Task: prediction intervals for new observations

strongly related to permutation tests

for making probabilistic predictions (not relying on very precise specification of probabilistic models)

CONFORMAL PREDICTION (GAMMERMAN, VOVK, VAPNIK 1998)

we know abs. nothing about F

$X_1, X_2, \dots, X_m \sim$ iid F (TRAINING SAMPLE)

$X_{m+1} \sim F$ (NEW OBSERVATION TO BE PREDICTED) } INDEPENDENT

VALIDITY → COVARAGE

- $C(X_1, \dots, X_m)$ IS A VALID PREDICTION SET (interval if in $\mathbb{R}$) FOR X_{m+1} IF:

$$\mathbb{P}(X_{m+1} \in C(X_1, \dots, X_m)) \geq 1-\alpha$$

We have 2 random objects: the new observation and the prediction set (which we consider as random as well)

→ JOINT PROBABILITY MEASURE OF $(X_1, \dots, X_m, X_{m+1})$ AND NOT THE UNIVARIATE PROBABILITY OF X_{m+1}!

COVERAGE / confidence level of the prediction set

Given a new dataset and a new prediction, the new dataset will generate a random set that will capture the new pred. with probability higher than $1-\alpha$.
Remember: the set is RANDOM

EFFICIENCY → PI LENGTH

- $C_1(X_1, \dots, X_m)$ IS MORE EFFICIENT THAN $C_2(X_1, \dots, X_m)$ IF:

a rule for building a pred. set

$$\mathbb{E}(|C_1(X_1, \dots, X_m)|) < \mathbb{E}(|C_2(X_1, \dots, X_m)|)$$

another rule to build a pred set

on average the length of the prediction intervals provided by the first rule is smaller than the length of the pred. int. provided by the second rule

A PARAMETRIC EXAMPLE: FISHER'S PREDICTION INTERVALS

$X_1, \dots, X_m \sim$ iid $N(\mu, \sigma^2)$
$X_{m+1} \sim N(\mu, \sigma^2)$ } INDEPENDENT, μ AND σ UNKNOWN

$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ } POINT-WISE PREDICTION

$\bar{X}$ — sample mean OF the training samples

new observ. X_{m+1}

$$\frac{X_{m+1} - \bar{X}}{S\sqrt{1+\frac{1}{m}}} \sim t(m-1)$$

$$\Downarrow$$

$$\mathbb{P}\left(\bar{X} - S\sqrt{1+\tfrac{1}{m}}\, t_{\alpha/2}(m-1) < X_{m+1} < \bar{X} + S\sqrt{1+\tfrac{1}{m}}\, t_{\alpha/2}(m-1)\right) = 1-\alpha$$

This is a random set just depending on the training sample s.t. the new observation is included in the random set with prob. $1-\alpha$

Formula that allows to build prediction intervals for new observations in a setting in which we know that data are gaussian

This can only be used if the data are gaussian (even if the sample size is large it needs to be gaussian)

For CI we're able to use the asymptotic CI, with PI we don't have asymptotic results

Our goal is to find a rule that provides valid and efficient prediction sets.

How the problem of constructing prediction intervals is tackled in the parametric framework?

in practice they cannot be computed since we don't know X_{n+1}

Nonparametric counterpart of the Fisher prediction intervals:

CONFORMAL PREDICTION INTERVALS

$X_1, \dots, X_m \sim \text{iid } F$, $X_{m+1} \sim F$ $\Big\rangle$ INDEPENDENT (CONTINUOUS F) $\Big\rangle$ AUGMENTED SAMPLE

random sample of size $n+1$ built by concatenating the training set with the new observation

LET US INDICATE WITH $\bar{X}_{-i} = \frac{1}{m}\left(X_1 + \dots + X_{i-1} + X_{i+1} + \dots + X_{m+1}\right)$, $i = 1, \dots, m, m+1$

we have $n+1$ sample means

LET US DEFINE THE $m+1$ NON-CONFORMITY SCORES:

$|X_i - \bar{X}_{-i}|$ $i = 1, \dots, m+1$

mismatch, in absolute terms, between the i-th observation and the sample mean computed on the remaining observ.

This quantity measures in some sense the non-conformity of X_i w.r.t. all the other data. It tells how extreme it the i-th data w.r.t. the remaining data. If this quantity is small then the i-th data is not strange w.r.t. the others.

$X_1, \dots, X_m, X_{m+1} \sim \text{iid } F \Rightarrow \underbrace{|X_1 - \bar{X}_{-1}|, \dots, |X_{m+1} - \bar{X}_{-(m+1)}|}_{\text{EXCHANGEABLE}}$

LET US DEFINE THE $m+1$ RANKS OF THE OBSERVATIONS ACCORDING TO THEIR CONFORMITY SCORES.

we sort the data according to the non conformity score - The datum which has the value 1 (R_1) is the datum which has the smallest non-conformity score

$$R_i = \sum_{j=1}^{m+1} \mathbb{1}\left(|X_j - \bar{X}_{-j}| \le |X_i - \bar{X}_{-i}|\right) \quad i = 1, \dots, m+1$$

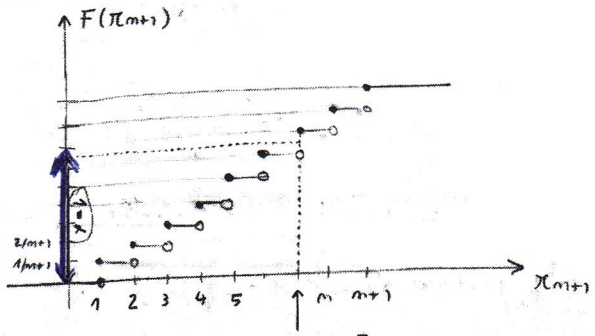
EXCHANGEABILITY OF SCORES (+ data are $\perp\!\!\!\perp$) $\Rightarrow$ $R_i \sim \text{UNIF}(\{1, \dots, m+1\})$ $i = 1, \dots, m+1$

we have a sequence of non conformity scores. All the sequence obtained by permuting the non conformity scores have exactly the same probability to be observed

IN PARTICULAR:

$R_{m+1} \sim \text{UNIF}(\{1, \dots, m+1\})$

the ranks are random variables since both the obs. and the new obs. are random



cumulative distr. function of the random variable that is the rank of the new observation (discrete uniform on $\{1, \dots, n+1\}$)

$\lceil (1-\alpha)(m+1) \rceil$ = superior α quantile

$$\mathbb{P}\left(R_{m+1} \le \lceil (1-\alpha)(m+1) \rceil\right) = \frac{\lceil (1-\alpha)(m+1) \rceil}{m+1} \ge 1-\alpha$$

The $\forall \alpha, \forall n$ $\forall F$

this event (observed or not) depends only on $X_1, \dots, X_{m+1}$, not on F

TO BUILD A PREDICTION SET WE JUST NEED TO RESTATE THE EVENT $\{R_{m+1} \leq \lceil (1-\alpha)(m+1) \rceil\}$ AS A FUNCTION OF $X_1, \dots, X_m$ AND OF THE POSSIBLE FUTURE VALUES x_{m+1} OF X_{m+1}:

$$C(X_1, \dots, X_m) := \left\{ x_{m+1} \in \mathbb{R} : x_{m+1} \leq \lceil (1-\alpha)(m+1) \rceil \right\}$$

still random because it's depending on the training set

possible future realization of X_{m+1}

MORE EXPLICITLY INTRODUCING THE AUGMENTED SAMPLE $(x_1, \dots, x_m, x_{m+1})$:

$$p(x_{m+1}) := \frac{1}{m+1} \sum_{j=1}^{m+1} \left(|x_j - \bar{x}_{-j}| \geq |x_{m+1} - \bar{x}_{-(m+1)}| \right) = \frac{(m+1) \cdot x_{m+1+1}}{m+1}$$

proportion of data in a specific realization having a non-conformative score that is longer or equal than the one associated to the new observation

→ P-VALUE OF $x_{m+1} \in \left\{ \frac{1}{m+1}, \frac{2}{m+1}, \dots, \frac{m}{m+1}, \frac{n+1}{n+1} \right\}$

$$C(X_1, \dots, X_m) = \left\{ x_{m+1} \in \mathbb{R} : p(x_{m+1}) \geq \alpha \right\}$$

(NB) LINK WITH PERMUTATION TESTS FOR TWO INDEPENDENT SAMPLES $(X_1, \dots, X_m)$ AND (X_{m+1})

MORE IN GENERAL NON-CONFORMITY SCORES CAN BE DEFINED THROUGH THE INTRODUCTION OF A NON-CONFORMITY MEASURE $A(\cdot, \cdot)$

$A(\{z_1, \dots, z_m\}, z_{m+1})$ MEASURING THE "NON-CONFORMITY" OF z_{m+1} TO THE SET $\{z_1, \dots, z_m\}$

FORMALLY $A: \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ AND SYMMETRIC WRT THE FIRST m ARGUMENTS (i.e. INVARIANT TO PERMUTATIONS OF THE FIRST m ARGUMENTS)

NON-CONFORMITY SCORES ARE CONSISTENTLY DEFINED AS:

$$A\left(\{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{m+1}\}, X_i\right) \quad i = 1, \dots, m+1$$

e.g. IN T PREDICTION INTERVALS

$$A(\{z_1, \dots, z_m\}, z_{m+1}) := |z_{m+1} - \bar{z}_{\cdot(m+1)}|$$

(NB) DIFFERENT NON-CONFORMITY MEASURES CAN BE TO TACKLE THE SAME PREDICTION PROBLEM.

Smart choices of non conformity measures will allow to obtain smaller prediction sets preserving the same validity.

(We can define an optimal non-conformity measure, but we would need the distributions)

The nonconformity measure plays the role of the test statistic in permutation tests

A POPULAR WAY TO BUILD NON-CONFORMITY MEASURES IS RELYING ON THE NOTION OF DISCREPANCY:

$$A\left((z_1, \dots, z_m), z_{m+1}\right) := \left| z_{m+1} - \hat{z}_{m+1}(z_1, \dots, z_m) \right|$$

non-conformity

in this way we're defining the scores as the PREDICTIVE RESIDUALS

i.e. THE PREDICTIVE RESIDUAL OF A NEW OBSERVATION FROM A PREDICTION OF IT BASED ON THE OTHER DATA

(NB) LINK WITH LEAVE-ONE-OUT CROSSVALIDATION.

(NB) WE HAVE VALID PREDICTION SET FOR ANY CHOICE OF $\hat{z}_{m+1}$

(NB) EFFICIENCY IS RELATED TO SMART CHOICES FOR $\hat{z}_{m+1}$

EX 1 UNIVARIATE $z_1, \dots, z_m, z_{m+1} \in \mathbb{R}$

[R-EXAMPLES. FULL CP.R]

$A = \left| z_{m+1} - \bar{z}_{-(m+1)} \right|$ ~~(T-INTERVALS)~~ → sample mean (for gaussian data this non-conformity measure is the best)

$A = \left| z_{m+1} - \text{MED}(z_1, \dots, z_m) \right|$ → useful if we suspect to have outliers

$A = \left| \log(z_{m+1}) - \frac{1}{m} \log(z_1 \cdot \dots \cdot z_m) \right|$ → if we suspect that the ε is not added but multiplied (!)

EX 2 MULTIVARIATE $z_1, \dots, z_m, z_{m+1} \in \mathbb{R}^p$

$A = \| z_{m+1} - \bar{z}_{-(m+1)} \|^2$

$A = (z_{m+1} - \bar{z}_{-(m+1)}) \, S^{-1}_{-(m+1)} \, (z_{m+1} - \bar{z}_{-(m+1)})'$ ← squared Mahalanobis' distance of the points to the sample mean

---

EX 3 REGRESSION

$z_1, \dots, z_m, z_{m+1} \in \mathbb{R}^2$ → WE EXPLORE POSSIBLE VALUES y_{m+1} OF Y_{m+1} FOR THE OBSERVED $(X_1, Y_1), \dots, (X_m, Y_m), X_{m+1}$

(X_1, Y_1) " (X_{m+1}, Y_{m+1}) } UNCONDITIONAL VALIDITY

discrepancy between the Y value observed for the $n+1$ observation and the predicted value for the new Y in correspondence of the new X, given a model build on the training set

$$A = \left| Y_{m+1} - \hat{Y}(X_{m+1})_{-(m+1)} \right|$$

- OLS
- RIDGE
- LASSO
- SPLINES
- KERNEL REGRESSION
- ...

NOT ALL NON-CONFORMITY MEASURES ARE BASED ON DISCREPANCY FROM A POINT-WISE PREDICTION. FOR EXAMPLE:

K-NEAREST NEIGHBOR CONFORMAL PREDICTION SET

$$A(\{z_1, \dots, z_n\}, z_{n+1}) = \sum_{j : j \text{ in the first } k \text{ neighbors of } z_{n+1}} d^2(z_{n+1}, z_j)$$

(NB) It can be used in metric spaces [link with distance based test statistic and to linkages]

ALGORITHMS

IN THE PRACTICE A TOLERANCE ε IS FIXED AND x_{m+1} IS EXPLORED ON A ε-SPACED GRID.

COMPUTATIONAL COSTS MAY BE LARGE IF:

- z_{m+1} BELONGS TO HIGH-DIMENSIONAL SPACE
- THE COMPUTATION OF $\hat{z}_{m+1}$ IS COMPUTATIONALLY EXPENSIVE

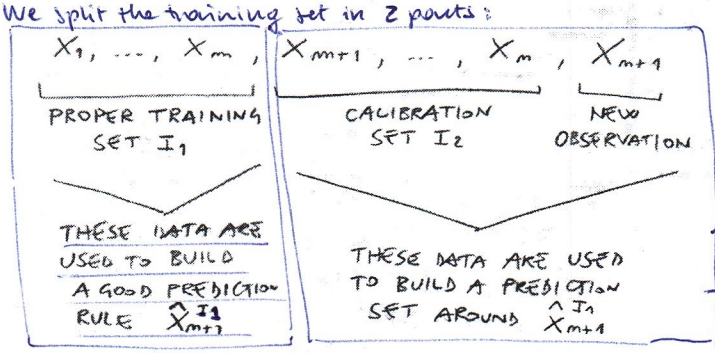
SOLUTIONS

- PARALLELIZATION
- SMART NON-CONFORMITY MEASURES (↓)

FROM FULL CONFORMAL PREDICTION TO SPLIT CONFORMAL PREDICTION

(PAPADOPOULOS, PROEDROU, VOVK, GAMMERMAN 2002)

We split the training set in 2 parts:



conformal predictions will be out of this

we'll use a std conformal approach based on discrepancy on the calibration set (we'll augment the calibration set with the new obs. and we'll use a non conformal approach on the calibr. set using a non-conform. measure which is (kind of) degenerative w.r.t. the calibration set that define the nonconformity score of one of the observation in the augmented calibration set as the mismatch with the prediction build from the first set.

THE $m-m+1$ NON-CONFORMITY SCORES ARE DEFINED AS:

$$A(\{z_{m+1}, \dots, z_n\}, z_{m+1}) = |z_{m+1} - \hat{X}_{m+1}^{I_1}|$$

degenerative because it's not taking into account $\{z_{m+1}, \dots, z_n\}$

AND THE $m-m+1$ RANKS AS:

$$R_i = \sum_{j=m+1}^{m+1} \mathbb{1}\left(|X_j - \hat{X}_{m+1}^{I_1}| \leq |X_i - \hat{X}_{m+1}^{I_1}|\right) \qquad i = m+1, \dots, m+1$$

AND THE P-VALUE OF x_{m+1} AS:

$$p(x_{m+1}) = \frac{1}{m-m+1} \sum_{j=m+1}^{m+1} \mathbb{1}\left\{|X_j - \hat{X}_{m+1}^{I_1}| \geq |x_{m+1} - \hat{X}_{m+1}^{I_1}|\right\}$$

AND THE PREDICTION SET AS:

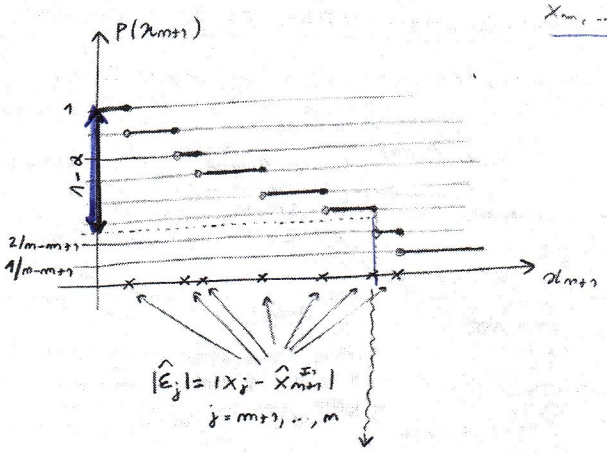
$$C(X_{m+1}, \dots, X_n) = \left\{x_{m+1} \in \mathbb{R} : p(x_{m+1}) \geq \alpha\right\}$$

ADVANTAGES

- $\hat{X}^{I_1}_{m+1}$ IS COMPUTED ONLY ONCE (COMPUTATIONALLY EXPENSIVE PREDICTION METHODS CAN BE USED)

- NON-CONFORMITY SCORES $|X_j - \hat{X}^{I_1}_{m+1}|$ $j = m+1, \dots, n$ ARE NOT AFFECTED BY THE VALUE OF X_{m+1} (i.e. SIMPLE RESIDUALS)

  ↳ $p(x_{m+1})$ CAN BE TRIVIALLY COMPUTED BY SORTING THE $n-m$ NON-CONFORMITY SCORES OF THE CALIBRATION SET $X_{m+1}, \dots, X_n$



THE $\lceil (1-\alpha)(n-m+1) \rceil$-TH ABSOLUTE RESIDUAL

SO THE PREDICTION SET IS SIMPLY DEFINED AS:

$$\left(\hat{X}^{I_1}_{m+1} - |\hat{\varepsilon}_{(\lceil (1-\alpha)(n-m+1) \rceil)}|;\ \hat{X}^{I_1}_{m+1} + |\hat{\varepsilon}_{(\lceil (1-\alpha)(n-m+1) \rceil)}| \right)$$

region in which the p-value is $\geq \alpha$

- train the prediction model once
- compute all the residuals ($n-m$ residuals)
- sort the residuals
- identify: $\hat{\varepsilon}_{(\lceil (1-\alpha)(n-m+1) \rceil)}$
- define the interval

DISADVANTAGES

- POSSIBLE LOSS IN EFFICIENCY WRT FULL CONFORMAL
- EXTRA RANDOMNESS DUE TO THE RANDOM SPLIT

SPLIT PROPORTIONS

- LARGER PROPER TRAINING SET → HIGHER QUALITY OF $\hat{X}^{I_1}_{m+1}$
- LARGER CALIBRATION SET → REDUCE THE IMPACT OF CEILING (less discretized p-value function)

LITERATURE SUGGESTS TO CHOOSE $|I_1|$ AND $|I_2|$ IN BETWEEN (66%, 33%) AND (75%, 25%) [R-EXAMPLE SPLIT CP.R]

trade-off (to stabilize ~~m~~ high or low)