Conditional Predictive Inference for High-Dimensional Stable Algorithms

Hannes Leeb and Lukas Steinberger

(University of Vienna, DataScience@UniVie)

Vienna University of Economics and Business
Nov. 6, 2020

Overview

Prediction of a response y_0 from a feature-vector x_0 given given an i.i.d. sample of feature/response pairs (x_i, y_i) is a fundamental task of statistical learning.

We study *prediction intervals* for y_0 that are based on empirical quantiles of leave-one-out residuals.

This task is easy in (classical) asymptotic settings where $\mathbb{E}[y_0 \| x_0]$ can be consistently estimated (Butler and Rothman 1980; Stine 1985; Schmoyer 1992; Olive 2007; and Politis 2013).

In other settings (large dimensions and/or model misspecification), resampling methods like the residual bootstrap do not perform well; cf. Bickel and Freedman (1983), Mammen (1996) and, recently, El Karoui and Purdom (2015).

For the proposed prediction intervals, we provide finite-sample and asymptotic performance bounds, without requiring that $\mathbb{E}[y_0 \| x_0]$ can be estimated consistently.

Leave-one-out prediction intervals

Consider a feature/response pair (x_0, y_0) with $x_0 \in \mathbb{R}^p$ and $y_0 \in \mathbb{R}$, and a training sample $T_n = (x_i, y_i)_{i=1}^n$, where the (x_i, y_i) are i.i.d. copies of (x_0, y_0). The goal is to predict y_0 from x_0 using T_n at level $1 - \alpha$.

Using a given prediction algorithm $\hat{m}_n(x_0) = \hat{m}_n(x_0, T_n)$, we proceed as follows:

- For each $i = 1, \ldots, n$, write $\hat{m}_n^{[i]}(\cdot)$ for the prediction algorithm computed from all but the i-th observation.
- Compute the leave-one-out residuals $\hat{u}_i = y_i - \hat{m}_n^{[i]}(x_i)$, $i = 1, \ldots, n$, the corresponding order statistics $\hat{u}_{(1)} \leq \cdots \leq \hat{u}_{(n)}$ and the empirical quantiles $\hat{q}_{\alpha/2} = \hat{u}_{(\lceil n\alpha/2 \rceil)}$ and $\hat{q}_{1-\alpha/2} = \hat{u}_{(\lceil n(1-\alpha/2) \rceil)}$.
- Compute the prediction interval

$$PI_\alpha(T_n, x_0) \quad = \quad \hat{m}_n(x_0) \; + \; (\hat{q}_{\alpha/2}, \; \hat{q}_{1-\alpha/2}].$$

CONDITIONAL COVERAGE PROBABILITY

Our goal is to control the *conditional coverage probability*

$$P(y_0 \in PI_\alpha(T_n, x_0) \| T_n).$$

We show that

$$\mathbb{E}_P \left| P(y_0 \in PI_\alpha(T_n, x_0) \| T_n) - (1-\alpha) \right|$$

is small, uniformly over a large class $\mathcal{P}$ of distributions P (details later), provided that

- the prediction algorithm is sufficiently stable so that $\hat{m}_n(\cdot) \approx \hat{m}_n^{[i]}(\cdot)$, and
- the prediction algorithm has bounded estimation error in probability, i.e., $\mathbb{E}[y_0 \| x_0] - \hat{m}_n(x_0) = O_P(1)$ (no consistency required).

With this, the unconditional coverage probability $P(y_0 \in PI_\alpha(T_n, x_0))$ is also close to $1-\alpha$.

THE CLASS OF DISTRIBUTIONS $\mathcal{P}$

We require the class $\mathcal{P}$ of distributions to satisfy the following condition.

(C1). Under every $P \in \mathcal{P}, \ldots$

- the feature/response pairs $(x_0, y_0), (x_1, y_1), \ldots$ are i.i.d.;
- the regression function $x \mapsto m_P(x) := \mathbb{E}_P[y_0 \| x_0 = x]$ exists;
- the error term $u_0 := y_0 - m_P(x_0)$ is independent of the regressor vector x_0 and has a Lebesgue density $f_{u,P}$ with $\|f_{u,P}\|_\infty < \infty$.

The stability condition on the prediction algorithm

Fix $\eta > 0$ and a class $\mathcal{P}$ of distributions as in (C1). A predictor $\hat{m}_n$ is *η-stable with respect to $\mathcal{P}$* if

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\left(\|f_{u,P}\|_\infty \left| \hat{m}_n(x_0) - \hat{m}_n^{[i]}(x_0) \right| \right) \wedge 1 \right] \quad \leq \quad \eta$$

for each $i = 1, \ldots, n$ (cf. Bousquet and Elisseeff, 2002).

A USEFUL LEMMA

Consider the (feasible) e.c.d.f. of the leave-one-out residuals, i.e.,

$$\hat{F}_n(s) \quad = \quad \hat{F}_n(s; T_n) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{u}_i \leq s\}$$

and the (infeasible) true (conditional) c.d.f. of the prediction error, i.e.,

$$\tilde{F}_n(s) \quad = \quad \tilde{F}_n(s; T_n) \quad = \quad P(y_0 - \hat{m}_n(x_0) \leq s \| T_n).$$

Then

$$\left| P(y_0 \in PI_\alpha(T_n, x_0) \| T_n) - \left(1 - \frac{\lfloor n\alpha/2) \rfloor + \lceil n\alpha/2 \rceil}{n}\right)\right| \quad \leq \quad 2\|\hat{F}_n - \tilde{F}_n\|_\infty.$$

In particular, if $\mathbb{E}_P\|\hat{F}_n - \tilde{F}_n\|_\infty$ is small, uniformly over $P \in \mathcal{P}$, then $E_P|P(y_0 \in PI_\alpha(T_n, x_0)\|T_n) - (1 - \alpha)|$ is small, uniformly over $P \in \mathcal{P}$.

Theorem 1

Assume that the class $\mathcal{P}$ of distributions satisfies (C1) and that the predictor $\hat{m}_n(\cdot)$ is symmetric and η-stable w.r.t. $\mathcal{P}$. Then, for each $P \in \mathcal{P}$, each $L > 1$ and each $\mu \in \mathbb{R}$, we have

$$
\begin{aligned}
\mathbb{E}_P \|\hat{F}_n - \tilde{F}_n\|_\infty \quad \leq \quad & P(|y_0 - m_P(x_0)| > L) \\
& + P(|m_P(x_0) - \hat{m}_n(x_0) - \mu| > L) \\
& + 3\left(L \|f_{u,P}\|_\infty \left(\frac{1}{2n} + 3\eta\right)\right)^{1/3} + \sqrt{\frac{1}{n} + 6\eta}.
\end{aligned}
$$

Asymptotics: Prediction with many variables

We study asymptotic settings where the dimension of the feature vector x_0 depends on n, i.e., $p = p_n$, so that $p_n/n \to \kappa \in (0, 1)$.

Our first result is an asymptotic adaptation of Theorem 1, which we then use to deal with more specific scenarii.

THEOREM 2

Let p_n be a sequence of positive integers and let $\mathcal{P}_n$ be as in (C1) with p_n replacing p. Moreover, suppose the following:

- The predictor $\hat{m}_n$ is symmetric and η_n-stable w.r.t. $\mathcal{P}_n$ with $\eta_n \to 0$.
- For each $P \in \mathcal{P}_n$, there exists $\sigma_P^2 \in (0, \infty)$ so that $\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \sigma_P \|f_{u,P}\|_\infty < \infty$.
- The scaled estimation errors $|m_P(x_0) - \hat{m}_n(x_0)|/\sigma_P$ and the scaled errors $|y_0 - m_P(x_0)|/\sigma_P$ both are $\mathcal{P}_n$-uniformly bounded.

Then

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}_P \|\hat{F}_n - \tilde{F}_n\|_\infty \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0.$$

In particular,

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}_P \left| P(y_0 \in PI_\alpha(T_n, x_0) \| T_n) - (1-\alpha) \right| \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0.$$

Regularized M-estimators

For a given convex loss function $\rho : \mathbb{R} \to \mathbb{R}$ and a fixed tuning parameter $\gamma \in (0, \infty)$ (both not depending on n), consider the estimator

$$\hat{\beta}_n^{(\rho)} \quad = \quad \operatorname{argmin}_{b \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x_i' b) + \frac{\gamma}{2} \|b\|_2^2.$$

These estimators are studied by El Karoui (2018) in a linear model $y_i = x_i' \beta + u_i$ allowing for heavy-tailed errors in an asymptotic setting where $p/n \to \kappa \in (0, 1)$.

Under the assumptions maintained in that reference, Theorem 2 applies.

James-Stein type predictors

We consider the predictor $\hat{m}_n(x_0) = x_0' \hat{\beta}_n(c)$, where $\hat{\beta}_n(c)$ is a James-Stein-type estimator

$$\hat{\beta}_n(c) = \begin{cases} \left(1 - \frac{c p_n \hat{\sigma}_n^2}{\hat{\beta}_n' X' X \hat{\beta}_n}\right)_+ \hat{\beta}_n, & \text{if } \hat{\beta}_n' X' X \hat{\beta}_n > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $c \in [0, 1]$ is a tuning-parameter, where $\hat{\beta}_n = (X'X)^\dagger X'Y$ and where $\hat{\sigma}_n^2 = \|Y - X'\hat{\beta}_n\|^2/(n - p_n)$.

For the classes $\mathcal{P}_n$ of underlying distributions, we consider families of nonlinear regression models where the feature-vectors are randomly scaled linear functions of i.i.d. variables, as described in (C2), which follows.

James-Stein type predictors

(C2). Fix finite constants $C_0 > 0$, $c_0 > 0$ and probability measures $\mathcal{L}_l$, $\mathcal{L}_w$ on $\mathbb{R}$, so that $\mathcal{L}_w$ has mean zero, unit variance and finite fourth moment, $\int s^2 \mathcal{L}_l(dx) = 1$ and $\mathcal{L}_l((-c_0, c_0)) = 0$. For each n, the following holds under each $P \in \mathcal{P}_n = \mathcal{P}_n(C_0, c_0, \mathcal{L}_l, \mathcal{L}_w)$:

- $(x_i, y_i) \in \mathbb{R}^{p_n+1}$ are i.i.d.
- The feature vector x_0 is distributed as

$$x_0 \quad \sim \quad l_0 \Sigma_P^{1/2}(w_1, \ldots, w_{p_n})',$$

  where the w_i are i.i.d. according to $\mathcal{L}_w$, $l_0 \sim \mathcal{L}_l$ is independent of the w_i and $\Sigma_P^{1/2}$ is the symmetric positive definite square root of a positive definite $p_n \times p_n$ matrix Σ_P.
- The response y_0 has mean zero and

$$y_0 \| x_0 \quad \sim \quad m_P(x_0) + \sigma_P v_0,$$

  where v_0 is independent of x_0, has a Lebesgue density, mean zero, unit variance and fourth moment bounded by C_0, with measurable regression function m_P satisfying $\mathbb{E}_P m_P(x_0) = 0$.

James-Stein type predictors

Theorem 3

For each n let $\mathcal{P}_n = \mathcal{P}_n(C_0, c_0, \mathcal{L}_l, \mathcal{L}_w)$ be as in (C2). For each $P \in P_n$, define β_P as the minimizer of $\mathbb{E}_p(y_0 - \beta' x_0)$ over $\beta \in \mathbb{R}^{p_n}$. Assume that $p_n/n \to \kappa \in (0,1)$; that the densities v_0 in (C2) are uniformly bounded; and that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{E}_P\left[\left(\frac{m_P(x_0) - x_0'\beta_P}{\sigma_P}\right)^2\right] \quad < \quad \infty.$$

Then Theorem 2 applies to the James-Stein type predictor $\hat{m}_n(x_0) = x_0'\hat{\beta}_n(c)$. (For $c = 0$, this also covers the OLS-predictor $x_0'\hat{\beta}_n$.)

INTERVAL LENGTH

We now turn to the length of the prediction interval $PI_\alpha(T_n, x_0)$, i.e.,

$$\hat{q}_{1-\alpha/2} - \hat{q}_{\alpha/2}.$$

For the classes $\mathcal{P}_n$ of underlying distributions, we consider families of parametric linear models indexed by the regression parameter $\beta_P \in \mathbb{R}^{p_n}$, by $\Sigma_P = \mathbb{E}x_0 x_0' \in \mathbb{R}^{p_n \times p_n}$ and $\sigma_P^2 = \mathbb{E}_P(y_0 - x_0'\beta_P)^2 \in (0, \infty)$. These classes are defined in (C3), which follows.

INTERVAL LENGTH

(C3). Fix a finite constant $c_0 > 0$ and probability measures $\mathcal{L}_l$, $\mathcal{L}_w$ and $\mathcal{L}_v$ on $\mathbb{R}$, so that $\mathcal{L}_w$ and $\mathcal{L}_v$ have zero mean, unit variance and finite fourth moments, and so that $\int s^2 \mathcal{L}_l(ds) = 1$ and $\mathcal{L}_l((-c_0, c_0)) = 0$. For each n, the following holds under each $P \in \mathcal{P}_n = \mathcal{P}_n(c_0, \mathcal{L}_l, \mathcal{L}_w, \mathcal{L}_v)$:

- $(x_i, y_i) \in \mathbb{R}^{p_n+1}$ are i.i.d.
- The feature vector x_0 is distributed as

$$x_0 \quad \sim \quad l_0 \Sigma_P^{1/2}(w_1, \ldots, w_{p_n})',$$

  where $w_1, \ldots, w_{p_n}$ are i.i.d. according to $\mathcal{L}_w$, where $l_0 \sim \mathcal{L}_l$ is independent of the w_i, and where $\Sigma_P^{1/2}$ is the symmetric square root of a positive definite $p_n \times p_n$ matrix Σ_P.
- The response y_0 satisfies

$$y_0 \| x_0 \quad \sim \quad x_0' \beta_P + \sigma_P v_0,$$

  where $\beta_P \in \mathbb{R}^{p_n}$, $\sigma_P \in (0, \infty)$, and where $v_0 \sim \mathcal{L}_v$ independent of x_0.

Interval length

Theorem 4

For each n let $\mathcal{P}_n = \mathcal{P}_n(c_0, \mathcal{L}_l, \mathcal{L}_w, \mathcal{L}_v)$ be as in (C3). If $p_n/n \to \kappa \in (0, 1)$, then the scaled empirical α-quantile $\hat{q}_\alpha/\sigma_P$ of the leave-one-out residuals $\hat{u}_i = y_i - x_i'\hat{\beta}_n^{[i]}$ based on the OLS estimator converges $\mathcal{P}_n$-uniformly in probability to the corresponding α-quantile of the distribution of

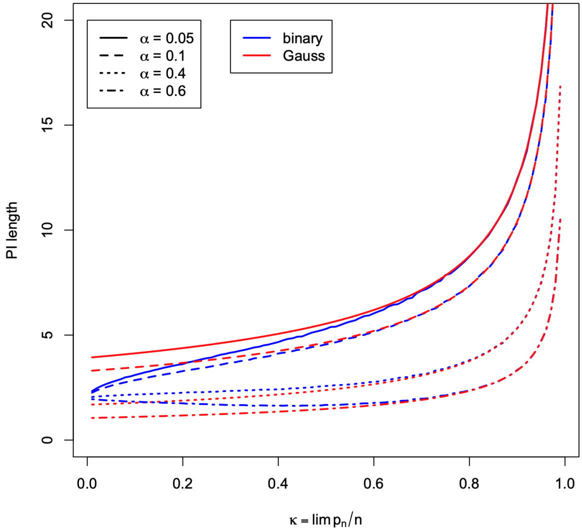
$$lN\tau + v,$$

where $l \sim \mathcal{L}_l$, $N \sim N(0, 1)$ and $v \sim \mathcal{L}_v$ are independent and where $\tau = \tau(\mathcal{L}, \kappa)$ is a constant.

This statement also holds in case $\kappa = 0$, provided that $\mathcal{L}_v$ has a continuous and strictly increasing c.d.f. and $p_n \to \infty$.

The constant κ satisfies $\kappa = 0$ if and only if $\tau(\mathcal{L}_l, \kappa) = 0$. Moreover, if $\mathcal{L}_l(\{-1, 1\}) = 1$, then $\tau(\mathcal{L}_l, \kappa) = \sqrt{\kappa/(1-\kappa)}$.

LIMIT OF $(\hat{q}_{1-\alpha/2} - \hat{q}_{\alpha/2})/\sigma_P$ WITH $\mathcal{L}_l(\{1\}) = 1$ AND FOR TWO CHOICES OF $\mathcal{L}_v$



Related methods

- Sample splitting: See last Figure.
- Jackknife+: A modification of the method proposed here by Barber et al. (2019). Controls unconditional coverage, even if predictor is not stable.
- Conformal prediction: Controls unconditional coverage, even if predictor is not stable. Cf. Vovk et al. (1999, 2005, 2009) as well as Lei et al. (2017, 2013) and Lei and Wasserman (2014).
- Tolerance regions: Give a confidence set for (x_0, y_0), from which a confidence set for y_0 can be obtained by cutting (so that efficiency is an issue). See Wilks (1941, 1942), Wald (1943) and Tukey (1947), and Krishnamoorthy and Mathew (2009) for an overview.

Extensions

- The requirement in (C1), that $u_0 = y_0 - m_P(x_0)$ is independent of x_0, is an issue. A relaxation of this is work in progress and is looking good so far.
- The requirement in (C1), that the density $f_{u,P}$ of u_0 satisfies $\|f_{u,P}\|_\infty < \infty$, can be replaced by a Hölder condition; the resulting theory becomes more complex.
- Our prediction intervals have constant width, independent of x_0. The construction of variable-width prediction intervals is being investigated.

References

- Barber, R.F., Candes, E.J., Ramdas, A. and Tibshirani, R.J. (2020): Predictive inference with the jackknife+. *Ann.Statist.* forthcoming.
- Bousquet, O. and A. Elisseeff (2002). Stability and generalization. *J. Mach. Learn. Res.* 2, 499526.
- El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields* 170(1-2), 95175.
- El Karoui, N. and E. Purdom (2015). Can we trust the bootstrap in high-dimension? The case of linear models. *J. Mach. Learn. Res.* 19 (2018) 1-66.
- Steinberger, L. and Leeb, H. (2020): Conditional predictive inference for high-dimensional stable algorithms. *arXiv preprint arXiv:1809.01412v2*
- Tukey, J. W. (1947): Non-parametric estimation ii. Statistically equivalent blocks and tolerance regions the continuous case. *Ann. Math. Statist.* 18(4), 529539.
- Vovk, V., Nouretdinov, I., and Gammerman, A. (2009): On-line predictive linear regression. *Ann. Statist.* 37(3), 15661590.