Conformal Prediction for Validity of Resampling Inference[*]

Arun Kumar Kuchibhotla[†]

August 1, 2020

University of Pennsylvania

Abstract

This note describes a deficiency of traditional proofs of consistency of resampling techniques for statistical inference and provides a simple solution based on conformal prediction.

1 Deficiency of Classical Consistency Guarantees

Suppose we are interested in inference for a "parameter" or "functional" $\theta_0 \in \mathbb{R}^p$ based on an estimator $\widehat{\theta} \in \mathbb{R}^p$ computed using data $Z_1, Z_2, \ldots, Z_n$. Assume that the estimator $\widehat{\theta}$ satisfies the asymptotic linear representation property, i.e.,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(Z_i) + r_n, \quad \text{such that} \quad \|r_n\| = o_p(1),$$

for some norm $\|\cdot\|$. The bootstrap and subsampling procedures for inference proceed as follows. For $1 \leqslant b \leqslant B$, compute bootstrapped estimators $\widehat{\theta}^{(b)}$ which means generating a bootstrap resample of the data and applying the algorithm that outputs $\widehat{\theta}$ to the resampled data. A bootstrap confidence region $\widehat{R}_n$ for θ_0 satisfies

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\{\sqrt{n}(\widehat{\theta}^{(b)} - \widehat{\theta}) \in \widehat{R}_n\} \geqslant 1 - \alpha. \tag{1}$$

One might, in practice, bootstrap a normalized statistic such as $n^{1/2}\mathrm{diag}(\widehat{\Sigma}_n)^{-1/2}(\widehat{\theta} - \theta_0)$. The discussion below holds readily for such a normalized bootstrap too. Traditionally consistency results for bootstrap prove

$$\sup_{A\in\mathcal{A}}\left|\mathbb{P}\left(\sqrt{n}(\widehat{\theta}^{(*)} - \widehat{\theta}) \in A\middle|\{Z_i\}\right) - \mathbb{P}\left(\sqrt{n}(\widehat{\theta} - \theta_0) \in A\right)\right| = o_p(1) \quad \text{as} \quad n \to \infty, \tag{2}$$

for a class of sets $\mathcal{A}$; here $\widehat{\theta}^{(*)}$ denotes a generic bootstrap estimator. For clarity, note that this is equivalent to

$$\sup_{A\in\mathcal{A}}\left|\int_A dP^*(\delta) - \int_A dP(\delta)\right| = o_P(1),$$

[*]This is a preliminary version of the paper.

[†]Department of Statistics. Email: `karun3kumar@gmail.com`.

where $P^*(\cdot)$ represents the probability measure of $n^{1/2}(\widehat{\theta}^{(*)} - \widehat{\theta})$ conditional on $\{Z_i\}$ and $P(\cdot)$ represents the probability measure of $n^{1/2}(\widehat{\theta} - \theta_0)$, that is, for any Borel set $B \subseteq \mathbb{R}^p$,

$$P^*(B) \;:=\; \mathbb{P}\left(\sqrt{n}(\widehat{\theta}^{(*)} - \widehat{\theta}) \in B \middle| \{Z_i\}_{i=1}^n\right) \quad \text{and} \quad \mathbb{P}(B) \;:=\; \mathbb{P}\left(\sqrt{n}(\widehat{\theta} - \theta_0) \in B\right).$$

It is clear that there is a gap between (1) and (2), because one cannot use just (2) to prove any validity guaranetee for $\widehat{R}_n$ obtained from (1). One simple reason for this is that (2) does not involve B while (1) does.

In order to clear this gap, one needs to prove that conditional on the data $\{Z_i\}$,

$$\sup_{A\in\mathcal{A}} \left| \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\sqrt{n}(\widehat{\theta}^{(b)} - \widehat{\theta}) \in A\} - \int_A dP^*(\delta) \right| = o_{p^*}(1), \quad \text{as} \quad B \to \infty. \tag{3}$$

(Ideally B grows with the sample size n.) The randomness o_{p^*} on the right hand side is through the randomness of the bootstrap samples conditional on $\{Z_i\}$. Combining (2) and (3), (asymptotic) validity guarantee for the confidence set $\widehat{R}_n$ in (1) follows:

$$\begin{aligned}
\int_{\widehat{R}_n} dP(\delta) &\geqslant \int_{\widehat{R}_n} dP^*(\delta) - o_p(1) \quad (\text{from } (2)) \\
&\geqslant \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\sqrt{n}(\widehat{\theta}^{(b)} - \widehat{\theta}) \in \widehat{R}_n\} - o_{p^*}(1) - o_p(1) \quad (\text{from } (3)) \\
&\geqslant 1 - \alpha - o_{p^*}(1) - o_p(1).
\end{aligned}$$

Because $\sqrt{n}(\widehat{\theta}^{(b)} - \widehat{\theta}), 1 \leqslant b \leqslant B$ are independent and identically distributed conditional on $\{Z_i\}$, proving (3) usually can be done through the results in empirical processes. If the VC dimension $\mathrm{VC}(\mathcal{A})$ of the class $\mathcal{A}$ of sets is finite, then Theorem 2 of Vapnik and Chervonenkis (1971) proves that

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} \left| \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\sqrt{n}(\widehat{\theta}^{(b)} - \widehat{\theta}) \in A\} - \int_A dP^*(\delta) \right| \geqslant \sqrt{\frac{16\mathrm{VC}(\mathcal{A})\log(3B)}{B}} \,\middle|\, \{Z_i\} \right) \leqslant \frac{1}{2B+1}, \tag{4}$$

by taking $\varepsilon = \sqrt{8\log((2B+1)^{\mathrm{VC}(\mathcal{A})})/B}$ in Theorem 2 of Vapnik and Chervonenkis (1971) and applying Theorem 9.3 of Györfi et al. (2006); also see the proof of Theorem 9.6 of Györfi et al. (2006) for a similar result. Inequality (4) implies (3) if $\mathrm{VC}(\mathcal{A}) = o(B/\log(B))$. The rate here cannot be improved, in general. For example, the VC dimension of the set of all rectangles in $\mathbb{R}^p$ with facets parallel to the coordinate axes is of order p (Györfi et al., 2006, Problem 9.2) and hence we need at least p bootstrap samples. This can be prohibitive in high-dimensional examples where p is much larger than the sample size n.

A Motivating Example. We now provide a relatively more concrete motivating example that emphasizes the need for resolving the gap mentioned above. In the high-dimensional case where p is allowed to grow much faster than n (e.g., $p = \exp(o(n^\gamma))$ for some $\gamma \in [0, 1]$), Chernozhukov et al. (2017) prove central limit theorem and bootstrap consistency results for the set of all hyper-rectangles. In this case the target of estimation can be thought as the population mean. The results of Chernozhukov et al. (2017) imply, under certain conditions, that mean zero independent random vectors $Z_1, \ldots, Z_n \in \mathbb{R}^p$ satisfy

$$\begin{aligned} \sup_{A \in \mathcal{A}^{\mathrm{re}}} \left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \in A \right) - \mathbb{P}\left(G \in A\right) \right| &\leqslant C \left(\frac{\log^7 p}{n} \right)^{1/6}, \\ \sup_{A \in \mathcal{A}^{\mathrm{re}}} \left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i^* \in A \middle| \{Z_i\} \right) - \mathbb{P}(G \in A) \right| &= O_p(1) \left(\frac{\log^5 p}{n} \right)^{1/6}. \end{aligned} \tag{5}$$

Here $\mathcal{A}^{\mathrm{re}}$ represents the set of all hyper-rectangles in $\mathbb{R}^p$ and $G \in \mathbb{R}^p$ represents a mean zero Gaussian random vector whose covariance matches that of $n^{-1/2} \sum_{i=1}^{n} Z_i$. These results have been improved in Chernozhukov et al. (2019) but the main message of all these results is that $\log p = o(n^{\gamma})$ for some $\gamma \in [0, 1]$ is enough for central limit theorem and bootstrap consistency to hold. The fact that $\mathrm{VC}(\mathcal{A}^{\mathrm{re}}) \asymp 2d$ implies from (4) that the number of bootstrap samples still has to satisfy $p = o(B)$ as $B \to \infty$. *Can we avoid requiring more bootstrap samples than the original sample size n?*

2 A Solution based on Conformal Prediction

The discussion above shows that constructing a set $\widehat{R}$ as in (1) may not in general have a validity guarantee unless B is very large, especially in high-dimensional settings. We now provide a solution to this problem which does not require proving (3). Instead, we directly aim to construct a set $\widehat{R}^*$ such that conditional on $\{Z_i\}$,

$$\int_{\widehat{R}^*} dP^*(\delta) \;=\; \mathbb{P}\left(\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) \in \widehat{R}^* \middle| \{Z_i\} \right) \;\geqslant\; 1 - \alpha, \tag{6}$$

where $\widehat{\theta}^*$ is a generic bootstrap estimator.

We now provide a computationally feasible to guarantee (6) irrespective of the dimension of the estimator $\widehat{\theta}$, based on conformal prediction. Conformal prediction (Balasubramanian et al., 2014) is a general technique that provides a prediction set for a future observation. Suppose $W_1, W_2, \ldots, W_m$ are exchangeable, then conformal prediction techniques can be used to construct a set $\widehat{S}$ such that

$$\mathbb{P}(W_{m+1} \in \widehat{S}) \geqslant 1 - \alpha, \tag{7}$$

whatever $m \geqslant 1$ and $\alpha \in [0, 1]$ maybe. This guarantee holds whenever W_{m+1} is exchangeable with $W_1, \ldots, W_m$. The probability in (7) is computed with respect to the randomness of W_{m+1} and of $(W_1, \ldots, W_m)$. In particular, if $W_1, \ldots, W_{m+1}$ are independent and identically distributed, then (7) is equivalent to

$$\mathbb{E}\left[\int_{\widehat{S}} dP_W(\delta) \right] \geqslant 1 - \alpha,$$

where the expectation is with respect to $(W_1, \ldots, W_m)$ and $P_W(\cdot)$ is a probability measure of W_{m+1}.

In case of bootstrap, conditional on $\{Z_i\}$, $T_1 = \sqrt{n}(\widehat{\theta}^{(1)} - \widehat{\theta})$, $\ldots$, $T_B = \sqrt{n}(\widehat{\theta}^{(B)} - \widehat{\theta})$ are independent and identically distributed. Applying the conformal prediction technique, one can obtain a set $\widehat{R}^{\dagger}$ such that

$$\mathbb{E}\left[\int_{\widehat{R}^{\dagger}} dP^*(\delta) \middle| \{Z_i\} \right] \;=\; \mathbb{P}\left(\sqrt{n}(\widehat{\theta}^{(B+1)} - \widehat{\theta}) \in \widehat{R}^{\dagger} \middle| \{Z_i\} \right) \geqslant 1 - \alpha. \tag{8}$$

The expectation in the first term here is with respect to the probability measure of $(T_1, \ldots, T_B)$ conditional on $\{Z_i\}$. This does not readily imply that $\widehat{R}^\dagger$ satisfies (6). We now use the guarantee (8) to construct a set $\widehat{R}^*$ satisfying (8). The basic idea is summarized in Equation (9).

$$\left.\begin{array}{rcccccccc}
\text{Bootstrap run } 1 & : & T_1^{(1)} & T_2^{(1)} & \cdots & T_B^{(1)} & \Rightarrow & \widehat{R}^\dagger_{1,B}(\alpha') \\
\text{Bootstrap run } 2 & : & T_1^{(2)} & T_2^{(2)} & \cdots & T_B^{(2)} & \Rightarrow & \widehat{R}^\dagger_{2,B}(\alpha') \\
\vdots & & \vdots & \vdots & \cdots & \vdots & & \\
\text{Bootstrap run } B' & : & T_1^{(B')} & T_2^{(B')} & \cdots & T_B^{(B')} & \Rightarrow & \widehat{R}^\dagger_{B',B}(\alpha')
\end{array}\right\} \quad \widehat{R}^* \;:=\; \bigcup_{b'=1}^{B'} \widehat{R}^\dagger_{b',B}(\alpha'). \tag{9}$$

In words, we generate B' many bootstrap datasets and obtain $\widehat{R}^\dagger_{b',B}(\alpha'), 1 \leqslant b' \leqslant B'$ satisfying (8) with α' (instead of α); the value of α' will be defined later. The final set $\widehat{R}^*$ is the union of $\widehat{R}^\dagger_{b',B}(\alpha')$.

Theorem 1. *Fix $\alpha, \delta \in [0,1]$. Let $\alpha' \in [0,1], B' \geqslant 1$ be any two numbers satisfying*

$$\alpha' + \sqrt{\frac{2\alpha' \log(1/\delta)}{B'}} + \frac{\log(1/\delta)}{B'} \leqslant \alpha. \tag{10}$$

If $\mathbb{E}[\int_{\widehat{R}^\dagger_{b',B}(\alpha')} dP^*(\delta) \big| \{Z_i\}] \geqslant 1 - \alpha'$ for all $1 \leqslant b' \leqslant B$, then $\widehat{R}^*$ defined in (9) satisfies

$$\mathbb{P}\left(\int_{\widehat{R}^*} dP^*(\delta) \geqslant 1 - \alpha \middle| \{Z_i\}\right) \geqslant 1 - \delta.$$

Proof. See Appendix A for a proof. ☐

The validity guarantee of Theorem 1 is finite sample. It does not require B or B' to diverge to infinity with the sample size; further it does not restrict the growth of the dimension p.

Inequality (10) is based on Bernstein's inequality and can be improved by using more refined concentration inequalities such as Bennett's (Theorem 3.1.7 of Giné and Nickl (2016)) or Benktus' (Theorem 1 of Bentkus (2002)). For practical implementation, we recommend the use of Bentkus' inequality because it is sharper than Bennett's concentration inequality.

The set $\widehat{R}^*$ in (9) can be replaced by a smaller set as follows. Fix $K \geqslant 0$ and define the set $\widehat{R}^\ddagger(K)$ by

$$\mathbb{1}\{x \in \widehat{R}^\ddagger\} \;\geqslant\; \frac{1}{B'}\sum_{b'=1}^{B'} \mathbb{1}\{x \in \widehat{R}^\dagger_{b',B}(\alpha')\} - \frac{K\log(1/\delta)}{B'} \quad \text{for all} \quad x \in \mathbb{R}^p. \tag{11}$$

It is clear that $\widehat{R}^\ddagger(K) \subseteq \widehat{R}^*$ for any $K > 0$. The union set $\widehat{R}^*$ is the smallest set satisfying (11) and the set $\widehat{R}^\ddagger$ reduces the set $\widehat{R}^*$ by only considering elements that belong to at least $B' - K\log(1/\delta)$ of the $\widehat{R}^\dagger_{b',B}(\alpha')$ sets. For this refined set $\widehat{R}^\ddagger(K)$, Theorem 1 does not hold readily. To restore validity, we use $\alpha' \in [0,1], B' \geqslant 1$ such that

$$\alpha' + \sqrt{\frac{2\alpha' \log(1/\delta)}{B'}} + \frac{(K+1)\log(1/\delta)}{B'} \leqslant \alpha. \tag{12}$$

For such a choice of $\alpha' \in [0,1]$ to exist, it is necessary that $B' > (K+1)\log(1/\delta)/\alpha$. We suggest using a small K so that $\widehat{R}^\ddagger(K)$ ignores such points in $\widehat{R}^*$ that only belong to one or two of the sets $\widehat{R}^\dagger_{b',B}(\alpha')$.

If $\widehat{R}^* \in \mathcal{A}$, then Theorem 1 combined with the (traditional) bootstrap consistency result (2) yields coverage validity for $\widehat{R}^*$. The assumption $\widehat{R}^* \in \mathcal{A}$ is crucial to applying (2), especially in high-dimensions where the "complexity" of $\mathcal{A}$ drastically impacts the rate of convergence in (2). Even if we construct the conformal prediction set $\widehat{R}^{\dagger}_{b',B}(\alpha')$ in such a way that they belong to $\mathcal{A}$, their union $\widehat{R}^*$ may not belong to $\mathcal{A}$; for example, take $\mathcal{A}$ to be the set of all hyper-rectangles. A natural example in high-dimensions where $\widehat{R}^* \in \mathcal{A}$ holds is $\mathcal{A} = \{\{x \in \mathbb{R}^p : \|x\|_\infty \leqslant t\} : t \geqslant 0\}$, the set of all hyper-cubes; the maximum norm here can be replaced by any other semi-norm. In many cases, one can find an element of $\mathcal{A}$ that contains $\widehat{R}^*$; for example, this is the case when $\mathcal{A}$ is the set of all hyper-rectangles.

3 A Concrete Application of Conformal Prediction

In this section, we provide a concrete application of the theory in previous section by constructing a specific conformal prediction region. Consider the problem of constructing a simultaneous confidence regions for a mean vector $\mu := (\mu_1, \mu_2, \ldots, \mu_p)^\top \in \mathbb{R}^p$. We have realizations of independent random vectors $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$ with mean $\mu \in \mathbb{R}^p$. There are many ways to construct simultaneous confidence regions:

Maximum Statistics. One can provide a single threshold for all coordinates of μ by bootstrapping the "max"-statistic:

$$\max_{1 \leqslant j \leqslant p} \frac{n^{1/2}|\bar{X}_j - \mu_j|}{\sigma_j},$$

where $\bar{X}_j$ represents the j-th coordinate of $\bar{X} = n^{-1}\sum_{i=1}^n X_i \in \mathbb{R}^p$ and $\sigma_j^2 = \mathrm{Var}(n^{1/2}(\bar{X}_j - \mu_j))$. This provides a confidence region of the form

$$\left\{\theta \in \mathbb{R}^p : \frac{n^{1/2}|\bar{X}_j - \mu_j|}{\sigma_j} \leqslant t_\alpha \quad \text{for all} \quad 1 \leqslant j \leqslant p\right\}.$$

Because the sets are hyper-cubes, the VC dimension of these sets is order 1 irrespective of what p is. Hence the empirical bootstrap distribution converges to the true bootstrap distribution, that is, (6) holds true, irrespective of what p is.

Pre-pivoted Statistics. The single threshold provides equal importance to all coordinates of $\mu \in \mathbb{R}^p$ and in some cases, there might be an importance ordering of μ_j's. Suppose we want a smaller confidence interval for μ_j than the confidence interval for μ_{j+1} for all $j \geqslant 1$. In this case, we can consider confidence regions of the type

$$\left\{\theta \in \mathbb{R}^p : \frac{n^{1/2}|\bar{X}_j - \mu_j|}{\sigma_j} \leqslant t_\alpha(j) \quad \text{for all} \quad 1 \leqslant j \leqslant p\right\}, \tag{13}$$

for some constants $t_\alpha(j)$ such that $t_\alpha(1) \leqslant t_\alpha(2) \leqslant \cdots \leqslant t_\alpha(p)$. A systematic way to obtain such increasing thresholds is by bootstrapping

$$\max_{1 \leqslant j \leqslant p} \bar{H}_j\left(H_j\left(\frac{n^{1/2}|\bar{X}_j - \mu_j|}{\sigma_j}\right)\right), \tag{14}$$

where $H_j(\cdot)$ is the cumulative distribution function (CDF) of $n^{1/2}|\bar{X}_j - \mu_j|/\sigma_j$ and $\bar{H}_j(\cdot)$ is the CDF of $\max_{1\leqslant k\leqslant j} H_k(n^{1/2}|\bar{X}_k - \mu_k|/\sigma_k)$. The quantile of (14) leads to confidence regions of the form (13) with increasing thresholds. The increasing thresholds follow from the fact that $\bar{H}_j(\cdot)$ are increasing in $1 \leqslant j \leqslant p$. The idea of considering the statistic (14) with $H_j(\cdot)$ and $\bar{H}_j(\cdot)$ is motivated by the idea of pre-pivoting from Beran (1987, 1988a,b).

In order to implement this idea with conformal prediction, we proceed as follows. For any bootstrap data $X_1^{(b)}, X_2^{(b)}, \ldots, X_n^{(b)}$ generated i.i.d. from the empirical distribution of $X_1, \ldots, X_n$, construct the bootstrap statistic

$$T_b := n^{1/2}(\mathrm{diag}(\widehat{\Sigma}))^{-1/2}\left(\bar{X}^{(b)} - \bar{X}\right),$$

where $\bar{X}^{(b)} = n^{-1}\sum_{i=1}^{n} X_i^{(b)} \in \mathbb{R}^p$ and $\widehat{\Sigma}$ is the sample covariance matrix based on $X_1, \ldots, X_n$, that is, $\widehat{\Sigma}_{jj} = (n-1)^{-1}\sum_{i=1}^{n}(X_{i,j} - \bar{X}_j)^2$. For bootstrap run 1, we have the "data" $T_b^{(1)}, 1 \leqslant b \leqslant B$. To construct $\widehat{R}_{1,B}^{\dagger}(\alpha')$ based on conformal prediction as follows:

1. Split the "data" $T_b^{(1)}, 1 \leqslant b \leqslant B$ into two parts

$$\mathcal{I}_1 := \{T_b^{(1)} : 1 \leqslant b \leqslant \lfloor B/2 \rfloor\} \quad \text{and} \quad \mathcal{I}_2 := \{T_b^{(1)} : \lfloor B/2 \rfloor + 1 \leqslant b \leqslant B\}.$$

2. Based on $\mathcal{I}_1$, construct estimators $\widehat{H}_j^{(1)}(\cdot), \widehat{\bar{H}}_j^{(1)}(\cdot)$ of $H_j(\cdot), \bar{H}_j(\cdot)$:

$$\widehat{H}_j^{(1)}(r) = \frac{1}{\lfloor B/2 \rfloor}\sum_{b=1}^{\lfloor B/2 \rfloor} \mathbb{1}\left\{|T_{b,j}^{(1)}| \leqslant r\right\}, \quad \widehat{\bar{H}}_j^{(1)}(r) = \frac{1}{\lfloor B/2 \rfloor}\sum_{b=1}^{\lfloor B/2 \rfloor} \mathbb{1}\left\{\max_{1\leqslant k\leqslant j} \widehat{H}_k(|T_{b,k}^{(1)}|) \leqslant r\right\}.$$

3. Apply conformal prediction to construct $\widehat{R}_{1,B}^{\dagger}(\alpha')$ as follows. Find the $(1+2/B)(1-\alpha')$-th quantile $\widehat{t}_{\alpha'}^{(1)}$ of

$$\max_{1\leqslant j\leqslant p} \widehat{\bar{H}}_j^{(1)}\left(\widehat{H}_j^{(1)}\left(T_{b,j}^{(1)}\right)\right), \quad \lfloor B/2 \rfloor + 1 \leqslant b \leqslant B. \tag{15}$$

The conformal prediction region is given by

$$\widehat{R}_{1,B}^{\dagger}(\alpha') \;:=\; \left\{\delta \in \mathbb{R}^p : |\delta_j| \leqslant t_{j,\alpha'}^{(1)}\right\}, \quad \text{where} \quad t_{j,\alpha'}^{(1)} \;:=\; (\widehat{\bar{H}}_j^{(1)})^{-1}\left((\widehat{H}_j^{(1)})^{-1}(\widehat{t}_{\alpha'}^{(1)})\right).$$

The procedure above is the split conformal method from Papadopoulos et al. (2002) and Lei et al. (2013); others versions of conformal prediction methods such as jackknife+ and CV+ from Barber et al. (2019) can also be used. As is well-known in the conformal literature, if we define $\widehat{t}_{\alpha'}^{(1)}$ as the quantile of randomized statistics in (15) randomized by adding $U_b \sim U(0, 10^{-8})$, then conformal prediction set $\widehat{R}_{1,B}^{\dagger}(\alpha')$ satisfies

$$1 - \alpha' \;\leqslant\; \mathbb{E}\left[\int_{\widehat{R}_{1,B}^{\dagger}(\alpha')} dP^*(\delta)\right] \;\leqslant\; 1 - \alpha' + \frac{2}{2+B}.$$

If we consider the set

$$\widehat{R}^* \;:=\; \left\{\delta \in \mathbb{R}^p : |\delta_j| \leqslant \max_{1\leqslant b'\leqslant B'} t_{j,\alpha'}^{(b')}\right\}, \tag{16}$$

then, for α', B' satisfying (10), we obtain

$$\mathbb{P}\left(\int_{\widehat{R}^*} dP^*(\delta) \geqslant 1-\alpha\right) \geqslant 1-\delta. \tag{17}$$

If the maximum in (16) is replaced by the $(B'-K\log(1/\delta))$-th quantile, then for α', B' satisfying (12) yields (17). Because inequalities (5) prove that the traditional bootstrap consistency (2) holds, we get a formal validity guarantee for $\widehat{R}^*$. The final $(1-\alpha)$ simultaneous confidence region for $\mu \in \mathbb{R}^p$ would be

$$\widehat{\mathrm{CI}}_n \ := \ \left\{\theta \in \mathbb{R}^p : \, n^{1/2}(\mathrm{diag}(\hat{\Sigma}))^{-1/2}(\bar{X}_n - \theta) \in \widehat{R}^*\right\}.$$

References

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications.* Newnes.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019). Predictive inference with the jackknife+. *arXiv:1905.02928.*

Bentkus, V. (2002). A remark on the inequalities of Bernstein, Prokhorov, Bennett, Hoeffding, and Talagrand. *Liet. Mat. Rink*, 42(3):332–342.

Bentkus, V., Kalosha, N., and Van Zuijlen, M. (2006). On domination of tail probabilities of (super) martingales: explicit bounds. *Lithuanian Mathematical Journal*, 46(1):1–43.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468.

Beran, R. (1988a). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83(403):679–686.

Beran, R. (1988b). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697.

Bhatia, R. and Davis, C. (2000). A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2019). Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:1912.10529.*

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression.* Springer Science & Business Media.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer.

Vapnik, V. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

APPENDIX

A Proof of Theorem 1

Because the bootstrap samples are independent conditional on $\{Z_i\}$, the random variables

$$\int_{\widehat{R}^{\dagger}_{1,B}(\alpha')} dP^*(\delta), \int_{\widehat{R}^{\dagger}_{2,B}(\alpha')} dP^*(\delta), \ldots, \int_{\widehat{R}^{\dagger}_{B',B}(\alpha)} dP^*(\delta) \; \in \; [0,1],$$

Setting

$$q_{\alpha'} := \mathbb{E}\left[\int_{\widehat{R}^{\dagger}_{b',B}(\alpha')} dP^*(\delta)\middle|\{Z_i\}\right],$$

Theorem 1 of Bhatia and Davis (2000) yields

$$\mathrm{Var}(\int_{\widehat{R}^{\dagger}_{b',B}(\alpha')} dP^*(\delta)\big|\{Z_i\}) \leqslant q_{\alpha'}(1-q_{\alpha'}) \leqslant \alpha'(1-\alpha'),$$

whenever $\alpha' < 1/2$. Hence, Bernstein's inequality (Theorem 3.1.7 of Giné and Nickl (2016)) implies that for all $u \geqslant 0$,

$$\mathbb{P}\left(\left|\frac{1}{B'}\sum_{b'=1}^{B'}\int_{\widehat{R}^{\dagger}_{b',B}(\alpha')} dP^*(\delta) - q_{\alpha'}\right| \geqslant \sqrt{\frac{2\alpha'(1-\alpha')u}{B'}} + \frac{u}{3B'}\middle|\{Z_i\}\right) \leqslant 2e^{-u}.$$

Bernstein's inequality here can be replaced by a more refined concentration inequality such as Theorem 1 of Bentkus (2002); see Bentkus et al. (2006, Section 9) for computation. Taking $u = \log(1/\delta)$ yields, with conditional (on $\{Z_i\}$) probability of at least $1-2\delta$,

$$\left|\frac{1}{B'}\sum_{b'=1}^{B'}\int_{\widehat{R}^{\dagger}_{b',B}(\alpha')} dP^*(\delta) - q_{\alpha'}\right| \leqslant \sqrt{\frac{2\alpha'(1-\alpha')\log(1/\delta)}{B'}} + \frac{\log(1/\delta)}{3B'}.$$

From the definition of $\widehat{R}^*$ in (9), it follows that

$$\begin{aligned}\int_{\widehat{R}^*} dP^*(\delta) \geqslant \max_{1\leqslant b'\leqslant B'}\int_{\widehat{R}^{\dagger}_{b',B}(\alpha')} dP^*(\delta) &\geqslant \frac{1}{B'}\sum_{b'=1}^{B'}\int_{\widehat{R}^{\dagger}_{b',B}(\alpha')} dP^*(\delta)\\ &\geqslant q_{\alpha'} - \sqrt{\frac{2\alpha'(1-\alpha')\log(1/\delta)}{B'}} - \frac{\log(1/\delta)}{3B'},\end{aligned}$$

with the conditional (on $\{Z_i\}$) probability of at least $1-\delta$. Hence if we take α' such that

$$1-\alpha' - \sqrt{\frac{2\alpha'\log(1/\delta)}{B'}} - \frac{\log(B')}{1/\delta} \geqslant 1-\alpha,$$

then we get with a conditional probability of at least $1-\delta$,

$$\int_{\widehat{R}^*} dP^*(\delta) \geqslant 1-\alpha.$$