# SL = MODEL SELECTION

In the supervised learning context, the prediction task consists in guessing the unknown label $y$ associated with a feature point $x$, based on a learning sample $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where each $(X_i, Y_i)$ is iid and distributed like a generic $(X, Y) \sim \mathbb{P}_{X,Y}$. The goal is to construct a function $f_n$ based on $\mathcal{L}_n$, where $f_n(X)$ stands as a guess for the unknown label $Y$. The predictive performance of $f_n$ is measured by means of a loss function $\ell$, such that $\ell(Y, f_n(X))$ incurs a cost for predicting $Y$ using $f_n(X)$. The risk of $f_n$ is defined as $R(f_n) = \mathbb{E}\{\ell(Y, f_n(X)) \mid \mathcal{L}_n\}$, and one usually is interested in evaluating the expected risk $\mathbb{E}_{\mathcal{L}_n} R(f_n)$. We refer the reader to the chapter SL=FOUNDATIONS for further information regarding loss functions, risk of a prediction rule, and related concepts.

A natural candidate to estimate $\mathbb{E}_{\mathcal{L}_n} R(f_n)$ in practice is the empirical risk

$$\hat{R}_n(f_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_n(X_i)).$$

↑ aka the TRAINING ERROR

Unfortunately, $\hat{R}_n(f_n)$ is a poor estimator of the true risk of $f_n$, since $\mathcal{L}_n$ is used twice: once for training the model, and once to evaluate its performance.

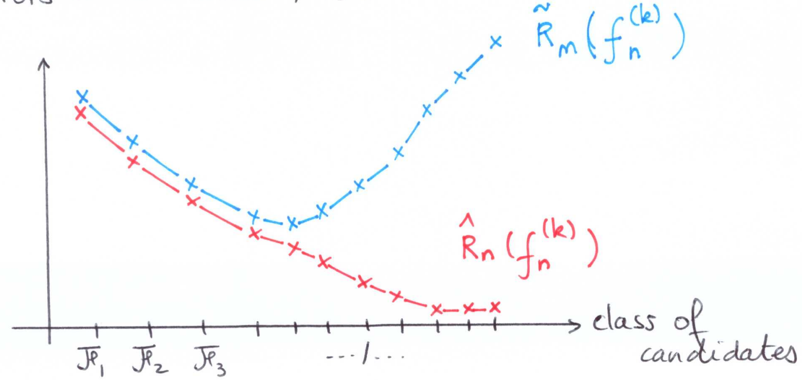
In the presence of large datasets, you may test the performance of $f_n$ on an additional independent test sample $\{(x_{n+1}, y_{n+1}), \dots, (x_{n+m}, y_{n+m})\}$, and use

$$\tilde{R}_m(f_n) := \frac{1}{m} \sum_{i=1}^{m} \ell(y_{n+i}, f_n(x_{n+i}))$$

as an estimator of $R(f_n)$.

↑ aka the TEST ERROR

The prediction rule is generally restricted to a class of candidate functions $\mathcal{F}$ [such as the class of linear predictors, of polynomials, ...]. We consider a family of classes $\{\mathcal{F}_k\}_{k \geq 1}$ of increasing order of "complexity" [we refrain from giving a precise definition of complexity]. For each $\mathcal{F}_k$, denote by $f_n^{(k)} \in \mathcal{F}_k$ the "best" element in $\mathcal{F}_k$, constructed from $\mathcal{L}_n$. Typically, the training and test errors behave like this:



With enough data, the dataset can be divided into 3 parts:

TRAINING 50%
Fit the models here

VALIDATION 25%
Compare the performance of each model, & Pick the one with the best predictive performance

TEST 25%
the test error of the selected model is computed here

With insufficient data, we may approximate the validation step by testing the model on a hold-out subset of the learning sample.

Ex: Cross-validation. ("Sample re-use" method)

Alternatively, we can correct the training error, and add an extra term which takes into account the class complexity, so that

$$\boxed{\hat{R}_n(f_n) + \text{Extra term} \approx \text{Test Error}}$$

Increases, as the class complexity increases.

↳ Recall the discussion on pages 45/46/47 in SL: FOUNDATIONS: we compare there the training error

$$\hat{R}_n(f_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_n(x_i)),$$

with the "in-sample" error

keeping $x_1, .., x_n$ fixed

$$\bar{R}_n(f_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i'}\{\ell(Y_i', f_n(x_i)) \mid \mathcal{L}_n\}.$$

where $Y_i' \mid X = x \overset{d}{=} Y_i \mid X = x$ iid

The "in-sample" error represents the true loss we should be expecting for the observed values $x_1, .., x_n$. The expected difference between $\bar{R}_n(f_n)$ and $\hat{R}_n(f_n)$ represents our optimism; that is by how much we are underevaluating the risk of $f_n$ when re-using the values $y_1, .., y_n$ a second time. In the context of a square loss, we show there that

$$\boxed{\mathbb{E}_{\mathcal{L}_n}\{\hat{R}_n(f_n)\} + \frac{2}{n}\sum_{i=1}^{n} \mathrm{cov}(Y_i, f_n(x_i)) = \mathbb{E}_{\mathcal{L}_n}\{\bar{R}_n(f_n)\}}$$

≈ what we have (training error)

correction ≡ the harder we fit the data (i.e. the more complex $\mathcal{H}$ is), the larger $f_n(x_i)$ correlates with $y_i$; increasing our optimism.

what we want (in-sample error)

In the context of ridge regression, lasso, and smoothing splines, the expression $\sum_{i=1}^{n} \mathrm{cov}(Y_i, f_n(x_i))$ was taken as a definition of the "effective degree of freedom" of the model, see SL: RR & LASSO & SL: SPLINES & SS. For example, it is shown there (page 19) that for the ridge estimator, $\sum_{i=1}^{n} \mathrm{cov}(Y_i, f_n(x_i)) = \sigma^2 \mathrm{Tr}(H_\lambda)$, where $H_\lambda = X(X^tX + \lambda I)^{-1}X^t$. In particular, for $\lambda = 0$ (no regularization), $\mathrm{Tr}(H_\lambda) = d$ = number of parameters to estimate in the linear regression model (with no intercept). The correction term reduces to $\frac{2\sigma^2 d}{n}$, and we consider the adjusted training error to be

$$\boxed{\hat{R}_n(f_n) + \frac{2\hat{\sigma}^2 d}{n}\,,}$$

increases with d,
decreases with n.

where $\hat{\sigma}^2$ is a consistent estimator of the noise variance in the linear model $y = X\beta + \varepsilon$. This coefficient is known as MALLOW'S $C_p$ in the litterature.

In this chapter, we discuss sample re-use techniques (such as cross validation) and analytical methods (such as Mallow's $C_p$, AIC, BIC) to approximate the validation step, and address the issue of model selection.

## I_ ANALYTICAL METHODS

### I.1. Akaike Information Criterion (AIC)

Akaike Information Criteria (AIC) is regarded as one of the first model selection criteria. It was introduced by Akaike in 1973 in his paper " Information Theory and Extension of the Maximum Likelihood Principle".

↳ Model estimation & selection at the same time

The model dimension is unknown and must be determined from the data.

Suppose that the data $\underline{z} = \overset{\text{iid}}{(z_1, \dots, z_n)}$ is generated from the true unknown density $f(\underline{z} \mid \theta_0)$ [$\theta_0$ = "true" parameter]

(In the SL context, $z_i$ are pairs of observations $(x_i, y_i)$).

Consider the class of candidates

$\mathcal{F}_k := \{ f(z \mid \theta_k),\ \theta_k \in \Theta_k \subset \mathbb{R} \}$, parametrized by the $k$-dimensional parameter $\theta_k$.

Let $\hat{\theta}_k = \hat{\theta}_k(\underline{z})$ be the MLE of $\theta$ based on observations $\underline{z}$.

We consider a finite collection of classes $\mathcal{F} = \{\mathcal{F}_{k_1}, \dots, \mathcal{F}_{k_L}\}$, assuming that each candidate and fitted model are distinguished by their dimension $k$; so that the model selection problem is equivalent to dimension determination.

We use the Kullback-Leibler (KL) divergence to evaluate how "far" the candidate model $f(\underline{z} \mid \theta_k)$ is from the true density $f(\underline{z} \mid \theta_0)$ [see chapter MS = MAXIMUM LIKELIHOOD ESTIMATION]

$$KL(\underset{f(\underline{z}\mid\theta_0)}{f} \,\|\, \underset{f(\underline{z}\mid\theta_k)}{f_k}) = \int f(z) \log\left(\frac{f(z)}{f_k(z)}\right) dz$$

$$= \underbrace{\int f(z) \log f(z)}_{\text{independent of } k} - \underbrace{\int f(z) \log f_k(z)\, dz}_{\text{" cross-entropy "}}$$

⇒ Minimizing $KL(f \| f_k)$ is equivalent to minimizing

$$\boxed{\begin{aligned} d(\theta_k) &:= -2 \int f(\underline{z} \mid \theta_0) \log\{f(\underline{z} \mid \theta_k)\}\, d\underline{z} \\ &= -2\, \mathbb{E}\{\log f(\underline{z} \mid \theta_k)\} \end{aligned}}$$

The function $d(\theta_k)$ is called the KULLBACH DISCREPANCY.

a random variable

↳ The measure $\boxed{d(\hat{\theta}_k) = -2\,\mathbb{E}\{\log f(\underline{z}|\theta)\}\big|_{\theta=\hat{\theta}_k}}$ reflects the amount of separation between the true generating model, and the fitted model $f(\underline{z}|\hat{\theta}_k)$. In practice, $d(\hat{\theta}_k)$ cannot be computed, as it requires the knowledge of the true density.

⇒ Akaike suggested to use the quantity $-2\log f(\underline{z}|\hat{\theta}_k(\underline{z}))$ as a (biased) estimator of $d(\hat{\theta}_k)$.

The bias adjustment is computed by considering the difference

$$(*) \quad \mathbb{E}\{d(\hat{\theta}_k)\} - \mathbb{E}\{-2\log f(\underline{z}|\hat{\theta}_k(\underline{z}))\}$$

Expected Kullbach discrepancy.
It can be rewritten

$$\mathbb{E}\{d(\hat{\theta}_k)\} = -2\,\mathbb{E}_{\underline{z}'}\mathbb{E}_{\underline{z}}\{\log f(\underline{z}|\hat{\theta}_k(\underline{z}'))\},$$

where $\underline{z}' = (z_1',\dots,z_n')$ is an independent sample, with the same distribution as $(z_1,\dots,z_n)$.

↳ compare with our estimator: the same sample is re-used twice.

We decompose (*) into 3 parts:

$$\begin{aligned}\mathbb{E}\{d(\hat{\theta}_k)\} = {} & \mathbb{E}\{-2\log f(\underline{z}|\hat{\theta}_k(\underline{z}))\} \\ & + \left[\mathbb{E}\{-2\log f(\underline{z}|\theta_0)\} - \mathbb{E}\{-2\log f(\underline{z}|\hat{\theta}_k(\underline{z}))\}\right] \\ & + \left[\mathbb{E}\{d(\hat{\theta}_k)\} - \mathbb{E}\{-2\log f(\underline{z}|\theta_0)\}\right]\end{aligned}$$

We evaluate the size of these terms

• Term $\mathbb{E}\{-2\log f(\underline{z}|\theta_0)\} - \mathbb{E}\{-2\log f(\underline{z}|\hat{\theta}_k(\underline{z}))\}$.

Consider the function $\varphi(\theta_0) := \log f(\underline{z}|\theta_0)$, and a second-order Taylor expansion around $\hat{\theta}_k$:

$$\varphi(\theta_0) \approx \varphi(\hat{\theta}_k) + [\nabla\varphi(\hat{\theta}_k)]^t(\theta_0-\hat{\theta}_k) + \frac{1}{2}(\theta_0-\hat{\theta}_k)^t H_\varphi(\hat{\theta}_k)(\theta_0-\hat{\theta}_k), \qquad (**)$$

We omit technical details

$=0$ since $\hat{\theta}_k$ is the MLE

with • $\nabla\varphi = \left(\frac{\partial\varphi}{\partial\theta_k^1},\dots,\frac{\partial\varphi}{\partial\theta_k^k}\right)^t$ = Gradient of $\varphi$

$\theta_k = (\theta_k^1,\dots,\theta_k^k) \in \Theta_k \subset \mathbb{R}^k$.

• $H_\varphi = \left(\frac{\partial^2\varphi}{\partial\theta_k^j\,\partial\theta_k^l}\right)_{j,l}$ = Hessian.

We introduce the observed Fisher information matrix (for n observations):

$$I(\hat{\theta}_k, \underline{z}) = -\left(\frac{\partial^2 \log f(\underline{z}|\theta_k)}{\partial\theta_k\,\partial\theta_k^t}\right)\Bigg|_{\theta_k=\hat{\theta}_k}$$

Taking $\mathbb{E}\{\cdots\}$ on both sides of (**) and multiplying by $-2$:

$$\boxed{\begin{aligned}&\mathbb{E}\{-2\log f(\underline{z}|\theta_0)\} - \mathbb{E}\{-2\log f(\underline{z}|\hat{\theta}_k(\underline{z}))\} \\ (1) \qquad & \approx \mathbb{E}\{(\hat{\theta}_k-\theta_0)^t I(\hat{\theta}_k,\underline{z})(\hat{\theta}_k-\theta_0)\}\end{aligned}}$$

Note that for (**) to hold, we need to assume that the true model $f(z|\theta_0)$ is a member of $\mathcal{F}_k$, so that $\theta_0 \in \Theta_k$ = the fitted model $f(z|\hat{\theta}_k)$ is either correctly specified or overfitted.

↳ Is this a problematic assumption?

- Term $\mathbb{E}\{d(\hat{\theta}_k)\} - \mathbb{E}\{-2\log f(\underline{z}\,|\,\theta_0)\}$

We proceed as before : consider a second-order Taylor expansion of $d(\hat{\theta}_k)$ around $\theta_0$.

Put, $\psi(\theta_0) := \mathbb{E}\{\log f(\underline{z}\,|\,\theta_0)\}$

. Expected Fisher Information Matrix :

$$I(\theta_k) := -\mathbb{E}\left\{\frac{\partial^2 \log f(\underline{z}\,|\,\theta_k)}{\partial\theta_k\,\partial\theta_k^t}\right\}$$

Then

$$\psi(\hat{\theta}_k) \approx \psi(\theta_0) \underset{\uparrow}{+} \frac{1}{2}(\theta_0-\hat{\theta}_k)^t \underset{\uparrow}{H_\psi(\theta_0)}(\theta_0-\hat{\theta}_k)$$

Gradient of $\psi$ vanishes at $\theta_0$.

Hessian Matrix of $\psi$.

(x2) $\downarrow$

$$\underbrace{\mathbb{E}\{-2\log f(\underline{z}\,|\,\hat{\theta}_k(\underline{z}))\}}_{= \, d(\hat{\theta}_k)} - \mathbb{E}\{-2\log f(\underline{z}\,|\,\theta_0)\} \approx (\hat{\theta}_k(\underline{z})-\theta_0)^t\, I(\theta_0)\,(\hat{\theta}_k(\underline{z})-\theta_0)$$

$\mathbb{E}(\ldots)$ $\downarrow$

$$\boxed{\begin{aligned}&\mathbb{E}\{d(\hat{\theta}_k)\} - \mathbb{E}\{-2\log f(\underline{z}\,|\,\theta_0)\}\\ &\qquad\simeq \mathbb{E}\{(\hat{\theta}_k-\theta_0)^t\, I(\theta_0)\,(\hat{\theta}_k-\theta_0)\}.\end{aligned}\quad (2)}$$

Combining (1) and (2) together, we get :

$$\begin{aligned}\mathbb{E}\{d(\hat{\theta}_k)\} \approx\ & \mathbb{E}\{-2\log f(\underline{z}\,|\,\hat{\theta}_k(\underline{z}))\}\\ &+ \mathbb{E}\{(\hat{\theta}_k-\theta_0)^t\, I(\hat{\theta}_k,\underline{z})\,(\hat{\theta}_k-\theta_0)\}\\ &+ \mathbb{E}\{(\hat{\theta}_k-\theta_0)^t\, I(\theta_0)\,(\hat{\theta}_k-\theta_0)\}\end{aligned}$$

Recall that under regularity assumptions on the true density,

$$n^{1/2}(\hat{\theta}_k-\theta_0)\,\underset{\uparrow}{I_1^{1/2}}(\theta_0) \xrightarrow{d} \mathcal{N}(0, I)$$

Fisher information matrix for one observation.

(see chapter MS = MAXIMUM LIKELIHOOD ESTIMATION)

Thus $n(\hat{\theta}_k-\theta_0)^t I(\theta_0)(\hat{\theta}_k-\theta_0) \underset{\uparrow}{\xrightarrow{d}} \chi^2(k)$

Convergence to $\chi^2(k)$ still holds if we replace $I(\theta_0)$ with a consistent estimator

In addition, under regularity conditions, convergence of moments are ensured as well. Since $\mathbb{E}\,\chi^2(k) = k$, we obtain:

$$\mathbb{E}\{(\hat{\theta}_k-\theta_0)^t I(\theta_0)(\hat{\theta}_k-\theta_0)\} \approx k$$

& $$\mathbb{E}\{(\hat{\theta}_k-\theta_0)^t I(\hat{\theta}_k,\underline{z})(\hat{\theta}_k-\theta_0)\} \approx k,$$

so that

$$\mathbb{E}\{d(\hat{\theta}_k)\} \approx \mathbb{E}\{-2\log f(\underline{z}\,|\,\hat{\theta}_k(\underline{z}))\} + 2k,$$

which motivates the definition of the AIC criterion :

$$\boxed{\mathrm{AIC} := -2\log \underset{\uparrow}{f}(\underline{z}\,|\,\hat{\theta}_k(\underline{z})) + \underset{\uparrow}{2k}}$$

$\leftarrow$ n large + $\nwarrow$ regularity conditions

goodness of fit

penalty

$\nearrow$ the application of the criterion does not require that the true model $\in \mathcal{F}_k$.

Remark : link with Mallow's $C_p$ coefficient introduced page 5, in the context of a linear regression model.

Observations $z_i$ corresponds to $y_i$ (given $x_i$), with $y_i\,|\,x_i \sim \mathcal{N}(x_i^t\beta_0, \sigma_0^2)$, iid. Denote by $(\hat{\beta}_k, \hat{\sigma}_k^2)$ the

MLEs of $(\beta_0, \sigma^2)$ in $\mathcal{F}_k$ (the true model $\in \mathcal{F}_k$)

Then

$$\log \underbrace{f(\underline{y} \mid \underline{x}, \hat{\beta}_k, \hat{\sigma}_k^2)}_{=:\, \hat{f}_k} = -\frac{n}{2} \log(2\pi \hat{\sigma}_k^2) - \frac{1}{2\hat{\sigma}_k^2} \underbrace{\sum_{i=1}^{n} (y_i - x_i^t \hat{\beta}_k)^2}_{\overset{\shortparallel}{n \hat{\sigma}_k^2}}$$

$\Rightarrow$ The goodness-of-fit term in the expression of the AIC is thus equal to $n(1 + \log 2\pi) + n \log \hat{\sigma}_k^2$. To compare the AIC coefficient with Mallow's $C_p = \hat{R}_n(f_n) + \frac{2\hat{\sigma}^2 d}{n}$, we consider $\hat{\sigma}^2$ = a consistent estimator of $\sigma^2$ [e.g. the one obtained when considering the largest model].

Minimizing $AIC = n(1 + \log 2\pi) + n \log \hat{\sigma}_k^2$ is thus equivalent to selecting the model minimizing

$$\underbrace{n(1 + \log 2\pi)}_{\text{remove}} + n \log \hat{\sigma}_k^2 \underbrace{- n \log \hat{\sigma}^2}_{\text{independent of } k} + 2k$$

$$\Leftrightarrow \text{minimizing} \quad n \log \frac{\hat{\sigma}_k^2}{\hat{\sigma}^2} + 2k \overset{(\bullet)}{\approx} n\left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}^2} - 1\right) + 2k$$

[expansion of the log around 1]

$$\approx \text{equivalent to minimizing} \quad n \frac{\hat{\sigma}_k^2}{\hat{\sigma}^2} + 2k = \frac{1}{\hat{\sigma}^2} \underbrace{\sum_{i=1}^{n} (y_i - x_i^t \hat{\beta}_k)^2}_{n \hat{R}_n(\hat{f}_k)} + 2k$$

$$\Leftrightarrow \text{minimizing} \quad \hat{R}_n(\hat{f}_k) + \frac{2\hat{\sigma}^2 k}{n} \leftarrow \text{Mallow's } C_p.$$

Whenever $(\bullet)$ is justified; which may fail to hold if the candidate model is underspecified $(\hat{\sigma}_k^2 \gg \hat{\sigma}^2)$

Note that in the case where the noise variance is known, the AIC and $C_p$ are (exactly) equivalent since in this case $AIC = n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} RSS + 2k$,

$$\Leftrightarrow \text{minimizing} \quad \frac{1}{n} RSS + \frac{2k\sigma^2}{n} = C_p.$$ ∎

The AIC criterion is more general than $C_p$, since it is applicable whenever a log-likelihood loss function is used. Moreover, the link with $C_p$ shows that AIC tends to select models that minimize the MSE.

## I.2. Corrected AIC

The derivation of the AIC relies on the large-sample properties of the MLE. The corrected AIC, denoted $AIC_c$, relaxes the assumption of large $n$, and derives an exact expression for the bias adjustment term ((*) page 7) in the context of linear regression.

- AIC = More general, but the bias estimation can be poor for small $n$.
- $AIC_c$ = Works in small sample situations, but requires a strong assumption on the model class.

Consider the (true) model $y = X_0 \beta_0 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$

The candidate model is $y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Assume that the true model is nested with the candidate models, so that $\beta_0 \in \mathbb{R}^{k_0}$, $\beta \in \mathbb{R}^k$, $0 < k_0 \leq k$, and the columns of $X_0$ are a subset of the columns of $X$. Put

$\hat\theta_k := (\hat\beta, \hat\sigma^2) =$ MLE of $(\beta, \sigma^2)$.

Still a problematic assumption.

The log-likelihood of the candidate model is

$$\log f(\underline{y} \mid \beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta)$$

$$-2\log f(\underline{y} \mid \beta, \sigma^2) = n\log 2\pi + n\log\sigma^2 + \frac{1}{\sigma^2}(y - X\beta)^t(y - X\beta)$$

We use this expression to evaluate the difference

$$\underbrace{\mathbb{E}\{d(\hat\theta_k)\}} - \underbrace{\mathbb{E}\{-2\log f(y \mid \hat\theta_k)\}}_{\begin{array}{l} \| \\ n\log 2\pi + n\,\mathbb{E}\log\hat\sigma^2 \\ \quad + \mathbb{E}\Big\{\frac{1}{\hat\sigma^2}\underbrace{(y - X\hat\beta)^t(y - X\hat\beta)}_{= n\hat\sigma^2}\Big\} \\ = n(1 + \log 2\pi) + \mathbb{E}(\log\hat\sigma^2)\end{array}} \qquad \text{(page 7)}$$

Using $y - X\beta = X(\beta_0 - \beta) + \varepsilon$, since $y = X_0\beta_0 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$

Add columns to $X_0$, and complete $\beta_0$ with zero coefficients to obtain vectors $\beta_0$ and $\beta$ of similar size.

Thus
$$(y - X\beta)^t(y - X\beta) = (\beta_0 - \beta)^t X^t X(\beta_0 - \beta) + (\beta_0 - \beta)^t X^t\varepsilon + \varepsilon^t X(\beta_0 - \beta) + \varepsilon^t\varepsilon$$

$$\Rightarrow \mathbb{E}\{-2\log f(y \mid \beta, \sigma^2)\} = n\log 2\pi + n\log\sigma^2 + n\frac{\sigma_0^2}{\sigma^2} + \frac{1}{\sigma^2}(\beta_0 - \beta)^t X^t X(\beta_0 - \beta).$$

We obtain

$$d(\hat\theta_k) = n\log 2\pi + n\log\hat\sigma^2 + n\frac{\sigma_0^2}{\hat\sigma^2} + \frac{1}{\hat\sigma^2}(\beta_0 - \hat\beta)^t X^t X(\beta_0 - \hat\beta).$$

Since $\hat\beta \sim \mathcal{N}(\beta_0, \sigma_0^2(X^tX)^{-1})$, $(\beta_0 - \hat\beta)^t \frac{X^tX}{\sigma_0^2}(\beta_0 - \hat\beta) \sim \chi^2(k)$, and $\hat\beta$ & $\hat\sigma^2$ are independent.

$$\Rightarrow \mathbb{E}\{d(\hat\theta_k)\} = n\log 2\pi + n\,\mathbb{E}\log\hat\sigma^2 + n^2\,\mathbb{E}\Big(\frac{\sigma_0^2}{n\hat\sigma^2}\Big) + n\,\mathbb{E}\Big(\underbrace{\frac{\sigma_0^2}{n\hat\sigma^2}}_{\frac{1}{\chi^2(n-k)}}\Big)\,\mathbb{E}\Big(\underbrace{(\beta_0 - \hat\beta)^t\frac{X^tX}{\sigma_0^2}(\beta_0 - \hat\beta)}_{\chi^2(k)}\Big)$$

independence

$\Big[$ For $U \sim \chi^2(k)$, $\mathbb{E}U = k$ and $\mathbb{E}\frac{1}{U} = \frac{1}{k-2}$ $\Big]$

$$\Rightarrow \mathbb{E}\{d(\hat\theta_k)\} = n\log 2\pi + n\,\mathbb{E}\log\hat\sigma^2 + \frac{n^2}{n-k-2} + \frac{nk}{n-k-2}$$
$$= n(1 + \log 2\pi) + n\,\mathbb{E}\log\hat\sigma^2 + \frac{2n(k+1)}{n-k-2}$$

Putting terms together, we see that

$$\mathbb{E}\{d(\hat\theta_k)\} = \mathbb{E}\{-2\log f(\underline{z} \mid \hat\theta_k(\underline{z}))\} + \underbrace{\frac{2n(k+1)}{n-k-2}}_{\begin{array}{c}\| \\ 2(k+1) + \frac{(k+1)(k+2)}{n-k-2}\end{array}}$$

Define

$$\boxed{\begin{aligned} AIC_c &= -2\log f(\underline{y} \mid \hat\beta, \hat\sigma^2) + \frac{2n(k+1)}{n-k-2} \\ &= AIC + \frac{(k+1)(k+2)}{n-k-2}\end{aligned}}$$

$k+1 =$ # parameters to estimate ($\beta \in \mathbb{R}^k$ & $\sigma^2$)

Remark = When $n$ is large compared to $k$, the penalty of $AIC_c$ is approximately $2(k+1)$, corresponding to the penalty of the AIC criterion.

## I.3. BAYESIAN INFORMATION CRITERION (BIC)

The Bayesian Information Criterion (BIC) was introduced by Schwartz (1978) as a competitor of the AIC.

BIC represents an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model.

Dataset : $z = (z_1, \dots, z_n)$.

The family of candidate models are $\mathcal{F}_{k_1}, \dots, \mathcal{F}_{k_L}$, where each $\mathcal{F}_k$ is uniquely parametrized by a vector $\theta_k \in \Theta_k \subset \mathbb{R}^k$

The MLE is denote $\hat{\theta}_k$.

Bayesian framework: put a prior on the class of candidates : the prior probability associated with $\mathcal{F}_k$ is denoted $\pi_k$. Moreover, the prior distribution on $\theta_k$ (given $\mathcal{F}_k$) is denoted $g(\theta_k \mid \mathcal{F}_k)$.

Strategy: Select the model which maximizes the posterior distribution

$$p(\mathcal{F}_k \mid z) = \frac{p(z \mid \mathcal{F}_k)\, p(\mathcal{F}_k)}{f(z)} \leftarrow = \pi_k$$

↖ The denominator is independent of $k$

$\Leftrightarrow$ maximizing $p(z \mid \mathcal{F}_k)\, p(\mathcal{F}_k) \leftarrow$ you may assume a constant prior

Consider the term $p(z \mid \mathcal{F}_k)$. We have

$$p(z \mid \mathcal{F}_k) = \underbrace{\int f(z \mid \theta_k, \mathcal{F}_k)\, g(\theta_k \mid \mathcal{F}_k)\, d\theta_k}$$

The exact computation of this integral is rarely possible. We introduce next a general tool to approximate such integrals.

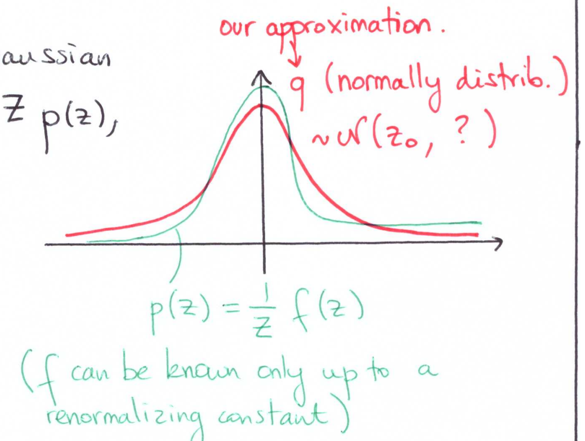
- Digression: LAPLACE APPROXIMATION

Goal : numerical evaluation / approximation of integrals $\int f(z)\,dz$, where $f > 0$.

The idea is to find a Gaussian approximation of $f(z) = Z\, p(z)$, that we denote $q$.

Let $z_0$ denote the mode of $p(z)$

$(\, p'(z_0) = 0 \,)$



The Gaussian distribution is such that its log is quadratic in its variables $\Rightarrow$ consider a Taylor expansion of $\log f(z)$ around its mode $z_0$:

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} A_f (z - z_0)^2, \text{ where } A_f = - \left.\frac{d^2 \log f(z)}{dz}\right|_{z=z_0}$$

↑ This approximation will be OK if $f$ concentrates around $z_0$.

Thus, $f(z) \approx f(z_0)\exp\left(-\frac{1}{2}A_f(z-z_0)^2\right)$

This expression generalizes well in multivariate settings:

$$f(\underline{z}) \approx f(\underline{z}_0)\exp\left(-\frac{1}{2}(\underline{z}-\underline{z}_0)^t\,\underline{\underline{A}}_f(\underline{z}-\underline{z}_0)\right), \quad \underline{z}\in\mathbb{R}^k$$

with $\underline{\underline{A}}_f = -\left.\frac{\partial^2 \log f(\underline{z})}{\partial \underline{z}\,\partial \underline{z}^t}\right|_{\underline{z}=\underline{z}_0}$,

and $q(\underline{z}) = \frac{|\underline{\underline{A}}_f|}{(2\pi)^{k/2}}\exp\left\{-\frac{1}{2}(\underline{z}-\underline{z}_0)^t\,\underline{\underline{A}}_f(\underline{z}-\underline{z}_0)\right\}$.

Thus,

$$\int f(\underline{z})d\underline{z} \approx f(\underline{z}_0)\int \exp\left\{-\frac{1}{2}(\underline{z}-\underline{z}_0)^t\,\underline{\underline{A}}_f(\underline{z}-\underline{z}_0)\right\}d\underline{z}$$

$$= f(\underline{z}_0)\frac{(2\pi)^{k/2}}{|\underline{\underline{A}}|^{1/2}}\underbrace{\int q(\underline{z})d\underline{z}}_{=1}.$$

↓

Summarizing:

$$\boxed{\int f(\underline{z})d\underline{z} \approx f(\underline{z}_0)\frac{(2\pi)^{k/2}}{|\underline{\underline{A}}_f|}, \quad \underline{\underline{A}}_f = -\left.\frac{\partial \log f(\underline{z})}{\partial \underline{z}\,\partial \underline{z}^t}\right|_{\underline{z}=\underline{z}_0},}$$

where $z_0$ = mode of $f$.

LAPLACE APPROXIMATION

In the derivation above, we did not attempt to evaluate the approximation error.

To derive the BIC coefficient, we need a more general expression of the Laplace approximation. Writing $\log f(\underline{z}) = n\,\ell(\underline{z})$ ($n$ = sample size in the context of model selection), we have

that $|\underline{\underline{A}}| = n^{k/2}|\underline{\underline{A}}_\ell|$, and it is possible to show that

$$\boxed{\int e^{n\ell(\underline{z})}d\underline{z} = e^{n\ell(\underline{z}_0)}\frac{(2\pi)^{k/2}}{n^{k/2}|\underline{\underline{A}}_\ell|} + O(n^{-1})}$$

This expression holds in more general situations as well, in particular in special cases where $\ell$ depends on $n$ itself.

Some control on the approximation error

End of digression. ■

Back to the formula on the top of page 16, the function we need to approximate is

$$h(\theta_k) := f(\underline{z}\,|\,\theta_k, \mathcal{F}_k)\, g(\theta_k\,|\,\mathcal{F}_k).$$

Consider

$$\log h(\theta_k) = \log f(\underline{z}\,|\,\theta_k, \mathcal{F}_k) + \log g(\theta_k\,|\,\mathcal{F}_k).$$

↳ mode is $\theta_k^* = \operatorname{argmax} h(\theta_k)$

↳ Hessian $A_h^* = -\left[\frac{\partial^2 \log h(\theta_k)}{\partial\theta_k\,\partial\theta_k^t}\right]_{\theta_k=\theta_k^*}$

Instead of considering the Laplace approximation of $\int h(\theta_k)d\theta_k$ around $\theta_k^*$, we notice that in $\log h(\theta_k)$, the term $\log f(\underline{z}\,|\,\theta_k, \mathcal{F}_k)$ dominates $\log g(\theta_k\,|\,\mathcal{F}_k)$ since $\log f(\underline{z}\,|\,\theta_k, \mathcal{F}_k) = \sum_{i=1}^{n}\log f(z_i\,|\,\theta_k, \mathcal{F}_k)$ = sum of $n$ terms, while $\log g(\theta_k \mid \mathcal{F}_k)$ remains constant as $n$ increases.

$\Rightarrow$ Consider • $\hat{\theta}_k = \underset{\theta_k}{\operatorname{argmax}} \log f(\underline{z} \mid \theta_k, \mathcal{F}_k)$

= mode & MLE

• $$I_{\hat{\theta}_k} = - \left[ \frac{\partial^2 \log f(\underline{z} \mid \theta_k, \mathcal{F}_k)}{\partial \theta_k \, \partial \theta_k^t} \right]_{\theta_k = \hat{\theta}_k}$$

= observed Fisher information matrix (for $n$ obs).

$\downarrow$

We make use of $\hat{\theta}_k$ & $I_{\hat{\theta}_k}$ to approximate $\int h(\theta_k)\, d\theta_k$. The use of $\hat{\theta}_k \,/\, I_{\hat{\theta}_k}$ instead of $\theta_k^* \,/\, A_h^*$ induces an approximation error of order $O(n^{-1/2})$, instead of $O(n^{-1})$ appearing on the top of page 18:

$$p(\underline{z} \mid \mathcal{F}_k) = \int h(\theta_k)\, d\theta_k$$
$$\approx f(\underline{z} \mid \hat{\theta}_k, \mathcal{F}_k)\, g(\hat{\theta}_k \mid \mathcal{F}_k) \frac{(2\pi)^{k/2}}{|I_{\hat{\theta}_k}|^{1/2}} + O(n^{-1/2}),$$

where $I_{\hat{\theta}_k} = n\, I'_{\hat{\theta}_k}$ $\leftarrow$ one observation.

$$\Rightarrow \log p(\underline{z} \mid \mathcal{F}_k) \approx \log f(\underline{z} \mid \hat{\theta}_k, \mathcal{F}_k) + \log g(\hat{\theta}_k \mid \mathcal{F}_k) + \frac{k}{2} \log 2\pi - \frac{k}{2} \log n - \frac{1}{2} \log |I'_{\hat{\theta}_k}|.$$

$(\times -2)$ $\downarrow$

$$-2 \log p(\underline{z} \mid \mathcal{F}_k) \approx \overbrace{-2 \log f(\underline{z} \mid \hat{\theta}_k, \mathcal{F}_k) + k \log n}^{\text{unbounded with } n} \underbrace{-2 \log g(\hat{\theta}_k \mid \mathcal{F}_k) - k \log(2\pi) + \log |I'_{\hat{\theta}_k}|}_{= O(1)} + O(n^{-1/2})$$

For model selection, we ignore the $O(1)$ and $O(n^{-1/2})$ terms, and we define

penalty

$$\boxed{\text{BIC} = -2 \log f(\underline{z} \mid \hat{\theta}_k, \mathcal{F}_k) + k \log n}$$

g.o.f. term

with a constant prior on $\mathcal{F}_k$, otherwise, add $-2 \log \pi_k$.

Does not need the specification of priors [such as $g(\theta_k \mid \mathcal{F}_k)$]

The penalty term in the expression of the BIC coefficient is heavier than the one appearing in the AIC.

$\Rightarrow$ BIC favour smaller models than AIC.

However, no criteria is universally superior to the other.

$\hookrightarrow$ The derivation of BIC indicates that it can be used to compare non-nested models (as opposed to the AIC).

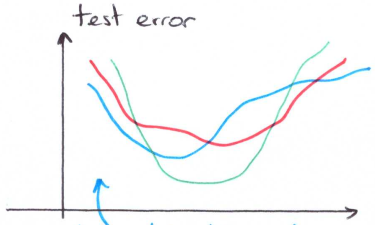
Usually, BIC is used when the goal is descriptive rather than predictive (recall the analogy between AIC and $C_p$: AIC picks models with small MSE). BIC is often used by frequentist practitioners, even if its derivation relies on large-sample Bayesian analysis. Is this justified?

## II. CROSS-VALIDATION

As opposed to analytical methods that compute the training error and then adjust it, cross-validation directly estimates the test-error. It is a very attractive approach to select the model with the best prediction accuracy, as it is very general.

(i) Validation = hold-out a subset of the training obs, and use it to estimate the test error.

TRAIN HERE | HOLD-OUT
TEST HERE



repeated estimation, for ≠ hold-out sets.

⚠ The validation estimate is highly variable, and depends heavily on which observations are put in the training & hold-out set.

(ii) LOOCV (Leave-One-Out-Cross-Validation).

In LOOCV, only one observation is used for validation:

→ Fit on $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$ and test on $x_n$, resulting in the error $\ell(y_n, \hat{y}_n)$

→ Repeat the procedure n times, each observation one after another is placed into the hold-out set, yielding a (highly variable) estimate $\ell(y_j, \hat{y}_j), j = 1, \dots, n$, where $\hat{y}_j$ is the predicted value associated with $x_j$, where the model was trained on $\{(x_1, y_1), \dots, (x_{j-1}, y_{j-1}), (x_{j+1}, y_{j+1}), \dots, (x_n, y_n)\}$.

The LOOCV estimate averages all the n test estimates:

$$\boxed{CV_{(n)} := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \hat{y}_i)}$$

↖ Very general: can be used with any loss function, and any kind of predictive model.

⊕ • Less bias than the validation procedure (almost all the data is used to fit the model)

• Tends not to overestimate the test error

• Always yields the same result: no randomness in the training & hold-out splits.

⊖ • Expensive to implement.

↳ A notable exception: linear regression $\hat{y} = Hy$.

Indeed, it is shown on page 36 of the lecture notes SL: LINEAR REGRESSION that

$$y - \hat{y}_i = (1 - h_{ii})(y_i - \tilde{y}_i)$$

i-th diagonal element

prediction of $y_i$ when obs $(x_i, y_i)$ is removed from the training set

$$\Rightarrow \boxed{CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2}$$

← square loss

↖ For free! We do not need to fit the linear regression model n times.

(iii) Generalized-Cross-Validation = approximates the LOOCV in linear regression settings, by replacing $h_{ii}$ by the mean value $\frac{1}{n} \mathrm{Tr}(H)$.

$$\boxed{GCV := \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{1 - \mathrm{Tr} H / n} \right)^2}$$

← Tr H = # parameters to estimate in LR. The formula can be used with other linear predictors as well

Making use of $\left(\frac{1}{1-x}\right)^2 \approx 1 + 2x$ for $x > 0$ small, we get

$$GCV \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \left(1 + \frac{2 \mathrm{Tr} H}{n}\right)$$

$$GCV \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{=RSS} + 2\frac{\operatorname{Tr} H}{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{=\hat{\sigma}^2}$$

$$= \frac{1}{n}\left(RSS + 2\hat{\sigma}^2 \operatorname{Tr} H\right)$$

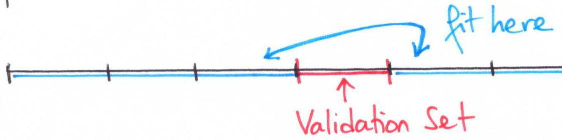
↖ For a linear regression model with normal error,

$$RSS = \hat{\sigma}^2\left(-2\log f(\underline{y}\mid\hat{\beta},\hat{\sigma}^2)\right) + \text{other terms}$$

$$= \frac{\hat{\sigma}^2}{n}\left(\underbrace{-2\log f(\underline{y}\mid\hat{\beta},\hat{\sigma}^2) + 2\operatorname{Tr} H}_{=AIC} + O(1)\right)$$

⇒ Minimizing the GCV is similar to minimizing the AIC (and $C_p$), in the context of linear regression.

(iv) K-fold CV: the training set is divided (randomly) into K groups (aka FOLDS) of approximately equal sizes.



⇒ Obtain K estimates of the test error

$$\boxed{CV_{(K)} := \frac{1}{K}\sum_{k=1}^{K}\tilde{R}^{(k)} \quad , \quad \tilde{R}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k}\ell(y_i,\hat{y}_i)}$$

↳ number of observations kept for training

↑ LOOCV is a special case, with K=n.

In practice K = 5, 10 folds are used.

The bias of K-fold CV is reduced compared to the validation approach. The bias is further reduced with LOOCV.

From a bias point of view, we prefer LOOCV. However, LOOCV has higher variance than $CV_{(K)}$. Indeed, the terms $\ell(y_i, \hat{y}_i)$ in LOOCV are highly positively correlated, since the n models are trained on almost all identical datasets (only one obs differs at a time). Thus

$$\operatorname{Var}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,\hat{y}_i)\right\} > \operatorname{Var} CV_{(K)}$$

↖ Need to add up all covariance terms.

The bias-variance trade-off !

Remark: When performing K-fold CV, you should **not** do any pre-processing of the data on the whole learning sample before testing the performance of the model on the hold-out set. Any pre-processing must be done on the (K-1) folds used for training.

Ex: You have a set of $d = 10^4$ predictors & you want to use PCA to reduce the dimensionality of the data, before applying a nearest neighbor classifier.

⊖ Wrong Way: Calculate the PC of the full training set, and use a selected few to apply K times the nearest neighbor classifier → you are cheating as information about the hold-out set was used during pre-processing

⊕ Right-Way: Divide the learning sample into folds. Compute the PCs on (K-1) folds, keep these components to test the nearest neighbour classifier on the hold-out set. The PCs in the hold-out set are computed for the principal directions calculated on the (K-1) folds used for training.