal e-prediction

$$\mathbb{E}\frac{N+1}{|\{n \in [N] : Z_n = z\}|+1} \le \frac{1}{P(Z=z)} \quad (13)$$

(see (10)) and its version

$$\mathbb{E}\frac{|\{n \in [N] : (X_n, Z_n) = (\tilde{x}, z)\}|+1}{|\{n \in [N] : (X_n, Y_n, Z_n) = (\tilde{x}, y, z)\}|+1} \le \frac{1}{P(Y=y \mid X=\tilde{x}, Z=z)} \quad (14)$$

that holds conditionally on $X_1, \dots, X_N, Z_1, \dots, Z_N$ (see (11), where X should be replaced by (X, Z)). Indeed, (12) can be obtained by multiplying (13) and (14).

Proof of Lemma 3

As in the previous proof, it suffices to prove (12). We have (13), and we also have (14) conditionally on (X_n, Z_n), $n \in [N]$ (which follows from the remark at the end of Sect. A.2). Combining (13) and (14) gives (12).

Proof of Lemma 4

The proof is modelled on that of Lemma 1 above. The ratio p_y/F_y is the weighted harmonic mean over (x, z) of $p_{x,y,z}/F_{x,y,z}$ taken with the weights $p_{x,y,z}/p_y$, where

$$p_{x,y,z} := P(x)P(z \mid \tilde{x})P(y \mid x, z)$$

and

$$F_{x,y,z} := \frac{\#x+1}{N+1}\,\frac{\#\tilde{x}z+1}{\#\tilde{x}+1}\,\frac{\#xyz+1}{\#xz+1}.$$

This follows from $F_y = \sum_{x,z} F_{x,y,z}$ and $p_y = \sum_{x,z} p_{x,y,z}$:

$$\frac{p_y}{F_y} = \frac{1}{\sum_{x,z} F_{x,y,z}/p_y} = \frac{1}{\sum_{x,z} \frac{p_{x,y,z}}{p_y}\frac{F_{x,y,z}}{p_{x,y,z}}}.$$

Therefore, it suffices to prove $\mathbb{E}(p_{x,y,z}/F_{x,y,z}) \le 1$ for fixed x, z, i.e.,

$$\mathbb{E}\left(\frac{\#x+1}{N+1}\,\frac{\#\tilde{x}z+1}{\#\tilde{x}+1}\,\frac{\#xyz+1}{\#xz+1}\right) \le \frac{1}{P(x)P(z \mid \tilde{x})P(y \mid x, z)}.$$

This inequality can be obtained by multiplying the following three ones:

$$\mathbb{E}\frac{\#x+1}{N+1} \le \frac{1}{P(x)}$$

(a version of (10)),

$$\mathbb{E}\frac{\#\tilde{x}z+1}{\#\tilde{x}+1} \le \frac{1}{P(z \mid \tilde{x})}$$

(a version of (11) that holds conditionally on $X_1, \ldots, X_N$), and

$$\mathbb{E}\frac{\#xyz+1}{\#xz+1} \le \frac{1}{P(y \mid x, z)}$$

(a version of (11) that holds conditionally on $X_1, \ldots, X_N, Z_1, \ldots, Z_N$).