

# Bootstrap

Wei Li

Syracuse University

Spring 2024

# OVERVIEW

Bootstrap methods

Bootstrap distribution

Nonparametric bootstrap, (semi)-parametric bootstrap

Bootstrap estimate of prediction error

Estimating degrees of freedom

Some discussions

# Bootstrap methods

# Bootstrap methods

Bradley Efron 1979.

“pull oneself up by one’s bootstraps” = “better oneself by one’s own effort.”

- ▶ The bootstrap is a general tool for assessing statistical accuracy.
  - ▶ expectation
  - ▶ variance (main application)
- ▶ As with cross-validation, the bootstrap can be used to estimate prediction error.
  - ▶ typically estimates well the expected prediction error  $\text{Err}$ .

# General ideas

The data are realizations of

$$Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} \mathbb{P}$$

$\mathbb{P}$  denotes an unknown distribution.

We denote a statistical procedure or estimator by

$$\hat{\theta}_n = S(Z_1, \dots, Z_n)$$

which is a (known) function  $S$  of the data  $Z_1, \dots, Z_n$ .

One typically would need to find out the

- ▶ sampling distribution of  $\hat{\theta}_n$ ,
- ▶ the expectation  $E(\hat{\theta}_n)$  or the variance  $\text{Var}(\hat{\theta}_n)$ .

If we knew the distribution  $\mathbb{P}$ :

- ▶ can simulate to obtain the distribution of any  $\hat{\theta}_n$  with arbitrary accuracy.

But we do not know what the distribution  $\mathbb{P}$ !

### Bootstrap:

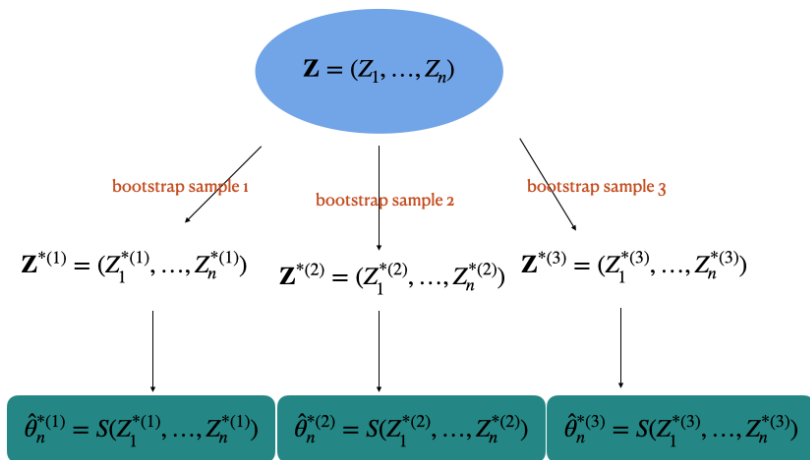
- ▶ use the empirical distribution  $\hat{\mathbb{P}}_n$  which places probability mass  $1/n$  on every data point  $Z_i, i = 1, \dots, n$ .
- ▶ simulate from  $\hat{\mathbb{P}}_n$  : generate simulated data

$$Z_1^*, \dots, Z_n^* \stackrel{i.i.d.}{\sim} \hat{\mathbb{P}}_n$$

i.e., generate  $n$  random drawings with **replacement** from the original data set  $\{Z_1, \dots, Z_n\}$ .

- ▶ such a simulated new data set is called a **bootstrap sample**.
- ▶ compute a bootstrap estimator  $\hat{\theta}_n^* = S(Z_1^*, \dots, Z_n^*)$  based on the bootstrap sample.
- ▶ We then repeat this many times (say obtain  $B$  bootstrap samples, then  $B$   $\hat{\theta}_n^*$ s )
- ▶ Get an approximate distribution for  $\hat{\theta}_n$  by the “histogram” of  $B$   $\hat{\theta}_n^*$ s.

# Bootstrap (an illustration)



$\hat{\theta}_n = S(\mathbf{Z})$  could be any quantity computed from the original data  
 $\mathbf{Z} = \{Z_1, \dots, Z_n\}$ .

The resampling with replacement is the key feature of bootstrap.

# (Nonparametric) bootstrap

The algorithm can be described as:

- ▶ 1. Generate a bootstrap sample

$$Z_1^*, \dots, Z_n^* \stackrel{i.i.d.}{\sim} \hat{\mathbb{P}}_n$$

That is obtain  $n$  random draws with replacement from the data set  $\{Z_1, \dots, Z_n\}$ .

- ▶ 2. Compute the *bootstrapped estimator* based on the bootstrap sample

$$\hat{\theta}_n^* = S(Z_1^*, \dots, Z_n^*)$$

- ▶ 3. Repeat steps 1 and 2 for  $B$  times to obtain

$$\hat{\theta}_n^{*(1)}, \dots, \hat{\theta}_n^{*(B)}.$$



# Bootstrap distribution

# Bootstrap distribution

The **bootstrap distribution** denoted by  $\mathbb{P}^*$ ,

- ▶ the conditional probability distribution which is induced by i.i.d. resampling (with replacement) of the data given the original data.

The bootstrap distribution of  $\theta_n^* = S(Z_1^*, \dots, Z_n^*)$  is the distribution which arises when resampling with  $\hat{\mathbb{P}}_n$  and applying the function  $S$  on such a bootstrap sample.

The bootstrap distribution of  $\theta_n^*$  can be described by Monte Carlo simulation:

$$\text{bootstrap expectation} \quad \mathbb{E}^* \left( \hat{\theta}_n^* \right) \cong \frac{1}{B} \sum_{b=1}^B \hat{\theta}_n^{*(b)}$$

$$\text{bootstrap variance} \quad \text{Var}^* \left( \hat{\theta}_n^* \right) \cong \frac{1}{B-1} \sum_{b=1}^B \left( \hat{\theta}_n^{*(b)} - \frac{1}{B} \sum_{b'=1}^B \hat{\theta}_n^{*(b')} \right)^2$$

$\alpha$  -quantile of the bootstrap distribution of  $\hat{\theta}_n^*$ :

empirical  $\alpha$  -quantile of  $\hat{\theta}_n^{*(1)}, \dots, \hat{\theta}_n^{*(B)}$

# Bootstrap consistency

The bootstrap is called to be **consistent** for  $\hat{\theta}_n$  if, for an increasing sequence  $a_n$ , for all  $x$

$$\mathbb{P} \left( a_n \left( \hat{\theta}_n - \theta \right) \leq x \right) - \mathbb{P}^* \left( a_n \left( \hat{\theta}_n^* - \hat{\theta}_n \right) \leq x \right) \xrightarrow{P} 0 \quad (n \rightarrow \infty)$$

In classical situations,  $a_n = \sqrt{n}$ .

In other words, if  $\hat{\theta}_n$  is asymptotically unbiased estimate some parameter  $\theta$ , the bootstrap consistency says that the sampling distribution of  $\hat{\theta}_n - \theta$  in  $\mathbb{P}$  and the bootstrap distribution of  $\hat{\theta}_n^* - \hat{\theta}_n$  in  $\mathbb{P}^*$  are asymptotically identical.

Such approximation may be reasonable when the distribution of  $\hat{\theta}_n - \theta$  is *pivotal*, that is the distribution does not depend on  $\theta$ .

# Estimating bias and variance

Under bootstrap consistency, the bias of  $\hat{\theta}_n$  may be approximated as

$$\begin{aligned} E(\hat{\theta}_n) - \theta &\approx E^*(\hat{\theta}_n^*) - \hat{\theta}_n \\ &\cong \frac{1}{B} \sum_{b=1}^B \hat{\theta}_n^{*(b)} - \hat{\theta}_n \end{aligned}$$

And

$$\text{Var}(\hat{\theta}_n) \approx \text{Var}^*(\hat{\theta}_n^*) \cong \frac{1}{B-1} \sum_{b=1}^B \left( \hat{\theta}_n^{*(b)} - \frac{1}{B} \sum_{b'=1}^B \hat{\theta}_n^{*(b')} \right)^2.$$

# Estimating confidence intervals

We can also construct the bootstrap confidence interval for  $\theta$ .

Recall that a  $(1 - \alpha)$  confidence interval for  $\theta$ , computed over  $z_1, \dots, z_n$ , is a random interval  $(L, U)$  satisfying

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

The bootstrap confidence interval for  $\theta$  is given by (why?)

$$\left(2\hat{\theta}_n - q_{1-\alpha/2}^*, 2\hat{\theta}_n - q_{\alpha/2}^*\right).$$

Here  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$ , are the  $\alpha/2$  and  $1 - \alpha/2$  are the bootstrap quantiles of  $\hat{\theta}_n^{*(1)}, \dots, \hat{\theta}_n^{*(B)}$ .

# Studentized bootstrap confidence intervals

In some cases, the distributions of  $(\hat{\theta}_n - \theta)/\widehat{\text{SE}}(\hat{\theta}_n)$  and  $(\hat{\theta}_n^* - \hat{\theta}_n)/\widehat{\text{SE}}(\hat{\theta}_n^*)$  could be close, where  $\widehat{\text{SE}}(\cdot)$  denote estimated standard errors. The so-called **studentized bootstrap confidence intervals** is obtained:

- ▶ repeat, for  $b = 1, \dots, B$  :
  - ▶ draw a bootstrap sample  $z_1^{*(b)}, \dots, z_n^{*(b)}$  from  $\{z_1, \dots, z_n\}$
  - ▶ recompute the statistic  $\hat{\theta}_n^{*(b)}$  based on  $z_1^{*(b)}, \dots, z_n^{*(b)}$
  - ▶ repeat, for  $m = 1, \dots, M$  :
    - ▶ draw a bootstrap sample  $z_1^{*(b,m)}, \dots, z_n^{*(b,m)}$  from  $\{z_1^{*(b)}, \dots, z_n^{*(b)}\}$
    - ▶ recompute the statistic  $\hat{\theta}_n^{*(b,m)}$  from  $\{z_1^{*(b,m)}, \dots, z_n^{*(b,m)}\}$
  - ▶ compute the sample standard deviation  $\hat{s}^{*(b)}$  of  $\hat{\theta}_n^{*(b,1)}, \dots, \hat{\theta}_n^{*(b,M)}$
  - ▶ compute  $(\hat{\theta}_n^{*(b)} - \hat{\theta}_n)/\hat{s}^{*(b)}$ .

From above we have a sample  $\{(\hat{\theta}_n^{*(b)} - \hat{\theta}_n)/\hat{s}^{*(b)} : b = 1, \dots, B\}$ , from which, we compute the quantiles  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$ .

The approximate  $1 - \alpha$  bootstrap confidence interval for  $\theta$  is given by

$$(\hat{\theta}_n - \widehat{\text{SE}}(\hat{\theta}_n)q_{1-\alpha/2}^*, \hat{\theta}_n - \widehat{\text{SE}}(\hat{\theta}_n)q_{\alpha/2}^*),$$

- $\widehat{\text{SE}}(\hat{\theta}_n)$  can be approximated with  $\text{Var}^*(\hat{\theta}_n^*)$  using bootstrap samples  $\{\hat{\theta}_n^{*(1)}, \dots, \hat{\theta}_n^{*(B)}\}$ .

# Nonparametric bootstrap, (semi)-parametric bootstrap



# Parametric bootstrap

Assume that the data are realizations from

$$Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} \mathbb{P}_\theta$$

where  $\mathbb{P}_\theta$  is given up to an unknown parameter (vector)  $\theta$ .

- ▶ estimate the unknown parameter  $\theta$  by  $\hat{\theta}_n$
- ▶ draw

$$Z_1^*, \dots, Z_n^* \stackrel{i.i.d.}{\sim} \mathbb{P}_{\hat{\theta}_n}$$

## Example (parametric regression)

- ▶  $Y_i = \beta^\top x_i + \varepsilon_i, (i = 1, \dots, n), \varepsilon_1, \dots, \varepsilon_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \theta = (\beta, \sigma^2).$
  - ▶ training data  $z = \{z_1, z_2, \dots, z_n\}$ , with  $z_i = (x_i, y_i) \ i = 1, 2, \dots, n.$
  - ▶  $\hat{\beta}, \hat{\sigma}$  denote the MLE estimates based on original data.
1. Simulate  $\varepsilon_1^*, \dots, \varepsilon_n^* \stackrel{i.i.d.}{\sim} N(0, \hat{\sigma}^2)$
  2. Construct

$$Y_i^* = \hat{\beta}^\top x_i + \varepsilon_i^*, i = 1, \dots, n$$

The parametric bootstrap regression sample is then

$$(x_1, Y_1^*), \dots, (x_n, Y_n^*)$$

where the predictors  $x_i$  are from the original data.

## nonparametric regression

Suppose  $Y = f(X) + \epsilon$ ,  $E(Y|X = x) = f(x) \approx \sum_{j=1}^M \beta_j h_j(x)$  where  $\text{var}(\epsilon) = \sigma^2$ .

$$\hat{\beta} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{y}$$

$$\widehat{\text{Var}}(\hat{\beta}) = (\mathbf{H}^\top \mathbf{H})^{-1} \hat{\sigma}^2$$

$$\hat{\sigma}^2 = \sum_{i=1}^n \left( y_i - \hat{f}(x_i) \right)^2 / n$$

- ▶ Let  $h(x)^\top = (h_1(x), h_2(x), \dots, h_M(x))$ .
- ▶  $\hat{f}(x) = h(x)^\top \hat{\beta}$
- ▶ standard error  $\widehat{\text{se}}(\hat{f}(x)) = \left( h(x)^\top (\mathbf{H}^\top \mathbf{H})^{-1} h(x) \right)^{\frac{1}{2}} \hat{\sigma}$ .
- ▶ The (biased) 95% pointwise confidence interval is  $\hat{f}(x) \pm 1.96 \widehat{\text{se}}(\hat{f}(x))$ .

## example: nonparametric bootstrap

Suppose we have  $n = 50$ . The nonparametric bootstrap works as in the following.

- ▶ We draw  $B$  datasets each of size  $n = 50$  with replacement from our training data, the sampling unit being the pair  $z_i = (x_i, y_i)$ .
- ▶ To each bootstrap dataset  $\mathbf{Z}^*$  we fit a cubic spline  $\hat{f}^*(x)$ .
- ▶ Using  $B = 200$  bootstrap samples, we can form a 95% pointwise confidence band from the percentiles at each  $x$ : we find the  $2.5\% \times 200 =$  fifth largest and smallest values at each  $x$ .

Generally, for  $x$ , obtain the upper and lower quantiles  $\hat{f}^*(x) - \hat{f}(x)$ , say  $R_{\alpha/2}, R_{1-\alpha/2}$ , then the pointwise confidence interval is given by  $(\hat{f}(x) - R_{1-\alpha/2}(x), \hat{f}(x) + R_{\alpha/2}(x))$ .

## example: semi-parametric bootstrap

Simulate new responses by adding Gaussian noise to the predicted values:

$$y_i^* = \hat{f}(x_i) + \varepsilon_i^*; \quad \varepsilon_i^* \sim N(0, \hat{\sigma}^2); \quad i = 1, 2, \dots, n$$

This process is repeated  $B$  times, where  $B = 200$  say. The resulting bootstrap datasets have the form  $(x_1, y_1^*), \dots, (x_n, y_n^*)$  and we recompute the splines on each.

The confidence intervals from this method will exactly equal the least squares intervals, as the number of bootstrap samples goes to infinity.

- the estimate based on bootstrap sample is

$$\hat{f}^*(x) = h(x)^\top (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{y}^* \text{ with distribution}$$

$$\hat{f}^*(x) \sim N\left(\hat{f}(x), h(x)^\top (\mathbf{H}^\top \mathbf{H})^{-1} h(x) \hat{\sigma}^2\right)$$

.

## Another version of bootstrap (residual bootstrap)

Suppose

$$Y_i = f(x_i) + \varepsilon_i$$
$$\varepsilon_1, \dots, \varepsilon_n \stackrel{i.i.d.}{\sim} \mathbb{P}_\varepsilon$$

where  $P_\varepsilon$  is unknown with expectation 0.

1. Estimate  $\hat{f}$  from the original data and compute the residuals  $r_i = Y_i - \hat{f}(x_i)$ .
2. Consider the centered residuals  $\tilde{r}_i = r_i - n^{-1} \sum_{i=1}^n r_i$ . In case of linear regression with an intercept, the residuals are already centered. Denote the empirical distribution of the centered residuals by  $\hat{\mathbb{P}}_{\tilde{r}}$ .
3. Generate

$$\varepsilon_1^*, \dots, \varepsilon_n^* \stackrel{i.i.d.}{\sim} \hat{\mathbb{P}}_{n, \tilde{r}}$$

Note that  $\hat{\mathbb{P}}_{n, \tilde{r}}$  is an estimate of  $\mathbb{P}_\varepsilon$ .

4. Construct the bootstrap response variables

$$Y_i^* = \hat{f}(x_i) + \varepsilon_i^*, i = 1, \dots, n$$

and the bootstrap sample is  $(x_1, Y_1^*), \dots, (x_n, Y_n^*)$ .

# Confidence band via bootstrap

Bootstrapping  $\sup_x |\hat{f}(x) - E(\hat{f}(x))|$ :

More generally, if  $\sigma^2 = \sigma^2(x)$  is not constant in  $x$ ,  $\hat{\sigma}^2(x)$  can be estimated as using the regression  $e_i^2 := (y_i - \hat{f}(x_i))^2$  v.s.  $x_i$ , and taking  $\hat{\sigma}^2(x_i) = \hat{e}_i^2$ .

One then obtain bootstrap (upper)  $\alpha$  quantile  $R_\alpha$  from

$$\frac{\sqrt{n} \sup_x |\hat{f}(x) - E(\hat{f}(x))|}{\hat{\sigma}(x)}$$

Then  $\hat{f}(x) \pm (R_\alpha \hat{\sigma}(x) / \sqrt{n})$  has a (supposedly) sup-norm coverage.

- ▶ still biased for the true  $f$  unless done with undersmoothing
- ▶ Or view this confidence band as for the smoothed version  $E(\hat{f}(\cdot))$  instead of  $f$ .

Bootstrap is to address the variance estimate, without undersmoothing, the above band is still biased. A even better approach is to use debiased estimator with the bootstrap.

## Bootstrap estimate of prediction error



If  $\hat{f}^{*(b)}(x_i)$  is the predicted value at  $x_i$ , from the model fitted to the  $b$ -th bootstrap dataset, our estimate of EPE is

$$\widehat{\text{Err}}_{\text{naive}} = \frac{1}{n} \sum_{i=1}^n \sum_{b=1}^B \frac{1}{B} L\left(y_i, \hat{f}^{*(b)}(x_i)\right)$$

- ▶ Repeat for  $b = 1, \dots, B$ :
  - ▶ Generate  $\left(X_1^{*(b)}, Y_1^{*(b)}\right), \dots, \left(X_n^{*(b)}, Y_n^{*(b)}\right)$  by resampling with replacement from the original data.
  - ▶ Compute the bootstrapped estimator  $\hat{f}^{*(b)}(\cdot)$  based on  $\left(X_1^{*(b)}, Y_1^{*(b)}\right), \dots, \left(X_n^{*(b)}, Y_n^{*(b)}\right)$
  - ▶ Evaluate  $\text{err}^{*(b)} = n^{-1} \sum_{i=1}^n L\left(Y_i, \hat{f}^{*(b)}(X_i)\right)$
- ▶ Approximate the bootstrap generalization error Err by

$$B^{-1} \sum_{b=1}^B \text{err}^{*(b)}$$

## Leave-one-out bootstrap estimate

Above estimate is not a good estimate in general. Tends to be overfitting.

The **leave-one-out bootstrap estimate** of prediction error is defined by

$$\widehat{\text{Err}}^{(1)} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|C^{-i}|} \sum_{b \in \{1, \dots, B\} \cap C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

- ▶  $C^{-i}$  is the set of indices of the bootstrap samples  $b$  that do not contain observation  $i$ ,
- ▶  $|C^{-i}|$  is the number of such samples.

The leave-one out bootstrap solves the overfitting problem suffered by  $\widehat{\text{Err}}_{\text{naive}}$ , but has the training-set-size bias mentioned in the discussion of cross-validation.

- ▶ Typically, the leave-one out bootstrap estimate will be biased upward.

# Bias

- ▶ Given a bootstrap sample by  $\mathbf{Z}^* = \{Z_1^*, \dots, Z_n^*\}$ , the out-of-bootstrap sample

$$\mathbf{Z}_{\text{out}}^* = \{Z_i : Z_i \notin \mathbf{Z}^*\}$$

The out-of-bootstrap estimate above can be written as:

$$\widehat{\text{Err}}^{(1)} = \frac{1}{B} \sum_{b=1}^B \frac{1}{|\mathbf{Z}_{\text{out}}^{*(b)}|} \sum_{i \in \mathbf{Z}_{\text{out}}^{*(b)}} L(y_i, \hat{f}^{*(b)}(x_i))$$

Note that  $\hat{f}^{*(b)}(\cdot)$  involves only data from  $\mathbf{Z}^{*(b)}$ , and  $(X_i, Y_i) \in \mathbf{Z}_{\text{out}}^*$ .

- ▶ The expected size of the out-of-bootstrap sample:  
 $E^*(|\mathbf{Z}_{\text{out}}^*|) \approx 0.368n$ .
- ▶  $\widehat{\text{Err}}^{(1)}$  is like a CV estimate that uses about 36.8% data points as test data, or about 63.2% data points as training data
- ▶ Unlike CV estimate, the training data in  $\widehat{\text{Err}}^{(1)}$  may have duplicates.

## The .632 estimator

The “.632 estimator” is designed to alleviate this bias:

$$\widehat{\text{Err}}^{(.632)} = .368 \cdot \overline{\text{err}} + .632 \cdot \widehat{\text{Err}}^{(1)}$$

The derivation of the .632 estimator is complex; intuitively it pulls the leave-one out bootstrap estimate down toward the training error rate, and hence reduces its upward bias.

Note that  $\overline{\text{err}} \leq \widehat{\text{Err}}^{(.632)} \leq \widehat{\text{Err}}^{(1)}$ .

- ▶ The .632 estimator works well in “light fitting” situations
- ▶ In the heavily-overfitting situations, one can further improve the .632 estimator: the .632+ estimator

# Estimating degrees of freedom

The so-called **effective d.f.** of a fitted model is

$$\text{d.f.}(\hat{f}) = \frac{\sum_{i=1}^n \text{cov}(\hat{y}_i, y_i)}{\sigma^2}, \quad \{x_i\}_i \text{ fixed}$$

We can estimate the covariance terms  $\text{cov}(\hat{y}_i, y_i)$  via the bootstrap.

After fitting  $\hat{y}_i = \hat{f}(x_i), i = 1, \dots, n$  using the original samples  $(x_i, y_i), i = 1, \dots, n$ , we record the (empirical) residuals

$$\hat{e}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

Then for  $b = 1, \dots, B$ , we repeat:

- ▶ obtain a bootstrap sample  $(x_i, \tilde{y}_i^{(b)}), i = 1, \dots, n$  according to

$$\tilde{y}_i^{(b)} = \hat{y}_i + \tilde{e}_i^{(b)}, \text{ where } \tilde{e}_i^{(b)} \stackrel{i.i.d.}{\sim} \{\hat{e}_1, \dots, \hat{e}_n\}, \quad i = 1, \dots, n$$

- ▶ re-estimate the estimator  $\hat{f}^{(b)}$  based on the sample  $(x_i, \tilde{y}_i^{(b)})_{i=1}^n$
- ▶ store  $\tilde{\mathbf{y}}^{(b)} = (\tilde{y}_1^{(b)}, \dots, \tilde{y}_n^{(b)})$ , and  $\hat{\mathbf{y}}^{(b)} = (\hat{f}^{(b)}(x_1), \dots, \hat{f}^{(b)}(x_n))$ .

With  $\tilde{\mathbf{y}}^{(b)} = (\tilde{y}_1^{(b)}, \dots, \tilde{y}_n^{(b)})$ , and  $\hat{\mathbf{y}}^{(b)} = (\hat{f}^{(b)}(x_1), \dots, \hat{f}^{(b)}(x_n))$ ,

we approximate the covariance of  $\hat{y}_i$  and  $y_i$  by the empirical covariance between  $\hat{y}_i^{(b)}$  and  $\tilde{y}_i^{(b)}$  over  $b = 1, \dots, B$ , i.e.

$$\text{Cov}(\hat{y}_i, y_i) \approx \frac{1}{B} \sum_{b=1}^B \left( \hat{y}_i^{(b)} - \frac{1}{B} \sum_{b'=1}^B \hat{y}_i^{(b')} \right) \left( \tilde{y}_i^{(b)} - \frac{1}{B} \sum_{b'=1}^B \tilde{y}_i^{(b')} \right).$$

Summing this up over  $i = 1, \dots, n$  yields the bootstrap estimate for degrees of freedom

$$\widehat{\text{d.f.}}(\hat{f}) \approx \frac{1}{\sigma^2} \sum_{i=1}^n \left( \frac{1}{B} \sum_{b=1}^B \left( \hat{y}_i^{(b)} - \frac{1}{B} \sum_{b'=1}^B \hat{y}_i^{(b')} \right) \left( \tilde{y}_i^{(b)} - \frac{1}{B} \sum_{b'=1}^B \tilde{y}_i^{(b')} \right) \right)$$

## Some discussions



# Correct way to do bootstrap for inference

The bootstrap procedure should be applied to the entire estimation process to obtain correct inference.

Suppose that we adaptively choose by cross-validation the number and position of the knots that define the  $B$ -splines, rather than fix them in advance. Denote by  $\lambda$  the collection of knots and their positions. Then the standard errors and confidence bands should account for the adaptive choice of  $\lambda$ .

The catch is with the bootstrap, we compute the  $B$ -spline smoother with an adaptive choice of knots for *each* bootstrap sample.

# Estimation of generalized error of tuned model

The  $0.632$  ( $0.632+$ ) estimators were proposed to assess the generalization error of a generic algorithm and can be directly applied to a (adaptively) tuned model:

specifically, for each bootstrap sample, the model can be fit and tuned (say by CV) with this bootstrap sample.

Alternatively, one can use Wang and Zou (2021) Honest leave-one-out cross-validation or nested cross validation to assess the performance of a tuned model.

# Bayesian bootstrap

The bootstrap discussed above is a “poor man’s” version of Bayesian bootstrap.

Bayesian bootstrap sample:

- ▶ Draw weights from a uniform Dirichlet distribution with the same dimension as the number of data points
- ▶ Sample from data accordingly to the probability defined by the Dirichlet weights
- ▶ Use the resampled data to calculate the statistics.