STAT 991: Topics in Modern Statistical Learning
Conformal Prediction for Dependent Data

Presented by
Kaifu Wang

University of Pennsylvania

February 18, 2022

Outline

Conformal Prediction under Exchangeability Assumption

\# Problem

- Given observations $Z_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^{T}$ and x_{T+1}, predict y_{T+1}.
- Exchangeability: the distribution of $(Z_{\text{obs}}, Z_{\text{test}})$ is invariant under any permutation of indices.

\# Algorithm

- Score any y by its (non)comformity to the obervation, which is any function that is equivariant with respect to permutations.
- For regression, typically defined with the residual: $\sum |y_i - \widehat{f}(x_i)|^p$
- The prediction region C_α includes all y_{T+1} that has a score greater than a significance level α.

\# Characterization

- Validity: $\mathbb{P}(Y_{t+1} \in C_\alpha(Z_{\text{obs}})) \geq 1 - \alpha$.
- Efficiency: $|C_\alpha \triangle C_\alpha^{\text{oracle}}| \to 0$ in probability as $T = |Z_{\text{obs}}| \to \infty$ under certain assumptions.

Conformal Prediction for Non-exchangeable Data

\# Problem

- $Z_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^T$ may break the exchangeability assumption.
- This could happen for time series data.

\# Intuitions

- Exchangeability requires invariance to **all** permutation, but is it necessary to ensure validity? (Section 2)
- Find (or learn) a **transformation** of the data to make it exchangeable. (Section 3)
- If the dependency is **weak**, conformal predictor may still be approximately valid given enough data. (Section 4)

Outline

Group Invariance Implies Validity

Problem

- The distribution of $Z = (Z_{\text{obs}}, Z_{\text{test}})$ is still but only invariant under *some* permutations $\Pi \subseteq S_{T+1}$.

Predictor

- With conformity score $S : \mathcal{Z}^{T+1} \to \mathbb{R}$, define the p-score as:

$$\widehat{p}(y_{\text{test}}) := \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbb{1}\{S(Z^{\pi}) \geq S(Z)\}$$

- Prediction region: $\mathcal{C}_{\alpha} := \left\{ y \in \mathcal{Y}^S : \widehat{p}(y) > \alpha \right\}$

Theorem 1 ([Chernozhukov et al., 2018])

Let $S^{(k_\alpha)}$ be the $\lfloor |\Pi| \alpha \rfloor^{\text{th}}$ largest element in $\{S(Z^{\pi}) : \pi \in \Pi\}$. Suppose the joint distribution of Z is invariant under any $\pi \in \Pi$ and S satisfies

$$S^{(k_\alpha)}(Z^{\pi}) \geq S^{(k_\alpha)}(Z) \quad \forall \pi \in \Pi \tag{1}$$

Then, $\mathbb{P}(Y_{\textit{test}} \in \mathcal{C}_{\alpha}) \geq 1 - \alpha$. In particular, if Π is a group, then (1) holds with equality.

Group Invariance Implies Validity: Proof

Proof. We have

$$
\begin{aligned}
\mathbb{P}(\widehat{p} \le \alpha) &= \mathbb{E}[\mathbb{1}\{S(Z) > S^{(k_\alpha)}(Z)\}] && \text{(definition)} \\
&= \frac{1}{|\Pi|}\mathbb{E}\left[\sum_{\pi \in \pi} \mathbb{1}\{S(Z^\pi) > S^{(k_\alpha)}(Z^\pi)\}\right] && \text{(assumption)} \\
&\le \frac{1}{|\Pi|}\mathbb{E}\left[\sum_{\pi \in \pi} \mathbb{1}\{S(Z^\pi) > S^{(k_\alpha)}(Z)\}\right] && \text{(equation (1))} \\
&\le \frac{|\Pi|\alpha}{|\Pi|} = \alpha && \text{(definition of } k_\alpha\text{)}
\end{aligned}
$$

In particular, if Π is a group, then every π has an inverse and hence $\{S(Z^{\pi'}) : \pi' \in \Pi\} = \{S((Z^\pi)^{\pi'}) : \pi' \in \Pi\}$. Therefore,

$$S^{(k_\alpha)}(Z^\pi) = S^{(k_\alpha)}(Z) \quad \forall \pi \in \Pi.$$

Outline

Problem

Model

- Consider a structured model:

$$Y_t = f(X_t) + \epsilon_t, \quad t = 1, 2, \ldots \tag{2}$$

- $\{X_t\}$ is a time series (possibly non-stationary).
- $f : \mathcal{X} \to \mathbb{R}$ is an unknown function (to be learnt).
- ϵ_t are independent and identically distributed with a CDF F, which is Lipschitz continuous with constant $L_F > 0$.

Ideas

- Transform the observation into an exchangeable sequence (the residuals).
- Precisely, we compute the quantiles of $\widehat{\epsilon}_i = y_i - \widehat{f}(x_i)$ where $\widehat{f}$ is an estimator of f .
- To make this idea work, we need $\widehat{f}$ to be a good estimator of f.

Constructing Prediction Interval

First we suppose the function f and the CDF of residual F is known. For any $\beta \in [0, \alpha]$, we can construct a valid *oracle* prediction for y_{T+1} region as

$$C_\alpha^{\mathrm{oracle}} = [f(x_{T+1}) + F^{-1}(\beta^*), f(x_{T+1}) + F^{-1}(1 - \alpha + \beta^*)]$$

Moreover, we can minimize the length of interval by setting

$$\beta^* := \underset{\beta \in [0,\alpha]}{\operatorname{argmin}}(F^{-1}(1 - \alpha + \beta) - F^{-1}(\beta))$$

The prediction region for Y_{T+1} is constructed as its empirical counterpart:

- Let $\widehat{f}_{-i}$ be a leave-one-out estimator of f.
- Compute the residuals: $\widehat{\epsilon}_i = y_i - \widehat{f}_{-i}(x_i)$ and their empirical CDF $\widehat{F}$.
- Choose $\widehat{\beta} = \operatorname{argmin}_{\beta \in [0,\alpha]}(\widehat{F}^{-1}(1 - \alpha + \beta) - \widehat{F}^{-1}(\beta)))$
- Construct $\widehat{C}_\alpha = [\widehat{f}_{-(T+1)}(x_{T+1}) + \widehat{F}^{-1}(\widehat{\beta}), \widehat{f}_{-(T+1)}(x_{T+1}) + \widehat{F}^{-1}(1 - \alpha + \widehat{\beta})]$

Validity

First we show the constructed prediction interval is approximately valid:

Theorem 2 ([Xu and Xie, 2021])

Suppose there exists a sequence $\{\delta_T\}_{T\geq 1}$ that converges to 0 such that

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{f}_{-t}(x_t) - f(x_t))^2 \leq \delta_T^2 \tag{3}$$

Then, for any T and $\alpha \in (0,1)$,

$$|\mathbb{P}(Y_{T+1} \in \widehat{C}_\alpha | X_{T+1} = x_{T+1}) - (1-\alpha)| \leq 24\sqrt{\frac{\log(16T)}{T}} + 4(L_F+1)\delta_T^{2/3} \tag{4}$$

Notice that the assumption in eqn (10) may fail to hold if the distribution of X_t changes dramatically at some point, in which case we are facing distribution shift.

Validity: Proof Sketch

1. First, we bound the result in terms of CDFs:

$$
\begin{aligned}
&|\mathbb{P}(Y_{T+1} \in \widehat{C}_\alpha | X_{T+1} = x_{T+1}) - (1-\alpha)| \\
=&|\mathbb{P}(\beta \le \widehat{F}(\widehat{\epsilon}_{T+1}) \le 1-\alpha+\beta) - \mathbb{P}(\beta \le F(\epsilon_{T+1}) \le 1-\alpha+\beta)| \\
\le& \cdots \le 4\left(\sup_x |\tilde{F}(x) - F(x)| + |\tilde{F}(x) - \widehat{F}(x)|\right)
\end{aligned}
$$

   where $\tilde{F}$ is the empirical CDF of the true residual.

2. Since ϵ_t are iid, so tehir empirical CDF approximates their true CDF:

$$\sup_x |\tilde{F}(x) - F(x)| \le \cdots \le \sqrt{\log(16T)/T} \tag{5}$$

3. By assumption, the LOO estimator of f has small error, so we can accurately estimate ϵ and its empirical CDF:

$$\sup_x |\tilde{F}(x) - \widehat{F}(x)| \le \cdots \le 24\sqrt{\log(16T)/T} + 4(L_F+1)\delta_T^{2/3} \tag{6}$$

Efficiency

Next we study the size of the prediction region.

Theorem 3 ([Xu and Xie, 2021])

Suppose the assumptions of Thm. 2 holds. Furthermore, assume there exists a sequence $\{\delta_T\}_{T\geq 1}$ *that converges to 0 such that*

$$|f(x_{T+1}) - \widehat{f}_{-(T+1)}(x_{T+1})| \leq \gamma_T$$

Lastly, assume $F^{-1}, \widehat{F}^{-1}, \widehat{F}$ *are Lipschitz continuous with constants* K, K', K''*. Then,*

$$|\mathcal{C}_\alpha^{\text{oracle}} \triangle \widehat{\mathcal{C}}_\alpha| \leq \gamma_T + 2(K + M_{T+1})(12\sqrt{\log(16T)/T} + 2C\delta_T^{2/3}) \tag{7}$$

where M_{T+1} *is a function that depends only on* K *and* K''*.*

Efficiency: Proof Sketch

1. For one part of the symmetric difference,

$$\begin{aligned}|\mathcal{C}_\alpha^{\text{oracle}} \triangle \widehat{\mathcal{C}}_\alpha|_{\text{right}} \leq\ & |f(X_{T+1}) - \widehat{f}_{-(T+1)}(X_{T+1})| \\ & + |\widehat{F}^{-1}(1-\alpha+\widehat{\beta}) - F^{-1}(1-\alpha+\widehat{\beta})| \\ & + |F^{-1}(1-\alpha+\widehat{\beta}) - F^{-1}(1-\alpha+\beta^*)|\end{aligned}$$

2. The first term is small by assumption.
3. The previous theorem's proof implies $\sup_x |\widehat{F}(x) - F(x)|$ is small. By the Lipschitz condition, the difference of inverses should also be bounded.
4. Since $\widehat{F}^{-1}$ converges to F^{-1}, we know $\widehat{\beta}$ converges to β^*.

Note: in the original paper, the author also considers the additional error brought by finding $\widehat{\beta}$ using grid search. We omit this step for simplicity.

A General Model-free Prediction Principle

Politis [Politis, 2015] proposed the following *Model-free Prediction Principle* for dependent data:

- Find a transformation H_t that maps the data onto a sequence $\epsilon_{1:t}^{(t)}$ that is iid with CDF F_t satisfying $F_t \to_{\mathrm{d}} F$.
- Suppose H_m is invertible. We can write

$$Y_t = g_t(Y_{1:t-1}, X_{1:t}, \epsilon_{1:t}^{(t)})$$

- To predict unobserved Y_{T+1}, we solve the equation:

$$Y_{T+1} = g_m\left(Y_{1:T}, X_{1:T+1}, \epsilon_{1:T+1}^{(T+1)}\right)$$

where ϵ_{T+1} is drawn from an estimated CDF F_{T+1}.

Outline

Outline

1. Introduction
2. Exchangeability with Respect to A Subset of Permutations
3. Exchangeability After A Transformation
4. No Exchangeablility, But Dependency Is Weak
   - α-Mixing
   - α-Mixing Residuals
   - α-Mixing Implies Approximate Ergodicity
5. Dealing with General Dependency: Adaptive Algorithms
6. Summary

Characterizing Dependency: α-Mixing Processes

- The α-mixing coefficient is a measure of dependency between two σ-fields. For a time series $\{u_t\}_t$, it is defined as:

$$\alpha(m) := \sup\{\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \\ : A \in \sigma(\{u_t : t \leq s\}), B \in \sigma(\{u_t : t \geq s+m\}), s \in \mathbb{N}\}$$

- $\{u_t\}$ is said to be **strongly mixing** (or mixing for short) if $\alpha(m) \to 0$ as $m \to \infty$. In words, the process forgets its history in the long run.

Theorem 4 ([Rio, 2017])

Let $\{u_t\}$ be strictly stationary with common CDF G and let $\alpha(m)$ be the sequence of its mixing coefficients. Let $\nu_n(x) = \sqrt{n}(\widehat{G}_n(x) - G(x))$. Then

$$\mathbb{E}\left(\sup_{x \in \mathbb{R}} |\nu_n(x)|^2\right) \leq \left(1 + 4\sum_{m=0}^{n-1} \alpha(m)\right)\left(3 + \frac{\log n}{2\log 2}\right)^2 \tag{8}$$

Mixing and Exchangeability

\# Example of Stationary and Mixing but Not Exchangeable Process

- Let $\{\epsilon_t\}_{t\geq 0}$ be iid Bernoulli variables and $X_t = \epsilon_t + \epsilon_{t-1}$ for all $t \geq 1$.
- $\{X_t\}$ is strictly stationary and strongly mixing since X_{t_0} and $X_{t\geq t_0+2}$ are independent.
- However, $\{X_t\}$ is not exchangeable. For example, $\mathbb{P}(X_{1:3} = (012)) > 0$ but $\mathbb{P}(X_{1:3} = (201)) = 0$.

\# Example of Exchangable but Not Mixing Process

- Let $\{u_t\}_{t\geq 0}$ be generated as $u_0 \sim \mathrm{Ber}(0.5)$ and $u_t \equiv u_0$ for all $t > 0$.
- $\{u_t\}_t$ is exchangeable since all of its elements are identical almost surely. However, it is not mixing since u_t is always remembers u_0.

In short, mixing and exchangeability do not imply each other. However, mixing provides a quantified characterization of dependence, which can be used to measure the difficulty of the prediction task.

Outline

α-Mixing Residuals

In this subsection, we aim to generalize the result in Section 3.
Recall the aforementioned model studied in [Xu and Xie, 2021]:

$$Y_t = f(X_t) + \epsilon_t, \quad t = 1, 2, \ldots \tag{9}$$

where $\{X_t\}$ is a time series (possibly non-stationary).

- We saw that when $\{\epsilon_t\}$ is an iid sequence we estimate its CDF well. This is the key to build valid and efficient prediction intervals.
- Theorem 4 tells that a good CDF estimator can also be obtained for α-mixing processes if $\sum_m \alpha(m)$ is small.
- Therefore, we can generalize Theorem 2 by relaxing the iid assumption to a mixing assumption.

α-Mixing Residuals

Theorem 5 (Generalized Theorem 2 [Xu and Xie, 2021])

Suppose there exists a sequence $\{\delta_T\}_{T\geq 1}$ that converges to 0 such that

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{f}_{-t}(x_t) - f(x_t))^2 \leq \delta_T^2 \tag{10}$$

and $\{\epsilon_t\}_{t=1}^{T+1}$ are stationary and stongly mixing with $\sum_{k\geq 0}\alpha(k) < M$. Then, for any T and $\alpha \in (0,1)$,

$$|\mathbb{P}(Y_{T+1} \in \widehat{C}_\alpha | X_{T+1} = x_{T+1}) - (1-\alpha)| \to 0 \text{ as } T \to \infty$$

This can be proved by replacing the estimation of the CDF in the proof of Theorem 2 with the α-mixing version (Theorem 4).
Using the same technique, we can also generalize the efficiency result (Theorem 3) to α-mixing residuals.

Outline

Problem

In this subsection, we aim to generalize the result in Section 2.

- We aim to make S-step-ahead predictions: given $Z_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^T$ and $x_{T+1:T+S}$, we predict $y_{\text{test}} = y_{T+1:T+S}$.
- We saw in [Chernozhukov et al., 2018] that if the joint distribution of $Z = (Z_{\text{obs}}, Z_{\text{test}})$ is invariant under a group of permutations Π, then we can construct a valid prediction region as

$$\mathcal{C}_\alpha := \left\{ y_{\text{test}} \in \mathcal{Y}^S \middle| \widehat{p} = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbb{1}\{S(Z^\pi) \geq S(Z)\} > \alpha \right\}$$

- Now we extend the result by relaxing the invariance condition and assuming the scores $\{S(Z^\pi)\}$ form a strongly mixing process.

Approximate Ergodicity Leads to Approximate Validity

To achieve validity, [Chernozhukov et al., 2018] proposes the following two sets of conditions for the oracle score and its approximation.

(E) Approximate Ergodicity

- With probability $1 - \gamma_{1n}$, the empirical CDF of the oracle score $\tilde{F}(x) := \sum_{\pi \in \Pi} \mathbb{1}\{S_*(Z^\pi) < x\}/|\Pi|$ is **approximately ergodic** for $F(x) := \mathbb{P}(S_*(Z) < x)$, namely

$$\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F(x) \right| \leq \delta_{1n}.$$

  In words, the average of the oracle scores in Π is close to the average in probability space.

(A) Approximating the Oracle: with probability $1 - \gamma_{1n}$:

- Average error is small: $\sum_\pi [S(Z^\pi) - S_*(Z^\pi)]^2/n \leq \delta_{2n}^2$
- Pointwise error is small: $|S(Z) - S_*(Z)|/n \leq \delta_{2n}$
- (Lipschitz) The PDF of $S_*(Z)$ is upper bounded by a constant D .

Approximate Ergodicity Leads to Approximate Validity

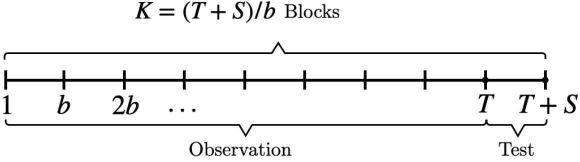
Theorem 6

If conditions (E) and (A) are satisfied, then

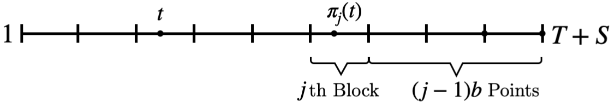
$$
\begin{aligned}
&|\mathbb{P}(Y_{\text{test}} \in \mathcal{C}_\alpha) - (1-\alpha)| \\
&\qquad \leq 6\delta_{1n} + 4\delta_{2n} + 2D\left(\delta_{2n} + 2\sqrt{\delta_{2n}}\right) + \gamma_{1n} + \gamma_{2n}.
\end{aligned}
$$

Designing Permutations Π for Dependent Data

1. We split the data into $K = (T+S)/b$ blocks (typically $b = S$).



2. For each block $j \in [K]$, define $\pi_j(t) := (t + (j-1)b - 1) \bmod (T+S) + 1$.



3. We let $\Pi_{\mathrm{NOB}} := \{\pi_j : j \in [K]\}$. Note that the original paper also considers a different type of permutation using overlapping blocks.
4. Using the language of residual, by doing so we compute the empirical residual for each block.

α-Mixing Implies Approximate Ergodicity

Theorem 7

Suppose that there exists a rearrangement of $\{S_*(Z^\pi) : \pi \in \Pi_{\mathrm{NOB}}\}$, *denoted by* $\{u_t\}_{t=1}^K$, *such that* $\{u_t\}_{t=1}^K$ *is stationary and strongly mixing with* $\sum_{k=1}^\infty \alpha(k) \leq M$. *Then, there exists a constant* $M' > 0$ *depending only on* M *such that*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\tilde{F}(X) - F(x)\right| \leq \delta_{1n}\right) \geq 1 - \gamma_n \tag{11}$$

where $\gamma_n = M'(\log K)^2 / (K\delta_{1n})$.

Proof. By Theorem 4, we have

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}} |\tilde{F}(x) - F(x)|^2\right] \leq \frac{1+4M}{K}\left(3 + \frac{\log K}{2\log 2}\right)^2$$

The result then follows by directly applying Markov's inequality.

Outline

ACI: Adaptive Conformal Inference

- [Gibbs and Candès, 2021] proposes an Adaptive Conformal Inference (ACI) algorithm for the online learning setting.
- ACI uses the historical miscoverage frequency to adaptively update the significance level α_t^* used in the construction of the prediction region:

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha_t - \mathbb{1}\{Y_t \notin C_{\alpha_t}^t\})$$

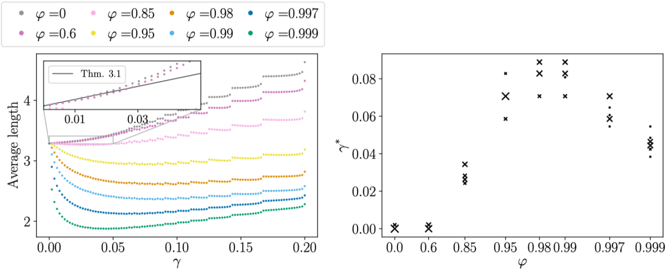
- ACI is shown to be asymptotically (in Cesàro sum) valid without any assumptions on the data generating distribution:

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{Y_t \notin C_{\alpha_t}^t\} - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \gamma}{T\gamma}$$

Efficiency of ACI

Recently[1] a newer paper [Zaffran et al., 2022] analyses ACI's efficiency. They found:

- In exchangeable cases, ACI degrades the efficiency linearly with γ compared to CP. In this case, a smaller γ is more perferable.
- For models with autoregressive residuals, there is a strictly positive γ^* that improves the efficiency. This shows adaptive methods can produce smaller intervals than CP.



[1] yesterday

AgACI: Parameter-free ACI

[Zaffran et al., 2022] also proposed two strategies based on ACI that avoid the issue of choosing γ.

- Naive strategy: at each step, choose the γ that achieved the best efficiency while ensuring validity in the past data.
- Online Expert Aggregation on ACI (AgACI): performs ACI with K different values of γ and use an aggregation rule to weight and combine the K prediction intervals.
- They show empirically that AgACI achieves valid coverage with good efficiency.

Outline

Summary

- Exchangeability with respect to all permutations is not necessary to achieve validity, any group invariance suffices.
- When we can transform the data to exchangeable or strongly-mixing data, we can construct prediction regions that have theoretical guarantees on validity and efficiency.
- ACL works for time series with general dependency; and is valid. It is also shown empirically to be efficient with a well chosen γ (possibly 0). AgACI avoids the choice of γ using online expert aggregation.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018).
Exact and robust conformal inference methods for predictive machine learning with dependent data.
In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732–749. PMLR.

Gibbs, I. and Candès, E. J. (2021).
Adaptive conformal inference under distribution shift.

Politis, D. (2015).
Model-Free Prediction and Regression.
Springer.

Rio, E. (2017).
Asymptotic Theory of Weakly Dependent Random Processes.
Springer.

Xu, C. and Xie, Y. (2021).
Conformal prediction interval for dynamic time-series.
In *ICML*.

Zaffran, M., Dieuleveut, A., F'eron, O., Goude, Y., and Josse, J. (2022).
Adaptive conformal predictions for time series.