Model Selection & Information Criteria: Akaike Information Criterion

Authors: M. Mattheakis, P. Protopapas

1 Maximum Likelihood Estimation

In data analysis the statistical characterization of a data sample is usually performed through a parametric *probability distribution* (or *mass function*), where we use a distribution to fit our data. The reason that we want to fit a distribution to our data is that it is easier to work with a model rather than data, and it is also more general. There are a lot of types of distributions for different types of data. Examples include, *Normal, exponential, Poisson, Gamma,* etc. A distribution is completely characterized by a set of parameters which we denote in vector form as $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, where k is the number of parameters. The goal is to estimate the distribution parameters in order to fit our data as best as it is possible. For instance, the method of *least squares* is a simple example that estimates the parameter set $\boldsymbol{\theta}$, but this is a method for a specific model. A more general and powerful method to find the optimal way to fit a distribution to data is *Maximum Likelihood Estimation* (MLE). This is the topic discussed in this section.

We can understand the idea of MLE through a simple example. We assume that we have a set of observations shown with the red circles in Fig. 1. We observe that most of the observations are arranged around a center, so we intuitively suggest the Normal distribution to fit this dataset (red curve in Fig. 1). The normal distribution is characterized by two parameters: the mean μ and the standard deviation σ, where, $\boldsymbol{\theta} = (\mu, \sigma)$. The

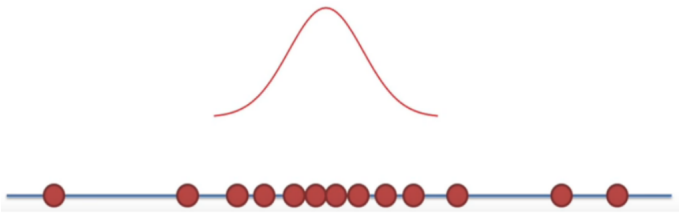

Figure 1: A data sample (red circles) that is represented by a normal distribution (red line). (The image was taken by "StatQuest, MLE" from youtube).

question is how can we estimate the parameters $\boldsymbol{\theta}$ of the normal distribution in order to maximize the likelihood of observing the data value. A straightforward way is to compute the likelihood for many values of parameters μ and σ, and find which set of $\boldsymbol{\theta} = (\mu, \sigma)$ maximizes the likelihood. This is schematically illustrated in Fig. 2.

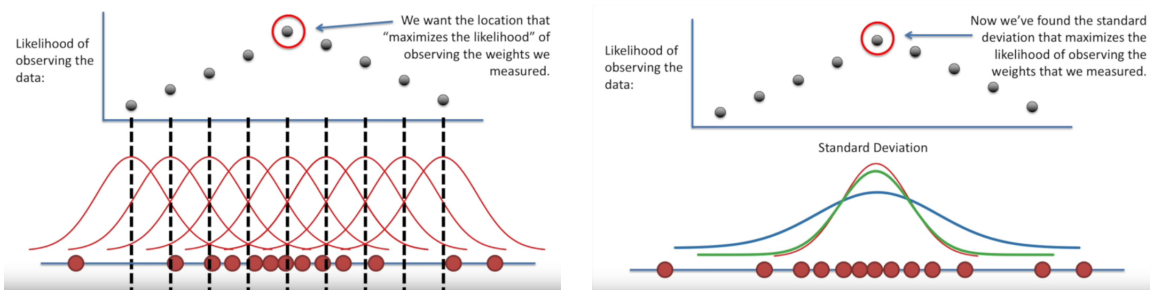

Figure 2: Maximum Likelihood Estimation (MLE): Scanning over the parameters μ and σ until the maximum value of likelihood is obtained. (The images were taken by "StatQuest, MLE" from youtube).

A formal method for estimating the optimal distribution parameters $\boldsymbol{\theta}$ is given by the MLE approach. Let us describe the main idea behind the MLE. We assume, conditional on $\boldsymbol{\theta}$, a parametric distribution $q(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y} = (y_1, ..., y_n)^T$ is a vector that contains n measurements (or observations). The likelihood is defined by the product:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} q(y_i|\boldsymbol{\theta}), \tag{1}$$

and gives a measure of how likely it is to observe the values of $\mathbf{y}$ given the parameters $\boldsymbol{\theta}$. Maximum likelihood fitting consists of choosing the distribution parameters $\boldsymbol{\theta}$ that maximizes the L for a given set of observations $\mathbf{y}$. It is easier and numerically more stable to work with the log-likelihood, since the product turns to summation as follows

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(q(y_i|\boldsymbol{\theta})\right), \tag{2}$$

where log here is the natural logarithm. In MLE we are able to use log-likelihood ℓ instead of the likelihood L because their derivatives become zero at the same point, since

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \log L = \frac{1}{L} \frac{\partial L}{\partial \boldsymbol{\theta}},$$

hence, both $L(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta})$ become maximum for the same set of parameters $\boldsymbol{\theta}$, which we will call $\boldsymbol{\theta}_{\text{MLE}}$, and thus,

$$\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MLE}}} = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MLE}}} = 0.$$

We present the basic idea of the MLE method through a particular distribution, the exponential. Afterwards, we present a very popular and useful workhorse algorithm that is based on MLE, the *Linear Regression* model with *normal error*. In this model, the optimal distribution parameters that maximize the likelihood can be calculated exactly (analytically). Unfortunately, for most distributions the analytic solution is not possible, so we use iterative methods (such as gradient descent) to estimate the parameters.

1.1 Exponential Distribution

In this section, we describe the *Maximum Likelihood Estimation* (MLE) method by using the *exponential distribution*. The exponential distribution occurs naturally in many real-world situations such as economic growth, the increasing growth rate of microorganisms, virus spreading, the waiting times between events (e.g. views of a streaming video in youtube), in nuclear chain reactions rates, in computer processing power (Moore's law), and in many other examples. Consequently, it is a very useful distribution. The exponential distribution is characterized by just one parameter, the so-called *rate parameter* λ, which is proportional to how quickly events happen, hence $\boldsymbol{\theta} = \lambda$. Considering that we have n observations that are given by the vector $\mathbf{y} = (y_1, ..., y_n)^T$, and assuming that these data follow the exponential distribution, then they can be described by the exponential probability density:

$$f(y_i|\lambda) = \begin{cases} \lambda e^{-\lambda y_i} & y_i \geq 0 \\ 0 & y_i < 0 \end{cases} . \tag{3}$$

The log-likelihood that corresponds to the exponential distribution density (3) is determined by the formula (2) and given by

$$\ell(\lambda) = \sum_{i=1}^{n} \log\left(\lambda e^{-\lambda y_i}\right) = \sum_{i=1}^{n} \left(\log(\lambda) - \lambda y_i\right). \tag{4}$$

Since we have only one distribution parameter (λ), we are maximizing the log-likelihood (4) with respect to λ, hence

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i = 0,$$

where the solution estimates the optimal parameter λ_{MLE} that maximizes the likelihood to be:

$$\lambda_{\text{MLE}} = \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^{-1} . \tag{5}$$

Inspecting the expression (5) we can observe that λ_{MLE} is the inverse of the mean of our data sample. This is a useful property of the rate parameter.

1.2 Linear Regression Model

Linear regression is a workhorse algorithm that is used in many scientific fields such as finance, the social sciences, the natural sciences and data science. We assume a dataset with n training data-points (y_i, x_i), for $i = 1, ..., n$, where y_i accounts to the i-th observation for the input point x_i. The goal of the linear regression model is to find a linear relationship between the quantitative response $\mathbf{y} = (y_1, ..., y_n)^T$ on the basis of the input (predictor) vector $\mathbf{x} = (x_1, ..., x_n)^T$. In Fig. 3 we illustrate the probabilistic interpretation of linear regression and the idea behind the MLE for linear regression model. In particular, we show a sample set of points (y_i, x_i) (red filled circles), and the corresponding prediction of

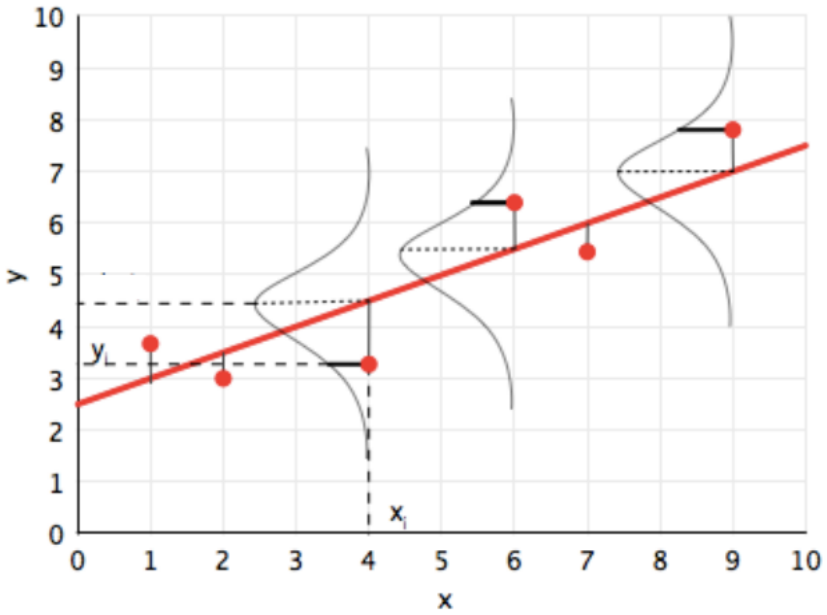

Figure 3: Linear regression model. The red filled circles show the data points (y_i, x_i) while the red solid line is the prediction of linear regression model.

the linear regression model at the same x_i (solid red line). We obtain the best linear model when the total deviation between the real y_i and the predicted values is minimized. This is achieved by maximizing the likelihood. As a result we can use the MLE approach.

The fundamental assumption of the linear regression model is that each y_i is normally (gaussian) distributed with variance σ^2 and with mean $\mu_i = \boldsymbol{\beta} \cdot \mathbf{x}_i = \mathbf{x}_i^T \boldsymbol{\beta}$, hence

$$\begin{aligned} y_i &= \sum_{j=0}^{k} x_{ij}\beta_j + \epsilon_i \\ &= \mathbf{x}_i \cdot \boldsymbol{\beta} + \epsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \end{aligned}$$

where ϵ_i is a gaussian random *error term* (or white stochastic noise) with zero mean and variance σ^2, i.e. $\epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma^2)$, k is the size of the coefficient vector $\boldsymbol{\beta}$ and now each $\mathbf{x}_i$ is a vector of the same size k; this statement is graphically demonstrated by the black gaussian curves in Fig. 3. The observation y_i is assumed to be given by the conditional normal distribution

$$y_i = q(y_i|\mu_i, \sigma^2) = \mathcal{N}(y_i|\mu_i, \sigma^2) = \mathcal{N}(y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2) \tag{6}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right). \tag{7}$$

Using the formulas (1) and (2) we write the likelihood for the normal distribution (6) as

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right), \tag{8}$$

and the corresponding log-likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)\right) \\ &= -\sum_{i=1}^{n}\left(\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma^2) + \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2, \end{aligned} \tag{9}$$

where the last term in Eq. (9) is called *loss* function. The MLE method requires the maximization of likelihood, hence we differentiate Eq. (9) with respect the distribution parameters $\boldsymbol{\beta}$ and σ^2, and set the equation to zero. Solving for the optimal parameters $\boldsymbol{\theta}_{\text{MLE}} = (\boldsymbol{\beta}_{\text{MLE}}, \sigma^2_{\text{MLE}})$, we obtain the standard formulas for the linear regression model:

$$\boldsymbol{\beta}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{10}$$

and

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_{\text{MLE}}\right)^2, \tag{11}$$

where the matrix $\mathbf{X}$ is called *the design matrix* and is created by stacking rows of $\mathbf{x}_i$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_{11} & \cdots & \mathbf{x}_{1v} \\ 1 & \mathbf{x}_{21} & \cdots & \mathbf{x}_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_{n1} & \cdots & \mathbf{x}_{nv} \end{pmatrix}. \tag{12}$$

2 Information Theory & Model Selection

2.1 KL Divergence

In the previous section we used MLE to estimate the parameters of a particular distribution in order to fit a given dataset of real observations. Two crucial questions that naturally arose regarding the learning of a model are: *how good do we fit the real data* and *what additional uncertainty have we introduced*. In other words, we would like to know how far our model is from "perfect accuracy". The answer to these crucial questions is given by the *Kullback-Leibler divergence* (KL) (also called *relative entropy*), which was introduced in 1951 in the context of information theory and shows the direct divergence between two distributions. In particular, the KL divergence is a measure of how one probability distribution is different from a second reference probability distribution. The KL divergence is a non-negative quantity and approaches zero when we expect similar, if not the same, behavior from two distributions.

We suppose that the data are generated by an unknown distribution $p(\mathbf{y})$, the "real" distribution which we wish to model. We try to approximate p with a parametric learning

model distribution $q(\mathbf{y}|\boldsymbol{\theta})$, which is governed by a set of adjustable parameters $\boldsymbol{\theta}$ that we have to estimate. The KL divergence is defined as:

$$\mathcal{D}_{\mathrm{KL}}(p \parallel q) = \sum_{i=1}^{n} p(y_i) \log\left(\frac{p(y_i)}{q(y_i|\boldsymbol{\theta})}\right) \tag{13}$$

$$= \int_{-\infty}^{\infty} p(\mathbf{y}) \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right) d\mathbf{y}, \tag{14}$$

where the formula (13) accounts for discrete variables, whereas in the continuous variables limit the KL divergence is given by (14). Note that the KL divergence is not a symmetrical quantity, that is to say $\mathcal{D}_{\mathrm{KL}}(p \parallel q) \neq \mathcal{D}_{\mathrm{KL}}(q \parallel p)$. In addition, we can easily check that the KL divergence between a distribution and itself is $\mathcal{D}_{\mathrm{KL}}(p \parallel p) = 0$.

We obtain another useful formula for the KL divergence by observing that the definitions (13) and (14) are essentially the discrete and continuous, respectively, expectation of $\log(p/q)$ conditional on the "real" distribution p, hence:

$$\mathcal{D}_{\mathrm{KL}}(p \parallel q) = \mathbb{E}_p\left[\log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)\right]$$

$$= \mathbb{E}_p\left[\log\left(p(\mathbf{y})\right) - \log\left(q(\mathbf{y}|\boldsymbol{\theta})\right)\right], \tag{15}$$

where $\mathbb{E}_p[.]$ denotes the expectation value conditional to p. We can show now that the KL divergence is always a non-negative quantity,

$$\mathcal{D}_{\mathrm{KL}}(p \parallel q) \geq 0, \tag{16}$$

with equality if, and only if, $p(\mathbf{y}) = q(\mathbf{y})$. We use Jensen's inequality for the expectation on a convex function $f(\mathbf{y})$:

$$\mathbb{E}[f(\mathbf{y})] \geq f(\mathbb{E}[\mathbf{y}]).$$

Hence, from (15) we read:

$$\mathcal{D}_{\mathrm{KL}}(p \parallel q) = \mathbb{E}_p\left[\log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)\right]$$

$$= \mathbb{E}_p\left[-\log\left(\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}\right)\right] \geq -\log\left(\mathbb{E}_p\left[\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}\right]\right) = 0,$$

where we used the fact that $-\log(.)$ is a strictly convex function. The last step of the proof above involves the definition of the conditional expectation value and the assumption of a normalized to one distribution q, such as:

$$\log\left(\mathbb{E}_p\left[\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}\right]\right) = \log\left(\int_{\mathbf{y}} p(\mathbf{y}) \frac{q(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} d\mathbf{y}\right)$$

$$= \log\left(\int_{\mathbf{y}} q(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}\right) = \log\left(\int_{\mathbf{y}} d\mathbf{y}\right) = 0.$$

In fact, since $-\log(.)$ is a strictly convex function, the equality in Eq. (16) only happens when $q(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y})$ for all $\mathbf{y}$.

2.2 Maximum Likelihood Justification

In section 1, we showed that the MLE is a powerful method used to estimate the optimal parameters $\boldsymbol{\theta}_{\text{MLE}}$ for which a parametric model distribution $q(\mathbf{y}|\boldsymbol{\theta})$ best fits the data that are given by a "real" distribution $p(\mathbf{y})$. Nevertheless, the MLE approach was not really derived, but emerged from our intuition. The KL divergence provides a way for a formal justification of the MLE method and is what we discuss here. In particular, we are seeking the parameters $\boldsymbol{\theta}$ that provide the best fit to the real distribution $p(\mathbf{y})$. In terms of KL divergence we want to minimize the KL divergence between $p(\mathbf{y})$ and $q(\mathbf{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We cannot do this directly because we do not know the real distribution $p(\mathbf{y})$ and thus, we cannot evaluate the integral (14) or work with the conditional expectation of Eq. (15). We suppose, however, that we have observed a finite set of training points y_i (for $i = 1, ..., n$) drawn from $p(\mathbf{y})$. Then the true distribution $p(\mathbf{y})$ can be approximated by a finite sum over these points given by the *empirical* distribution:

$$p(\mathbf{y}) \simeq \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{y} - y_i), \tag{17}$$

where δ is the Dirac function. Using the approximation (17) into the integral (14) yields:

$$\begin{aligned}\mathcal{D}_{\text{KL}}(p \parallel q) &\simeq \int_{-\infty}^{\infty} p(\mathbf{y}) \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right) d\mathbf{y} \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\delta(\mathbf{y} - y_i)\log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)d\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p(y_i)}{q(y_i|\boldsymbol{\theta})}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\log p(y_i) - \log q(y_i|\boldsymbol{\theta})\right), \end{aligned} \tag{18}$$

where we used the property of the delta function: $\int_{-\infty}^{\infty}\delta(x - x_0)f(x)dx = f(x_0)$. We want to minimize the Eq. (18) with respect to $\boldsymbol{\theta}$. We observe that the first term in Eq. (18) is independent of $\boldsymbol{\theta}$ and the second term is the negative log-likelihood. Thus, minimizing the expression (18) essentially means maximizing $\sum_{i=1}^{n}\log q(y_i|\boldsymbol{\theta})$ as the MLE states.

2.3 Model Comparison

The KL divergence can be used to compare two different model distributions $q(\mathbf{y}|\boldsymbol{\theta})$ and $r(\mathbf{y}|\boldsymbol{\theta})$ in order to check which model fits better to the real data given by $p(\mathbf{y})$. Using Eq. (15) we have:

$$\begin{aligned}\mathcal{D}_{\text{KL}}(p \parallel q) - \mathcal{D}_{\text{KL}}(p \parallel r) &= \mathbb{E}_p\left[\log\left(p(\mathbf{y})\right) - \log\left(q(\mathbf{y}|\boldsymbol{\theta})\right)\right] - \mathbb{E}_p\left[\log\left(p(\mathbf{y})\right) - \log\left(r(\mathbf{y}|\boldsymbol{\theta})\right)\right] \\ &= \mathbb{E}_p\left[\log\left(r(\mathbf{y}|\boldsymbol{\theta})\right) - \log\left(q(\mathbf{y}|\boldsymbol{\theta})\right)\right] = \mathbb{E}_p\left[\log\left(\frac{r(\mathbf{y}|\boldsymbol{\theta})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)\right]. \end{aligned} \tag{19}$$

We read from Eq. (19) that in order to compare two different models with distributions $q(\mathbf{y}|\boldsymbol{\theta})$ and $r(\mathbf{y}|\boldsymbol{\theta})$, respectively, we just need the sample average of the logarithm of the

ratio r/q conditional to p. Moreover, we can use the approximation (18) to compare the two model distributions in terms of likelihood, hence:

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(p \parallel q) - \mathcal{D}_{\mathrm{KL}}(p \parallel r) &= \frac{1}{n}\sum_{i=1}^{n}(\log p(y_i) - \log q(y_i|\boldsymbol{\theta})) - \frac{1}{n}\sum_{i=1}^{n}(\log p(y_i) - \log r(y_i|\boldsymbol{\theta})) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\log r(y_i|\boldsymbol{\theta}) - \log q(y_i|\boldsymbol{\theta})) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{r(y_i|\boldsymbol{\theta})}{q(y_i|\boldsymbol{\theta})}\right) = \frac{1}{n}\log\left(\frac{\prod_{i=1}^{n} r(y_i|\boldsymbol{\theta})}{\prod_{i=1}^{n} q(y_i|\boldsymbol{\theta})}\right) \\ &= \frac{1}{n}\log\left(\frac{L_r(\mathbf{y}|\boldsymbol{\theta})}{L_q(\mathbf{y}|\boldsymbol{\theta})}\right),\end{aligned}$$

where the ratio inside the brackets of the last term is the likelihood ratio for r and q distributions, respectively, and can be used to test the goodness of fit. Let us point out that the real distribution p has been eliminated.

2.4 Akaike Information Criterion

We have seen that MLE provides a mechanism for estimating the optimal parameters of a model with specific dimension (number of parameters k) and str ucture (distribution model). However, MLE does not say anything about the number of parameters k that should be used to optimize the predictions. *Akaike Information Criterio* (AIC) is introduced in 1973 and provides a framework in which the optimal model dimension is also unknown and must be estimated from the data. Thus, AIC proposes a method where both model estimation (optimal parameters $\boldsymbol{\theta}_{\mathrm{MLE}}$) and selection (optimal number of parameters k) are simultaneously accomplished. The idea behind the AIC is that by continuing to add parameters to a model it fits a little bit better, but there is a trade off with *overfitting* and actually we begin losing information about the real data. Hence, AIC represents a trade off between the number of parameters k that we add and the increase of error; the less information a model loses, the higher the quality of that model.

AIC is derived as an asymptotic approximation of KL divergence $\mathcal{D}_{\mathrm{KL}}(p \parallel q)$ between the model generating the data ("real" model) and the fitting candidate model of the interest, which are described by the distributions $p(\mathbf{y})$ and $q(\mathbf{y}|\boldsymbol{\theta})$, respectively. As we discussed in Sec. 2.2, the KL divergence cannot be estimated directly since we do not know the real distribution $p(\mathbf{y})$ that generates the data. AIC serves as an estimator of the expected KL divergence $\mathcal{D}_{\mathrm{KL}}(p \parallel q)$ and is justified in a very general framework; thus, it offers a crude estimator of the expected KL divergence. In particular, in instances where the sample size n is large and the dimension of the model (k) is relatively small, AIC serves as an approximated unbiased estimator. On the other hand, when k is comparable to the sample size n the AIC is characterized by a large negative bias and subsequently, its effectiveness as criterion is reduced.

We suppose that we have a set of parameters $\boldsymbol{\theta}_{\mathrm{MLE}}$ that maximizes the likelihood, but its size k is yet unknown. The AIC criterion provides a way to estimate how many

parameters (the size of $\boldsymbol{\theta}_{\text{MLE}}$) should be used in order to maximize the likelihood. More specifically, in previous sections we present a method, in the context of MLE, to estimate the parameters $\boldsymbol{\theta}_{\text{MLE}}$ that maximize the likelihood for a given number of parameters. We are going further by considering that we know the $\boldsymbol{\theta}_{\text{MLE}}$ and seek the optimal number of parameters k. For instance, in polynomial regression models, where the response variable $\mathbf{y}$ is approximated by a k-th order polynomial such as

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij},$$

we would like to estimate the optimal k. We essentially want to *select the model* that best describes our data (a different k denotes a different model). In the particular case of polynomial regression models, when k is smaller than the optimal more parameters would improve the prediction. On the other hand, when k becomes larger than the optimal parameter dimension we have *overfitting* and thus, the model cannot make good predictions. Besides the polynomial model, in which different order k corresponds to different model, we can compare between similar distribution models, like a mixture of Gaussians, or between different families of distribution models.

Suppose that we have some models $\mathcal{M}_1, ..., \mathcal{M}_k$, where each one is a set of densities given by the model distribution $q(\mathbf{y}|\boldsymbol{\theta}^{(j)})$. Let $\boldsymbol{\theta}^{(j)}_{\text{MLE}}$ be the MLE parameters that maximize the likelihood of the empirical distribution for the model j, hence

$$\left.\frac{\partial \ell(\boldsymbol{\theta}^{(j)})}{\partial \boldsymbol{\theta}^{(j)}}\right|_{\boldsymbol{\theta}^{(j)}_{\text{MLE}}} = \left[\frac{\partial}{\partial \boldsymbol{\theta}^{(j)}} \frac{1}{n} \sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}^{(j)})\right]_{\boldsymbol{\theta}^{(j)}_{\text{MLE}}} = 0 \tag{20}$$

The model with smallest KL divergence $\mathcal{D}_{\text{KL}}\left(p \parallel \hat{q}_j\right)$ should be the best model, where from Eq. (14) the KL divergence is

$$\mathcal{D}_{\text{KL}}\left(p \parallel \hat{q}_j\right) = \int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} - \int p(\mathbf{y}) \log q_j(\mathbf{y}|\boldsymbol{\theta}^{(j)}_{\text{MLE}}) d\mathbf{y}, \tag{21}$$

where $\hat{q}_j = q_j(\mathbf{y}|\boldsymbol{\theta}^{(j)}_{\text{MLE}})$. The first term in Eq. (21) is independent of the model j and its parameters $\boldsymbol{\theta}^{(j)}$. So, minimizing the KL divergence $\mathcal{D}_{\text{KL}}\left(p \parallel \hat{q}_j\right)$ over j essentially means maximizing the second term of Eq. (21), which we call K_j and define as:

$$K_j = \int p(\mathbf{y}) \log q_j(\mathbf{y}|\boldsymbol{\theta}^{(j)}_{\text{MLE}}) d\mathbf{y}. \tag{22}$$

We need to estimate K_j. We may use the empirical distribution approach, as in section 2.2, to obtain

$$\bar{K}_j = \frac{1}{n} \sum_{i=1}^{n} \log q_j(y_i|\boldsymbol{\theta}^{(j)}_{\text{MLE}}) = \frac{\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}})}{n}, \tag{23}$$

where $\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}})$ is the maximum log-likelihood for the j-th model and n denotes the number of the observations. However, Akaike noticed that this estimate is very biased because the

data are being used twice: first for the MLE to get $\boldsymbol{\theta}^{(j)}_{\text{MLE}}$ and second to evaluate the integral. He showed that the bias is related to the dimension of parameters and approximately given by k_j/n, where k_j is the dimension of the parameters for the j-th model. As we prove in the end of this section, the integral (22) asymptotically yields

$$\begin{aligned} K_j &= \bar{K}_j - \frac{k_j}{n} \\ &= \frac{\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}})}{n} - \frac{k_j}{n}. \end{aligned}$$

By using the last result, we define the Akaike information criterion as

$$\text{AIC}(j) = 2nK_j \tag{24}$$

$$= 2\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}}) - 2k_j. \tag{25}$$

We notice that maximizing the AIC(j) is the same as maximizing K_j over j. The multiplication factor $2n$ is introduced for historical reasons and does not actually play any role in the maximization of the AIC. Finally, we point out a very important feature: the goal of AIC is to select the best model for a given dataset without assuming that the "true" data sample, or the data generating process p, is in the family of the fitting models from which are selecting. Hence, AIC is a very general and powerful *selection model* tool.

In the followings, we present the derivation of AIC, which is a long derivation and requires asymptotic analysis and some further assumptions. The key point in this derivation is to estimate the deviation of the empirical formula (23) from the correct (22). For simplicity, we focus on a single model and drop the subscript j, hence we need to estimate the difference $\bar{K} - K$. First of all, we assume that $\boldsymbol{\theta}_{\text{MLE}}$ maximize the likelihood of the empirical distribution p, but it is not the correct optimal parameters set for our model distribution q, in other words $\boldsymbol{\theta}_{\text{MLE}}$ maximizes the $\bar{K}$ of Eq. (23) but not the K of Eq. (22). Furthermore, we suppose that $\boldsymbol{\theta}_0$ is a set of parameters that maximizes the model likelihood and, in turn, the K. Since $\boldsymbol{\theta}_0$ is an extrema of the log-likelihood of the model distribution q and $\boldsymbol{\theta}_{\text{MLE}}$ is in the neighborhood of $\boldsymbol{\theta}_0$, we expand the expressions (22), (23) around $\boldsymbol{\theta}_0$. First of all let us define some useful formulas.
The log-likelihood for the model distribution q:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}). \tag{26}$$

The *score function*, which is $k \times 1$ vector distribution:

$$s(y|\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}). \tag{27}$$

The *Hessian*, which is a $k \times k$ matrix of the second derivatives:

$$H(y|\boldsymbol{\theta}) = \nabla \nabla^T \ell(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta_\mu \partial \theta_\nu} \sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}). \tag{28}$$

The *Fisher Information matrix*:

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_p\left[H(y|\boldsymbol{\theta})\right], \tag{29}$$

and the summation:

$$\mathcal{I}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} H(y_i|\boldsymbol{\theta}), \tag{30}$$

where in the large population limit ($n \to \infty$) the Fisher information matrix (29) and the summation formula (30) approximately become equal, hence in this derivation we consider:

$$\mathcal{I}(\boldsymbol{\theta}) \simeq \mathcal{I}_n(\boldsymbol{\theta}). \tag{31}$$

Let us make some assumptions here regarding the "real" ideal parameter set $\boldsymbol{\theta}_0$:
By model construction and the central limit theorem (CLM), the score function is considered to follow a normal distribution as:

$$\mathbb{E}_p\left[s(\mathbf{y}|\boldsymbol{\theta}_0)\right] = 0, \tag{32}$$
$$\text{var}\left[s(\mathbf{y}|\boldsymbol{\theta}_0)\right] = V. \tag{33}$$

We define the sum:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} s(y_i|\boldsymbol{\theta}_0), \tag{34}$$

where by CLT we read:

$$\sqrt{n}S_n = \mathcal{N}(S_n|0, V). \tag{35}$$

Since $\boldsymbol{\theta}_{\text{MLE}}$ is in the neighborhood of $\boldsymbol{\theta}_0$, it is reasonable to consider that $(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)$ is a random vector obtained by a normal distribution as,

$$Z = \sqrt{n}\left(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0\right), \tag{36}$$
$$Z_i = \mathcal{N}\left(Z_i|0, V_Z\right)). \tag{37}$$

We intuitively suspect that Z and S_n are correlated and thus, V_Z can be expressed in terms of V. Along the derivation we will find an approximately relationship between V and V_Z.

Furthermore, we prove two mathematical properties that we are going to use in the derivation. Assuming a constant $k \times k$ matrix A with elements a_{ij} and a $1 \times k$ random column vector $\mathbf{b}$ with mean $\boldsymbol{\mu}$ and with symmetric covariant matrix Σ of elements σ_{ij}, we have the following identities:

$$\text{var}\left[A\mathbf{b}\right] = A\ \text{var}\left[\mathbf{b}\right]A^T. \tag{38}$$

$$\mathbb{E}\left[\mathbf{b}^T A\mathbf{b}\right] = \text{Tr}\left[A\Sigma\right] + \boldsymbol{\mu}^T A\boldsymbol{\mu}, \tag{39}$$

where Tr [.] denotes the trace. The second identity reads:
Proof of Eq. (38) :

$$\begin{aligned}\text{var}\left[A\mathbf{b}\right] &= \mathbb{E}\left[A(\mathbf{b}-\boldsymbol{\mu})\left(A(\mathbf{b}-\boldsymbol{\mu})\right)^T\right] \\ &= \mathbb{E}\left[A(\mathbf{b}-\boldsymbol{\mu})\left(\mathbf{b}-\boldsymbol{\mu}\right)^T A^T\right] \\ &= A\mathbb{E}\left[(\mathbf{b}-\boldsymbol{\mu})\left(\mathbf{b}-\boldsymbol{\mu}\right)^T\right]A^T \\ &= A\ \text{var}\left[\mathbf{b}\right]A^T. \blacksquare\end{aligned}$$

Proof of Eq. (39) :

$$\begin{aligned}\mathbb{E}\left[\mathbf{b}^T A\mathbf{b}\right] &= \sum_i\sum_j a_{ij}b_i b_j = \sum_i\sum_j a_{ij}\mathbb{E}\left[b_i b_j\right] \\ &= \sum_i\sum_j a_{ij}\left(\sigma_{ij}+\mu_i\mu_j\right) \\ &= \sum_i [A\Sigma]_{ii} + \boldsymbol{\mu}^T A\boldsymbol{\mu} \\ &= \text{Tr}\left[A\Sigma\right] + \boldsymbol{\mu}^T A\boldsymbol{\mu},\end{aligned}$$

where we used the linearity of the expectation and the covariant formula for two dependent variables:

$$\mathbb{E}\left[X\cdot Y\right] = \text{Cov}[X\cdot Y] + \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]. \blacksquare$$

We proceed with the expansion of "real" and empirical KL divergences K and $\bar{K}$ of Eqs. (22) and (23), respectively. We first expand Eq. (22) around $\boldsymbol{\theta}_0$:

$$\begin{aligned}K &= \int p(\mathbf{y})\log q(\mathbf{y}|\boldsymbol{\theta}_{\text{MLE}})d\mathbf{y} = \int p(\mathbf{y})\left[\log q(\mathbf{y}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MLE}}} d\mathbf{y} \\ &\simeq \int p(\mathbf{y})\left(\log q(\mathbf{y}|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log q(\mathbf{y}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0} + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\left[\nabla\nabla^T\log q(\mathbf{y}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)_{\boldsymbol{\theta}_{\text{MLE}}} d\mathbf{y} \\ &= \int p(\mathbf{y})\left(\log q(\mathbf{y}|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T s(\mathbf{y}|\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T H(\mathbf{y}|\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)_{\boldsymbol{\theta}_{\text{MLE}}} d\mathbf{y} \\ &= \int p(\mathbf{y})\left(\log q(\mathbf{y}|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}_{\text{MLE}}-\boldsymbol{\theta}_0)^T s(\mathbf{y}|\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}_{\text{MLE}}-\boldsymbol{\theta}_0)^T H(\mathbf{y}|\boldsymbol{\theta}_0)(\boldsymbol{\theta}_{\text{MLE}}-\boldsymbol{\theta}_0)\right)d\mathbf{y} \\ &= \mathbb{E}_p\left[\log q(\mathbf{y}|\boldsymbol{\theta}_0)\right] + \frac{1}{\sqrt{n}}Z^T\mathbb{E}_p\left[s(\mathbf{y}|\boldsymbol{\theta}_0)\right] + \frac{1}{2n}Z^T\mathbb{E}_p\left[H(\mathbf{y}|\boldsymbol{\theta}_0)\right]Z \\ &= K_0 - \frac{1}{2n}Z^T\mathcal{I}(\boldsymbol{\theta}_0)Z,\end{aligned} \tag{40}$$

where the second term is dropped out because of Eq. (32) and K_0 is defined as:

$$K_0 = \mathbb{E}_p\left[\log q(\mathbf{y}|\boldsymbol{\theta}_0)\right] = \int p(\mathbf{y})\log q(\mathbf{y}|\boldsymbol{\theta}_0)d\mathbf{y}. \tag{41}$$

We proceed with the expansion of Eq. (23) around $\boldsymbol{\theta}_0$:

$$\begin{aligned}
\bar{K}_j &= \frac{1}{n}\sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}_{\text{MLE}}) = \left[\frac{1}{n}\sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_{\text{MLE}}} \\
&\simeq \left[\frac{1}{n}\sum_{i=1}^{n} \left(\log q(y_i|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T s(y_i|\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T H(y_i|\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)\right]_{\boldsymbol{\theta}_{\text{MLE}}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}_{\text{MLE}}-\boldsymbol{\theta}_0)^T \frac{1}{n}\sum_{i=1}^{n} s(y_i|\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}_{\text{MLE}}-\boldsymbol{\theta}_0)^T \frac{1}{n}\sum_{i=1}^{n} H(y_i|\boldsymbol{\theta}_0)\,(\boldsymbol{\theta}_{\text{MLE}}-\boldsymbol{\theta}_0) \\
&= K_0 + A_n + \frac{1}{\sqrt{n}} Z^T S_n - \frac{1}{2n} Z^T \mathcal{I}_n(\boldsymbol{\theta}_0) Z. \qquad (42)
\end{aligned}$$

In the last step we used that for an arbitrary function $f(\mathbf{y})$:

$$\mathbb{E}\left[f(\mathbf{y})\right] \simeq \frac{1}{n}\sum_i f(y_i),$$

and thus, by using the CLT the expectation can be approximately written as

$$\mathbb{E}\left[f(\mathbf{y})\right] = \frac{1}{n}\sum_i f(y_i) + \xi_n,$$

where ξ_n is a gaussian random term with zero mean. In this sense, we define:

$$\begin{aligned}
A_n &= \frac{1}{n}\sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta}_0) - K_0, \\
\mathbb{E}_p\left[A_n\right] &= 0. \qquad (43)
\end{aligned}$$

Using the approximatios (31) and (43) the difference $\bar{K} - K$ of Eqs. (40) and (42) reads:

$$\bar{K} - K \simeq A_n + \frac{1}{\sqrt{n}} Z^T S_n, \qquad (44)$$

and its expectation

$$\begin{aligned}
\mathbb{E}_p\left[\bar{K} - K\right] &= \mathbb{E}_p\left[A_n\right] + \frac{1}{\sqrt{n}}\mathbb{E}_p\left[Z^T S_n\right] \\
&= \frac{1}{\sqrt{n}}\mathbb{E}_p\left[Z^T S_n\right]. \qquad (45)
\end{aligned}$$

We are seeking for a bias term, so we interested in Eq. (45). As we mentioned when we define Z, there should be a correlation between S_n and Z, let us work on that by expanding

S_n around $\boldsymbol{\theta}_{\text{MLE}}$, we have:

$$
\begin{aligned}
S_n &= \frac{1}{n}\sum_{i=1}^{n} s(y_i|\boldsymbol{\theta}_0) = \left[\frac{1}{n}\sum_{i=1}^{n} s(y_i|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0} \\
&= \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log q(y_i|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0} = \left[\frac{1}{n}\frac{\partial}{\partial \boldsymbol{\theta}}\sum_{i=1}^{n} \log q(y_i|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0} \\
&\simeq \left[\frac{1}{n}\frac{\partial}{\partial \boldsymbol{\theta}}\sum_{i=1}^{n} \left(\log(y_i|\boldsymbol{\theta}_{\text{MLE}}) + (\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MLE}})^T s(y_i|\boldsymbol{\theta}_{\text{MLE}}) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MLE}})^T H(y_i|\boldsymbol{\theta}_{\text{MLE}})(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MLE}})\right)\right]_{\boldsymbol{\theta}_0} \\
&= \left[\frac{1}{2n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MLE}})^T H(y_i|\boldsymbol{\theta}_{\text{MLE}})(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MLE}})\right]_{\boldsymbol{\theta}_0} \\
&= \left[\frac{1}{n}\sum_{i=1}^{n} H(y_i|\boldsymbol{\theta}_{\text{MLE}})(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MLE}})\right]_{\boldsymbol{\theta}_0} \\
&= \frac{1}{n}\sum_{i=1}^{n} H(y_i|\boldsymbol{\theta}_{\text{MLE}})(\boldsymbol{\theta}_0-\boldsymbol{\theta}_{\text{MLE}})
\end{aligned}
$$

hence,

$$
S_n \simeq \frac{1}{\sqrt{n}}\mathcal{I}(\boldsymbol{\theta}_{\text{MLE}})Z, \tag{46}
$$

where we used the the MLE property (20) to drop out the score function term. Now we can estimate the correlation between S_n and Z by taking the variances of Eq. (46) and using the distribution functions (35), (37) and the identity (38):

$$
\begin{aligned}
\text{var}\left[\mathcal{I}Z\right] &= \text{var}\left[\sqrt{n}S_n\right] \\
\mathcal{I}^T \text{var}\left[Z\right]\mathcal{I} &= V \\
\mathcal{I}^T V_Z \mathcal{I} &= V \\
V_Z &= \mathcal{I}^{-1} V \mathcal{I}^{-1}.
\end{aligned} \tag{47}
$$

Getting back in the central Eq. (44) and using the relationship (46) and the identity (39), we obtain:

$$
\begin{aligned}
\mathbb{E}_p\left[\bar{K}-K\right] &= \frac{1}{\sqrt{n}}\mathbb{E}_p\left[Z^T S_N\right] \\
&= \frac{1}{n}\mathbb{E}_p\left[Z^T \mathcal{I} Z\right] \\
&= \frac{1}{n}\text{Tr}\left[\mathcal{I}V_Z\right] + \mathbb{E}_p\left[Z^T\right]\mathcal{I}\,\mathbb{E}_p\left[Z\right] \\
&= \frac{1}{n}\text{Tr}\left[\mathcal{I}\mathcal{I}^{-1}V\mathcal{I}^{-1}\right] + 0 \\
&= \frac{1}{n}\text{Tr}\left[V\mathcal{I}^{-1}\right],
\end{aligned}
$$

which is the bias term, thus we get

$$K \simeq \bar{K} - \frac{1}{n}\mathrm{Tr}\left[V\mathcal{I}^{-1}\right]. \tag{48}$$

In the last step of this long derivation we take the limit that our model is correct, i.e. $\boldsymbol{\theta}_{\mathrm{MLE}} = \boldsymbol{\theta}_0$. In this limit, the Fisher information matrix $\mathcal{I}$ is equal to the variance of the score function, namely $\mathcal{I} = V$. Subsequently, the term $V\mathcal{I}^{-1}$ is the $k \times k$ identity matrix $\mathbf{I}$ with trace equal to the parameters dimension k, hence $\mathrm{Tr}\,[\mathbf{I}] = k$. As a result, we obtain the correct biased estimator of the KL divergence which is the AIC:

$$K \simeq \bar{K} - \frac{k}{n}. \tag{49}$$

The derivation requires many approximations and assumptions, and thus AIC is a very crude criterion. Nevertheless, it is still a very useful tool based on a very clever idea and inspires new more efficient information criteria that demand fewer assumptions, such as the *corrected Akaike Information Criterion* and the *Bayesian Information Criterion*.

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*, 8th ed. Springer (2008).

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 8th ed., Springer (2017).

[3] A. Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley (2015).