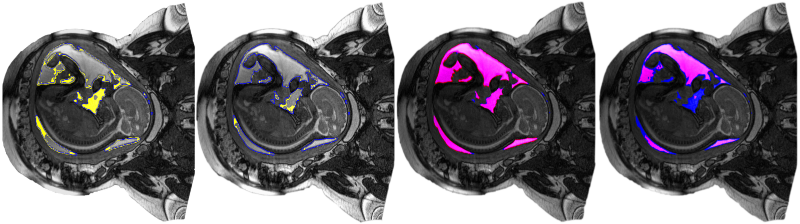
Split Conformal Prediction for Dependent Data

Paulo Orenstein

September 28th, 2022

IMPA



Joint work with Roberto Imbuzeiro Oliveira, Thiago Ramos, João Vitor Romano and others

Agenda

- Motivation: the need for uncertainty quantification
- Solution: split conformal prediction, with a single crucial assumption
- Extending split CP to dependent data: new results
- In practice: effect of dependency is negligible
- Conclusion: further directions

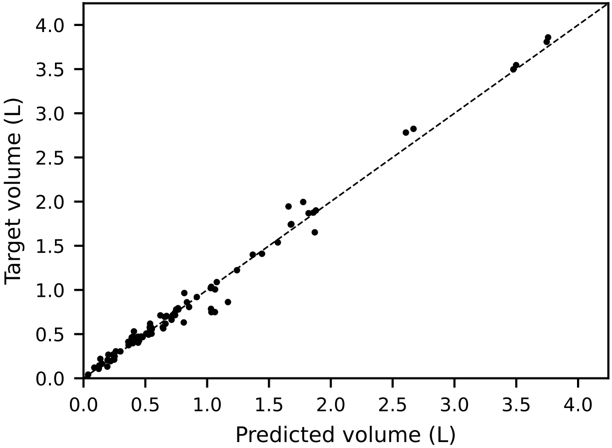
Video with blue solid.

Motivation

- Dr Heron Werner (DASA): "Given fetal MRI images, can we predict the amount of amniotic fluid"?
  - abnormal volume indicates pregnancy pathologies
  - usual measurements are imprecise or subjective
  - estimation is manually done by trained physician, taking hours to days
- Goal: accurate algorithm for volume estimation, in seconds
- How: segment each layer in the MRI using U-Net, count voxel size for volume
- Results: $\sim 92\%$ Dice accuracy in under 5 seconds

Video with estimates.

Results


Target volume (L)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Predicted volume (L)

Problem: uncertainty quantication

- Can we really trust the results?
- In medicine, uncertainty quantification is crucial; best guess is 2.80L but...
  - "I'm 90% sure the true AF volume is between 2.72L and 2.88L"
  - "I'm 90% sure the true AF volume is between 1.90 and 3.70L"
- How can we provide valid predictive intervals for black-box prediction methods?

Given data $\{(X_i, y_i)\}_{i=1}^n$ to train any prediction method $\hat{\mu}$ and any level $\alpha \in (0, 1)$, can we construct a prediction set $C_{1-\alpha}(x)$ such that, for a new point (X_{n+1}, y_{n+1}),

$$\mathbb{P}[y_{n+1} \in C_{1-\alpha}(X_{n+1})] \geq 1 - \alpha?$$

(For us, X_i is an MRI exam, y_i is the fluid volume, $\hat{\mu}$ is a U-Net, C is a rule specifying a volume interval for X_i.)

Conformal Prediction

- Conformal Prediction was proposed by Vladimir Vovk[*]
- Provides valid predictive sets for any level $\alpha \in (0, 1)$ and any model $\hat{\mu}$
- Many recent variations and extensions, from regression to classification settings[†]
- We will consider the most popular incarnation: split CP[‡]
- Important assumption: data $(X_i, y_i)_{i=1}^n$ is exchangeable (which is implied by iid)

[*] Vovk, Gammerman, and Shafer. "Algorithmic learning in a random world", Springer (2005).
[†] Angelopoulos and Bates, "A Gentle Introduction to Conformal Prediction", arXiv (2021).
[‡] Lei, G'Sell, Rinaldo, Tibshirani, and Wasserman, "Distribution-free predictive inference for regression", JASA (2018).

Split Conformal Prediction: Setup

- Split the data: $\{(X_i, y_i)\}_{i \in I_{tr}}$, $\{(X_j, y_j)\}_{j \in I_{cal}}$, $\{(X_k, y_k)\}_{k \in I_{test}}$, with sizes $n_{tr}, n_{cal}, n_{test}$
- Train predictive method $\hat{\mu}_{tr} : \mathcal{X} \to \mathcal{Y}$
- Discrepancy scores $\hat{s}_{tr} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (e.g., $\hat{s}_{tr}(x, y) = |y - \hat{\mu}(x)|$)
- Calibrate quantile: if $\hat{s}_j = \hat{s}_{tr}(X_j, y_j)$ for $j \in I_{cal}$,

$$\hat{q}_{1-\alpha} := \hat{q}_{1-\alpha}\left(\{\hat{s}_j\}_{j \in I_{cal}}\right) = \underset{t \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n_{cal}} \sum_{j \in I_{cal}} \mathbb{I}_{[\hat{s}_j \leq t]} \geq 1 - \alpha \right\}$$

- Prediction set:

$$C_{1-\alpha}(x) = \{y \in \mathcal{Y} \ : \ \hat{s}_{tr}(x, y) \leq \hat{q}_{(1+1/n_{cal})(1-\alpha)}\}.$$

Split Conformal Prediction: Results

Marginal coverage

Given exchangeable data $\{(X_i, y_i)\}_{i=1}^n$ and level $1-\alpha \in (0,1)$, consider the calibrated quantile $\hat{q}_{(1+1/n_{cal})(1-\alpha)}$ and define

$$C_{1-\alpha}(x) = \{y \in \mathcal{Y} \ : \ \hat{s}_{\text{tr}}(x,y) \leq \hat{q}_{(1+1/n_{cal})(1-\alpha)}\}.$$

Then, for any single test data point (X_k, y_k), $k \in I_{\text{test}}$,

$$\mathbb{P}[y_k \in C_{1-\alpha}(X_k)] \geq 1-\alpha.$$

Additionally, if $\hat{s}_j$ are almost surely distinct, then $\mathbb{P}[y_k \in C_{1-\alpha}(X_k)] \leq 1-\alpha + 1/(n_{\text{cal}}+1)$.

Proof sketch: since data is exchangeable, $\hat{s}_j$ are also exchangeable. Consider the $1-\alpha$ quantile of $\{\hat{s}_j\}_{j\in I_{\text{cal}}} \cup \{\hat{s}_k\}$; the probability of $\hat{s}_k$ being bigger than the quantile must be bigger than $1-\alpha$. Issue: can't use $\hat{s}_k$ for the quantile, but can you can assume it's infinite:

$$\hat{s}_k > \hat{q}_{1-\alpha}(\{\hat{s}_j\}_{j\in I_{\text{cal}}} \cup \{\hat{s}_k\}) \iff \hat{s}_k > \hat{q}_{1-\alpha}(\{\hat{s}_j\}_{j\in I_{\text{cal}}} \cup \{\infty\}).$$

So: $\mathbb{P}[\hat{s}_k \leq \hat{q}_{(1+1/n_{\text{cal}})(1-\alpha)}(\{\hat{s}_j\}_{j\in I_{\text{cal}}})] = \mathbb{P}[\hat{s}_k \leq \hat{q}_{(1-\alpha)}(\{\hat{s}_j\}_{j\in I_{\text{cal}}} \cup \{\infty\})] \geq 1-\alpha.$ $\square$

Split Conformal Prediction: Results

Empirical coverage

If the data $\{(X_i, y_i)\}_{i=1}^n$ is iid, then for any $\varepsilon > 0$ there exists $c_\varepsilon > 0$ such that

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}}\sum_{k\in I_{\text{test}}} \mathbb{I}_{[y_k\in C_{1-\alpha}(X_k)]} \geq 1-\alpha-\varepsilon\right] \geq 1 - e^{-c_\varepsilon n_{\text{test}}}.$$

So, empirically over the entire test set, $C_{1-\alpha}$ approximates the $1-\alpha$ quantile (with a penalty).

Conditional coverage

If the data $\{(X_i, y_i)\}$ is iid and $\mathcal{A} \subset \mathcal{X}$ has finite VC dimension, then for any $A \in \mathcal{A}$ where $\mathbb{P}[X_k \in A]$ is not too small,

$$\mathbb{P}\left[y_k \in C_{1-\alpha}(X_k; K) \mid X_k \in A\right] \geq 1-\alpha-\varepsilon.$$

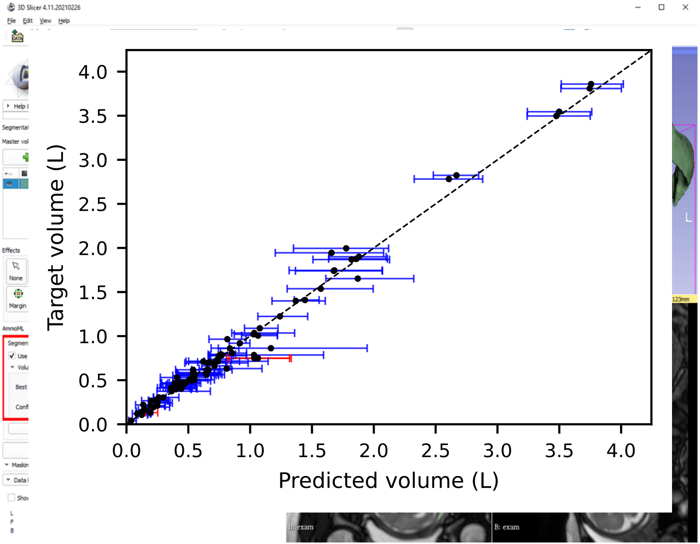
Thus, split CP can guarantee coverage even if conditioned on some events.

Split Conformal Prediction: General Tool

- Provides valid coverage and finite-sample statistical guarantees
- Works for any exchangeable data $\{(X_i, y_i)\}_{i=1}^n$, any model $\hat{\mu}$
- Simple to implement, computationally cheap
- Arbitrary discrepancy score $\hat{s}_{\text{tr}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$:
  - residuals: $\hat{s}_{\text{tr}}(x, y) = |y - \hat{\mu}(x)|$
  - conditional likelihood: $\hat{s}_{\text{tr}}(x, y) = -\log \hat{p}(y|x)$
  - conformalized quantile: $\hat{s}_{\text{tr}}(x, y) = \max\{\hat{\mu}_{\alpha/2}(x) - y, y - \hat{\mu}_{1-\alpha/2}(x)\}$
- Many more generalizations: e.g., prediction masks*

*Bates, Angelopoulos, Lei, Malik, and Jordan, "Distribution-free, risk-controlling prediction sets"

Results: Split CP



But severe limitation: without exchangeability theory falls apart

(For us, there could be some dependency across exams.)

Dealing with Dependence

- Recent interest in independent data with distributional drift*
- Our work†: rebuild split conformal prediction without exchangeability
- Intuition: see how data CDF concentrates when exchangeability is replaced by looser conditions:

$$\mathbb{P}[y_k \in C_{1-\alpha+\eta}(X_k)] \geq 1-\alpha, \text{ so } \mathbb{P}[y_k \in C_{1-\alpha}(X_k)] \geq 1-\alpha-\eta,$$

  where η is an added penalty due to non-exchangeability
- Tools: concentration inequalities and decoupling properties

*Barber, Candès, Ramdas and Tibshirani. "Conformal prediction beyond exchangeability", arXiv (2022).
†Oliveira, O., Ramos, Romano, "Split Conformal Prediction for Dependent Data", arXiv (2022).

Theoretical Results

- Assumptions on data:
  - Stationarity: $(Z_t, \ldots, Z_m) \stackrel{d}{=} (Z_{t+k}, \ldots, Z_{t+m+k})$
  - β-mixing: $\beta(a) = \|\mathbb{P}_{-\infty:0,a:\infty} - \mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty}\|_{TV} \xrightarrow{a\to\infty} 0$
- Data is time-invariant and asymptotically independent
- Examples: Markov chains, renewal processes, AR(1)
- Main theoretical tool: Blocking technique*

*Yu, "Rates of Convergence of Empirical Processes of Stationary Mixing Sequences", Annals of Probability (1994)

Main Theoretical Results

Marginal coverage

Suppose that $\{(X_i, y_i)\}_{i=1}^n$ is stationary β-mixing. Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, for $k \in I_{\text{test}}$,

$$\mathbb{P}[y_k \in C_{1-\alpha}(X_k)] \geq 1 - \alpha - \eta,$$

with $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{tr}} + \delta_{\text{cal}}$, where $\varepsilon_{\text{tr}} = \beta(k - n_{\text{tr}})$.

Empirical coverage

Suppose that $\{(X_i, y_i)\}_{i=1}^n$ is stationary β-mixing. Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, $\delta_{\text{test}} > 0$:

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{k \in I_{\text{test}}} \mathbb{I}_{[y_k \in C_{1-\alpha}(X_k)]} \geq 1 - \alpha - \eta\right] \geq 1 - \delta_{\text{cal}} - \delta_{\text{test}},$$

with $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$.

The Details

- $F_{\text{cal}} = \left\{(a, m, r) \in \mathbb{N}_+^3 : 2ma = n_{\text{cal}} - r + 1, \delta_{cal} > 4(m-1)\beta(a) + \beta(r)\right\}$
- $F_{\text{test}} = \left\{(a, m, s) \in \mathbb{N}_+^3 : 2ma = n_{\text{test}} - s, \delta_{test} > 4(m-1)\beta(a) + \beta(n_{\text{cal}})\right\}$
- $\tilde{\sigma}(a) = \sqrt{1/4 + (2/a)\sum_{j=1}^{a-1}(a-j)\beta(j)}$
- $\varepsilon_{\text{cal}} = \inf_{(a,m,r)\in F_{\text{cal}}} \left\{\tilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{cal}}-r+1}\log\left(\frac{4}{\delta_{\text{cal}}-4(m-1)\beta(a)-\beta(r)}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{cal}}-4(m-1)\beta(a)-\beta(r)}\right) + \frac{r-1}{n_{\text{cal}}}\right\}$
- $\varepsilon_{\text{test}} = \inf_{(a,m,s)\in F_{\text{test}}} \left\{\tilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{test}}}\log\left(\frac{4}{\delta_{\text{test}}-4(m-1)\beta(a)-\beta(n_{\text{cal}})}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{test}}-4(m-1)\beta(a)-\beta(n_{\text{cal}})}\right) + \frac{s}{n_{\text{test}}}\right\}$

Conditional Theoretical Results

Marginal coverage, conditional version

Suppose that $\{(X_i, y_i)\}_{i=1}^n$ is stationary β-mixing. Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, for any $k \in I_{\text{test}}$ and $K \in \mathcal{K}$ (with $\text{VC}(\mathcal{K}) = d, \mathbb{P}[X_k \in K] > \gamma$),

$$\mathbb{P}[y_k \in C_{1-\alpha}(X_k; K) \mid X_k \in K] \geq 1 - \alpha - \eta,$$

with $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$.

Empirical coverage, conditional version

Suppose that $\{(X_i, y_i)\}_{i=1}^n$ is stationary β-mixing. Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, $\delta_{\text{test}} > 0$ and $K \in \mathcal{K}$:

$$\mathbb{P}\left[\inf_{K \in \mathcal{K}} \frac{1}{n_{\text{test}}(K)} \sum_{k \in I_{\text{test}}(K)} \mathbb{I}_{[y_k \in C_{1-\alpha}(X_k; K)]} \geq 1 - \alpha - \eta\right] \geq 1 - \delta_{\text{cal}} - \delta_{\text{test}},$$

with $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$.

The Details

- $G_{\text{cal}} = \left\{(a, m, r) \in \mathbb{N}_+^3 : 2ma = n_{\text{cal}} - r + 1, \delta_{cal} > 16(m-1)\beta(a) + \beta(r)\right\}$

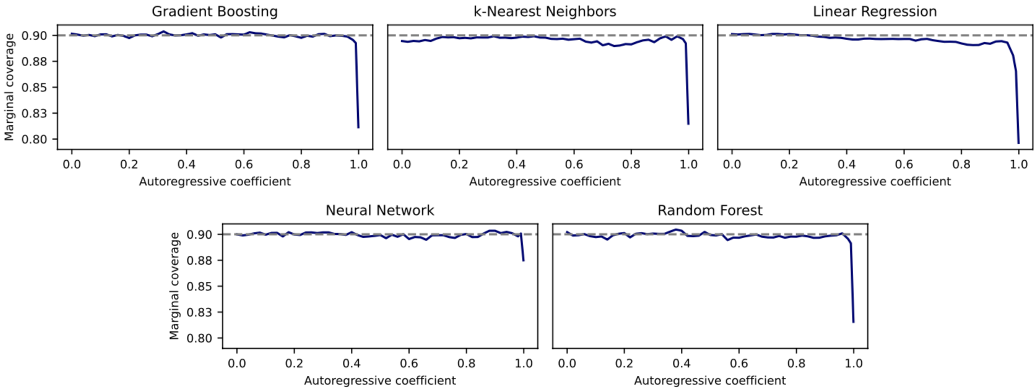
- $G_{\text{test}} = \left\{(a, m, s) \in \mathbb{N}_+^3 : 2ma = n_{\text{test}} - s, \delta_{test} > 8(m-1)\beta(a) + \beta(n_{\text{cal}})\right\}$

- $\varepsilon_{\text{cal}} = \inf_{(a,m,r)\in G_{\text{cal}}} \left\{ \frac{1}{\gamma} \left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + \frac{2(r-1)}{n_{\text{cal}}} + 2\sqrt{\frac{1}{2m} \log\left(\frac{16}{\delta_{\text{cal}} - 16(m-1)\beta(a) - \beta(r)} \right)} \right) \right\}$

- $\varepsilon_{\text{test}} = \inf_{(a,m,s)\in G_{\text{test}}} \left\{ \frac{1}{\gamma} \left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + \frac{2s}{n_{\text{test}}} + 2\sqrt{\frac{1}{2m} \log\left(\frac{8}{\delta_{\text{test}} - 8(m-1)\beta(a) - \beta(n_{\text{cal}})} \right)} \right) \right\}$

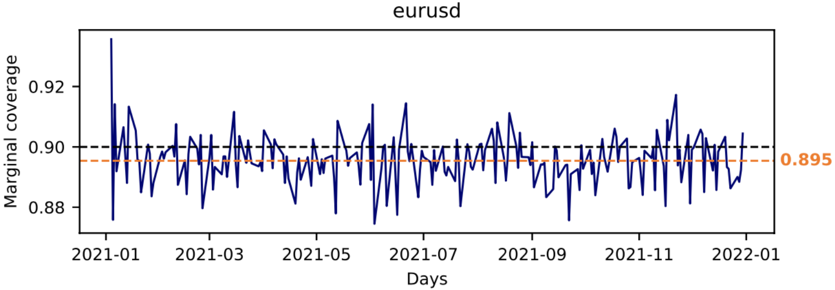
Application: Autoregressive Process

- For every 11 points in AR(1) time series, predict the following point
- Get predictive set via split conformal quantile regression



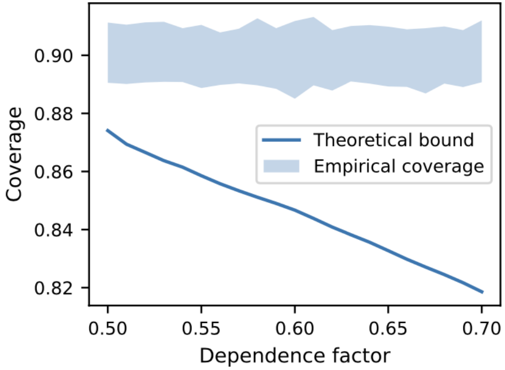
Application: Finance

- Time series with EUR/USD spot exchange rate; predictions with boosting
- Sliding window of 1000 training points, 500 calibration points and 1 test point
- Get predicitive set via split conformal quantile regression



Application: Empirical Coverage

- Two-state hidden Markov model
- Gradient boosting model with 1000 training points, 15000 calibration points and 15000 test points
- Average over 1000 simulations to ascertain empirical coverage: $\frac{1}{n_{\text{test}}} \sum_{k \in I_{\text{test}}} \mathbb{I}_{[y_k \in C_{1-\alpha}(X_k)]}$



Conclusion

- Uncertainty quantification is crucial for the deployment of ML systems.
- Conformal prediction is a set of tools that yield marginal, empirical and conditional coverage.
- It traditionally requires little beyond exchangeability; we show it works even for dependent data.
- Our results can be extended beyond stationarity and to non-split CP (e.g., rank-one-out, risk-controlling prediction sets).
- There is much more theory and algorithms to be developed on top of it.

References

- Vovk, Gammerman, Shafer, *Algorithmic Learning in a Random World*. Springer, 2005
- Lei, G'Sell, Rinaldo, Tibshirani, Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association* , vol. 113, no. 523, pp. 1094–1111, 2018
- Angelopoulos, Bates, "A Gentle Introduction to Conformal Prediction and Distribution-free Uncertainty Quantification," *arXiv*, 2021
- Csillag, Monteiro, Ramos, Romano, Schuller, Seixas, Oliveira, O., "AmnioML: Amniotic Fluid Segmentation and Volume Prediction with Uncertainty Quantification," in submission, 2022
- Barber, Candès, Ramdas, Tibshirani. "Conformal Prediction Beyond Exchangeability," *arXiv*, 2022
- Oliveira, O., Ramos, Romano, "Split Conformal Prediction for Dependent Data", *arXiv*, 2022