Statistical learning and deep learning: theoretical background and hands-on sessions

Lezione 3

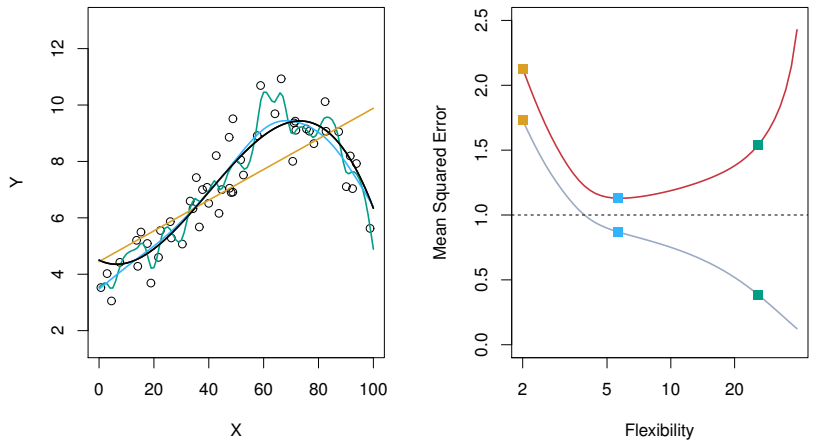
S. Biffani

IBBA/CNR

27 gennaio, 2023

Metodi di campionamento 1

- perchè campionare ?
- *cross-validation*
  - stima del *test error*
  - scelta del livello di flessibilità
- *bootstrap*
  - accuratezza di un parametro
  - metodi *ensemble* (bagging, boosting)



Metodi di campionamento 2

- Validation Set Approach

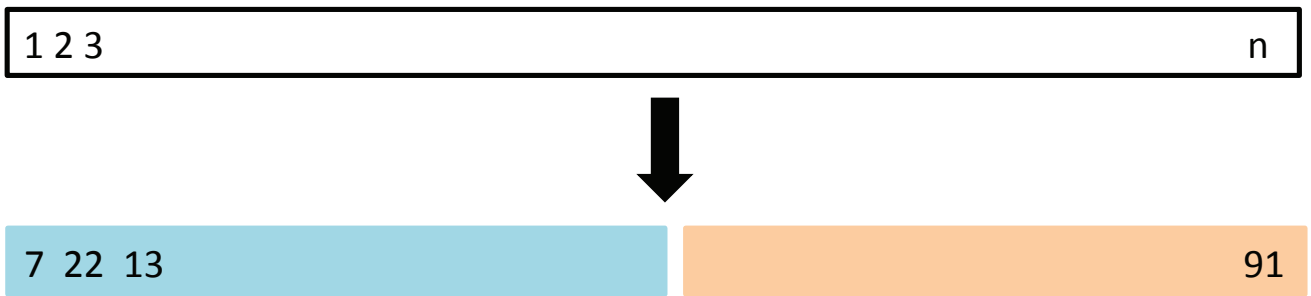


Figura 1: Validation Set Approach: divido in 2 parti il train data. Con una sviluppo il modello e con l'altra lo testo

Metodi di campionamento 3

Validation Set Approach - Esempio Velocità e Potenza

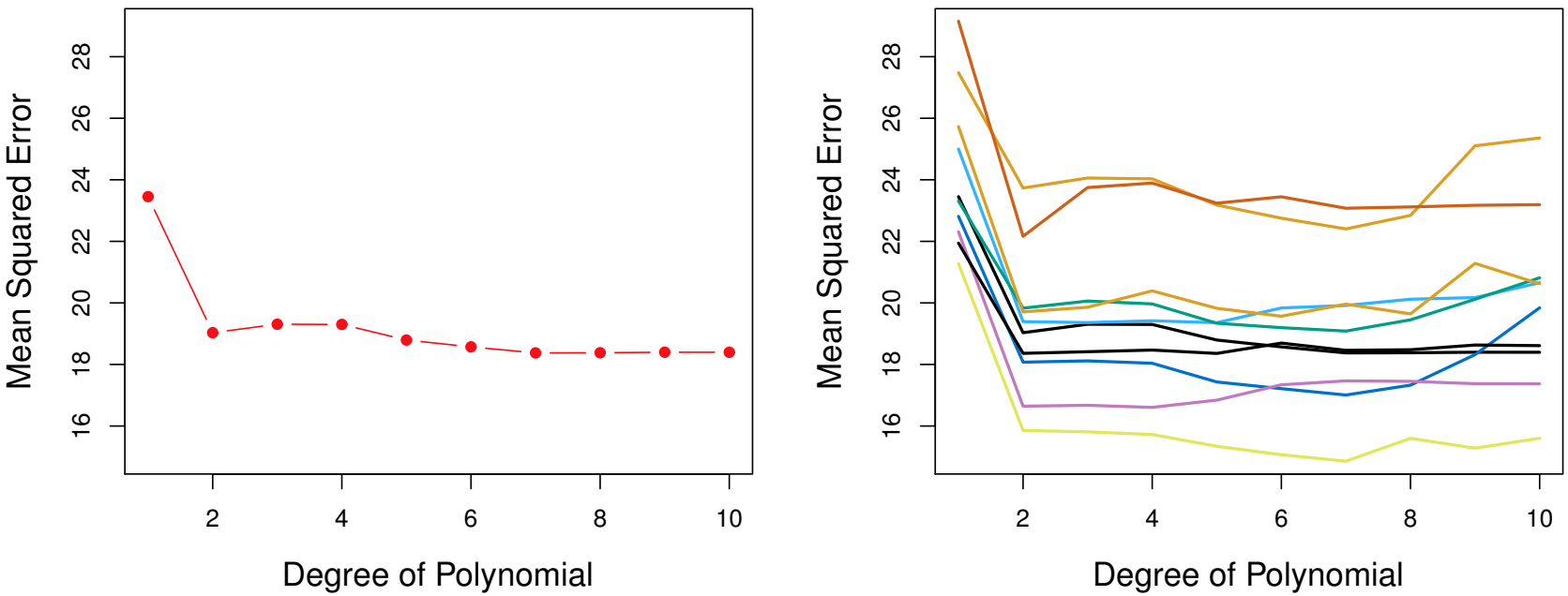


Figura 2: Test MSE:Guardiamo la figura a SINISTRA

Metodi di campionamento 4

Validation Set Approach - Esempio Velocità e Potenza

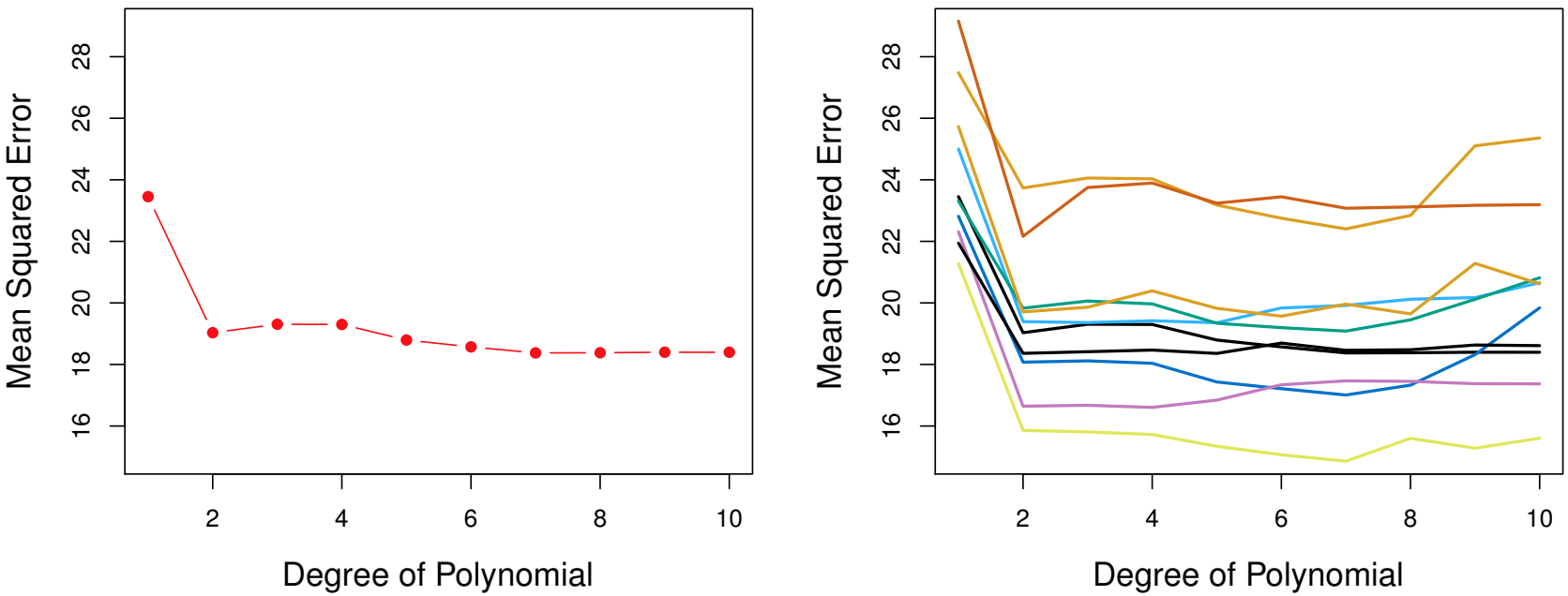


Figura 3: Test MSE:Ora Guardiamo la figura a destra

Metodi di campionamento 4

Validation Set Approach

- PRO: molto semplice da usare
- CONTRO:
  - anche se ripetuto potrebbe fornire una stima molto variabile del *test MSE* (a causa della randomizzazione nella creazione dei 2 subset)
  - il *test MSE* tende ad essere *sovrastimato*

Metodi di campionamento 5

Leave-One-Out Cross-Validation

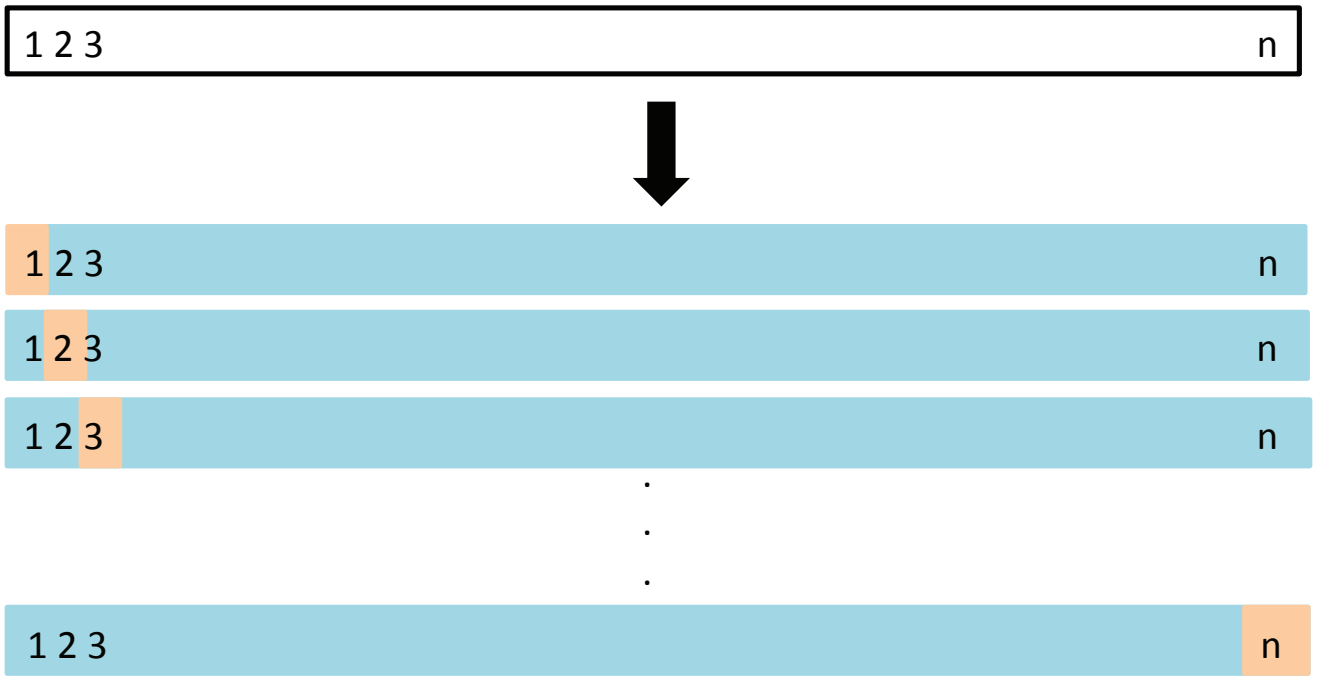
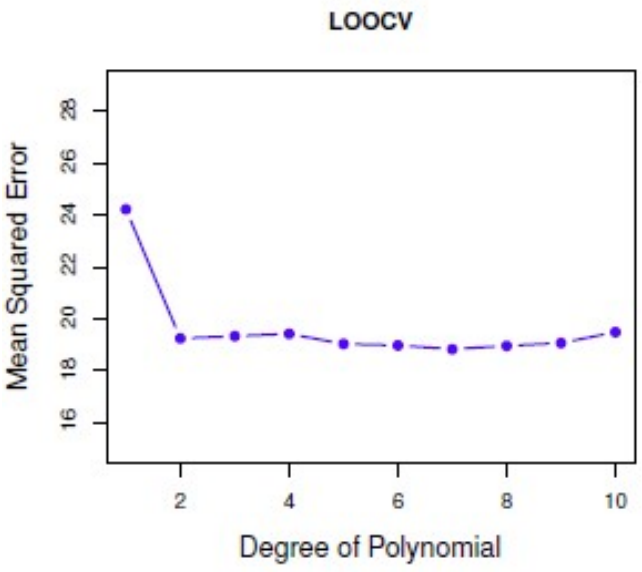


Figura 4: LOOCV

Metodi di campionamento 6

Leave-One-Out Cross-Validation



test error = media dei singoli *test error*

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(MSE_i)$$

n = numero di **records**

PRO:

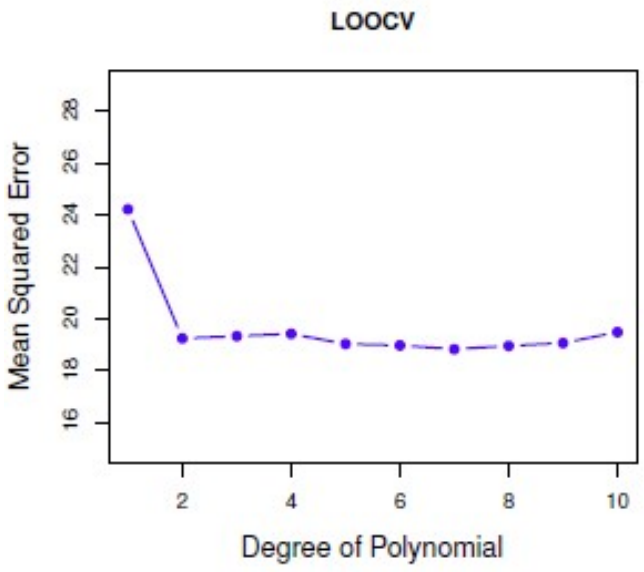
1. minor bias (# dati $>$ nel *train data*)
2. stima più stabile (non c'è randomizzazione)

Metodi di campionamento 7

Leave-One-Out Cross-Validation

test error = media dei singoli *test error*

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(MSE_i)$$



CONTRO:

1. time consuming (se # dati molto grande)
2. *train data* più correlati (effetto su compromesso bias-variance)
3. metodo oggi *deprecated*

Metodi di campionamento 8

k-Fold Cross-Validation

- evoluzione del LOOCV

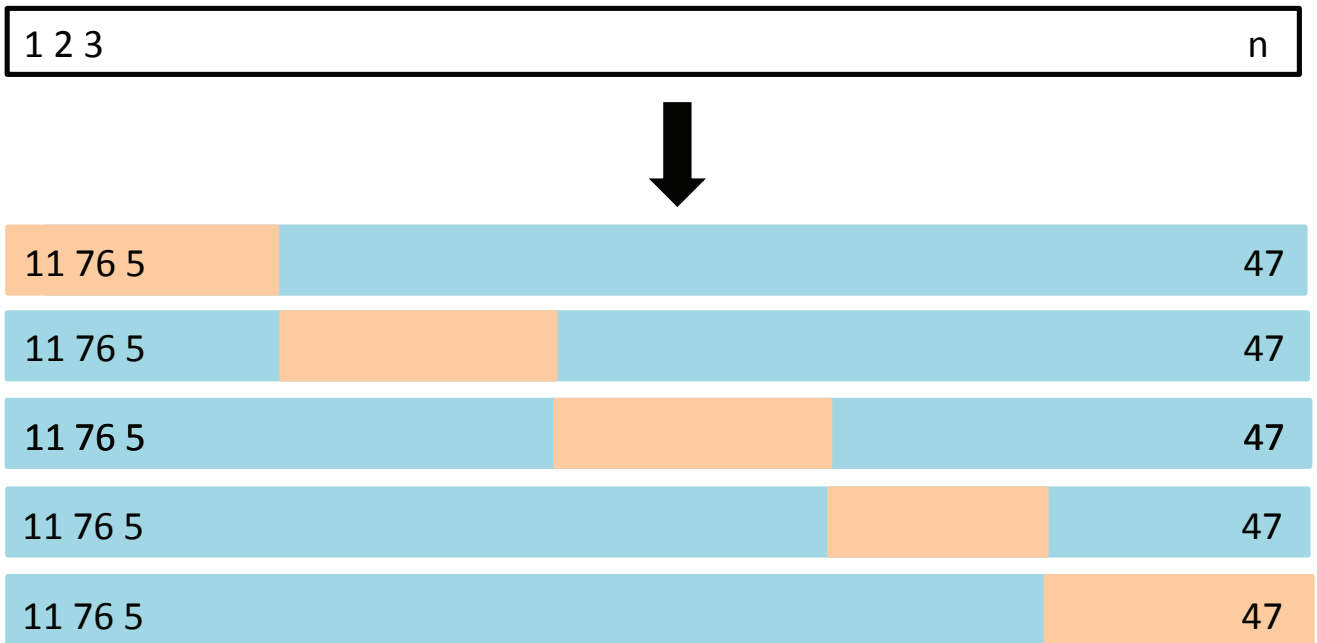
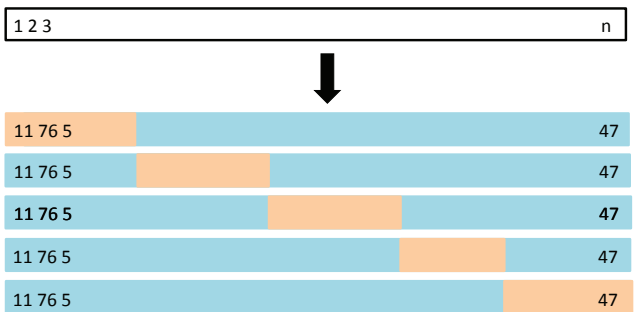
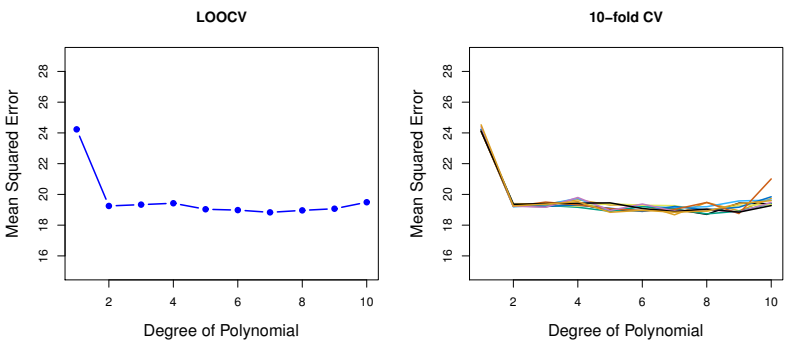


Figura 5: Esempio di 5-fold CV

Metodi di campionamento 9





k *test errors* = media delle k stime

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}(MSE_i)$$

k = numero di **folds**

PRO:

1. meno time consuming del LOOCV
2. stima accurate del *test MSE*

Metodi di campionamento 10

Dati Simulati: Vero test MSE, LOOCV MSE e 10-fold MSE

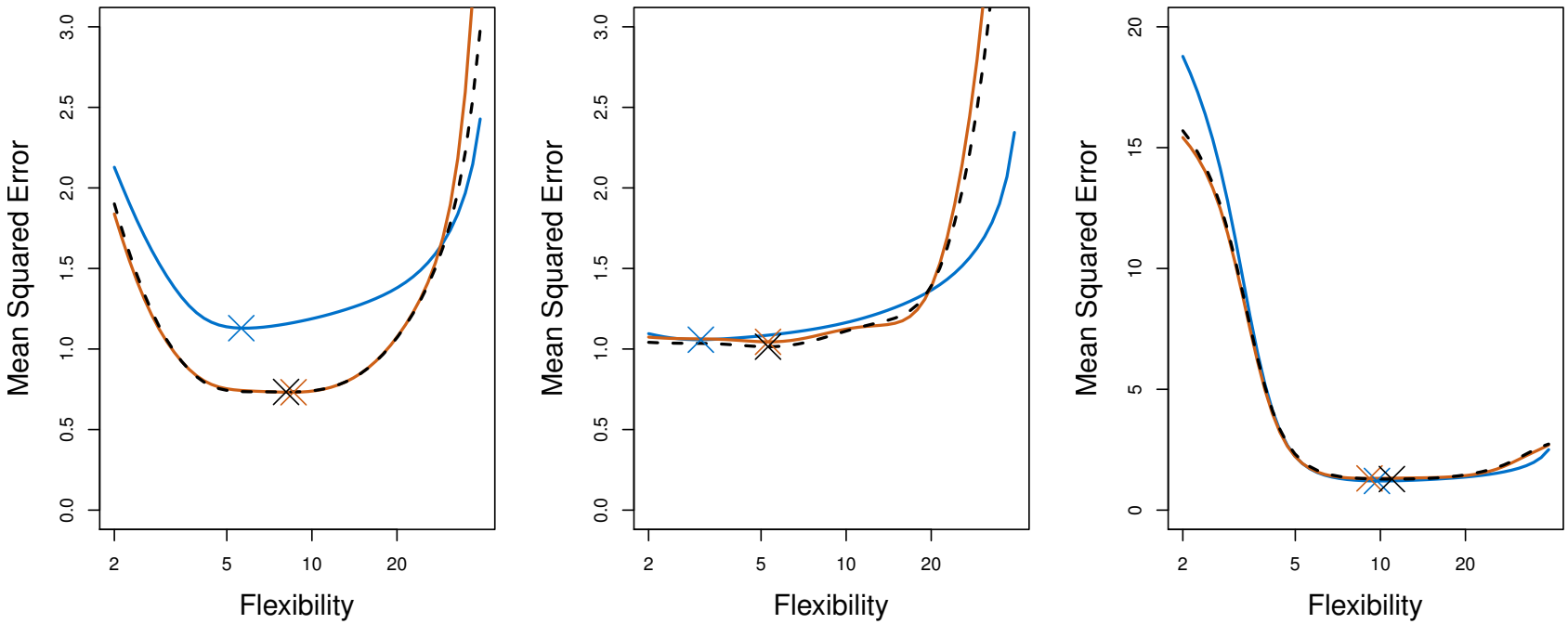


Figura 6: Effetto del metodo di campionamento sulla stima del Test MSE

Metodi di campionamento 10

k-fold CV: Compromesso tra Bias e Varianza

- **validation set**: test error sovrastimato + alta variabilità
- **LOOCV**: data set correlati. La media di quantità altamente correlate ha maggiore variabilità!
- **Esempio:** generiamo 2 set di 5000 campioni ognuno composto da 2 valori (e.g. MSE_i e MSE_{i-1}). Nel primo set usiamo una correlazione pari a .9, nel secondo pari a 0

Metodi di campionamento 11

```
library(mvtnorm)
library(tidyverse)
set.seed(9876)
rho <- 0.9
n_sims <- 5000
sigma_corr = matrix(c(1, rho, rho, 1), nrow = 2, ncol = 2)
sigma_uncorr <- diag(2)
```

generiamo i 2 set di dati:

```
samples_corr <- as.data.frame(rmvnorm(n_sims,
                                      mean = c(0, 0),
                                      sigma = sigma_corr))

samples_uncorr <- as.data.frame(rmvnorm(n_sims,
                                        mean = c(0, 0),
                                        sigma = sigma_uncorr))
```

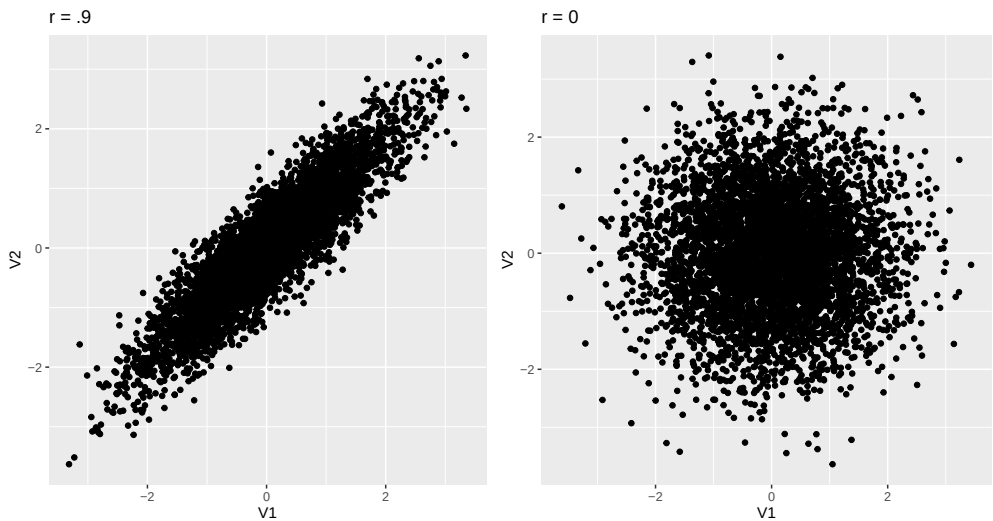
Metodi di campionamento 11

Verifichiamo la distribuzione

```
library(patchwork)
ggplot(samples_corr, aes(x = V1, y = V2)) +
    geom_point() +
    coord_fixed()+
  ggtitle('r = .9')-> CORRS

ggplot(samples_uncorr, aes(x = V1, y = V2)) +
    geom_point() +
    coord_fixed()+
  labs(title='r = 0')-> UNCORRS

CORRS | UNCORRS -> g
g
```



Metodi di campionamento 12

Calcoliamo le 2 medie, visualizziamo la distribuzione e calcoliamo la varianza

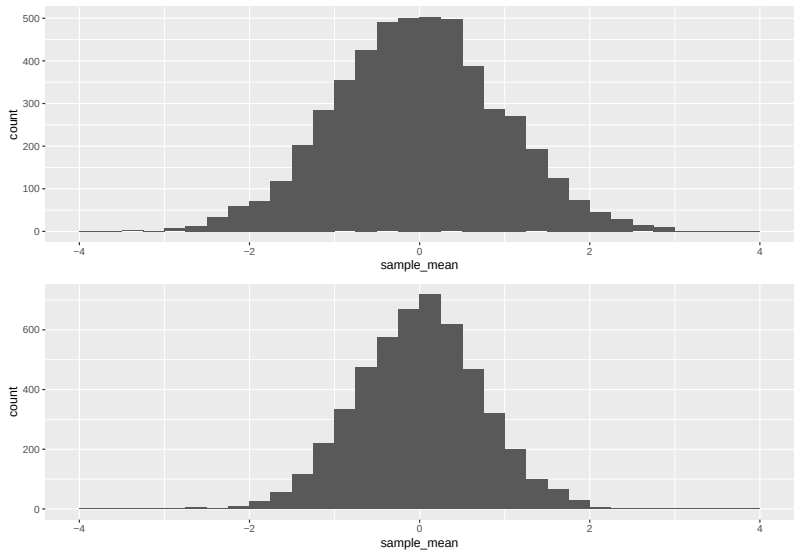
```
library(patchwork)
### R = .9
samples_corr2 <- samples_corr %>%
    rowwise() %>%
    mutate(sample_mean = mean(c(V1, V2)))

ggplot(samples_corr2, aes(x = sample_mean)) +
    xlim(-4, 4) +
    geom_histogram(binwidth = 0.25,
                   boundary = 0)-> hist1
### R = 0
samples_uncorr2 <- samples_uncorr %>%
    rowwise() %>%
    mutate(sample_mean = mean(c(V1, V2)))

ggplot(samples_uncorr2, aes(x = sample_mean)) +
    xlim(-4, 4) +
    geom_histogram(binwidth = 0.25,
                   boundary = 0)-> hist2

### varianze
var_corr <- var(samples_corr2$sample_mean)
var_uncorr <- var(samples_uncorr2$sample_mean)
```

```
hist1/hist2
```




```
var_corr
```

```
## [1] 0.9543948
```

```
var_uncorr
```

```
## [1] 0.5169131
```

Metodi di campionamento 13

Cross-Validation con variabili qualitative:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} Err_i$$

dove $Err_i = I(y_i \neq \hat{y}_i)$

Metodi di campionamento 14

- bootstrap

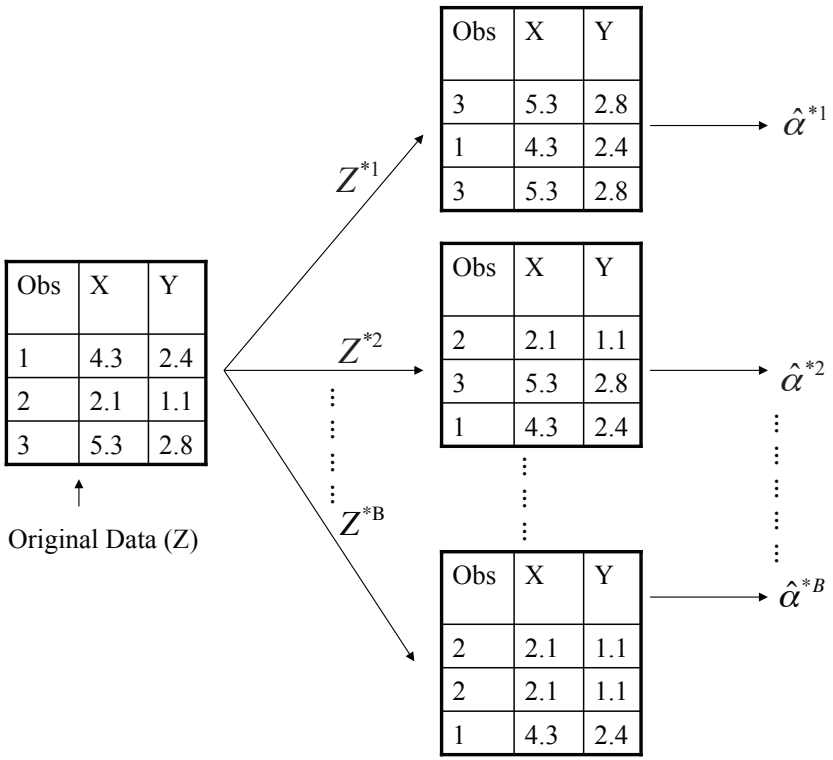


Figura 8: Bootstrap