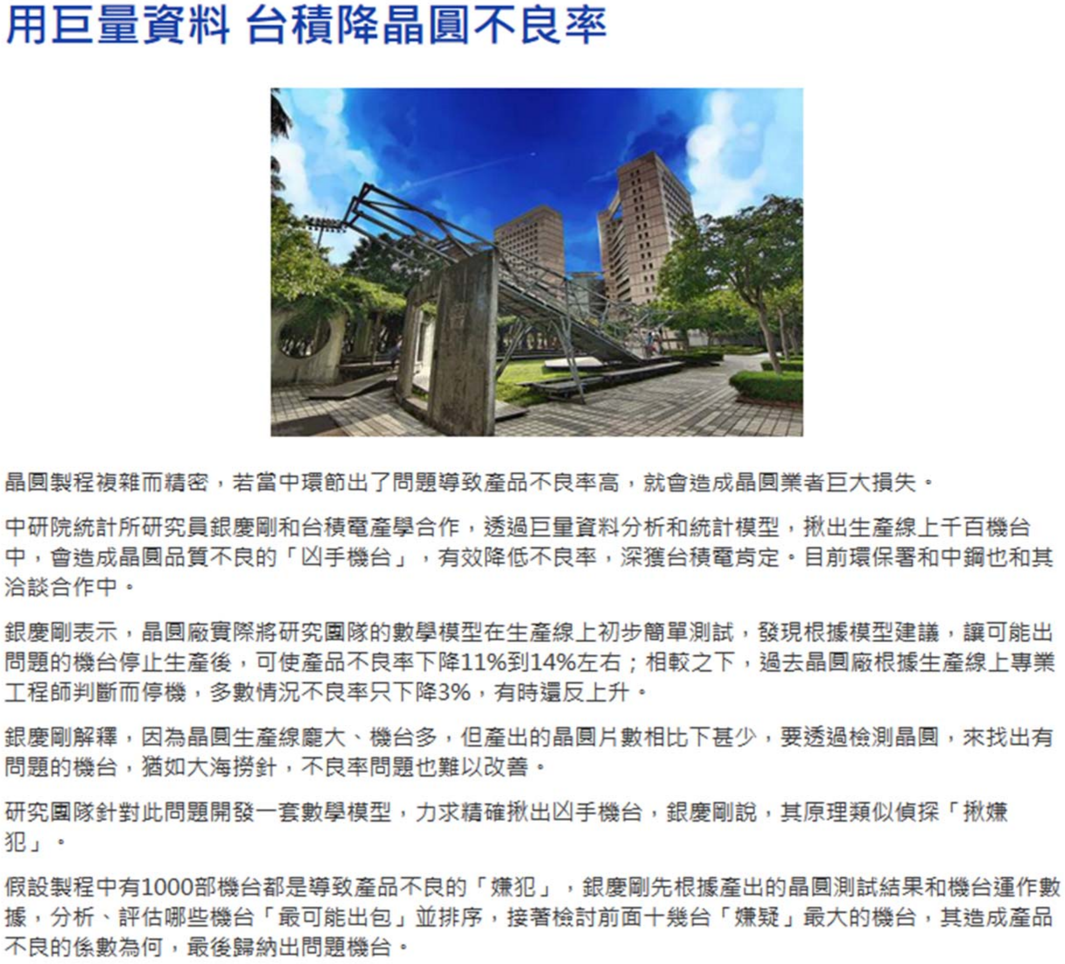
Model Selection

Also known as

- Feature selection
- Variables selection
- Attribute selection
- Variable subset selection

…



用巨量資料 台積降晶圓不良率

晶圓製程複雜而精密，若當中環節出了問題導致產品不良率高，就會造成晶圓業者巨大損失。

中研院統計所研究員銀慶剛和台積電產學合作，透過巨量資料分析和統計模型，揪出生產線上千百機台中，會造成晶圓品質不良的「凶手機台」，有效降低不良率，深獲台積電肯定。目前環保署和中鋼也和其洽談合作中。

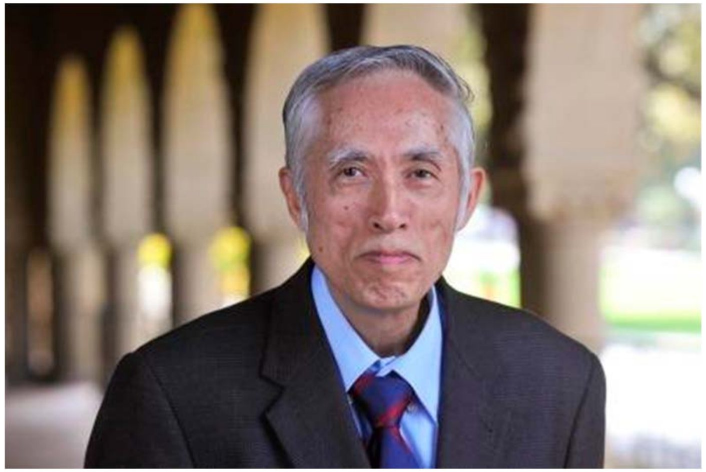
銀慶剛表示，晶圓廠實際將研究團隊的數學模型在生產線上初步簡單測試，發現根據模型建議，讓可能出問題的機台停止生產後，可使產品不良率下降11%到14%左右；相較之下，過去晶圓廠根據生產線上專業工程師判斷而停機，多數情況不良率只下降3%，有時還反上升。

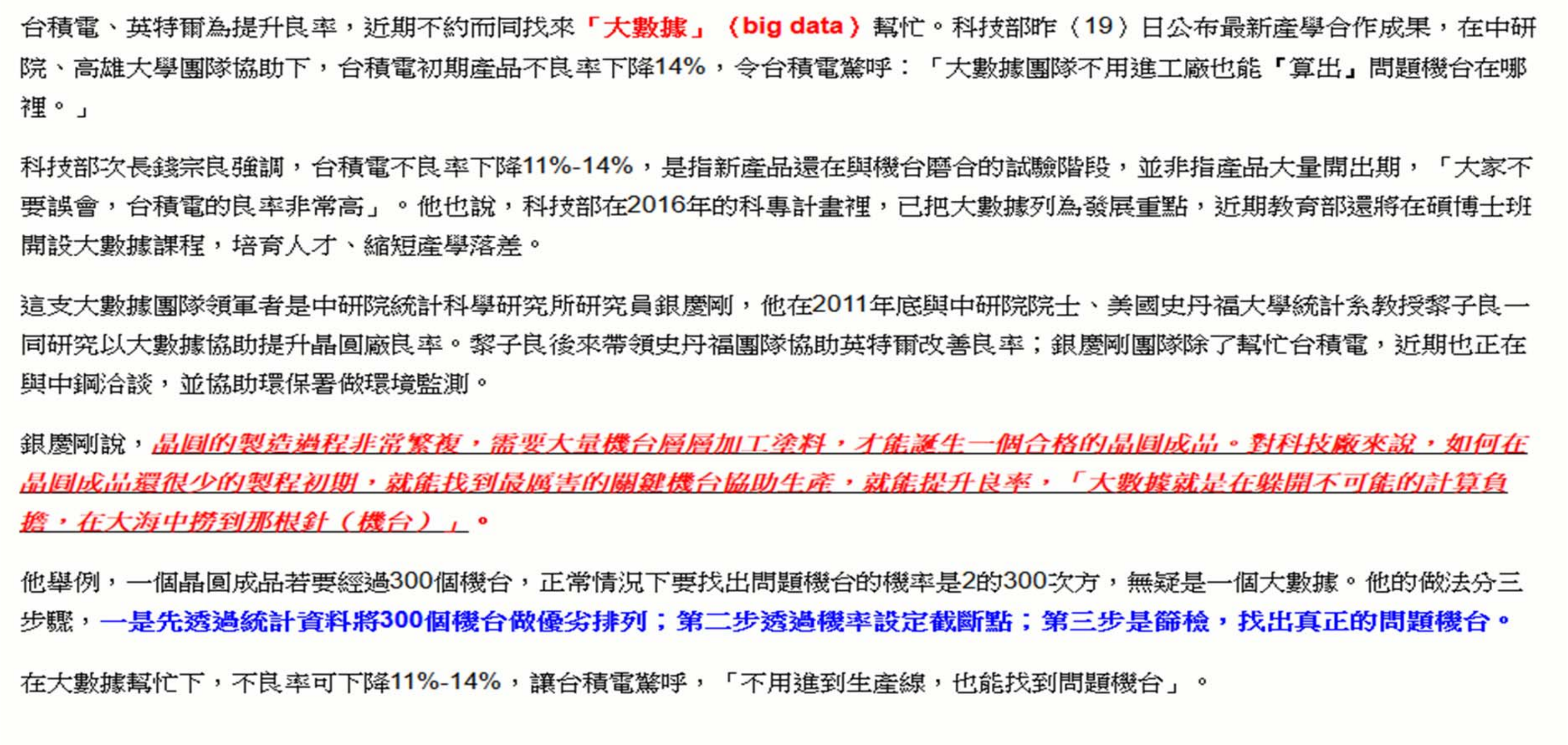
銀慶剛解釋，因為晶圓生產線龐大、機台多，但產出的晶圓片數相比下甚少，要透過檢測晶圓，來找出有問題的機台，猶如大海撈針，不良率問題也難以改善。

研究團隊針對此問題開發一套數學模型，力求精確揪出凶手機台，銀慶剛說，其原理類似偵探「揪嫌犯」。

假設製程中有1000部機台都是導致產品不良的「嫌犯」，銀慶剛先根據產出的晶圓測試結果和機台運作數據，分析、評估哪些機台「最可能出包」並排序，接著檢討前面十幾台「嫌疑」最大的機台，其造成產品不良的係數為何，最後歸納出問題機台。

https://news.ltn.com.tw/news/life/breakingnews/1161356
https://www.most.gov.tw/most/attachments/cfe65ec3-076a-412b-b960-0ede03885de3
http://statweb.stanford.edu/~ckirby/lai/
http://www.genehope.com.tw/modules/tadnews/index.php?nsn=2787





台積電、英特爾為提升良率，近期不約而同找來**「大數據」（big data）**幫忙。科技部昨（19）日公布最新產學合作成果，在中研院、高雄大學團隊協助下，台積電初期產品不良率下降14%，令台積電驚呼：「大數據團隊不用進工廠也能**『算出』**問題機台在哪裡。」

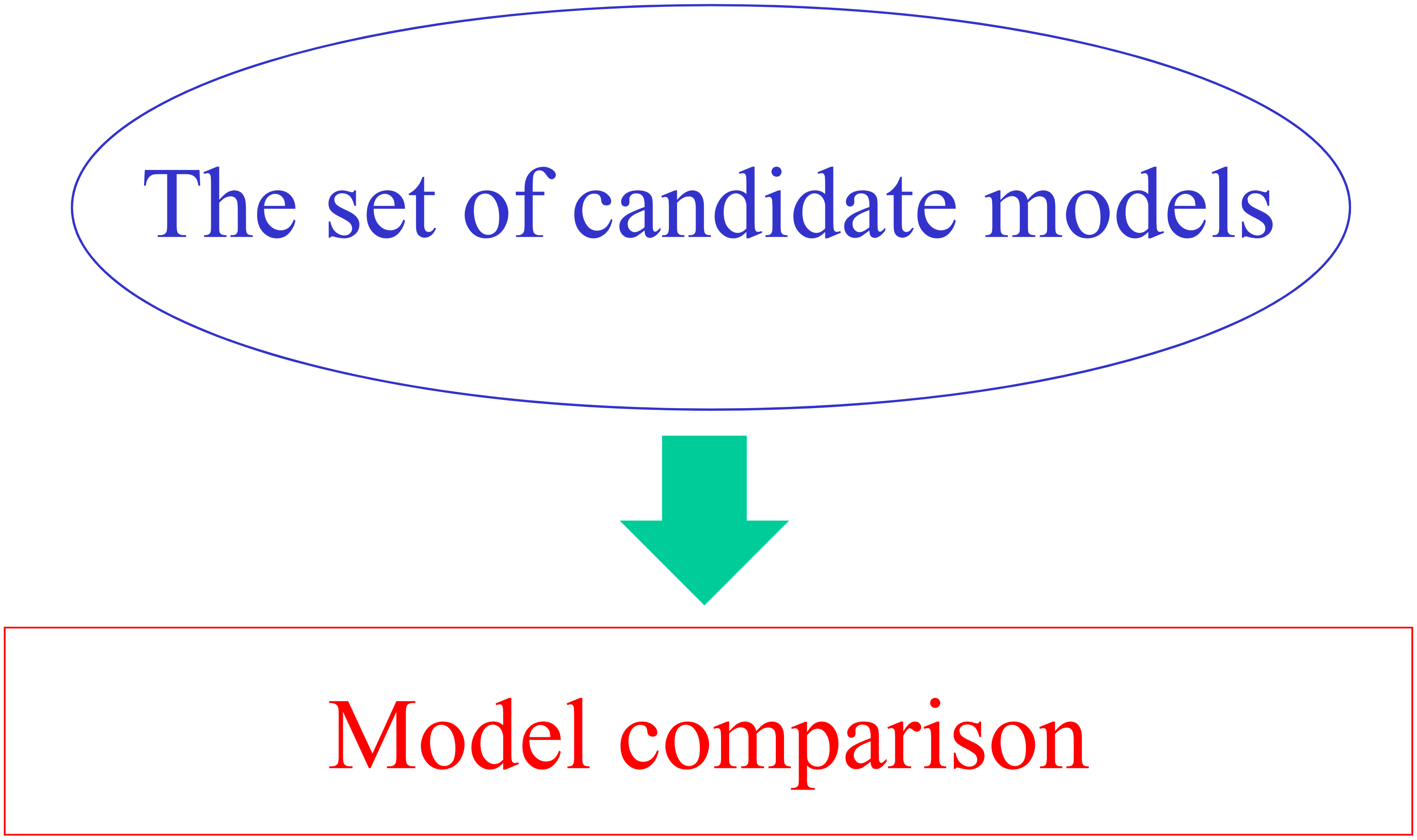
科技部次長錢宗良強調，台積電不良率下降11%-14%，是指新產品還在與機台磨合的試驗階段，並非指產品大量開出期，「大家不要誤會，台積電的良率非常高」。他也說，科技部在2016年的科專計畫裡，已把大數據列為發展重點，近期教育部還將在碩博士班開設大數據課程，培育人才、縮短產學落差。

這支大數據團隊領軍者是中研院統計科學研究所研究員銀慶剛，他在2011年底與中研院院士、美國史丹福大學統計系教授黎子良一同研究以大數據協助提升晶圓廠良率。黎子良後來帶領史丹福團隊協助英特爾改善良率；銀慶剛團隊除了幫忙台積電，近期也正在與中鋼洽談，並協助環保署做環境監測。

銀慶剛說，***晶圓的製造過程非常繁複，需要大量機台層層加工塗料，才能誕生一個合格的晶圓成品。對科技廠來說，如何在晶圓成品還很少的製程初期，就能找到最厲害的關鍵機台協助生產，就能提升良率，「大數據就是在躲開不可能的計算負擔，在大海中撈到那根針（機台）」***。

他舉例，一個晶圓成品若要經過300個機台，正常情況下要找出問題機台的機率是2的300次方，無疑是一個大數據。他的做法分三步驟，**一是先透過統計資料將300個機台做優劣排列；第二步透過機率設定截斷點；第三步是篩檢，找出真正的問題機台。**

在大數據幫忙下，不良率可下降11%-14%，讓台積電驚呼，「不用進到生產線，也能找到問題機台」。



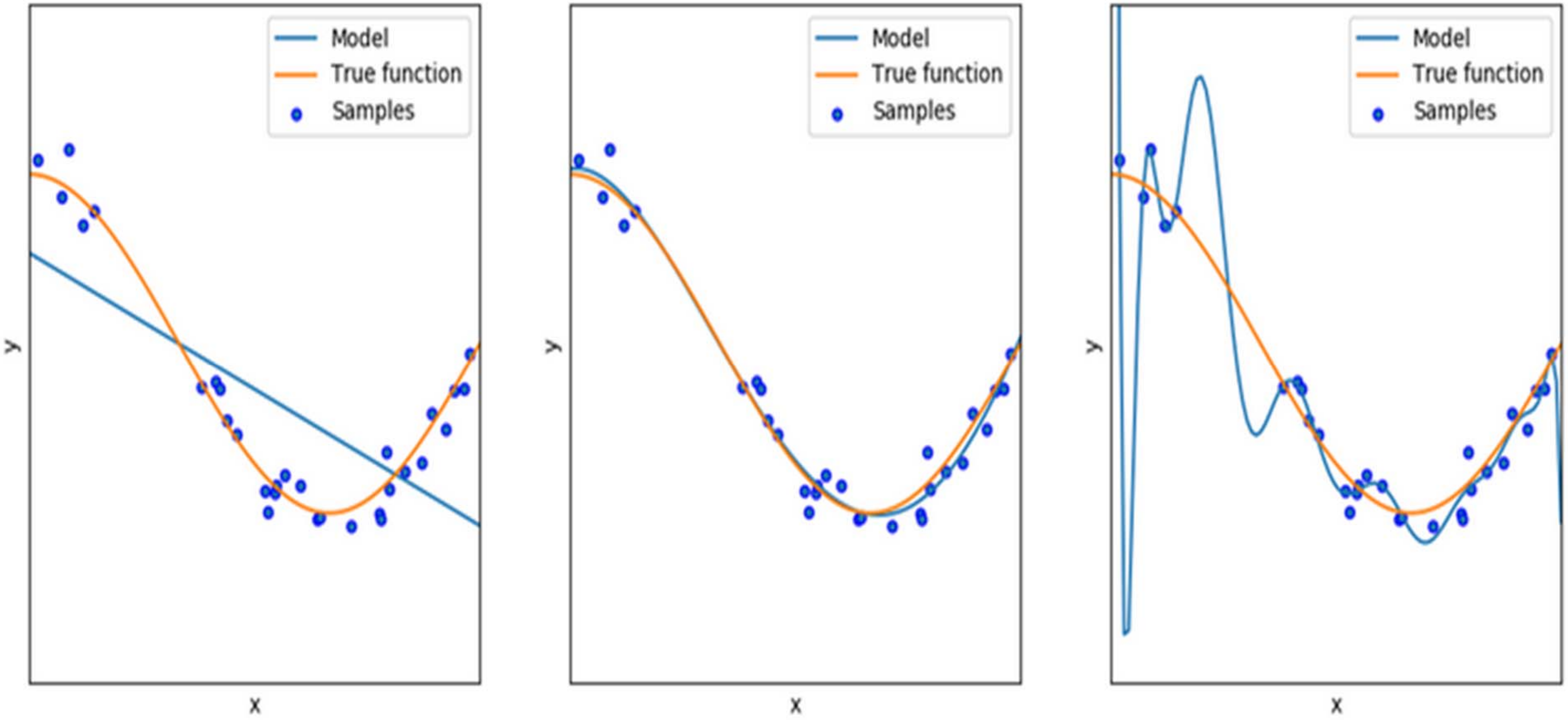
The set of candidate models

Model comparison

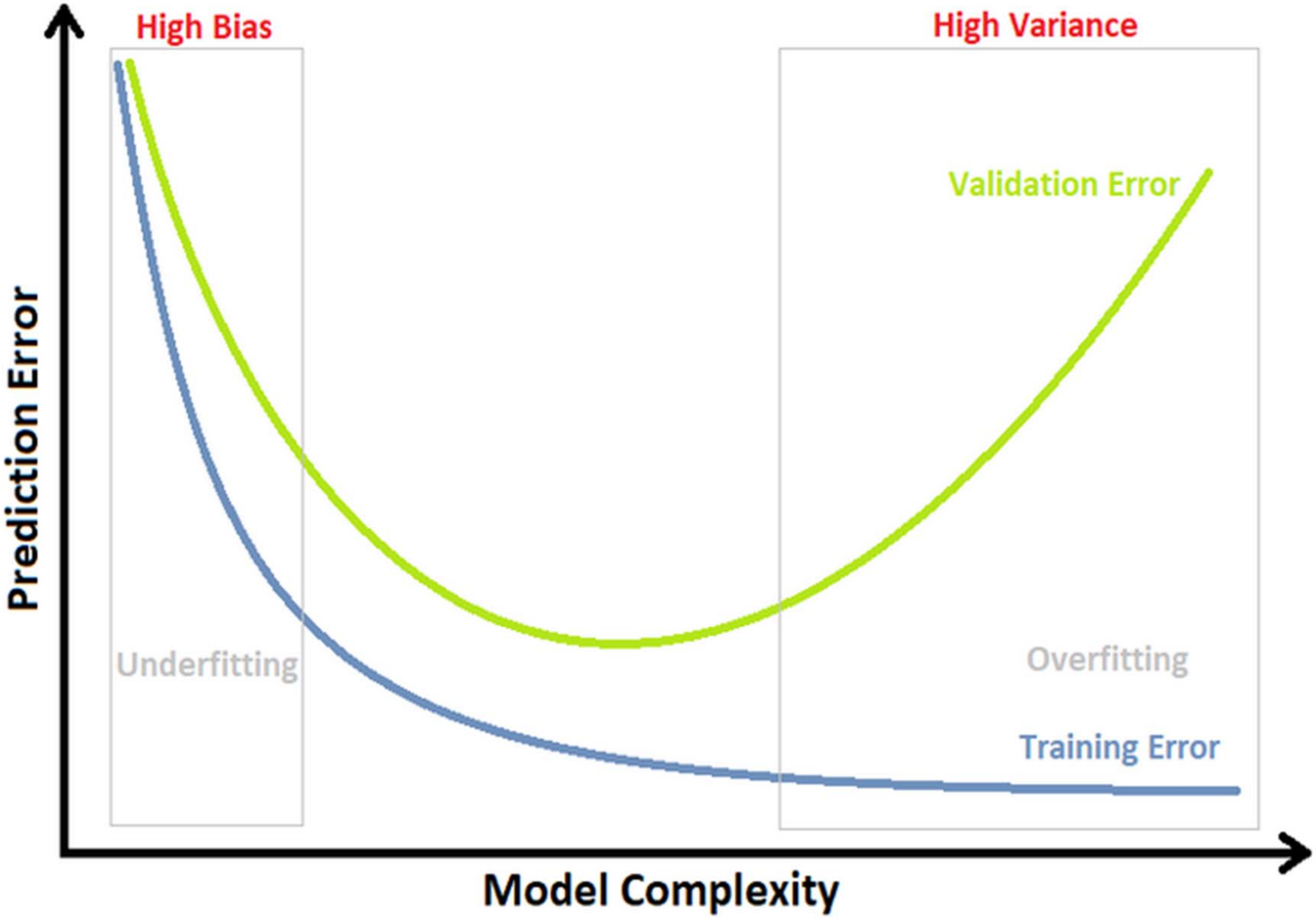
How to select the “best” model?

- the parsimonious model, i.e. the smallest correct model, avoid under-fitting and over-fitting
- the model that will have the lowest error for prediction, i.e. the model has the lower mean squared prediction error

under-fitting and over-fitting



https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html



https://jason-chen-1992.weebly.com/home/-bias-variance-tradeoff

An ideal classifier

- Accurate able to classify unseen cases with low error
- Intuitively provide insight into the roles
- Unbiased inference about predictors variables should be correct and unbiased
- Fast classification rule should be reasonably quick to construct

The number of candidate models

Assume that all relevant variables that could be included, plus perhaps a few irrelevant ones, are included in the set of k possible independent variables.

$p = 1,$ the number of candidate models $= 2 - 1 = 1$

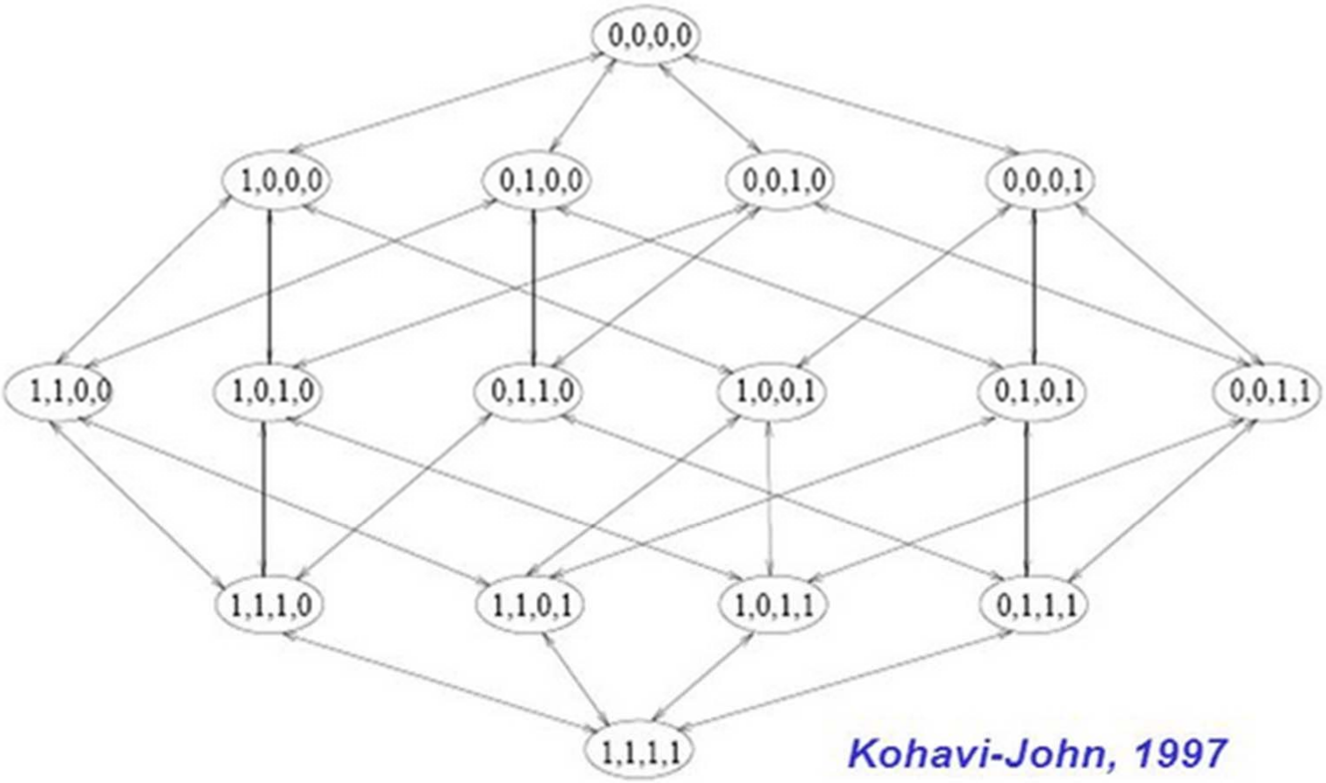
$p = 2,$ the number of candidate models $= 2^2 - 1 = 3$

$p = 10,$ the number of candidate models $= 2^{10} - 1 = 1023$

$p = 20$ (1048575), 30 or more...

How to choose from a large number, ex. $2^{100} (\doteq 10^{30})$, of candidate models?

Possible variable subsets



https://slideplayer.com/slide/4829839/

Conventional methods of model selection

- Testing-based methods: forward selection, backward elimination, and stepwise procedure
- Information-based methods: AIC, BIC or HQ, etc.
- Prediction-based methods: Mallows' C_p, cross-validation (CV)
- A criterion is required for deciding if one subset is better than another.

Consider

$$y_i = \beta_0 + \boldsymbol{x}_i{}' \boldsymbol{\beta} + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, i = 1, \ldots, n$$

where $\boldsymbol{x}_i$ is a p-dimensional explanatory vector,

$\boldsymbol{\beta}$ is the coefficient vector, and

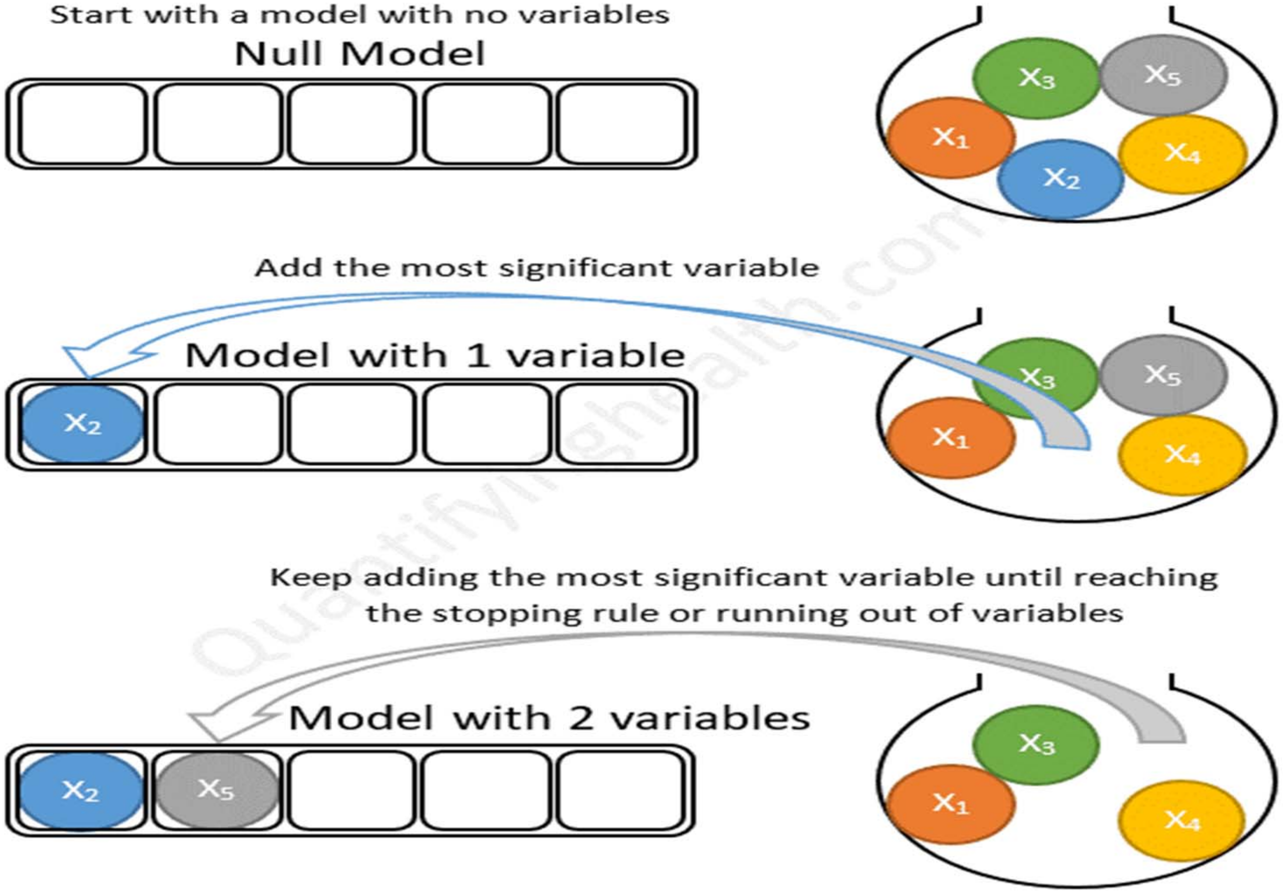
ε_i are i.i.d. with zero mean and a common variance σ^2.

i.e. $E\left(\varepsilon_i\right) = 0$ and $E\left(\varepsilon_i^2\right) = \sigma^2 > 0$.

Forward selection

1. Start with a null model. The null model has no predictors.
2. Fit p simple linear regression models, each with one of the variables in and the intercept. So basically, you just search through all the single – variable models the best one (the one that results in the lowest residual sum of squares).
3. Search through the remaining p minus 1 variables and find out which variable should be added to the current model to best improve the residual sum of squares.
4. Continue until some stopping rule is satisfied, for example when all remaining variables have a p – value above some threshold, ex: 0.1 or 0.15 etc.

Forward stepwise selection example with 5 variables:



Start with a model with no variables

Null Model

Add the most significant variable

Model with 1 variable

Keep adding the most significant variable until reaching the stopping rule or running out of variables

Model with 2 variables

https://quantifyinghealth.com/stepwise-selection/

1. Let M_0 denote the null model. i.e. $\mathrm{M}_0 = \{\ \}$.

2. For $k = 1$, choose the highest correlated and significant variable, $x_{(1)}$, on simple regression model.

$$x_{(1)} = \underset{j \in \{1,\ldots,p\}}{\arg\max} \left(1 - \frac{\sum_{i=1}^{n} \left(y_i - x_{i,j} \hat{\beta} \right)^2}{\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i{}' \hat{\boldsymbol{\beta}} \right)^2 / \left(n - p - 1 \right)} \right)$$

$$= \underset{j \in \{1,\ldots,p\}}{\arg\max} \frac{SSR_{(j)}}{SSE / \left(n - p - 1 \right)}$$

3. For k $= 2, \cdots, K \le p-1$,

$$x_{(k)} = \underset{j \in \{1,\ldots,p\} \backslash \{(1),\ldots,(p-1)\}}{\arg\max} \frac{\left(\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i{'}_{((1),\ldots,(j))} \hat{\boldsymbol{\beta}}_{((1),\ldots,(j))} \right)^2 - \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i{'} \hat{\boldsymbol{\beta}} \right)^2 \right) / (p-k)}{\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i{'} \hat{\boldsymbol{\beta}} \right)^2 / (n-p-1)}$$

$$= \underset{j \in \{1,\ldots,p\} \backslash \{(1),\ldots,(k-1)\}}{\arg\max} \frac{\left(SSE_{((1),\ldots,(j))} - SSE \right) / (p-k)}{SSE / (n-p-1)}$$

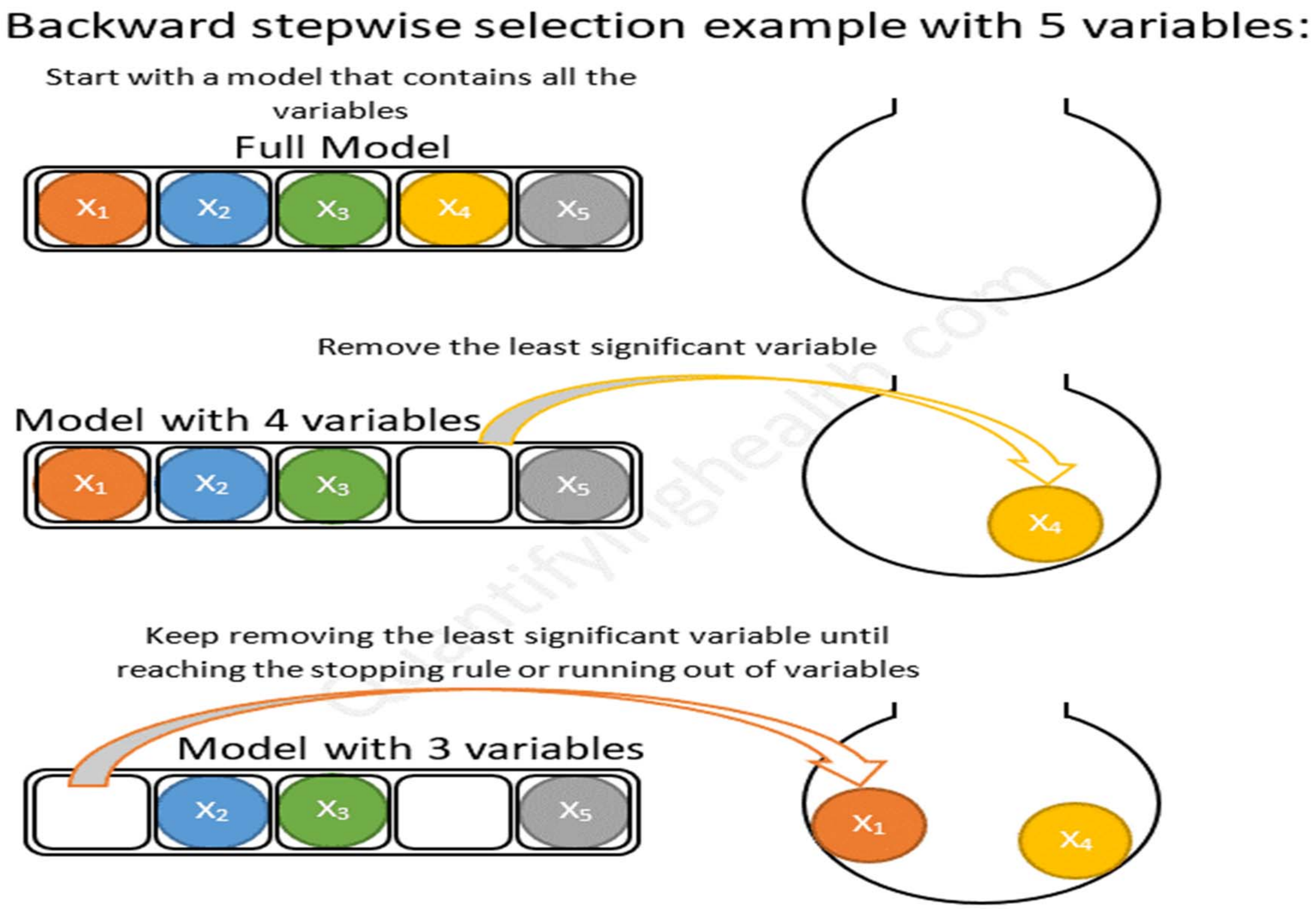
$$= \underset{j \in \{1,\ldots,p\} \backslash \{(1),\ldots,(k-1)\}}{\arg\max} \{\text{partial F test statistics}\}$$

and partial F test statistics is significant.

Thus, $\mathrm{M}_K = \{x_{(1)}, \ldots, x_{(K)}\}$ is the selected model.

Backward selection

1. Start with all variables in the model.
2. Remove the variable with the largest $p-$value. That is, the variable that is the least statistically significant. i.e. the removed variable is with the largest $p-$value.
3. Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant $p-$value defined by some significance threshold.



Backward stepwise selection example with 5 variables:

Start with a model that contains all the variables

Full Model

Remove the least significant variable

Model with 4 variables

Keep removing the least significant variable until reaching the stopping rule or running out of variables

Model with 3 variables

https://quantifyinghealth.com/stepwise-selection/

1. Let M_p denote the full model. i.e. $\mathrm{M}_p = \left\{x_1, ..., x_p\right\}$.

2. For $k = 1$, remove the least significant variable, $x_{(1)}$. The model is $\mathrm{M}_{p-1} = \left\{x_1, ..., x_p\right\} \setminus \left\{x_{(1)}\right\}$.

$$x_{(1)} = \arg\min_{j \in \{1,\ldots,p\}} \left(1 - \frac{\sum_{i=1}^{n} \left(y_i - x_{i,j} \hat{\beta} \right)^2}{\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \right)^2 / \left(n - p - 1 \right)} \right)$$

$$= \arg\min_{j \in \{1,\ldots,p\}} \frac{SSR_{(j)}}{SSE / \left(n - p - 1 \right)}$$

3. For k $= 2, \cdots, K \le p-1$,

$$x_{(k)} = \underset{j \in \{1,\ldots,p\} \setminus \{(1),\ldots,(k-1)\}}{\arg\min} \frac{\left(\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_{i\ ((1),\ldots,(j))}' \hat{\boldsymbol{\beta}}_{((1),\ldots,(j))} \right)^2 - \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \right)^2 \right) / (p-k)}{\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \right)^2 / (n-p-1)}$$

$$= \underset{j \in \{1,\ldots,p\} \setminus \{(1),\ldots,(k-1)\}}{\arg\min} \frac{\left(SSE_{((1),\ldots,(j))} - SSE \right) / (p-k)}{SSE / (n-p-1)}.$$

Thus, $\mathrm{M}_K = \{x_1, \ldots, x_p\} \setminus \{x_{(1)}, \ldots, x_{(p-K)}\}$ is the remaining model.

Should you use forward or backward selection?

- Where forward stepwise is better?
- Where backward stepwise is better?

Where forward stepwise is better?

- Applied in settings where the number of variables under consideration is **larger than** the sample size!

- This is because forward selection starts with a null model and proceeds to add variables one at a time, it DOES NOT have to consider the full model.
- In practical, it will only consider models with number of variables less than:

a. the sample size (for linear regression)

b. the number of events (for logistic regression)

Where backward stepwise is better?

- Starting with the full model has the advantage of considering the effects of all variables **simultaneously**.

- This is especially important in case of collinearity because backward stepwise may be forced to keep them all in the model unlike forward selection where none of them might be entered.
- Unless the number of candidate variables > sample size (or number of events), use a backward stepwise approach.

Stepwise selection

- Start on forward selection
- At every step, backward elimination
- Continue until a stopping rule is reached.
- No significant variable added and no insignificant variable removed.

Stepwise selection on R

null=lm(y~1) ; full = lm(y~X); # X: all variables

- step(null, scope=list(lower=null, upper=full), direction=“forward”)
- step(full, scope=list(upper=full), direction=“backward”)
- step(null, scope=list(upper=full), direction=“both”)

Information criterion

- The Akaike information criterion (AIC, Akaike (1974))
- The Bayesian information criterion (BIC(SBC), Schwarz (1978))
- The Hann-Quinn criterion (HQ)
- AICc, AICu, QAIC, QAICc
- …etc

AIC

- $AIC(J_k) = -2\log\left(\hat{L}_n\right) + 2k$

  $L_n = f(\boldsymbol{y} \mid \boldsymbol{\theta_k})$, likelihood function

  k : the number of variables included on model
- On regression assumptions (normal distribution),

$$AIC(J_k) = n\log\left(\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i(J_k)\right)^2}{n}\right) + 2k$$

$$= n\log\left(\frac{SSE(J_k)}{n}\right) + 2k.$$

BIC

- $BIC(J_k) = -2\log\left(\hat{L}_n\right) + k\log(n)$

  $L_n = f(\boldsymbol{y} \mid \boldsymbol{\theta}_k)$, likelihood function

  k : the number of variables included on model
- On regression assumptions (normal distribution),

$$BIC(J_k) = n\log\left(\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i(J_k)\right)^2}{n}\right) + k\log(n)$$

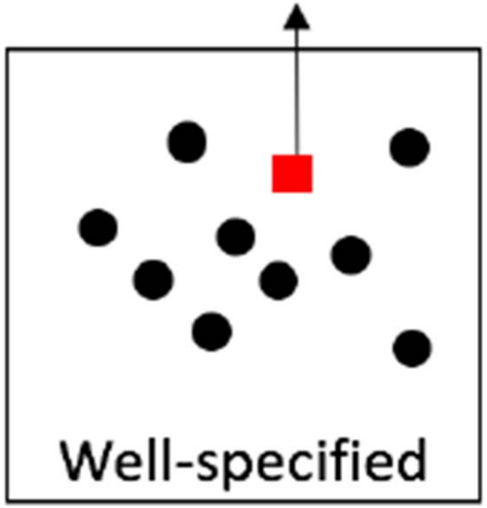
$$= n\log\left(\frac{SSE(J_k)}{n}\right) + k\log(n).$$

Other penalty

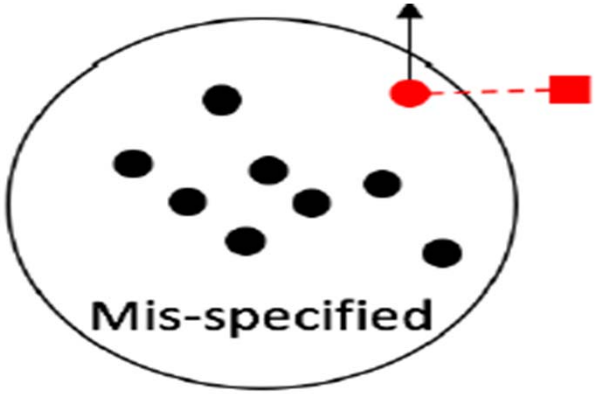
- $AIC_c\left(J_k\right) = AIC + \dfrac{2k^2 + 2k}{n-k-1}$
- $\mathrm{HQ}_c\left(J_k\right) = -2\log\left(\hat{L}_n\right) + 2k\log\left(\log\left(n\right)\right)$
- $\mathrm{Q}AIC_c\left(J_k\right) = \dfrac{-2\log\left(\hat{L}_n\right)}{c} + 2k, c : \mathrm{VIF}$

AIC and BIC

- BIC is **consistent** in selection.

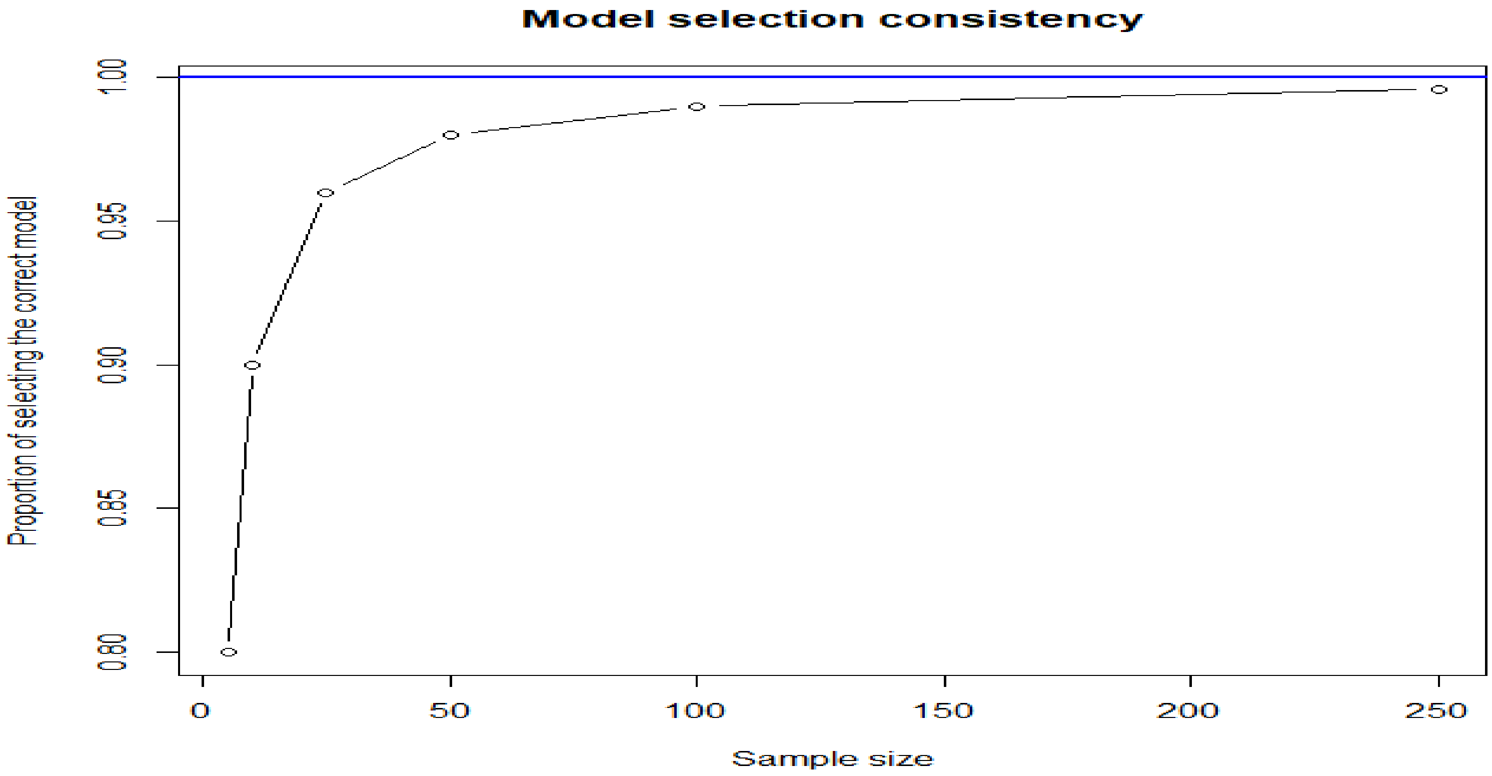


- AIC performs well in an asymptotic **efficiency**.



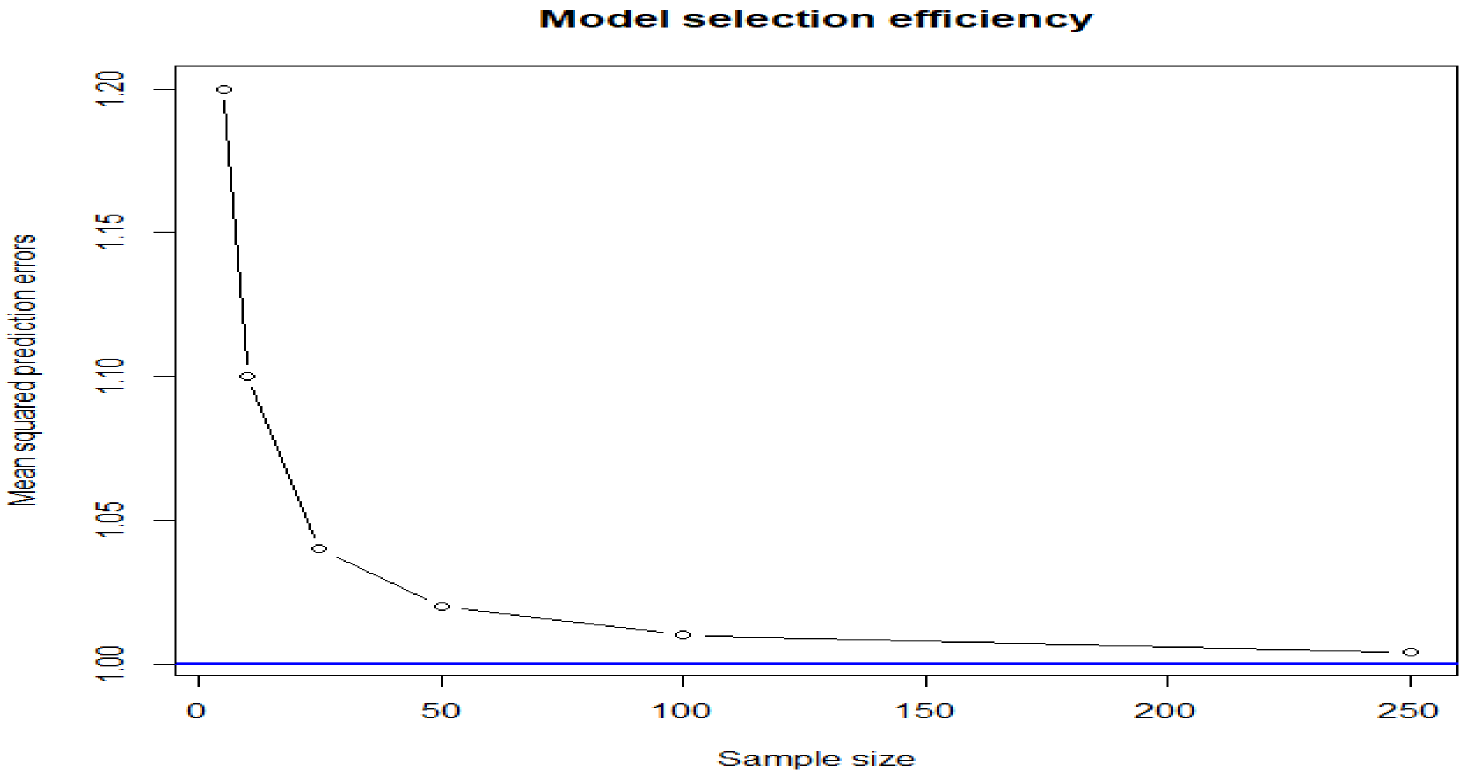
Model selection consistency

- Select all relevant variables and no irrelevant variables, with probability approaching 1



Model selection efficiency

- Select the model that has the lower prediction error. i.e. the model that predicts better at future observations



AIC or BIC?

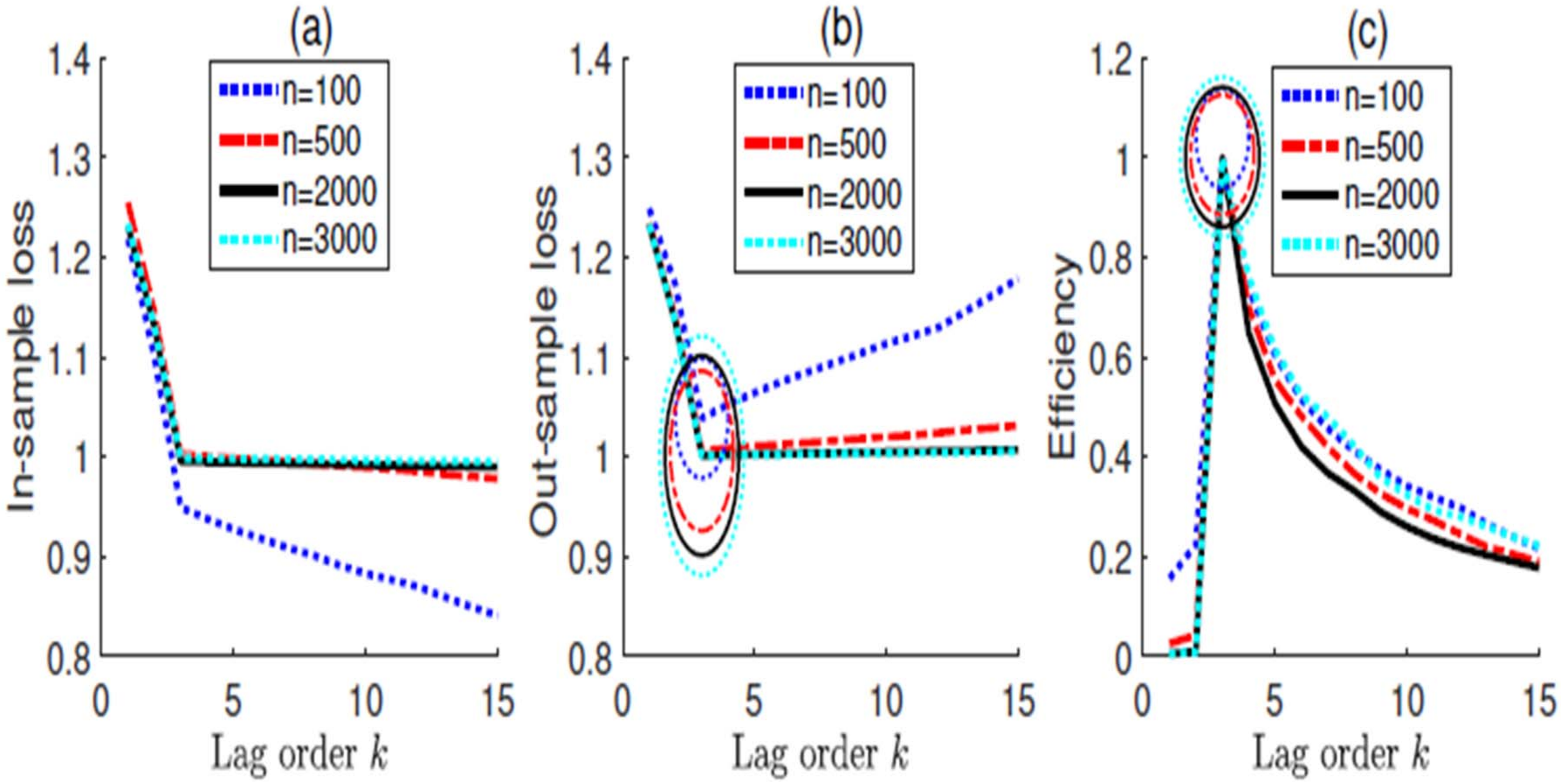


Fig. 1: Parametric framework: the best predictive performance is achieved at the true order 3

AIC or BIC?

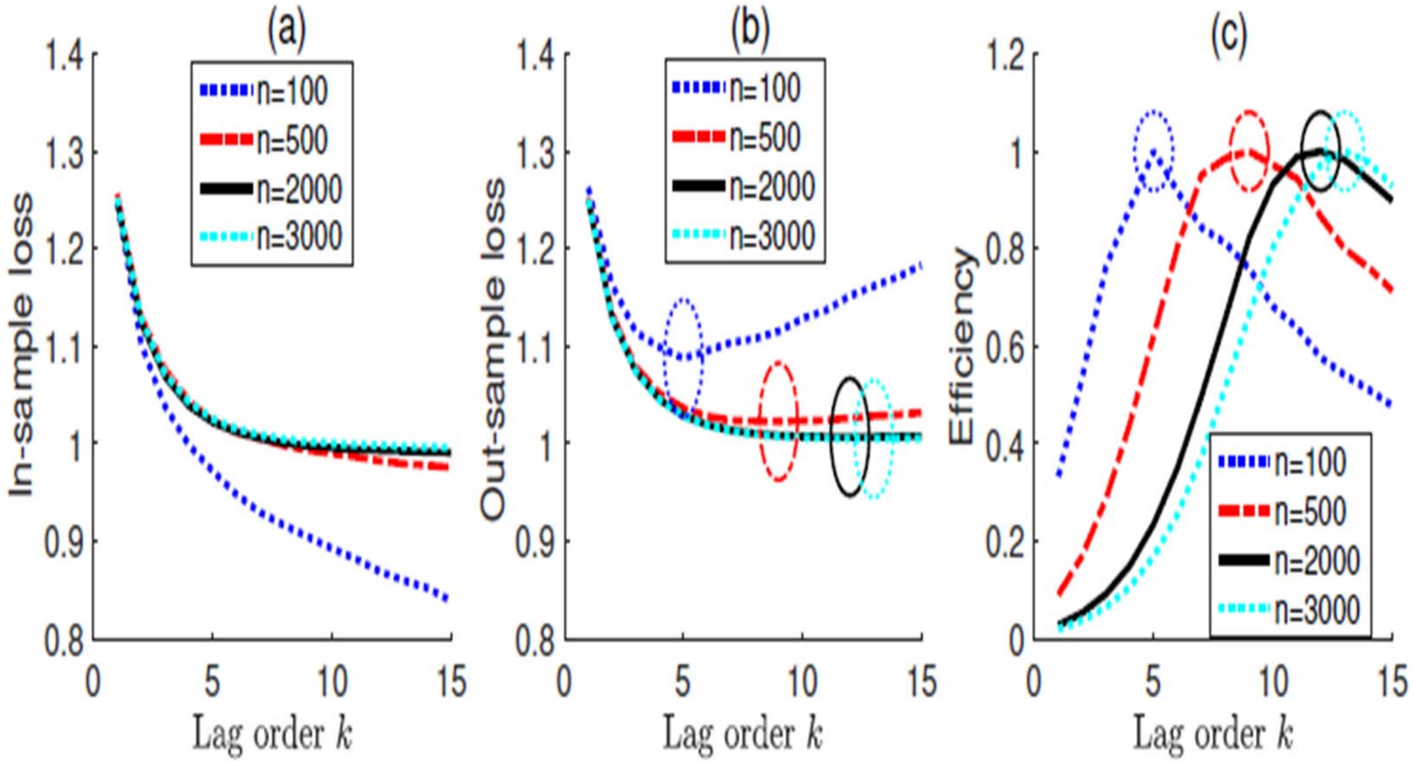


Fig. 2: Nonparametric framework: the best predictive performance is achieved at an order that depends on the sample size

Other criteria

- Mallows's C_p

$$C_p(J_k) = \frac{SSR_k}{MSE_p} - n + 2k$$

note that: equivalent to AIC in the special case of normal linear regression

- PRESS

$$\text{PRESS}(J_k) = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i(i)}\right)^2$$

note that:

$\hat{y}_{i(i)}$: prediction for i-th using all data except i-th observation

Cross Validation (CV)

- Split $\{y_i, \boldsymbol{X}_i, i = 1,\ldots,n\}$ into a training sample $\{y_i, \boldsymbol{x}_i, i \in s^c\}$ and a testing sample $\{y_i, \boldsymbol{X}_i, i \in s\}$, where $s \cup s^c = \{1,\ldots,n\}$ and $s \cap s^c = \varnothing$. Denote $\#(s^c) = n_c$ and $\#(s) = n_v$.
- For a given candidate J_α, its CV value is given by

$$\frac{1}{n_v \binom{n}{n_v}} \sum_{\text{all}(s,s^c)\text{combinations}} \sum_{i \in s} \left(y_i - \hat{y}_{i,s^c,J_\alpha} \right)^2$$

where $\hat{y}_{i,s^c,J_\alpha} = \boldsymbol{X}_{i,J_\alpha} \left(\boldsymbol{X}^T_{s^c,J_\alpha} \boldsymbol{X}_{s^c,J_\alpha} \right)^{-1} \boldsymbol{X}^T_{s^c,J_\alpha} y_{s^c}$.

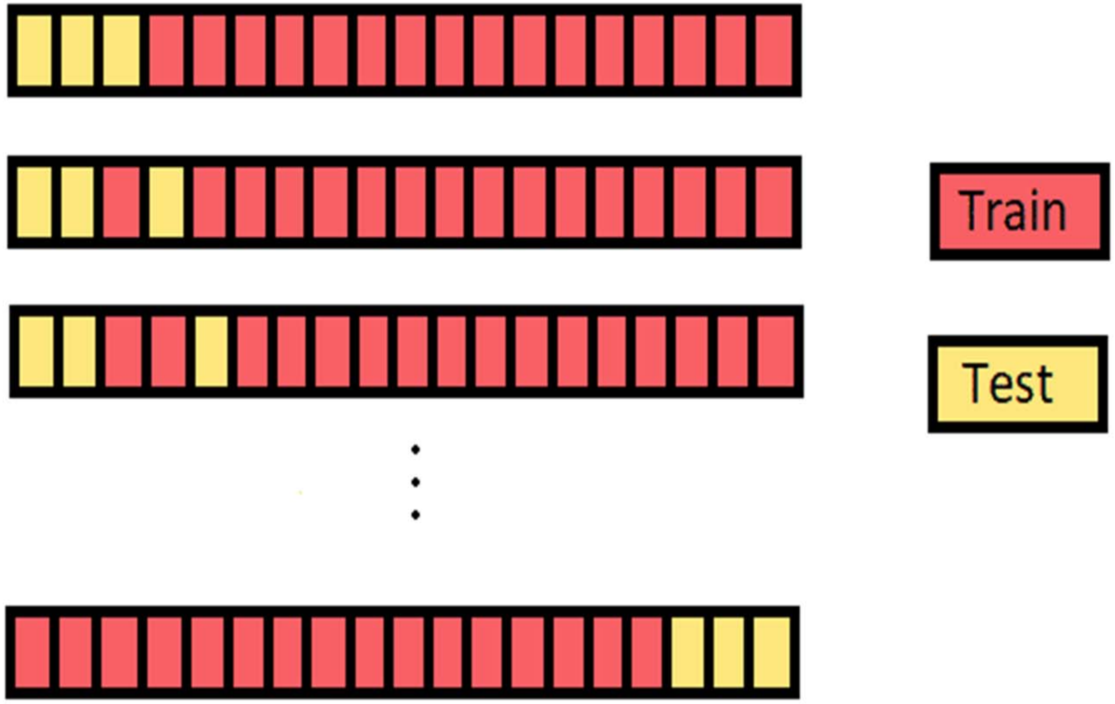
$\hat{y}_{i(i)}$: prediction for i-th using all data except i-th observation

- On specific model, evaluate Training => Testing sample => average squared prediction error

- (1) Fitting a model using training data, not all sample; (2) validating the fitted model using testing data

- The selected model is with the minimum average squared prediction error.

All (training, testing) combinations?



$$n_v = 0.3n, \binom{n}{n_v} = ?$$

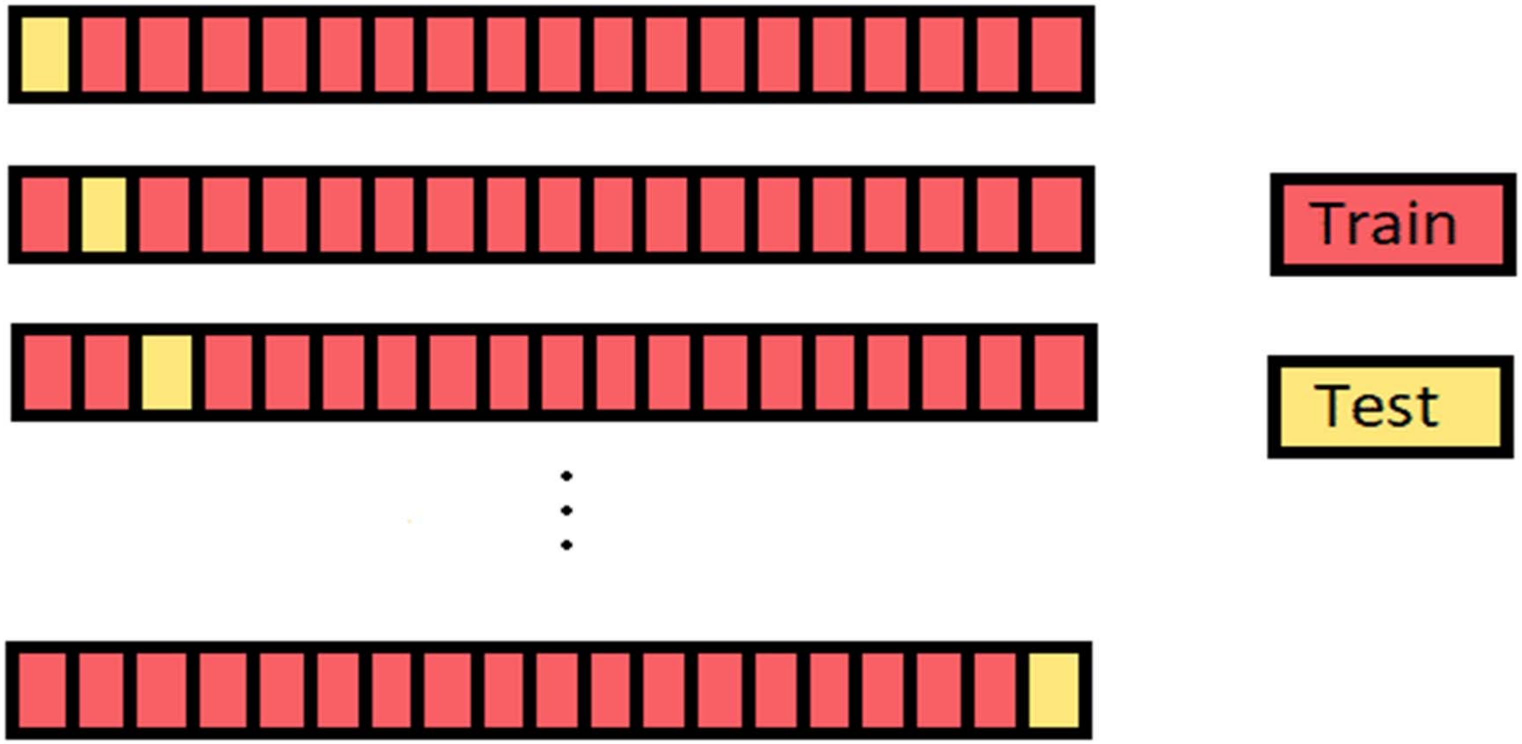
$$\binom{50}{15} \doteq 2.25\times10^{12}$$

$$\binom{100}{30} \doteq 2.93\times10^{25}$$

Infeasible!

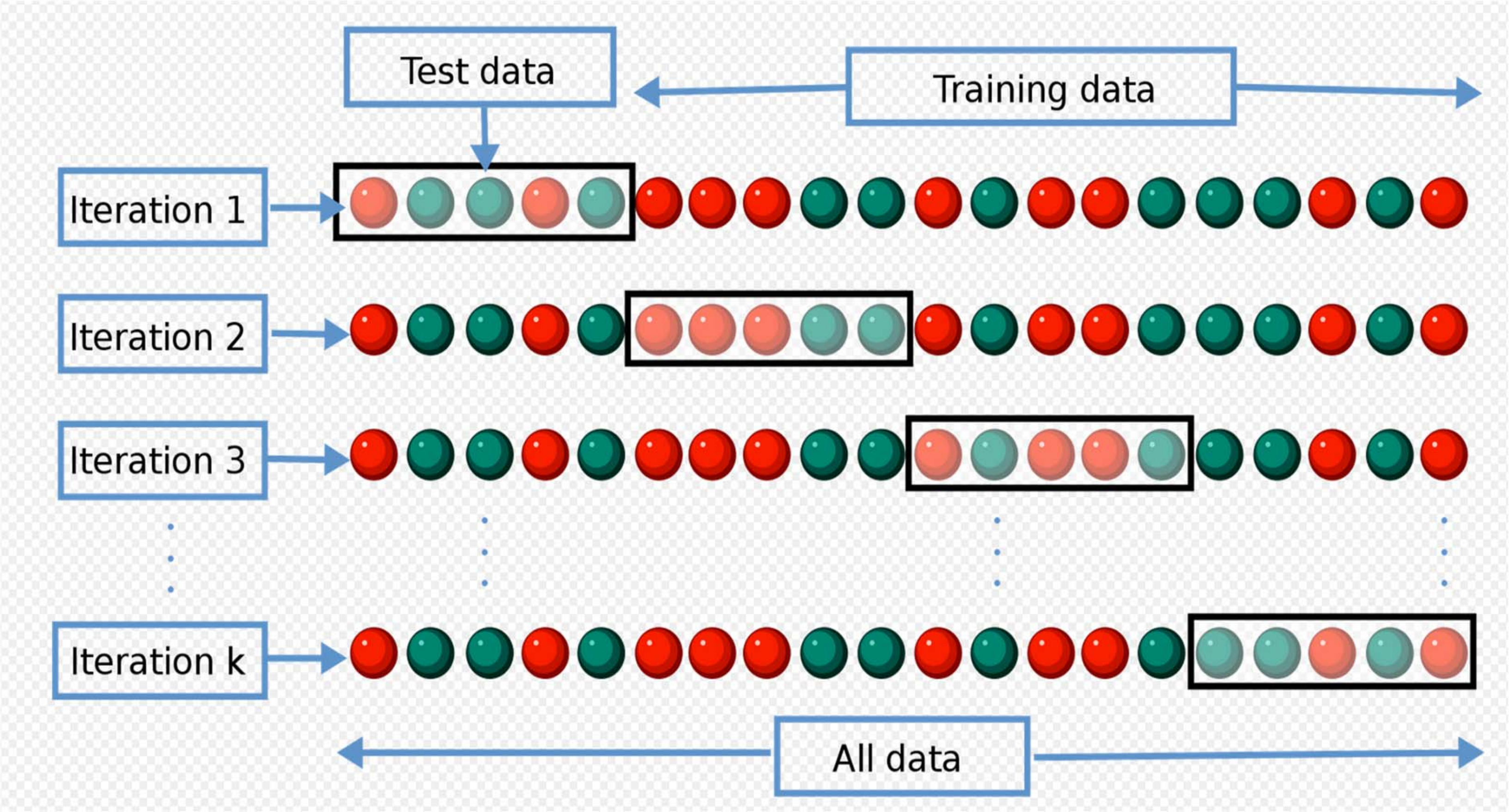
https://aiaspirant.com/cross-validation/

Leave-one-out CV

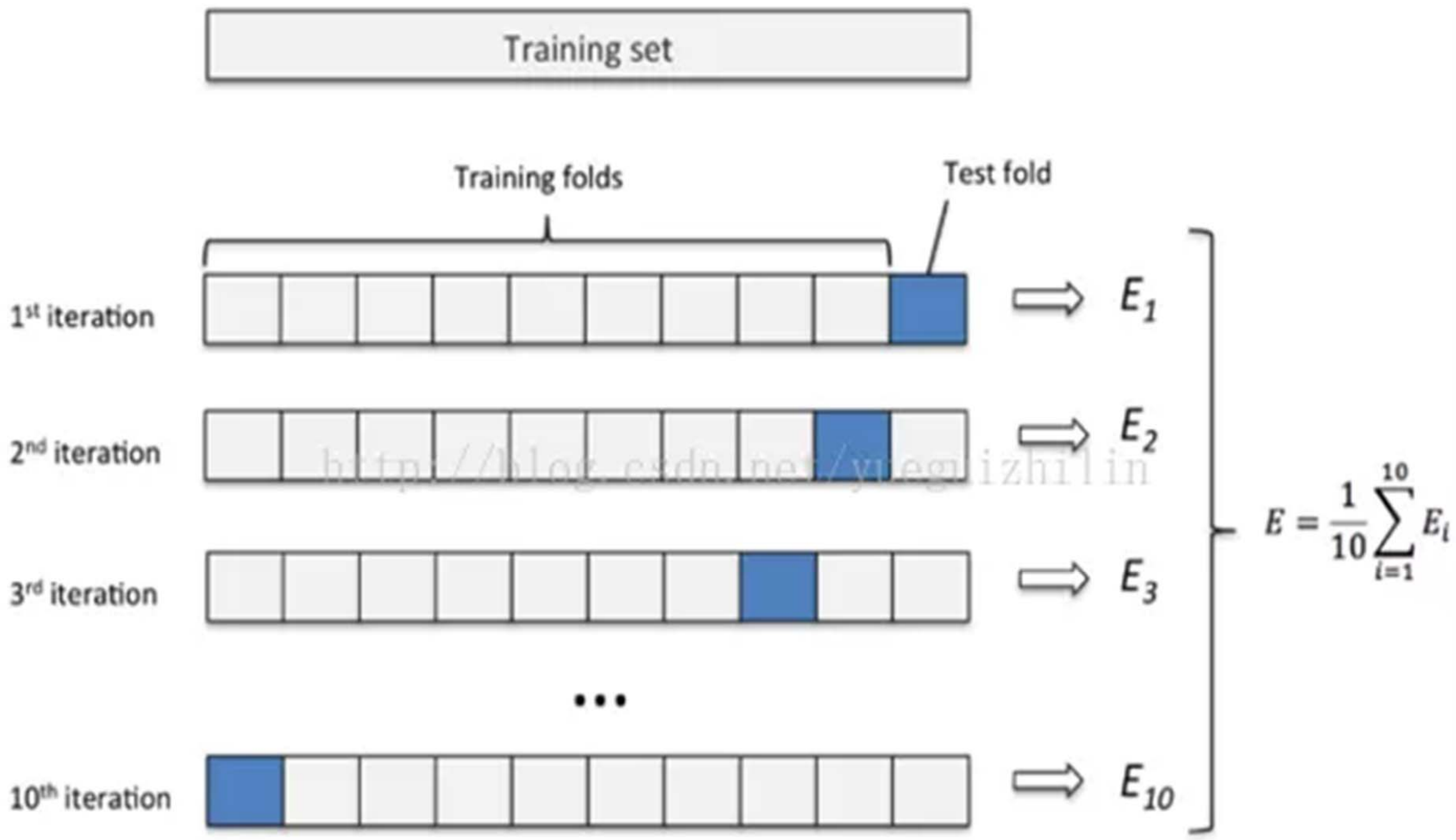


https://aiaspirant.com/cross-validation/

k-folds CV



https://en.wikipedia.org/wiki/Cross-validation_(statistics)



https://www.itread01.com/content/1547033769.html

$$J_{\mathrm{all}(s,s^c)\mathrm{combinations}} = J_{\mathrm{Leave-one-out\ CV}}?$$

$$\mathrm{or} = J_{\mathrm{k-folds\ CV}}?$$

Some facts about CV

- When $n_v = 1$, we call the CV as delete-1 CV, whose performance is asymptotically equivalent to AIC (Stone, 1977).

  i.e. $J_{\text{Leave-one-out CV}} \sim J_{AIC}$

- For delete-n_v CV with $n_v > 1$, Shao (1993) showed that subset selection based on CV is consistent, provided $\dfrac{n_v}{n} \to 1$ and $n_c \to \infty$.

  i.e. $J_{\text{Leave-}n_v\text{-out CV}} = J_{BIC}$

Some facts about CV

- A Monte Carlo CV (MCCV): Shao (1993) suggested a MCCV

$$\hat{\Gamma}_{J_\alpha,n}^{MCCV} = \frac{1}{n_v b}\sum_{s\in R}\left\|\boldsymbol{y}_s - \hat{\boldsymbol{y}}_{J_\alpha,s^c}\right\|^2,$$

where R is a collection of randomly drawing b subsets of 1, . . . , n

that have size n_v. By assuming $\frac{n_v}{n} \to 1$ and $\frac{n^2}{bn_c^2} \to 0$, he

showed that MCCV is consistent.

Least absolute shrinkage and selection operator (Lasso)

- Shrinkage method like ridge, with subtle but important differences, L-2 ridge penalty is replaced by the L-1 lasso penalty

- Lasso estimate:

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

- The equivalent Lagrangian form:

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_{lasso} \sum_{j=1}^{p} |\beta_j| \right\} .$$

Lasso

- Ridge regression does a proportional shrinkage
- “Soft Thresholding”: used in the context of wavelet-based smoothing

  Lasso translates each coefficient by a constant factor , truncating at zero
- “Hard Thresholding”

  Best-subset selection drops all variables with coefficients smaller than the *M*-th largest

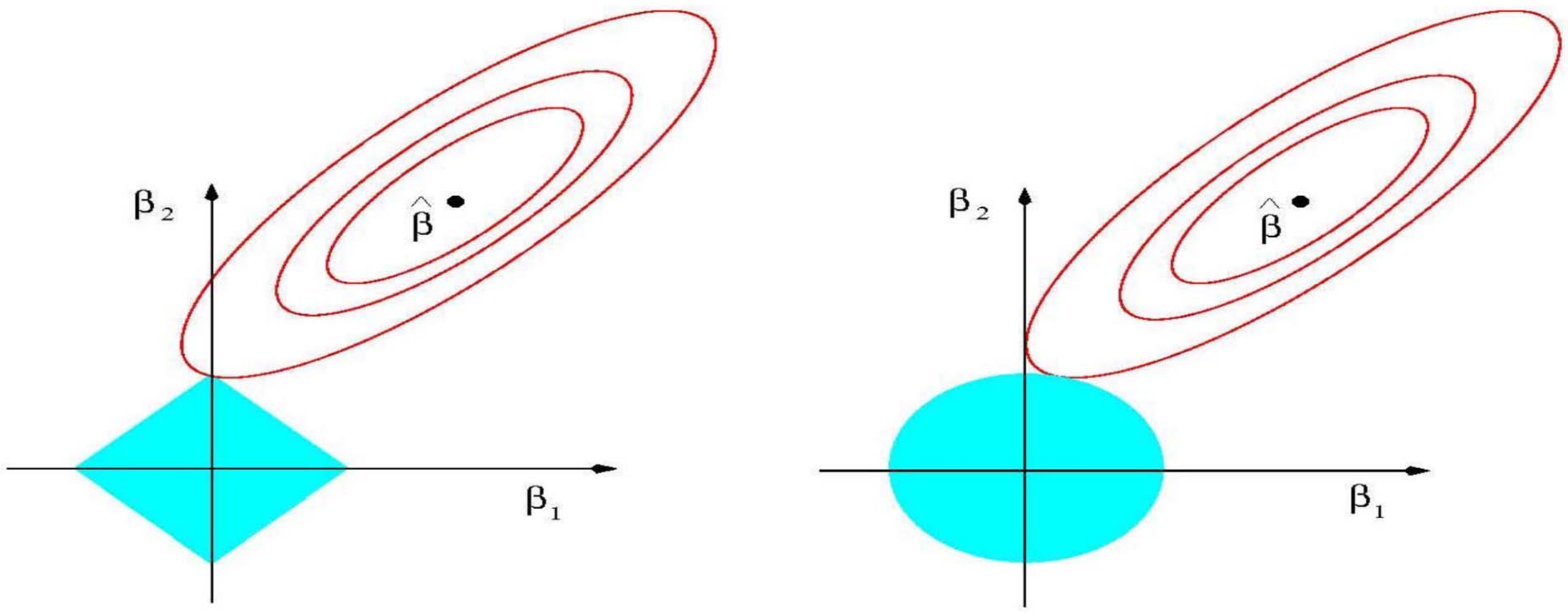


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

https://stats.stackexchange.com/questions/348308/graphical-interpretation-of-lasso

Lasso on R

- package(glmnet)
- glmnet(x, y, alpha =1, family=“gaussian”)

# alpha = 0: ridge regression

- cv.glmnet(x, y, alpha =1, family=“gaussian”)$lambda.min

References

- C.-K. Ing (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series, *Annals of Statistics*, **35**, 1238-1277.

- W. Liu and Y.-H. Yang (2011). Parametric or Nonparametric? A Parametricness Index for Model Selection. *Annals of Statistics*, **39**, 2074-2102.

- Y.L. Zhang and Y.-H. Yang (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, **187**, 95-112.

- J. Ding, V. Tarokh, Y.-H. Yang (2018). Model Selection Techniques: An Overview, *IEEE Signal Processing Magazine*, November, 1-19.

High-dimensional data with time dependent error

- C.-K Ing and T. L. Lai (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models, *Statistica Sinica,* **21**, 1473-1513.

- H.-L. Hsu, C.-K. Ing and H. Tong (2019). On model selection from a finite family of possibly misspecified time series models, *Annals of Statistics*, **47**, 1061-1087.

- H.-T. Chiou, M. Guo and C.-K. Ing (2020). Variable Selection for High-Dimensional Regression Models with Time Series and Heteroscedastic Errors, *Journal of Econometrics*, **216**, 118-136.