Contents

1
Model Selection

In this chapter we consider parametric models, with special emphasis on multiple regression models. In very general terms, consider a family of models $\{M_\gamma, \gamma \in \Gamma\}$. Given data, the goal is to select the best model in terms of some criterion, such as predictive power of future observations, or best fit. Sometimes the goal seems to be stated as that of choosing the "true" model that generated the data. The complexity of real data suggests that this goal is not a realistic. Even if a true model exists, it will generally not be on our list of models, and one must settle for a good and useful approximation to reality. Models are used for inference, prediction and discrimination, and as tools for understanding the data's structure and the relations between variables. Even if one can conceive of a true full model, in the presence of a data set of given size, selecting a smaller "parsimonious" model may result in better prediction or inference. Some of the discussion below is based on Claeskens and Hjort (2008), Burnham and Anderson (2002), and Konishi and Kitagawa (2007), texts dedicated to model selection.

As a first example, consider multiple regression with k potential covariates. Let Γ denote the collection of all subsets $\{1, \ldots, k\}$, and for $\gamma \in \Gamma$ let M_γ be the multiple regression model that uses the covariates $\{X_j\}$ for $j \in \gamma \subseteq \{1, \ldots, k\}$.

For a second example consider a density $f(x)$, or a regression function, $r(x) = \mathrm{E}(Y \mid X = x)$ for x in some interval. Assume that $f(x)$, or $r(x)$, can be represented in the form $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$, where $\phi_1, \phi_2, \ldots$ is a given basis for the class of density or regression functions being considered. One can estimate the unknown parameters β_j; however, an estimator such as $\widehat{f}(x) = \sum_{j=1}^{\infty} \widehat{\beta}_j \phi_j(x)$ is usually inefficient due the presence of a large number of unknown parameters, and one would do better by selecting a finite subset of basis functions among $\phi_1, \phi_2, \ldots$, and estimate only their corresponding parameters.

Use of models such as regression lead to questions of the connection between correlation and causality. We discuss it briefly here. Suppose we observe a sample of (X, Y) values from a given distribution or population. We may find that the variables are positively correlated, say. This does not necessarily mean that a higher X causes a higher Y, and that it advisable to increase X in order to increase Y if increasing Y is desirable. In this chapter we discuss prediction, say of Y from X, under the given distribution from which the sample was taken; causality does not enter.

For a simple example, suppose we study the relation between a certain health variable Y, and a variable indicating taking a certain vitamin, X. A person's health condition and vitamin taking may be related in different ways and for different reasons, such as their awareness of their own health, which affects both Y and X, even in situations where the vitamin is useless. In such a case X is an *endogenous* variable We may succeed in predicting Y from X in the given population, but may fail under other distributions, including the one we create artificially by advising on taking the vitamin, and thus changing X.

An experiment in which subjects are allocated by the researcher into groups taking the vitamin or not, rather than their own self-selection, makes X *exogenous* if the allocation is independent of health conditions, say by using *random allocation.* From such an experiment, if the only difference between the groups is in taking the vitamin or not, one may be able to infer causality.

In this section we study prediction or forecasting rather than causal inference. The goal is to establish criteria and methods for model selection. Here we discuss some general principles and apply them to some particular models; further examples will be given in other specialized sections.

1.1 Stepwise selection

In this section we describe two procedures for selecting covariates for multiple linear regression models,both of which are implemented by simple steps. The procedures described here are known as *backward elimination and forward selection.* It appears in every statistical package of linear regression, and is used very often. With some natural adjustments, it is also used for non-linear models, such as Logistic Regression.

Given a set of potential independent variables from which we want to extract a subset for use in a forecasting model, the backward elimination and forward procedures can be described as follows. We either start with all potential variables in the model and proceed backwards, eliminating one variable at a time, or we begin with no variables in the model and proceed forward, selecing one variable to add. At each step, we perform the following calculations: for each variable currently in the model, compute the F-statistic described in order to decide if it should stay in the model. The resulting F-statistics is called "F-to-remove"; for each variable not in the model, compute the F-statistic to decide whether it should be added, and reports it as its"F-to-enter" statistic. At the next step, enter the variable with the highest F-to-enter statistic, or remove the variable with the lowest F-to-remove statistic, in accordance with certain specified control parameters which define how large (small) should the F-to-enter (F-to-remove) be in order to enter (remove) a variable.

Such procedures often lead to a reasonable choice of a subset of variables, without having to explore all subsets, a procedure that may be too lengthy. However, there is no proof or principle the guarantees a good choice is made by these two procedures, and indeed, they may return different final models. Multiple testing is performed, which may lead to models that are too large, and the relation between the significance of single variables as tested, and the best prediction model, is not clear.

1.2 Mallows' C_p

Let X be a $n \times k$ matrix of covariates and consider the model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathrm{Var}(\mathbf{Y} \mid X) = \sigma^2 I$. This model is assumed to be correct, or more realistically, a good approximation. However, for a given sample size, a model that does not use all available variables may yield better predictions than the full model. This may be due to the large number of parameters in the full model, and in particular if the columns of X are nearly dependent, high variance of the estimators are obtained due to multicollinearity. In general, models that are too large may overfit the given data, and provide poor prediction.

We therefore consider using subsets $\gamma \subseteq \{1, \dots, k\}$ and models that consist of the variables with indices in γ, $\mathbf{Y} = X_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}$, where X_γ is the matrix formed by the columns of X with indices in γ, and $\boldsymbol{\beta}_\gamma$ formed in the same manner from $\boldsymbol{\beta}$. Let

$$\widehat{\mathbf{Y}}_\gamma = X_\gamma \widehat{\boldsymbol{\beta}}_\gamma \quad \text{where} \quad \widehat{\boldsymbol{\beta}}_\gamma = (X_\gamma^\mathsf{T} X_\gamma)^{-1} X_\gamma^\mathsf{T} \mathbf{Y}. \tag{1.1}$$

Suppose we predict new unobserved variables Y_i^*, independent of the observed Y_i, but having the same p-vector of covariates $\mathbf{X}_i$ (the ith row of X) using only the variables in γ, $\mathbf{X}_{i\gamma}$, that is, using the predictor $\widehat{\mathbf{Y}}_{i\gamma} = \mathbf{X}_{i\gamma}\widehat{\boldsymbol{\beta}}_\gamma$. A natural 'cross-validation' type measure of the quality of prediction would be the loss $L(\gamma) = \sum_{i=1}^n (Y_i^* - \widehat{\mathbf{Y}}_{i\gamma})^2$. Note also that under normality, $-L(\gamma)$ is proportional to the log-likelihood of the Y_i^*'s, so if we minimize $L(\gamma)$ we are maximizing the estimated likelihood of new data from the same distribution. More specifically, assuming that conditioned on the covariate vector X_i, Y_i and $Y_i^* \sim N(\mu_i, \sigma^2 I)$ independently, the likelihood of the Y_i^*'s is

$$\mathcal{L}(\mathbf{Y} \mid X) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-||\mathbf{Y}^* - \boldsymbol{\mu}||^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_i (Y_i^* - \mu_i)^2/2\sigma^2}, \tag{1.2}$$

and minimizing the sum of squares $L(\gamma)$ is equivalent to maximizing the likelihood $\mathcal{L}(\mathbf{Y}^* \mid X)$ over subsets γ with $\mu_i = \mathbf{X}_{i\gamma}\widehat{\boldsymbol{\beta}}_\gamma$.

We do not observe the additional values Y_i^*, so $L(\gamma)$ is not observed. Instead we consider the expected prediction risk

$$R(\gamma) = \mathrm{E}\{\sum_{i=1}^n (Y_i^* - \widehat{Y}_{i\gamma})^2\}, \tag{1.3}$$

where here and below, all expectations are conditioned on X, so that X is considered fixed. It is natural to select the model with the smallest $R(\gamma)$. We next discuss estimation of $R(\gamma)$. The problem is that $R(\gamma)$ contains unobserved quantities, the Y_i^*'s, and all we observe is the sample $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$.

On the basis of the sample $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ we define the sample error sum of squares for the subset γ to be

$$SS(\gamma) = \sum_{i=1}^n (Y_i - \widehat{Y}_{i\gamma})^2. \tag{1.4}$$

This statistic, often denoted by SSE_γ or SSE_p[1] where $p = |\gamma|$, the size of γ, is generally an underestimate of $R(\gamma)$ since $\widehat{\mathbf{Y}}_{i\gamma}$ is obtained by minimizing this very sum for the given Y_i's for any fixed γ. In order to assess the bias in estimating $R(\gamma)$ by $SS(\gamma)$ one can observe that, using the independence of $\{Y_i^*, i = 1, \ldots, n\}$ and $\{Y_i, i = 1, \ldots, n\}$,

$$R(\gamma) - \mathrm{E}[SS(\gamma)] = 2\sum_{i=1}^n \mathrm{Cov}(Y_i, \widehat{Y}_{i\gamma}), \tag{1.5}$$

see Exercise 1.4.1.

We now compute the above covariance; we suppress the subscript γ in the following calculations. We have $\widehat{Y}_i = \mathbf{X}_i(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{Y}$. Using the formula $\mathrm{Cov}(U, \mathbf{a}^\mathsf{T}\mathbf{V}) = \mathrm{Cov}(U, \mathbf{V})\mathbf{a}$, where $\mathbf{a}$ denote a constant column vector, U a random variable, and $\mathbf{V}$ a random column vector, and noting that $\mathrm{Var}(Y_i, \mathbf{Y}) = (0, \ldots, 0, \sigma^2, 0, \ldots, 0) = \sigma^2\mathbf{e}_i$, the standard basis vector in direction i, we obtain

$$\begin{aligned}\mathrm{Cov}(Y_i, \widehat{Y}_{iA}) &= \mathrm{Cov}(Y_i,\, \mathbf{X}_i(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{Y}) \\ &= \sigma^2\mathbf{e}_i X(X^\mathsf{T}X)^{-1}\mathbf{X}_i^\mathsf{T} = \sigma^2\mathbf{X}_i(X^\mathsf{T}X)^{-1}\mathbf{X}_i^\mathsf{T}.\end{aligned}$$

[1] The letters SSR or RSS are also used

Summing, we obtain

$$\sum_{i=1}^{n} \mathrm{Cov}(Y_i, \widehat{Y}_{i\gamma}) = \sigma^2 \sum_{i=1}^{n} \mathbf{X}_i (X^\top X)^{-1} \mathbf{X}_i^\top$$
$$= \sigma^2 \mathrm{tr}(X(X^\top X)^{-1} X^\top) = \sigma^2 \mathrm{tr}(I_{|\gamma|}) = \sigma^2 |\gamma|, \tag{1.6}$$

see Exercise 1.4.2, where $|\gamma|$ is the size of γ, and $I_{|\gamma|}$ is the identity matrix of this dimension.

We conclude, using (1.5), that an unbiased estimator of $R(\gamma)$ is

$$\widehat{R}(\gamma) = SS(\gamma) + 2|\gamma|\widehat{\sigma}^2 \tag{1.7}$$

whenever $\widehat{\sigma}^2$ is an unbiased estimator of σ^2 such as S^2 computed from the full model, and $SS(\gamma)$ is given by (1.4).

With $|\gamma| = p$, the expression on the right-hand side of (1.7), sometimes expressed in other equivalent forms, is called Mallows C_p, and since it is an estimate of the prediction risk (1.3), from one point of view we should look for a set γ that minimizes $\widehat{R}(\gamma)$.

A common formulation that leads to the same result is given by writing $Y_i^* = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i^*$ with ε_i^* independent of the data, and then

$$R(\gamma) = \mathrm{E}\sum(\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i^* - \mathbf{X}_{i\gamma}\widehat{\boldsymbol{\beta}}_\gamma)^2 = \mathrm{E}\sum(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_{i\gamma}\widehat{\boldsymbol{\beta}}_\gamma)^2 + n\sigma^2. \tag{1.8}$$

If instead of the unobserved $R(\gamma)$, we adopt as our criterion for prediction error the expression

$$\mathrm{E}\sum(\mathrm{E}(Y_i^* \mid X) - \widehat{Y}_{i\gamma})^2 = \mathrm{E}\sum(\mathrm{E}(Y_i \mid X) - \widehat{Y}_{i\gamma})^2 = \mathrm{E}\sum(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_{i\gamma}\widehat{\boldsymbol{\beta}}_\gamma)^2,$$

we can estimate it using (1.7) and (1.8) by $SS(\gamma) + 2|\gamma|\sigma^2 - n\sigma^2$. Dividing by σ^2 and plugging again its estimator S^2, we obtain the statistic

$$C_\gamma = SS(\gamma)/S^2 - n + 2|\gamma|,$$

see Mallows (1973), and which is equivalent to (1.7), and minimized by the same γ.

One should expect $\mathrm{E}SS(\gamma)/(n-p) \geq \sigma^2$ with near equality if the set γ provides a good model. Note that for the full model $S^2 = ||\mathbf{y} - X\widehat{\boldsymbol{\beta}}||^2/(n-k)$ is an unbiased estimator of σ^2. If the set γ provides a good model then $SS(\gamma)/(n-p) = ||\mathbf{y} - X\widehat{\boldsymbol{\beta}}_\gamma||^2/(n-p)$ will have a numerator close to that of S^2, and larger otherwise due to the non-inclusion of explanatory covariates, which makes the fit worse. Since $n - k < n - p$ the inequality follows.

Therefore we should expect to have $SS(\gamma)/S^2 \geq n - p$ and therefore $C_p \geq p$, with near equality for a good choice of γ. This suggests looking for a model with C_γ close to $|\gamma|$, or slightly larger, rather than minimizing C_γ, which may sometimes be much smaller than p.

Note that C_γ depends on γ through two terms. The first, $SS(\gamma)$ measures lack of fit, and the second, $2|\gamma|$, can be seen as a penalty for the model size.

The need to balance these two terms is a consequence of the tension between bias and variance, which is ubiquitous in statistical estimation problems.

1.3 Akaike's AIC

The *Kullback-Leibler Divergence*, also called a distance, or a relative entropy, between two densities f and g is defined as

$$D(g||f) = \int g(y) \log\left(\frac{g(y)}{f(y)}\right) dy, \tag{1.9}$$

with an obvious analog in the discrete case. It is clearly non symmetric in f and g, and and one can show that it does not satisfy the triangle inequality. In particular, it is not a metric. However it does measure the discrepancy between g and f and provides a notion of distance that is relevant in statistics. This integral may not be finite, but it is always nonnegative, see Exercise 1.4.3, and is well defined.

To gain some intution about the Kullback-Leibler, or K-L Divergence, consider g as the 'true' density generating a random variable Y, so that we have

$$D(g||f) = E\left[\log\left(\frac{g(Y)}{f(Y)}\right)\right].$$

As $\log(x)$ is a concave function, Jensen's inequality shows that

$$\log E\left(\frac{g(Y)}{f(Y)}\right) \le E\left[\log\left(\frac{g(Y)}{f(Y)}\right)\right]. \tag{1.10}$$

As $\log(x)$ is strictly convex, the lower bound is achieved, that is, the K-L divergence is minimized, if and only if $g(Y)/f(Y)$ is a constant on the support of g, which, as f is a density, implies that $f(y) = g(y)$. As we may write

$$D(g||f) = E\left[\log g(Y)\right] - E\left[\log f(Y)\right],$$

for g any fixed density function, we likewise see that the negative of the second term,

$$A(f) = \int g(y) \log f(y) dy, \tag{1.11}$$

is maximized when f and g are equal.

Given a parametric family of densities $f(y;\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta$, and the data generated by a "true" unknown density g, maximum likelihood estimation can be seen as an approximation to the problem of minimizing the K-L divergence over $\boldsymbol{\theta} \in \Theta$. First, for the parametric $f(\,\cdot\,;\boldsymbol{\theta}))$, write (1.11) as

$$A(\boldsymbol{\theta}) = \int g(y) \log f(y;\boldsymbol{\theta}) dy. \tag{1.12}$$

As (1.11) is maximized at the true density, we would like to choose θ which maximizes (1.12). However, g is unknown and hence (1.12) cannot be calculated.

Yet, given the sample $Y_1, \ldots, Y_n$ from g, the Law of Large Numbers implies that $A(\boldsymbol{\theta})$ is the limit of the expectation $n^{-1}\sum_{i=1}^{n} \log f(Y_j;\, \boldsymbol{\theta})$. So as an approximation to maximizing $A(\boldsymbol{\theta})$, we instead maximize the latter expression, resulting in the MLE. Thus, using data to find the model that has minimum K-L distance to g, that is, the K-L 'projection' of g to the family of models, that is minimizing the K-L divergence between the true model and the parametric one, is, in the words of Akaike (1973) "a natural extension of the classical maximum likelihood principle". See also Akaike (1974).

Our goal is to find a measure that allows us to compare different models, and then to estimate this measure. The Akaike's Information Criterion, or AIC, is based on attempting to select the model that is the nearest to the truth as measured by the K-L divergence. Given models $f_i(y;\boldsymbol{\theta})$ over parameter spaces Θ_i and $\widehat{\boldsymbol{\theta}}_i$ an estimator of the parameter in the ith model, the idea discussed in the following sections is to select a model that maximizes $\int g(y) \log f(y\,; \widehat{\boldsymbol{\theta}})dy$. Besides minimizing the K-L divergence, as explained in the following section, this criterion is equivalent to choosing a model on the basis of given data in order to maximize the expected likelihood of future observations.

1.3.1 Background towards the definition of AIC

Before describing Akaike's criterion we introduce some definitions and make other preparations. Let $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} \int g(y) \log f(y\,;\boldsymbol{\theta})dy$, where $\{f(y\,;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ is a given a family of densities, so that $f(y\,;\boldsymbol{\theta}_0)$ is a type of projection of g to the family. We assume that the maximum $\boldsymbol{\theta}_0$ is unique, and is an interior point of the parameter space Θ, is a d-dimensional subset of $\mathbb{R}^d$.

In view of the discussion in the previous section, it makes sense to propose $A(\boldsymbol{\theta})$ of (1.12) as a measure of the quality of a model and propose how to estimate it. However, since we do not know $\boldsymbol{\theta}_0$ it may make more sense to estimate $A(\boldsymbol{\theta})$ by as best as possible by a suitable $\widehat{\boldsymbol{\theta}}$ and consider instead the measure

$$B = \int g(y) \log f(y\,; \widehat{\boldsymbol{\theta}})dy; \tag{1.13}$$

in fact $\int g(y) \log g(y)dy - B$ is the distance between g and the model in the family that will actually be used. Choosing according to A reflects a search for the best theoretical model, while B is related to choosing the best model to be use with estimated parameters. Following most of the literature we will take this approach, define $\widehat{\boldsymbol{\theta}}$ as the MLE, show that it approaches $\boldsymbol{\theta}_0$, and discuss estimation of B. When comparing models, the larger B, the better the model.

Some regularity conditions on $f(y\,;\boldsymbol{\theta})$ are needed, in the spirit of those required for maximum likelihood estimation, which we assume are in force. Given a sample of i.i.d. variables $Y_1, \ldots, Y_n$ from the unknown density g, let the maximum likelihood estimate be given by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \log f(Y_k; \boldsymbol{\theta}),$$

By the consistency of the MLE, as $n \to \infty$ we have $\widehat{\boldsymbol{\theta}} \overset{p}{\to} \boldsymbol{\theta}_0$, and by a uniform law of large numbers $\frac{1}{n}\sum_{k=1}^{n} \log f(Y_k; \boldsymbol{\theta}) \overset{p}{\to} \int g(y) \log f(y; \boldsymbol{\theta}) dy$ so that that as $n \to \infty$, $A(\widehat{\boldsymbol{\theta}}) \overset{p}{\to} A(\boldsymbol{\theta}_0)$.

Note that if $Y_1^*, \ldots, Y_n^* \sim g$ is an independent copy of the original sample distributed as $Y^* \sim g$, then

$$B = \mathrm{E}[\log f(Y^*\,; \widehat{\boldsymbol{\theta}}) \mid \widehat{\boldsymbol{\theta}}] = \mathrm{E}\left[\frac{1}{n}\sum_{k=1}^{n} \log f(Y_k^*\,; \widehat{\boldsymbol{\theta}}) \,\middle|\, \widehat{\boldsymbol{\theta}}\right].$$

Thus, preferring models having larger values of B amounts to maximizing the likelihood, or expected likelihood, of new observations from distribution g, from the model with $\theta = \widehat{\theta}$, which will be close to θ_0 for large samples. As we cannot calculate B due to g being unknown, we consider estimating $E[B]$, or equivalently, look for an estimator D for which $\mathrm{E}[D - B] = 0$. If new independent observations are available then one can use D of the form either

$$D = \frac{1}{n}\sum_{k=1}^{n} \log f(Y_k^*\,; \widehat{\boldsymbol{\theta}}) \quad \text{or likewise} \quad D = \frac{1}{n}\sum_{k=1}^{n} \log f(Y_k\,; \widehat{\boldsymbol{\theta}}^*), \qquad (1.14)$$

with $\widehat{\boldsymbol{\theta}}^*$ being the MLE based on the Y_k^*'s. However we do not assume that such observations are given.

Next, one might simply consider estimating $E[B]$ by

$$C = \frac{1}{n}\sum_{k=1}^{n} \log f(Y_k\,; \widehat{\boldsymbol{\theta}}).$$

However, C is a positively biased estimator for $E[B]$, having larger bias for larger models, since $\widehat{\boldsymbol{\theta}}$ is chosen to maximize expression of C. To see that C is so biased, note that by the defining property of the MLE, assuming uniqueness here, we have, with positive probality, that

$$\frac{1}{n}\sum_{i=k}^{n} \log f(Y_k; \widehat{\boldsymbol{\theta}}) > \frac{1}{n}\sum_{k=1}^{n} \log f(Y_k; \widehat{\boldsymbol{\theta}}^*),$$

showing that $E[C] > E[D]$,and thus $E[C] > E[B]$. In the next section we evaluate the bias

$$E[C] - E[B] = \mathrm{E}\left[\frac{1}{n}\sum_{k=1}^{n}\log f(Y_k\,;\widehat{\boldsymbol{\theta}})\right] - \mathrm{E}\int g(y)\log f(y\,;\widehat{\boldsymbol{\theta}})dy, \quad (1.15)$$

after somewhat lengthy but instructive calculations. As C is computable from the data we can obtain our criteria, such as Aikaike's Information Criteria as in eq:AIC.d.gamma, by attempting to correct for the known bias.

Akaike's approach chooses the model that maximizes an unbiased estimator of $\mathrm{E}[D]$ for D of (1.14). Since likelihood is random, it is natural to take expectation, and consider $\mathrm{E}[D]$ which coincides with $\mathrm{E}[B]$ of (1.13). This method can also be interpreted in cross-validation terms, since a good model is one that maximizes the likelihood of new data. Thus, it is not surprising that the resulting AIC criterion below is closely related to cross-validation, see, Stone (1977).

1.3.2 Evaluation of the bias

Letting $A = A(\theta_0)$, As $C - B = (C - A) - (B - A)$, we evaluate the bias $\mathrm{E}[C - B]$ by evaluating $\mathrm{E}[C - A]$ (which is of interest of its own due to the above discussion) and $\mathrm{E}[B - A]$, and taking the difference. In the following we let ∂_θ stand for taking the vector of partial derivatives, and likewise ∂_θ^2 produces a matrix of second partial derivatives. Beginning with $C - A$ we make a Taylor expansion around the MLE $\widehat{\boldsymbol{\theta}}$, with the first derivative vanishing at the MLE, and obtain

$$\begin{aligned} E[C - A] &= \frac{1}{n}\sum_{k=1}^{n}\log f(Y_k\,;\widehat{\boldsymbol{\theta}}) - \frac{1}{n}\sum_{k=1}^{n}\log f(Y_k\,;\boldsymbol{\theta}_0) \\ &\approx -\frac{1}{2}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})^{\mathsf{T}}\left[\frac{1}{n}\sum_{k=1}^{n}\partial_\theta^2\log f(Y_k\,;\widehat{\boldsymbol{\theta}})\right](\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}). \quad (1.16) \end{aligned}$$

Note that the expectation of the left-hand side of (1.16) is exactly $\mathrm{E}[C - A]$.

Expanding the integral form as in (1.13), now at $\boldsymbol{\theta}_0$, again the first derivative term vanishes, and we similarly obtain

$$\begin{aligned} E[B - A] &= \int g(y)[\log f(y\,;\widehat{\boldsymbol{\theta}}) - \log f(y\,;\boldsymbol{\theta}_0)]dy \\ &\approx \frac{1}{2}\int g(y)(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})^{\mathsf{T}}\left[\partial_\theta^2\log f(y\,;\boldsymbol{\theta}_0)\right](\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})dy. \quad (1.17) \end{aligned}$$

In order to continue, we study the asymptotic distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. The arguments are similar to those for the MLE except that now the data is not generated by $f(y\,;\boldsymbol{\theta}_0)$, but rather by the density g. Expanding the gradient of the first sum on the left hand side of (1.16) around $\boldsymbol{\theta}_0$ we have

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\sum_{k=1}^{n} \partial_\theta \log f(Y_k\,;\widehat{\boldsymbol{\theta}}) \\ &\approx \frac{1}{n}\sum_{k=1}^{n} \partial_\theta \log f(Y_k\,;\boldsymbol{\theta}_0) + \left[\frac{1}{n}\sum_{k=1}^{n} \partial_\theta^2 \log f(Y_k\,;\boldsymbol{\theta}_0)\right](\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).\end{aligned}$$

Thus,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \approx -\left[\frac{1}{n}\sum_{k=1}^{n} \partial_\theta^2 \log f(Y_k\,;\boldsymbol{\theta}_0)\right]^{-1} \frac{1}{\sqrt{n}}\sum_{k=1}^{n} \partial_\theta \log f(Y_k\,;\boldsymbol{\theta}_0). \tag{1.18}$$

Since by definition $\boldsymbol{\theta}_0$ maximizes $\mathrm{E}[\log f(Y_k\,;\boldsymbol{\theta})]$, the summand on the right hand side of (1.18) has expectation zero, and variance-covariance matrix

$$K = \mathrm{E}\left\{\frac{\partial_\theta f(Y_k\,;\boldsymbol{\theta}_0)\partial_\theta f(Y_k\,;\boldsymbol{\theta}_0)^T}{f^2(Y_k\,;\boldsymbol{\theta}_0)}\right\}, \tag{1.19}$$

where $Y_k \sim g$.

To handle the first term on the right-hand side of (1.18) we have, see Exercise 1.4.4,

$$J = -\mathrm{E}\left\{\partial_\theta^2 \log f(Y_k\,;\boldsymbol{\theta}_0)\right\} = -\mathrm{E}\left\{\frac{\partial_\theta^2 f(Y_k\,;\boldsymbol{\theta}_0)}{f(Y_k\,;\boldsymbol{\theta}_0)}\right\} + K, \tag{1.20}$$

which is obtained by straightforward differentiation under the expectation sign. When $g(y) = f(y\,;\boldsymbol{\theta}_0)$, the first term on the right-hand side of (1.20) vanishes, see Exercise 1.4.4, and we obtain the well known Fisher information identity $J = K$. This is the case in standard MLE theory.

Returning to (1.18), the average inside the inverse in the first term converges in probability to its expectation J. Invoking the Central Limit Theorem to the second term and applying Slutsky's Lemma yields

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \overset{\mathcal{D}}{\to} N(\mathbf{0}, J^{-1}KJ^{-1}). \tag{1.21}$$

If $g(y) = f(y\,;\boldsymbol{\theta}_0)$, the above asymptotic distribution is the usual $N(\mathbf{0}, K^{-1})$.

We now return to (1.16). Using that $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and a uniform law of large numbers, the sum converges to J, so under suitable conditions

$$\begin{aligned}\mathrm{E}(C-A) &\approx \frac{1}{2n}\mathrm{E}\{n(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}})^\mathsf{T} J(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}})\} = \frac{1}{2n}\mathrm{tr}\mathrm{E}\{n(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}})^\mathsf{T} J\} \\ &= \frac{1}{2n}\mathrm{tr}(J^{-1}KJ^{-1}J) = \frac{1}{2n}\mathrm{tr}(J^{-1}K).\end{aligned} \tag{1.22}$$

Turning to (1.17) it is easy to see, Exercise 1.4.5, that the same approximation yields the same result,

$$E[B-A] \approx -\frac{1}{2n}\mathrm{tr}(J^{-1}K). \tag{1.23}$$

Putting all these results together we finally end our calculations with the desired approximation of the bias

$$\mathrm{E}[C-B] \approx \frac{1}{n}\mathrm{tr}(J^{-1}K). \tag{1.24}$$

If the family $\{f(y\,;\boldsymbol{\theta})\}$ contains g, then we must have $g(y) = f(y\,;\boldsymbol{\theta}_0)$ and $J = K$, yielding $\mathrm{tr}(J^{-1}K) = d$. Hence d may be a good approximation for $\mathrm{tr}(J^{-1}K)$ for models that include densities not too far from g. Most users of the AIC follow Akaike, and use d as the approximate bias rather than $\mathrm{tr}(J^{-1}K)$. The application of (1.24) with or without this approximation is discussed in the next section.

1.3.3 Selecting a model by the AIC

Now with Γ some at most countable index set, consider a class of data generating models M_γ given by densities $f_\gamma(y\,;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_\gamma$, $\gamma \in \Gamma$ where Θ_γ denotes the parameter space for the model, M_γ, having dimension d_γ. We can parameterize the family with a single parameter, say, $\eta = (\gamma, \boldsymbol{\theta})$. Again, we observe a sample $Y_1, \ldots, Y_n$, generated by the true model with density g, which may, or may not belong to the above class of densities.

Our goal is to choose an $(\gamma, \boldsymbol{\theta})$ so that the model $f_\gamma(y;\boldsymbol{\theta})$ provides a good approximation to g in the sense of minimizing the K-L divergence, or equivalently, maximizing

$$\int g(y)\log f_\gamma(y\,;\widehat{\boldsymbol{\theta}}_\gamma)dy \quad \text{where} \quad \widehat{\boldsymbol{\theta}}_\gamma = \arg\max_{\boldsymbol{\theta}\in\Theta_\gamma} \frac{1}{n}\sum_{k=1}^{n}\log f_\gamma(Y_k;\boldsymbol{\theta}), \tag{1.25}$$

the MLE under the model M_γ.

Adjusting for the bias as given (1.24) we obtain we see we prefer the model that would maximize

$$\sum_{k=1}^{n}\log f_\gamma(Y_k\,;\widehat{\boldsymbol{\theta}}) - \mathrm{tr}(J_\gamma^{-1}K_\gamma) \tag{1.26}$$

where K_γ and J_γ are as in (1.19) and (1.20), applied to the model M_γ, with parameter space of dimension d_γ. When the sample estimates

$$\widehat{K}_\gamma = \frac{1}{n}\sum_{k=1}^{n}\partial_\theta \log f_\gamma(Y_k\,;\widehat{\boldsymbol{\theta}}_\gamma)\partial_\theta \log f_\gamma(Y_k\,;\widehat{\boldsymbol{\theta}}_\gamma)^T,$$

$$\widehat{J}_\gamma = \frac{1}{n}\sum_{k=1}^{n}\partial_\theta^2 \log f_\gamma(Y_k\,;\widehat{\boldsymbol{\theta}}_\gamma)$$

are plugged into (1.25) we obtain the TIC, Takeuchi (1976).

With the approximation d_γ for $\mathrm{tr}(J_\gamma^{-1}K_\gamma)$, the criterion becomes choosing γ which maximizes Akaike's Information Criterion, which is given by

$$\mathrm{AIC} = \sum_{k=1}^{n} \log f_\gamma(Y_k\,; \widehat{\boldsymbol{\theta}}_\gamma) - d_\gamma. \tag{1.27}$$

There is a tradition of multiplying the above quantity by -2, perhaps in connection with Wilks' Theorem, and minimizing over γ. It is clear from the form (1.27) how larger models, those that may overfit, are penalized.

1.4 Exercises

Exercise 1.4.1 *Prove* (1.5). *Note that from the first term, $R(\gamma)$, one obtains the product of expectations part of the covariance using independence of Y_i^* and Y_i, whereas the expectation of the product is obtained from $\mathrm{E}L(\gamma)$. Other expressions in $R(\gamma) - \mathrm{E}L(\gamma)$ cancel.*

Exercise 1.4.2 *Prove* (1.6). *For the penultimate equality use the formula* $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Exercise 1.4.3 *Prove that the Kullback-Liebler divergence given by* (1.9) *is non-negative for any two densities f and g Hint: Use* (1.10).

Exercise 1.4.4 *Prove* (1.20) *and show that the first term on its right-hand side vanishes.*

Exercise 1.4.5 *Show that* (1.23) *follows from an argument similar to the one for* (1.22).

References

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle." In *Second International Symposium on Information Theory, B. N. Petrov, and F. Csaki, (eds.) Akademiai Kiado, Budapest*, 267–281.

Akaike, H. (1974), "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on*, 19, 716–723.

Burnham, K.P. and D.R. Anderson (2002), *Model selection and multi-model inference: a practical information-theoretic approach*. Springer.

Claeskens, G. and N.L. Hjort (2008), *Model Selection and Model Averaging*, volume 27. Cambridge University Press.

Konishi, S. and G. Kitagawa (2007), *Information criteria and statistical modeling*. Springer.

Mallows, C.L. (1973), "Some comments on c p." *Technometrics*, 15, 661–675.

Stone, M. (1977), "An asymptotic equivalence of choice of model by cross-validation and akaike's criterion." *Journal of the Royal Statistical Society. Series B (Methodological)*, 44–47.

Takeuchi, K. (1976), "Distributions of information statistics and criteria for adequacy of models." *Math. Sci*, 153, 12–18.