UNIVERSITY OF OXFORD

DEPARTMENT OF STATISTICS

Split Conformal Prediction under Data Contamination

Gesine Reinert

Department of Statistics
University of Oxford
and
The Alan Turing Institute

with Jase Clarkson, Wenkai Xu and Mihai Cucuringu.

COPA 2024
The 13th Symposium on Conformal and Probabilistic Prediction with Applications

Split Conformal Prediction

(*Gammerman, Vovk and Vapnik (1998)*)
Data points $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, with $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$

Model $\hat{f} : \mathcal{X} \to \mathcal{Y}$

Example: $\hat{f}$ predicts that Y is of class $i \in \{1, \ldots, K\}$ when $X = x$ is observed

Aim: For an observed X_{n+1} obtain a $(1-\alpha)$-probability prediction set for a test datapoint $Z_{n+1} = (X_{n+1}, Y_{n+1})$

On-line setting: Y_i's are predicted successively, each one is revealed before the next one is predicted.

Tool: (non-conformity) score function $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The smaller, the better.

Example: classification

Suppose y_i is a perhaps non-numerical label for x_i. We observed "calibration data" $(x_i.y_i), i = 1, \ldots n$ and now we observe x.

Nearest-neighbour method:
find the x_i which is closest to the observed x
use the label of x_i as predicted label for y.

We could use as score

$$S(x, y) = \frac{\min\{|x_i - x| : 1 \leq i \leq n, y_i = y\}}{\min\{|x_i - x| : 1 \leq i \leq n, y_i \neq y\}}$$

comparing the distance of x to old objects with the same label to its distance to old objects with a different label.

Procedure

Intuition: predict y for which the corresponding score is "typical".

Compute the score for each calibration data-point $S_i = S(X_i, Y_i)$, take the order statistics $S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(n)}$, set

$$\hat{q} = S_{(i)} \text{ where } i = \lceil (1-\alpha)(n+1) \rceil.$$

Use as the prediction set

$$\widehat{C}_n\left(X_{n+1}\right) = \left\{y \in \mathcal{Y} : S\left(X_{n+1}, y\right) \leqslant \hat{q}\right\}.$$

If the data are exchangeable then

$$1-\alpha \leqslant \mathbb{P}\left(Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right) = \lceil (1-\alpha)(n+1) \rceil (n+1)^{-1} \leqslant 1-\alpha+(n+1)^{-1}.$$

[redacted]

Equivalent: Estimate the prediction set boundary $\hat{q}$ as

$$\hat{q} = Q_{1-\alpha}\left(\sum_{i=1}^{n} \delta_{S_i} + \delta_{+\infty}\right)$$

where δ_x is point mass at x and for a probability measure μ on $\mathbb{R}$,

$$Q_{1-\alpha}(\mu) = \inf\{x : \mu((-\infty, x]) \geq 1 - \alpha\}.$$

Extension to non-exchangeable situation: *Barber et al. (2023)*

Assumes that the data come from the same distribution.

What if not?

The Huber contamination model

Huber (1964, 1965)

Let $\epsilon \in [0, 1)$. Suppose that the calibration data are sampled i.i.d from a mixture model

$$\tilde{Z}_i = (X_i, Y_i) \sim (1 - \epsilon)\pi_1 + \epsilon\pi_2,$$

where π_1, π_2 are two distribution functions over $\mathcal{X} \times \mathcal{Y}$.

Then the scores $\tilde{S}(X_i, Y_i)$ are also distributed as a mixture,

$$\tilde{S}_i = \tilde{S}(X_i, Y_i) \sim \tilde{\Pi},$$

giving the standard i.i.d. setting, but for the contaminated distribution.

[redacted]

Split conformal prediction for the standard setting gives

$$\mathbb{P}(\tilde{S}_{n+1} \leqslant \tilde{q}) \geqslant 1 - \alpha \text{ for } \tilde{S}_{n+1} \sim \tilde{\Pi}$$

and $\tilde{q}$ the quantile for the mixture distribution.

Aim: a $(1-\alpha)$-probability prediction set for a "clean" test datapoint $Z_{n+1} = (X_{n+1}, Y_{n+1}) \sim \pi_1$

Theoretical guarantees

Recall: in the i.i.d. setting,

$$1-\alpha \leqslant \mathbb{P}\left(Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right) \leqslant 1-\alpha+(n+1)^{-1}.$$

Barber et al. (2023): In the Huber contamination model with $\tilde{Z}_i=(X_i, Y_i) \sim (1-\epsilon)\pi_1+\epsilon\pi_2$, and $Z_{n+1} \sim \pi_1$,

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right) \geqslant 1-\frac{\alpha}{1-\epsilon}.$$

They consider a slightly more general contamination model and relax the exchangeability assumption.

Theoretical guarantees continued

Sesia et al. (2024): Classification problem, K labels, i.i.d. observations, with latent labels Y_i and possibly contaminated observed labels $\tilde{Y}_i$
Let $n_k = |\{i \in 1, \ldots, n : \tilde{Y}_i = k\}|$, set $S_k(i) = \{S(X_i, k), i = 1, \ldots, n\}$,

$$\hat{q}_k = S_k(i) \text{ where } i = \lceil (1-\alpha)(n_k + 1) \rceil$$

and

$$\widehat{C}_{n,k}(X_{n+1}) = \{y \in \mathcal{Y} : S(X_{n+1}, k) \leqslant \hat{q}_k\}.$$

Then, for *label-conditional coverage*, if $Y_i = \tilde{Y}_i$ almost surely,

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,k}(X_{n+1}) \,|\, Y_{n+1} = k\right) \geqslant 1 - \alpha.$$

Sesia et al. (2024):

Notation: conditional distribution functions

$$F_\ell^k(t) = \mathbb{P}(S(X,k) \leq t | Y = \ell)$$

$$\tilde{F}_\ell^k(t) = \mathbb{P}(S(X,k) \leq t | \tilde{Y} = \ell);$$

coverage inflation factor

$$\Delta_k(t) = F_k^k(t) - \tilde{F}_k^k(t)$$

Then

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,k}\left(X_{n+1}\right) | Y_{n+1} = k\right) \geqslant 1 - \alpha + \mathbb{E}\Delta_k(\hat{q}_k).$$

If all scores are distinct: matching upper bound with an additive factor $(n+1)^{-1}$.

Our theoretical guarantees

Notation: $\tilde{\Pi} = (1-\epsilon)\Pi_1 + \epsilon \Pi_2$ has cumulative distribution function (cdf)

$$\tilde{F} = (1-\epsilon)F_1 + \epsilon F_2$$

where F_1, F_2 are cdfs over the scores computed from each mixture component. Under the mixture model, when $(X_{n+1}, Y_{n+1}) \sim \pi_1$, with $\mathbb{P}_1$ indicating this,

$$\begin{aligned} (1-\alpha) - \epsilon \mathbb{E}[F_2(\tilde{q}) - F_1(\tilde{q})] &\leqslant \mathbb{P}_1\left(Y_{n+1} \in \widehat{C}_n(X_{n+1})\right) \\ &\leqslant (1-\alpha) + \frac{1}{n+1} + \epsilon \mathbb{E}[F_1(\tilde{q}) - F_2(\tilde{q})] \end{aligned}$$

and $\mathbb{E}[F_1(\tilde{q}) - F_2(\tilde{q})]$ can be replaced by the Kolmogorov distance $d_K(\Pi_1, \Pi_2)$.

Example: Gaussian linear regression

$$
\begin{aligned}
Y &= \beta^T X + E,\\
E &\sim (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(0,\sigma_2^2),
\end{aligned}
$$

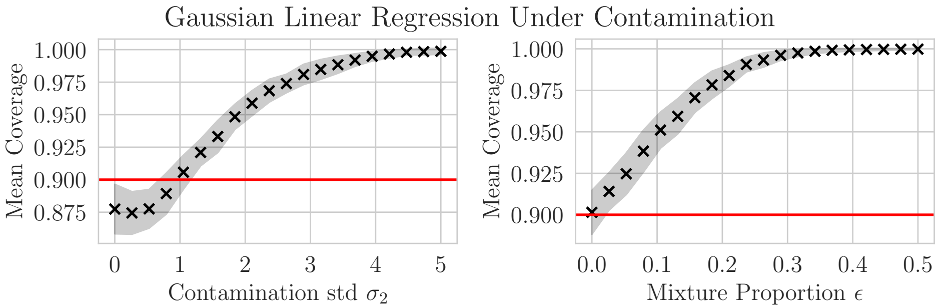
where β is known; use $S(X,Y) = |Y - \beta^T X|$. Then with $\sigma_1 = 1$,

$$
F_i(x) = \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma_i}\right), \quad x \geqslant 0.
$$

for $i = 1, 2$.

[redacted]

Coverage: $\mathbb{P}_1\left(Y_{n+1} \in \widehat{C}_n(X_{n+1})\right)$



Left: vary the standard deviation of the corruption σ_2 from 0 to 5, keeping $\epsilon = 0.2$.
Right: vary the mixing proportion ϵ from 0 to 0.5, keeping $\sigma_2 = 3.0$.

Classification under label noise

K classes; $X_i \sim F_X$, and $Y_i \sim F_{Y|X}$; Y denotes a true label and $\tilde{Y}$ an observed label.
We assume that
* labels are corrupted with probability $\epsilon \in (0, \frac{1}{2})$, independently of the conditional distribution $X|Y$
* $P_{ji} = P_{ji}(\epsilon) = \mathbb{P}(Y = j|\tilde{Y} = i)$ gives an invertible matrix
* for all $q \in \mathbb{R}$, $i \in \{1, \dots, K\}$,

$$\max_{c:c\neq i} \mathbb{P}(S(X, c) \leqslant q|Y = i) \leqslant \mathbb{P}(S(X, i) \leqslant q|Y = i).$$

Proposition: [Over-coverage] Then

$$\mathbb{P}_1(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geqslant 1 - \alpha.$$

Example: Uniform noise

Assume that the corrupting noise chooses one of the K labels uniformly at random, regardless of the true label, so that a corrupted label Y^c follows the uniform distribution on $[K]$ (this is a *randomised response model*).

Assume that the true label Y also follows the uniform distribution on $[K]$ (but in contrast to Y^c it contains a signal on X). Then

$$P^{-1} = \frac{1}{1-\epsilon} I - \frac{\epsilon}{K(1-\epsilon)} 1 1^\intercal$$

and the proposition applies (for suitable scoring functions).

Aim: Amend conformal prediction to reduce the over-coverage.

CRCP: Contamination Robust Conformal Prediction

Recall: F_1 is the true cdf and $\tilde{F}$ is the observable cdf (with contamination).

Set $g(q) := F_1(q) - \tilde{F}(q)$, and $i = \lceil (1-\alpha)(n+1) \rceil$. Then our proposition can be rephrased as

$$\mathbb{P}_1(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geqslant 1 - \alpha + \mathbb{E}[g(S_{(i)})].$$

Idea If we knew $\mathbb{E}g(S_{(j)}), j = 1, \ldots, n$, then we could instead take $i = i_c$ such that

$$i_c = \lceil (1 - \alpha - \mathbb{E}g(S_{(i_c)}))(n+1) \rceil$$

and $\tilde{q}_c = S_{(i_c)}$. Then using $\tilde{q}_c$ instead of $\tilde{q}$,

$$\mathbb{P}_1(Y_{n+1} \in \widehat{C}_n(X_{n+1})) = \lceil (1 - \alpha - \mathbb{E}g(S_{(i_c)}))(n+1) \rceil (n+1)^{-1} + \mathbb{E}[g(S_{(i_c)})] \geqslant 1 - \alpha.$$

But...

we do not know $\mathbb{E}g(S_{(j)}), j = 1, \dots, n$. Instead:

* estimate $g(q)$ by $\hat{g}_n(q)$,
* bound $\mathbb{E}[|g(S_{(i)}) - \hat{g}_n(S_{(i)})|] \leq C(n, \epsilon)$;
* instead of $\lceil (1-\alpha)(n+1) \rceil$, take $i = i_c$ as

$$i_c = \lceil (1 - \alpha - \hat{g}_n(S_{(i)}) + C(n, \epsilon))(n+1) \rceil.$$

Then

$$\mathbb{P}_1(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geqslant 1 - \alpha + \mathbb{E}[g(S_{(i)}) - \hat{g}_n(S_{(i)})] - C(n, \epsilon).$$

We call this Contamination Robust Conformal Prediction (CRCP).

Theoretical guarantee:

Set $w_i^{(1)} = P_{i,i}^{-1} P_i - \tilde{P}_i$ and $w_{ij}^{(2)} = P_i P_{ji}^{-1}$, and $b(n,j) = (1-\tilde{P}_j)^n + \sqrt{\frac{\pi}{n\tilde{P}_j}}$. Then

$$\mathbb{E}[|\hat{g}(S_{(i)}) - g(S_{(i)})|] \leqslant C(n,\epsilon) = \sum_{i=1}^{K} \left(|w_i^{(1)}| b(n,i) + \sum_{i \neq j} |w_{ij}^{(2)}| b(n,j) \right).$$

Note: $C(n,\epsilon) \to 0$ when $n \to \infty$.

Idea of the proof: Using that the corruption is independent of the clean distribution, write $F_1(q)$ in terms of $\tilde{F}$ which in turn can be estimated from the data.

The Dvoretzky-Kiefer-Wolfowitz inequality is used to control this approximation.

[redacted]

In detail: For $\tilde{F}(q;i,j) = \mathbb{P}(S(X,i) \leqslant q | \tilde{Y} = j)$ (and similar notion $F_1(q;i,j)$) we have

$$\tilde{F}(q,i,j) \quad = \quad \sum_{k=1}^{K} \mathbb{P}(Y = k | \tilde{Y} = j) \mathbb{P}(S(X,i) \leq q \mid \tilde{Y} = j, Y = k) = \sum_{k=1}^{K} P_{kj} F_1(q,i,k).$$

Thus, $F_1(q) = \tilde{F}(q) P^{-1}$. We estimate $\tilde{F}(q,i,j)$ by its empirical version

$$\tilde{F}_n(q,i,j) = \frac{\sum_{\ell=1}^{n} 1(S(X_\ell, i) \leq q) 1(y_\ell = j)}{\sum_{\ell=1}^{n} 1(y_\ell = j)}$$

and $g(q) = F_1(q) - \tilde{F}(q)$ by

$$\hat{g}_n(q) = \sum_{i=1}^{K} \sum_{j=1}^{K} \left(P_i P_{ji}^{-1} \tilde{F}_n(q,i,j) - \sum_{i=1}^{K} \tilde{P}_i \tilde{F}_n(q,i,i) \right).$$

Selected experiments

CIFAR-10N (*Wei et al, 2022*):
60,000 images, 10 classes, 6000 images per class
50,000 training images, 10,000 test images
images labelled by independent workers

`Clean`: is CIFAR-10, noise rate 0%
`Aggr`: noise rate 9.03%
`R2`: noise rate 18.12%
`Worst`: noise rate 40.21%.

Aim: 90% coverage

CP:

	Coverage	Size
`Clean`	0.900 ± 0.005	1.507 ± 0.019
`Aggr`	0.940 ± 0.003	2.003 ± 0.027
`R2`	0.977 ± 0.002	3.177 ± 0.066
`Worst`	0.990 ± 0.001	5.473 ± 0.078

CRCP:

	Coverage	Size
`Clean`	0.909 ± 0.005	1.507 ± 0.019
`Aggr`	0.899 ± 0.005	1.550 ± 0.019
`R2`	0.903 ± 0.006	1.658 ± 0.021
`Worst`	0.917 ± 0.009	2.189 ± 0.093

Connection with adaptive conformal classification

Sesia et al. (2024) have a very similar procedure, which is a key ingredient in what they call *adaptive conformal classiication*, for slightly different conformal prediction problems:

* label-conditional coverage
* marginal coverage
* calibration-conditional coverage.

They give very nice theoretical guarantees and also very nice extensive simulation studies.

There are some differences in the assumption, but the key difference is in $C(n, \epsilon)$.

Example: Uniform noise (randomised response model)

The corrupting noise chooses one of the K labels uniformly; the true labels are also uniform. Then

$$C(n,\epsilon) = 2\frac{\epsilon}{(1-\epsilon)}\frac{(K-1)}{K}\left\{\left(1-\frac{1}{K}\right)^n + \sqrt{\frac{\pi K}{n}}\right\}$$

whereas Sesia et al. (2024) get, with n_* the smallest number of observations in a class,

$$c(n) + 2(K-1)\frac{\epsilon}{(1-\epsilon)}\frac{1}{\sqrt{n_*}}\min\left\{K^2\sqrt{\frac{\pi}{2}}, \frac{1}{\sqrt{n_*}} + \sqrt{\frac{\log(2K^2)+\log(n_*)}{2}}\right\}$$

where $c(n) \to 0$ with n. So $C(n,\epsilon)$ tends to 0 faster with n.

Discussion

CRCP coverage is close to the desired 90% whereas CP over-covers

The CRCP intervals are narrower than the CP intervals and hence more precise

Contamination can affect coverage and CRCP can ameliorate it.

Future:

Investigate repercussions with *Sesia et al. (2024)* more thoroughly.

Run on CIFAR-10H and compare to the adaptive conformal prediction methods from *Sesia et al. (2024)*

CRCP for regression.