

STK-IN4300

Statistical Learning Methods in Data Science

Riccardo De Bin

debin@math.uio.no

STK4030: lecture 2

1/ 38

Linear Regression Models and Least Squares: recap

Consider:

- continuous outcome Y , with $Y = f(X) + \epsilon$;
- linear regression $f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

We know:

- $\hat{\beta} = \operatorname{argmin}_{\beta} RSS(\beta) = (X^T X)^{-1} X^T y$;
- $\hat{y} = X \hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{\text{hat matrix H}} y$;
- $\operatorname{Var}(\hat{\beta}) = (X^T X)^{-1} \sigma^2$
 - $\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^N (y_i - \hat{y}_i)^2$;

When $\epsilon \sim N(0, \sigma^2)$,

- $\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$;
- $(N - p - 1) \hat{\sigma}^2 \sim \sigma^2 \chi_{N-p-1}^2$.

STK4030: lecture 2

3/ 38

Outline of the lecture

- Linear Methods for Regression
 - Linear Regression Models and Least Squares
 - Subset selection
- Model Assessment and Selection
 - Bias, Variance and Model Complexity
 - The Bias–Variance Decomposition
 - Optimism of the Training Error Rate
 - Estimates of In-Sample Prediction Error
 - The Effective Number of Parameters
 - The Bayesian Approach and BIC

STK4030: lecture 2

2/ 38

Linear Regression Models and Least Squares: Gauss – Markov theorem

The least square estimator $\hat{\theta} = a^T (X^T X)^{-1} X^T y$ is the

B est ← smallest error (MSE)

L inear ← $\hat{\theta} = a^T \beta$

U nbaised ← $E[\hat{\theta}] = \theta$

E stimator

Remember the error decomposition,

$$E[(Y - \hat{f}(X))^2] = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\operatorname{Var}(\hat{f}(X))}_{\text{variance}} + \underbrace{E[\hat{f}(X) - f(X)]^2}_{\text{bias}^2};$$

mean square error (MSE)

then, any estimator $\tilde{\theta} = c^T Y$, s.t. $E[c^T Y] = a^T \beta$, has

$$\operatorname{Var}(c^T Y) \geq \operatorname{Var}(a^T \hat{\beta})$$

STK4030: lecture 2

4/ 38

Linear Regression Models and Least Squares: hypothesis testing

To test $H_0 : \beta_j = 0$, we use the Z-score statistic,

$$z_j = \frac{\hat{\beta}_j - 0}{sd(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{[j,j]}}}$$

- When σ^2 is unknown, under H_0 ,

$$z_j \sim t_{N-p-1},$$

where t_k is a Student t distribution with k degrees of freedom.

- When σ^2 is known, under H_0 ,

$$z_j \sim N(0; 1).$$

To test $H_0 : \beta_j, \beta_k = 0$,

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p - 1)},$$

where $_1$ and $_0$ refer to the larger and smaller models, respectively.

Subset selection: variable selection

Why choosing a **sparser** (less variables) **model**?

- prediction **accuracy** (smaller variance);
- **interpretability** (easier to understand the model);
- **portability** (easier to use in practice).

Classical approaches:

- forward selection;
- backward elimination;
- stepwise and stepback selection;
- best subset technique.
- stagewise selection.

Subset selection: classical approaches

Forward selection:

- **start** with the **null model**, $Y = \beta_0 + \epsilon$;
- among a set of possible variables, **add** that which reduces the unexplained variability the most
 - e.g.: after the first step, $Y = \beta_0 + \beta_2 X_2 + \epsilon$;
- **repeat iteratively** until a certain **stopping criterion** (p-value larger than a threshold α , increasing AIC, ...) is met.

Backward elimination:

- **start** with the **full model**, $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$;
- **remove** the variable that contributes the least in explaining the outcome variability
 - e.g.: after the first step, $Y = \beta_0 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$;
- **repeat iteratively** until a **stopping criterion** (p-value of all remaining variable smaller than α , increasing AIC, ...) is met.

Subset selection: classical approaches

Stepwise and stepback selection:

- **mixture** of forward and backward selection;
- **allow both adding and removing** variables at each step;
 - starting from the **null** model: **stepwise** selection;
 - starting from the **full** model: **stepback** selection.

Best subset:

- compute all the **2^p possible models** (each variable in/out);
- choose the model which **minimizes** a loss function (e.g., AIC).

Stagewise selection:

- similar to the forward selection;
- at each step, the specific regression coefficient is updated **only** using the information related to the corresponding variable;
 - slow to converge in low-dimensions;
 - turned out to be effective in high-dimensional settings.

Model Assessment and Selection: introduction

- **Model Assessment:** **evaluate** the performance (e.g., in terms of prediction) of a selected model.
- **Model Selection:** **select** the best model for the task (e.g., best for prediction).
- **Generalization:** a (prediction) model must be valid in broad **generality**, not specific for a specific dataset.

Bias, Variance and Model Complexity: categorical variables

Similar story for the **categorical** variables:

- G = target variable \rightarrow takes K values in \mathcal{G} ;

Typical choices for the loss function in this case are:

- $L(Y, \hat{f}(X)) = \mathbb{1}(G \neq \hat{G}(X))$, the **0-1 loss**;
- $L(Y, \hat{f}(X)) = -2 \log \hat{p}_G(X)$, the **deviance**.
- $\log \hat{p}_G(X) = \ell(\hat{f}(X))$ is **general** and can be used for every kind of outcome (binomial, Gamma, Poisson, log-normal, ...)
- the factor -2 is added to make the loss function equal to the squared loss in the Gaussian case,

$$L(\hat{f}(X)) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (Y - \hat{f}(X))^2 \right\}$$

$$\ell(\hat{f}(X)) = -\frac{1}{2} (Y - \hat{f}(X))^2$$

Bias, Variance and Model Complexity: definitions

Define:

- Y = target variable;
- X = input matrix;
- $\hat{f}(X)$ = prediction rule, trained on a training set \mathcal{T} .

The error is measured through a **loss function**

$$L(Y, \hat{f}(X))$$

which **penalizes differences** between Y and $\hat{f}(X)$.

Typical choices for continuous outcomes are:

- $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$, the **quadratic loss**;
- $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$, the **absolute loss**.

Bias, Variance and Model Complexity: test error

The **test error** (or generalization error) is the prediction error over an **independent test** sample

$$\text{Err}_{\mathcal{T}} = E[L(Y, \hat{f}(X)) | \mathcal{T}]$$

where both X and Y are drawn randomly from their joint distribution.

The **specific training set** \mathcal{T} used to derive the prediction rule is **fixed** \rightarrow the test error refers to the error for this specific \mathcal{T} .

In general, we would like to minimize the **expected prediction error** (expected test error),

$$\text{Err} = E[L(Y, \hat{f}(X))] = E[\text{Err}_{\mathcal{T}}].$$

Bias, Variance and Model Complexity: training error

- We would like to get Err , but we **only have information** on the single training set (we will see later how to solve this issue);
- our goal, therefore, is to **estimate** $\text{Err}_{\mathcal{T}}$.

The **training error**

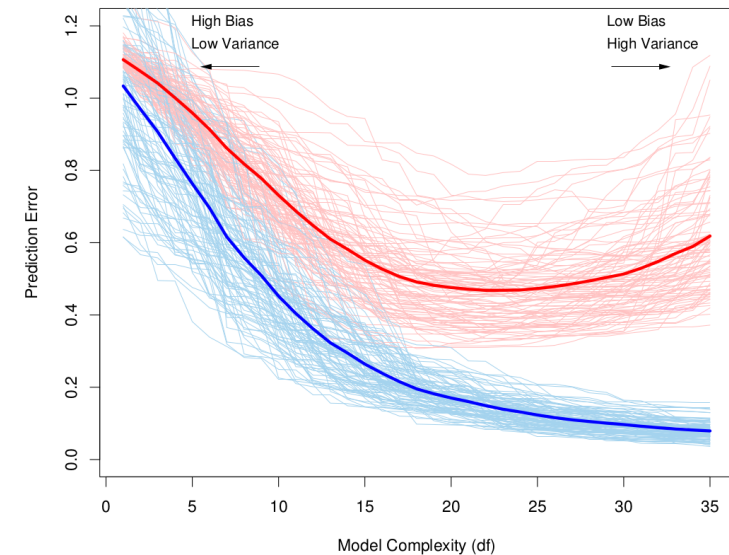
$$\bar{\text{err}} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i)),$$

is **NOT** a good estimator of $\text{Err}_{\mathcal{T}}$.

We do not want to minimize the training error:

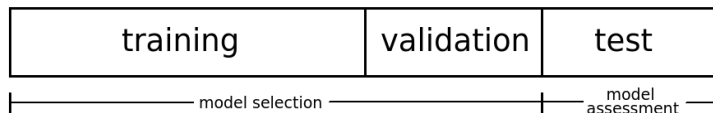
- increasing the **model complexity**, we can always decrease it;
- overfitting issues:
 - model specific for the training data;
 - generalize** very poorly.

Bias, Variance and Model Complexity: prediction error



Bias, Variance and Model Complexity: data split

In an ideal (= a lot of data) situation, the best option is **randomly splitting** the data in three **independent** sets,



- training set:** data used to **fit** the model(s);
- validation set:** data used to **identify** the best model;
- test set:** data used to **assess the performance** of the best model (must be completely ignored during model selection).

NB: it is extremely important to use the sets **fully independently**!

Bias, Variance and Model Complexity: data split

Example with k-nearest neighbour:

- in the **training set:** **fit** kNN with **different values of k** ;
- in the **validation set:** **select** the model with best performance (**choose k**);
- in the **test set:** **evaluate** the prediction error of the model with the **selected k** .

Bias, Variance and Model Complexity: data split

How to **split** the data in three set? There is not a general rule.

The book's suggestion:

- **training set:** 50%;
- **validation set:** 25%;
- **test set:** 25%.

We will see later what to do when there are not enough data;

- difficult to say when the data are "enough".

The Bias–Variance Decomposition: k-nearest neighbours

For the kNN regression:

$$\begin{aligned}\text{Err}(x_0) &= E_Y[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + \left[f(x_0) - \frac{1}{k} \sum_{\ell=1}^k f(x_\ell) \right]^2 + \frac{\sigma_\epsilon^2}{k}\end{aligned}$$

Note:

- the number of neighbour is **inversely related** to the complexity;
- **smaller k** \rightarrow smaller bias, larger variance;
- **larger k** \rightarrow larger bias, smaller variance.

The Bias–Variance Decomposition: computations

Consider $Y = f(X) + \epsilon$, $E[\epsilon] = 0$, $\text{Var}[\epsilon] = \sigma^2$. Then

$$\begin{aligned}\text{Err}(x_0) &= E[(Y - \hat{f}(X))^2 | X = x_0] \\ &= E[Y^2] + E[\hat{f}(x_0)^2] - 2E[Y\hat{f}(x_0)] \\ &= \text{Var}[Y] + f(x_0)^2 + \text{Var}[\hat{f}(x_0)] + E[\hat{f}(x_0)]^2 - 2f(x_0)E[\hat{f}(x_0)] \\ &= \sigma^2 + \text{bias}^2(\hat{f}(x_0)) + \text{Var}[\hat{f}(x_0)] \\ &= \text{irreducible error} + \text{bias}^2 + \text{variance}\end{aligned}$$

Remember that:

- $E[Y] = E[f(X) + \epsilon] = E[f(X)] + E[\epsilon] = f(X) + 0 = f(X)$;
- $E[Y^2] = \text{Var}[Y] + E[Y]^2 = \sigma^2 + f(X)^2$;
- $\hat{f}(X)$ and ϵ are uncorrelated.

The Bias–Variance Decomposition: linear regression

For linear regression, with a **p -dimensional** β (regression coefficients) estimated by **least squares**,

$$\begin{aligned}\text{Err}(x_0) &= E_Y[(Y - \hat{f}_p(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [f(x_0) - E[f_p(x_0)]]^2 + \|h(x_0)\|^2 \sigma_\epsilon^2\end{aligned}$$

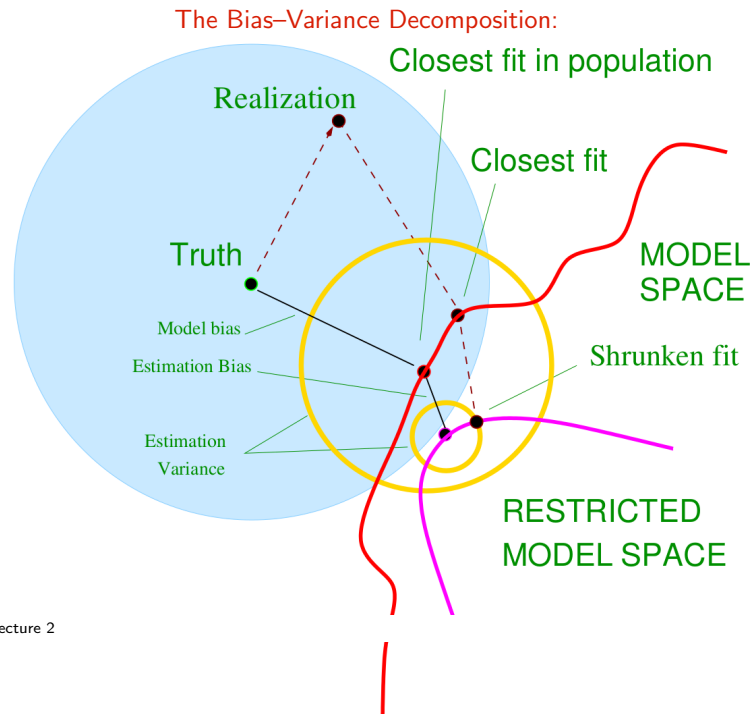
where $h(x_0) = X(X^T X)^{-1} x_0$,

- $\hat{f}_p(x_0) = x_0^T (X^T X)^{-1} X^T y \rightarrow \text{Var}[\hat{f}_p(x_0)] = \|h(x_0)\|^2 \sigma_\epsilon^2$.

In average,

$$\frac{1}{N} \sum_{i=1}^N \text{Err}(x_i) = \sigma_\epsilon^2 + \frac{1}{N} \sum_{i=1}^N [f(x_i) - E[f_p(x_i)]]^2 + \frac{p}{N} \sigma_\epsilon^2,$$

so the model complexity is **directly related** to p .



Optimism of the Training Error Rate: definitions

Being a little bit more formal,

$$\text{Err}_{\mathcal{T}} = E_{X_0, Y_0} [L(Y_0, \hat{f}(X_0)) | \mathcal{T}]$$

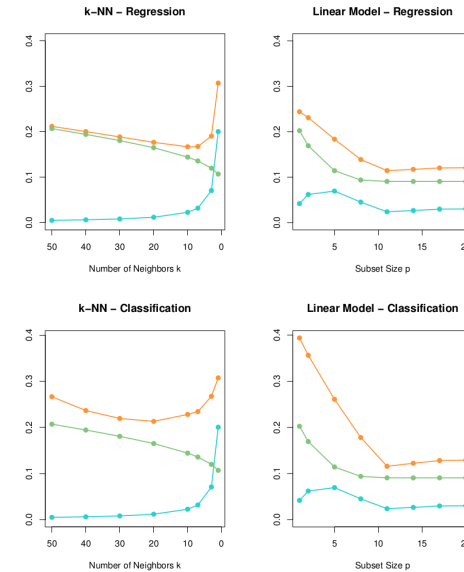
where:

- (X_0, Y_0) are from the new test set;
- $\mathcal{T} = \{(x_1, y_1) \dots (x_n, y_n)\}$ is fixed.

Taking the expected value over \mathcal{T} , we obtain the **expected error**

$$\text{Err} = E_{\mathcal{T}} \left[E_{X_0, Y_0} [L(Y_0, \hat{f}(X_0)) | \mathcal{T}] \right].$$

The Bias–Variance Decomposition: example



Optimism of the Training Error Rate: definitions

We said that the **training error**,

$$\bar{\text{err}} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i)),$$

is **NOT** a good estimator of $\text{Err}_{\mathcal{T}}$:

- **same data** used both for training and test;
- a fitting method tends to **adapt** to the specific dataset;
- the result is a **too optimistic evaluation of the error**.

How to measure this optimism?

Optimism of the Training Error Rate: optimism and average optimism

Let us define the **in-sample error**,

$$\text{Err}_{\text{in}} = \sum_{i=1}^N E_{Y_0}[L(Y_{i0}, \hat{f}(x_i)) | \mathcal{T}],$$

i.e., the error computed w.r.t. new values of the outcome on the **same** values of the **training points** $x_i, i = 1, \dots, N$.

We define **optimism** the difference between Err_{in} and $\bar{\text{err}}$,

$$\text{op} := \text{Err}_{\text{in}} - \bar{\text{err}}.$$

and the **average optimism** its expectation,

$$\omega := E_Y[\text{op}].$$

NB: as the training points are fixed, the expected value is taken w.r.t. their outcomes.

Optimism of the Training Error Rate: optimism and average optimism

For a reasonable number of **loss functions**, including 0-1 loss and squared error, it can be shown that

$$\omega = \frac{2}{N} \sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i),$$

where:

- Cov stands for covariance;
- \hat{y}_i is the prediction, $\hat{y}_i = \hat{f}(x_i)$;
- y_i is the actual value.

Therefore:

- optimism depends on **how much** y_i affects its own prediction;
- the **"harder"** we fit the data, the **larger** the value of $\text{Cov}(\hat{y}_i, y_i)$ → the **larger** the optimism.

Optimism of the Training Error Rate: optimism and average optimism

As a consequence,

$$E_Y[\text{Err}_{\text{in}}] = E_Y[\bar{\text{err}}] + \frac{2}{N} \sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i).$$

When \hat{y}_i is obtained by a linear fit of d inputs the expression simplifies. For the linear additive model $Y = f(X) + \epsilon$,

$$\sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i) = d\sigma_\epsilon^2,$$

and

$$E_Y[\text{Err}_{\text{in}}] = E_Y[\bar{\text{err}}] + 2\frac{d}{N}\sigma_\epsilon^2. \quad (1)$$

Therefore:

- optimism **increases linearly** with the number of predictors;
- it **decreases linearly** with the training sample size.

Optimism of the Training Error Rate: estimation

Methods we will see:

- C_p , AIC, BIC **estimate the optimism** and **add** it to the **training error** (work when estimates are linear in their parameters);
- cross-validation and bootstrap directly estimate the **expected error** (work in general).

Further notes:

- **in-sample error** is in general **NOT** of interest;
- when doing model selection/find the right model complexity, we are more interested in the **relative difference in error** rather than the absolute one.

Estimates of In-Sample Prediction Error: C_p

Consider the general form of the in-sample estimates,

$$\widehat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \hat{\omega}.$$

Equation (1),

$$E_Y[\text{Err}_{\text{in}}] = E_Y[\bar{\text{err}}] + 2 \frac{d}{N} \sigma_\epsilon^2,$$

in the case of linearity and square errors, leads to the C_p statistics,

$$C_p = \bar{\text{err}} + 2 \frac{d}{N} \hat{\sigma}_\epsilon^2,$$

where:

- $\bar{\text{err}}$ is the **training error** computed by the square loss;
- d is the **number of parameters** (e.g., regression coefficients);
- $\hat{\sigma}_\epsilon^2$ is an estimate of the **noise variance** (computed on the full model, i.e., that having the smallest bias).

Estimates of In-Sample Prediction Error: AIC

Similar idea for **AIC** (Akaike Information Criterion):

- we start from equation (1);
- more general by using a **log-likelihood approach**,

$$-2E[\log p_{\hat{\theta}}(Y)] \approx -\frac{2}{N} E \left[\sum_{i=1}^N \log p_{\hat{\theta}}(y_i) \right] + 2 \frac{d}{N}$$

Note that:

- the result holds **asymptotically** (i.e., $N \rightarrow \infty$);
- $p_{\hat{\theta}}(Y)$ is the **family of densities** of Y , indexed by θ ;
- $\sum_{i=1}^N \log p_{\hat{\theta}}(y_i) = \ell(\hat{\theta})$, the **maximum likelihood estimate**.

Examples:

- logistic regression, $\text{AIC} = -\frac{2}{N} \ell(\hat{\theta}) + 2 \frac{d}{N}$;
- linear regression, $\text{AIC} \propto C_p$.

Estimates of In-Sample Prediction Error: AIC

To find the best model, we choose that with the **smallest AIC**:

- straightforward in the simplest cases (e.g., linear models);
- more attention must be devoted in more complex situations
 - issue of finding a **reasonable measure** for the **model complexity**;

Usually **minimizing the AIC** is **not** the best solution to find the value of the **tuning parameter**

- cross-validation works better in this case.

The Effective Number of Parameters

Generalize the concept of **number of predictors** to extend the previous approaches to more complex situations.

Let

- $y = (y_1, \dots, y_n)$ be the outcome;
- $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ be the prediction.

For linear methods,

$$\hat{y} = Sy$$

where S is a $N \times N$ matrix which

- depend on X
- does **NOT** depend on y .

The Effective Number of Parameters

The **effective number of parameters** (or effective degrees of freedom) is defined as

$$\text{df}(S) := \text{trace}(S);$$

- $\text{trace}(S)$ is the **sum of the diagonal elements** of S ;
- we should **replace** d with $\text{trace}(S)$ to obtain the correct value of the criteria seen before;
- if $y = f(X) + \epsilon$, with $\text{Var}(\epsilon) = \sigma_\epsilon^2$, then $\sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i) = \text{trace}(S)\sigma_\epsilon^2$, which motivates

$$\text{df}(\hat{y}) = \frac{\sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i)}{\sigma_\epsilon^2}.$$

The Bayesian Approach and BIC: BIC

The **BIC** (Bayesian Information Criterion) is an alternative criterion to AIC,

$$\frac{1}{N} \text{BIC} = -\frac{2}{N} \ell(\hat{\theta}) + \log N \frac{d}{N}$$

- similar to AIC, with **log N** instead of 2;
- if $N > e^2 \approx 7.4$, BIC tends to favor **simpler** models than AIC.
- For the Gaussian model,

$$\text{BIC} = \frac{N}{\sigma_\epsilon^2} \left[\overline{\text{err}} + (\log N) \frac{d}{N} \sigma_\epsilon^2 \right].$$

The Bayesian Approach and BIC: motivations

Despite similarities, AIC and BIC come from **different ideas**. In particular, BIC comes from the **Bayesian model selection** approach. Suppose

- \mathcal{M}_m , $m = 1, \dots, M$ be a set of **candidate models**;
- θ_m be their correspondent parameters;
- $Z = (x_1, y_1), \dots, (x_N, y_N)$ be the training data.

Given the **prior distribution** $Pr(\theta_m | \mathcal{M}_m)$ for all θ_m , the posterior is

$$\begin{aligned} Pr(\mathcal{M}_m | z) &\propto Pr(\mathcal{M}_m) \cdot Pr(Z | \mathcal{M}_m) \\ &\propto Pr(\mathcal{M}_m) \cdot \int_{\Theta_m} Pr(Z | \mathcal{M}_m, \theta_m) \cdot Pr(\theta_m | \mathcal{M}_m) d\theta_m. \end{aligned}$$

The Bayesian Approach and BIC: motivations

To **choose between** two models, we compare their posterior distributions,

$$\frac{Pr(\mathcal{M}_m | z)}{Pr(\mathcal{M}_\ell | z)} = \underbrace{\frac{Pr(\mathcal{M}_m)}{Pr(\mathcal{M}_\ell)}}_{\text{prior preference}} \cdot \underbrace{\frac{Pr(Z | \mathcal{M}_m)}{Pr(Z | \mathcal{M}_\ell)}}_{\text{Bayes factor}}$$

- usually the first term on the right hand side is **equal to 1** (same prior probability for the two models);
- the choice between the models is based on the **Bayes factor**.

Using some algebra (including the Laplace approximation), we find

$$\log Pr(Z | \mathcal{M}_m) = \log Pr(Z | \hat{\theta}_m, \mathcal{M}_m) - \frac{d_m}{2} \log N + O(1).$$

where:

- $\hat{\theta}_m$ is the maximum likelihood estimate of θ_m ;
- d_m is the number of free parameters in the model \mathcal{M}_m .

The Bayesian Approach and BIC: motivations

Note:

- If the loss function is $-2 \log Pr(Z|\hat{\theta}_m, \mathcal{M}_m)$, we find again the expression of BIC;
- selecting the model with **smallest BIC corresponds** to selecting the model with the **highest posterior** probability;
- in particular, note that,

$$\frac{e^{-\frac{1}{2}BIC_m}}{\sum_{\ell=1}^M e^{-\frac{1}{2}BIC_\ell}}$$

is the **probability of selecting** the model m (out of M models).

The Bayesian Approach and BIC: AIC versus BIC

For **model selection**, what to choose between AIC and BIC?

- there is **no clear winner**;
- BIC leads to a **sparser** model;
- AIC tends to be better for **prediction**;
- BIC is **consistent** ($N \rightarrow \infty$, $\Pr(\text{select the true model}) = 1$);
- for finite sample sizes, BIC tends to select a model which is **too sparse**.