PS5841

Data Science in Finance & Insurance

Bootstrap & Bagging

Yubo Wang

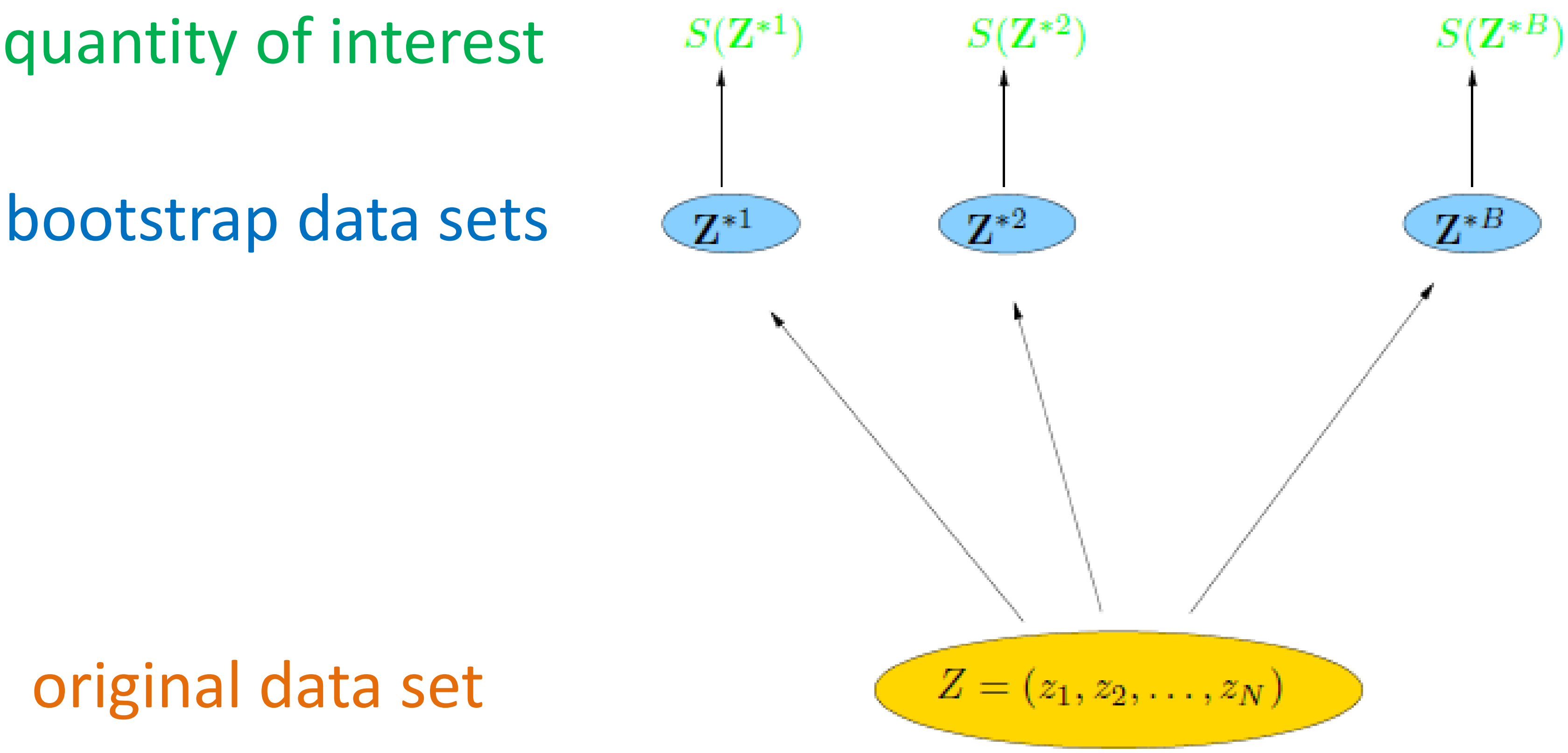
Autumn 2021

Bootstrap

The Bootstrap

- Emulate the process of obtaining new samples without generating additional samples
- Obtain distinct data sets by repeatedly sampling, with replacement, observations from the original data set
- Works well if the original data set is a good representation of the population
  - Risk of “garbage in, garbage out”

The Bootstrap



quantity of interest

$S(\mathbf{Z}^{*1})$ $S(\mathbf{Z}^{*2})$ $S(\mathbf{Z}^{*B})$

bootstrap data sets

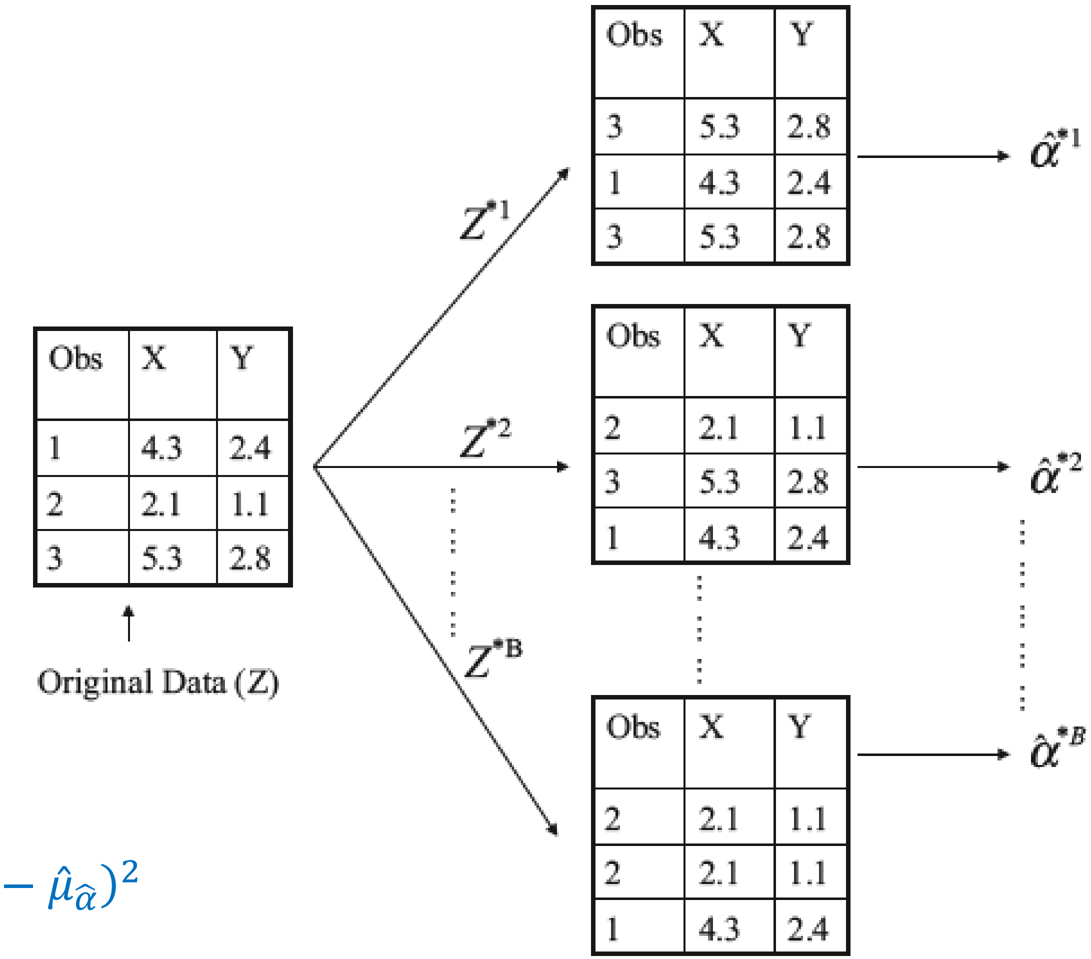
$\mathbf{Z}^{*1}$ $\mathbf{Z}^{*2}$ $\mathbf{Z}^{*B}$

original data set

$Z = (z_1, z_2, \ldots, z_N)$

Bootstrap Data Sets

$$\hat{\mu}_{\hat{\alpha}} = \frac{1}{B}\sum_{r=1}^{B} \hat{\alpha}^{*r}$$



Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

$$\hat{\sigma}^2_{\hat{\alpha}} = \frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \hat{\mu}_{\hat{\alpha}}\right)^2$$

Example (1)

- $X \sim \mathcal{N}(\mu_X = 0, \sigma_X^2 = 10^2)$
- Training Set $\mathrm{Z} = (x_1, x_2, \dots, x_{N=100})$
- Quantity of interest $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\texttt{true\_mean}(\bar{X}) = \mu_X = 0$$

$$\texttt{true\_se}(\bar{X}) = \frac{\sigma_X}{\sqrt{N}} = 1$$

Example (2)

- $X \sim \mathcal{N}(\mu_X = 0, \sigma_X^2 = 10^2)$
- Training Set $\mathrm{Z} = (x_1, x_2, \dots, x_{N=100})$
- Quantity of interest $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Generate B=1000 true training sets

$$\texttt{true\_sampling\_mean}(\bar{X}) = \frac{1}{B} \sum_{r=1}^{B} \overline{x}^{(r)}$$

$$\texttt{true\_sampling\_se}(\bar{X})$$

$$= \left[\frac{1}{B-1} \sum_{r=1}^{B} \left(\bar{x}^{(r)} - \texttt{true\_sampling\_mean}(\bar{X})\right)^2\right]^{\frac{1}{2}}$$

Example (3)

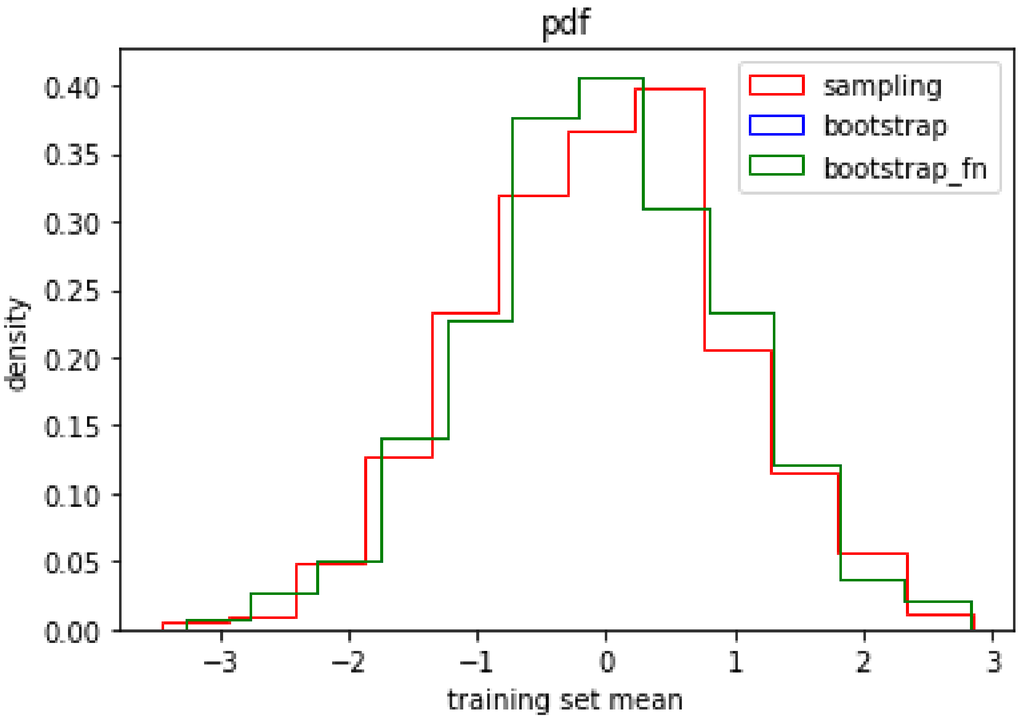
- $X \sim \mathcal{N}(\mu_X = 0, \sigma_X^2 = 10^2)$
- Training Set $\mathrm{Z} = (x_1, x_2, \dots, x_{N=100})$
- Quantity of interest $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Generate B=1000 bootstrap training sets

$$\texttt{boot\_sampling\_mean}(\bar{X}) = \frac{1}{B}\sum_{r=1}^{B} \bar{x}^{*(r)}$$

$$\texttt{boot\_sampling\_se}(\bar{X})$$

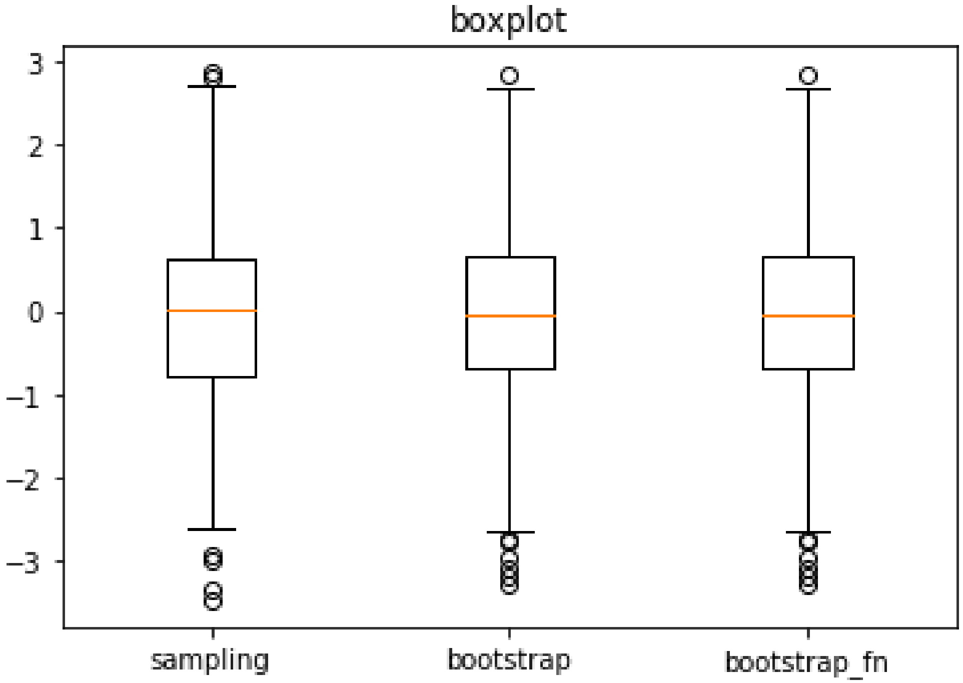
$$= \left[\frac{1}{B-1}\sum_{r=1}^{B}\left(\bar{x}^{*(r)} - \texttt{boot\_sampling\_mean}(\bar{X})\right)^2\right]^{\frac{1}{2}}$$

Example (4)



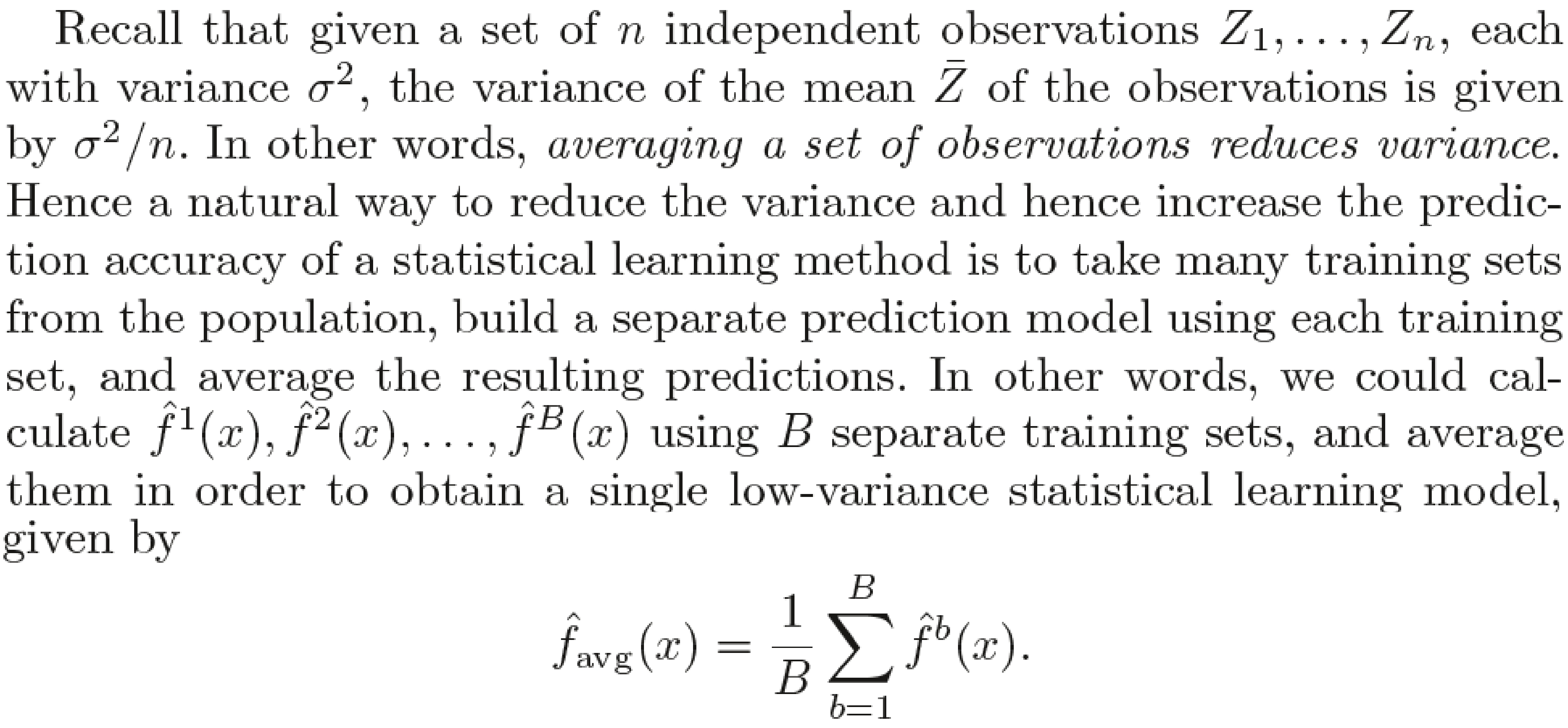
```
theoretical mean & se          =   0.000000,  1.000000
"true" sampling mean & se      =  -0.045906,  1.009381
"bootstrap" sampling mean & se =  -0.033890,  1.005676
```

Example　(5)



Bagging

Averaging to reduce variance in SL



Recall that given a set of n independent observations $Z_1, \ldots, Z_n$, each with variance σ^2, the variance of the mean $\bar{Z}$ of the observations is given by σ^2/n. In other words, *averaging a set of observations reduces variance.* Hence a natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. In other words, we could calculate $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ using B separate training sets, and average them in order to obtain a single low-variance statistical learning model, given by

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x).$$

Bagging

Of course, this is not practical because we generally do not have access to multiple training sets. Instead, we can bootstrap, by taking repeated samples from the (single) training data set. In this approach we generate B different bootstrapped training data sets. We then train our method on the bth bootstrapped training set in order to get $\hat{f}^{*b}(x)$, and finally average all the predictions, to obtain

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$

This is called bagging.

That was