Chapter 4

Model Selection Criteria

4.1 Introduction and Examples

We now consider criteria for choosing the best statistical model, from a selection of models available, to describe data. Consider the following examples:

Example 4.1.

Consider a *survival analysis* problem, where $Y_1, \ldots, Y_n$ are the ages at death of n individuals. The aim is to model the survival distribution, from which $Y_1, \ldots, Y_n$ is considered as an i.i.d. sample. Some popular models are:

1. $\mathcal{M}_1$: the suvival times are an observed exponential random sample; density $p(y;\theta) = \theta e^{-\theta y}\mathbf{1}_{[0,+\infty)}$.
2. $\mathcal{M}_2$: the survival times are from a Gamma distribution; $p(y;\alpha,\beta) = \frac{b^a}{\Gamma(a)} y^{a-1}e^{-by}$.
3. $\mathcal{M}_3$; the survival times are from a log-normal distribution; $\log Y \sim N(\mu, \sigma^2)$.

Note that for this example, $\mathcal{M}_1 \subset \mathcal{M}_2$; if θ is the parameter for Model 1, this corresponds to $(a, b) = (1, \theta)$ in Model 2. Model 3, however, is an entirely different two-parameter model. □

Example 4.2.

Suppose that $Y_1, Y_2, \ldots, Y_n$ is data from a stationary time series, which is modelled by an AR(p) process, where p is unknown; that is:

$$Y_t - \phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p} = \epsilon_t$$

where $(\epsilon_t)_{t\in\mathbb{Z}}$ are uncorrelated, each with the same variance σ^2. The parameters are: $(\phi_1, \ldots, \phi_p; \sigma^2)$.

Here, the models are nested and the aim is to choose p. □

Example 4.3.

Gaussian-mixture models:

$$p(y) = \sum_{j=1}^{n} \pi_j \phi(y; \mu, \Sigma_j)$$

where $\phi(.; \mu, \Sigma)$ is the multivariate Gaussian kernel with mean vector μ and covariance matrix Σ; $\sum_{j=1}^{n} \pi_j = 1$, $\pi_j \geq 0$ for each j, the number n, the quantities $\pi_1, \ldots, \pi_n$, $\mu_1, \ldots, \mu_n$ and $\Sigma_1, \ldots, \Sigma_n$ are unknown and have to be chosen to give the best model. □

In some situations, we have a sequence of models $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \mathcal{M}_3 \subseteq \ldots$, where the parameter space for model $\mathcal{M}_i$ is a subspace of the parameter space for model $\mathcal{M}_{i+1}$. The aim of model selection is simply to choose which parameters, from a selection of parameters available, gives the best model.

In other cases, we are comparing models that are completely different.

4.2 Nested Models: Multiple Linear Regression

Recall that in 'Statistics', we considered Gaussian models of the form

$$Y = X\beta + \epsilon$$

where Y is an n-vector, X is an $n \times r$ design matrix of (known) covariates, β is an r vector of (unknown) parameters and $\epsilon \sim N(0, \sigma^2 I_n)$ is a vector of i.i.d. Gaussian errors, each with (unknown) variance σ^2. Recall that $\widehat{\beta}_{OLS} = \widehat{\beta}_{ML} = (X'X)^{-1}X'Y$, and that $\widehat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2)$.

The *fitted values* are given by the vector $\widehat{Y} = X\widehat{\beta}$ and the *residuals* (or errors) are $R := Y - \widehat{Y}$. The error sum of squares is

$$Q_{\text{res}} := R'R.$$

Suppose there are $p + q$ covariates to choose from and we would like to decide whether to take all of them, or only the first q covariates. We then compare two models:

$$\begin{cases} I & Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_j \\ II & Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=p+1}^{q} x_{ij}\beta_j + \epsilon_j \end{cases}$$

Are the additional parameters necessary?

For model I, $r = 1 + p$ and X_1 (the design matrix) is an $n \times p + 1$ matrix; the first column (which corresponds to the parameter β_0) is $\mathbf{1}_n$. Let $\widetilde{X}_1$ denote the $n \times p$ matrix with entries x_{ij} for $1 \leq i \leq n$, $1 \leq j \leq p$ and $\widetilde{X}_2$ the $n \times q$ matrix with entries x_{ij+p} for $1 \leq i \leq n$ and $1 \leq j \leq q$. Then the design matrix for Model I is $X_1 = (\mathbf{1}_n | \widetilde{X}_1)$ while the design matrix for Model II is $X_2 = (\mathbf{1}_n | \widetilde{X}_1 | \widetilde{X}_2)$.

In 'Statistics' we showed that if $H_0 : \beta_{p+1} = \ldots = \beta_{p+q} = 0$ is true, then

$$F := \frac{(Q_{\text{res},I} - Q_{\text{res},II})/q}{Q_{\text{res},II}} \sim F_{p,n-(p+q+1)}.$$

This gives a model selection criterion; for significance level α, we include variables $\beta_{p+1}, \ldots, \beta_{p+q}$ if $F > F_{p,n-(p+q+1);\alpha}$ and we do not include them if $F < F_{p,n-(p+q+1);\alpha}$.

In this situation, we have a *nested model* where it is known that the errors are i.i.d. normal.

4.3 Asymptotic Log Likelihood Ratio and Wald Test

A standard test statistic is the *likelihood ratio* test statistic:

$$\Lambda(x) = \frac{\sup_{\theta\in\Theta_0} L(\theta; x)}{\sup_{\theta\in\Theta} L(\theta, x)}.$$

For i.i.d. sampling, if $\Theta \subseteq \mathcal{R}^k$ and $\Theta_0 \subseteq \mathcal{R}^p$, with $\Theta_0 \subset \Theta$, then under broad hypotheses (covered in the course Statistics), we have the asymptotic result:

$$-2 \lim_{n\to+\infty} \log \Lambda(x) \simeq \chi^2_{k-p}.$$

The important result for multivariate analysis is the multi-dimensional setting:

Theorem 4.4. *Let $\underline{X} = (X_1, \ldots, X_n)$ be a random sample from a regular parametric family with p.d.f. or p.m.f. $p(x, \underline{\theta})$ where $\underline{\theta} \in \Theta \subseteq \mathbb{R}^k$ is a k-dimensional parameter vector (that is, there are p* free *parameters) where the parametrisation is identifiable. Let $\log L(\underline{\theta}, x) = \log p(x, \underline{\theta})$ denote the log likelihood function for a single observation and $\log L(\underline{\theta}, \underline{x}) = \sum_{j=1}^n \log L(\underline{\theta}, x_j)$ the log likelihood function for an observed random sample $(x_1, \ldots, x_n)$. Let $\widehat{\underline{\theta}}_n$ denote the MLE based on $(X_1, \ldots, X_n)$. Assume that the following conditions hold:*

1. *$\widehat{\underline{\theta}}_n$ is* consistent*,*
2. *$\log L(\underline{\theta}, \underline{x})$ is twice differentiable in $\underline{\theta}$,*
3. $$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_j} \log L(\theta, \underline{X}) \right] = 0 \qquad j = 1, \ldots, p,$$
4. $$\mathbb{E}_\theta \left[|\nabla_\theta \log L(\theta, \underline{X})|^2 \right] < +\infty$$
5. *$\sum_{ij} \mathbb{E}_\theta \left[\left| \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta, \underline{X}) \right| \right] < +\infty$*
6. *the information matrix $I(\theta)$ is non singular for each $\theta \in \Theta$.*

Let Θ_0 denote a specification of the first $k-p$ parameters, where $0 \leq p \leq k$. For $\theta \in \Theta_0$, the distribution of the statistic

$$-2\log\lambda(\underline{X}) \overset{n\to+\infty}{\longrightarrow} \chi^2_{k-p}.$$

4.4 Wald Test and Rao Test

Theorem 4.4, applied to the problem of determining whether a reduced model may be appropriate, givs rise to two variants that are frequently found in software; the *Wald test* and the *Rao test*.

The Wald Test Let $k_{p,\alpha}$ denote the value such that $\mathbb{P}(V > k_{p,\alpha}) = \alpha$ for $V \sim \chi^2_p$. For a p dimensional parameter space $\Theta \subseteq \mathbb{R}^p$ and a parameter vector vector $\theta \in \Theta$, assume that $\sqrt{n}\left(\widehat{\underline{\theta}}_n - \theta\right) \overset{n\to+\infty}{\longrightarrow}_{(d)} N(\underline{0}, I^{-1}(\underline{\theta}))$ where $I(\underline{\theta})$ denotes the Fisher information matrix for a single observation. *Wald's test* rejects $H_0 : \underline{\theta} = \underline{\theta}_0$ ($\theta_0 \in \mathbb{R}^p$ - a single point) in favour of $H_1 : \underline{\theta} \neq \underline{\theta}_0$ when

$$W_n(\underline{\theta}_0) := n(\widehat{\underline{\theta}} - \underline{\theta}_0)^t I(\underline{\theta}_0)(\widehat{\underline{\theta}} - \underline{\theta}_0) \geq k_{p,\alpha}.$$

This test has asymptotic level α.

This may be extended to subsets of the parameter space. The following is left as an exercise: Let $\underline{\theta} = \binom{\underline{\theta}^{(1)}}{\underline{\theta}^{(2)}}$ where $\underline{\theta}^{(1)}$ is a $p-r$ vector and $\underline{\theta}^{(2)}$ an r vector. Let

$$I^{-1} = \begin{pmatrix} I^{11}_{-} & I^{12}_{-} \\ I^{21}_{-} & I^{22}_{-} \end{pmatrix}.$$

Let

$$\widetilde{W}_n(\underline{\theta}^{(1)}_0) = n(\widehat{\underline{\theta}}^{(1)} - \widehat{\underline{\theta}}^{(1)})^t (I^{11}_{-})^{-1}(\widehat{\underline{\theta}}^{(1)} - \widehat{\underline{\theta}}^{(1)}).$$

Then, under the null hypothesis $H_0 : \theta^{(1)} = \theta^{(1)}_0$ (a specification of $p-r$ parameters)

$$\widetilde{W}_n(\underline{\theta}^{(1)}_0) \overset{n\to+\infty}{\longrightarrow}_{(d)} \chi^2_{p-r}.$$

The Rao Score Test Let W be a k-random vector satisfying $W \sim N(\underline{0}, \Sigma)$ where Σ is of rank k then $W^t\Sigma^{-1}W \sim \chi^2_k$. Let

$$\underline{\psi}_n(\underline{\theta}) := \nabla\rho_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} \nabla_\theta \log L(\underline{\theta}, X_j)$$

where $\log L$ is the log likelihood function for a single observation. Then, provided the likelihood satisfies the assumptions of Theorem 4.4, if $\underline{\theta} = \underline{\theta}_0$,

$$\sqrt{n}\underline{\psi}_n(\underline{\theta}_0) \longrightarrow_{(d)} N(0, I(\underline{\theta}_0)).$$

It follows that if $I(\underline{\theta}_0)$ is $p \times p$ and positive definite, then

$$R_n(\underline{\theta}_0) := n\underline{\psi}(\underline{\theta}_0)^t I^{-1}(\underline{\theta}_0)\underline{\psi}_n(\underline{\theta}_0) \longrightarrow_{(d)} \chi^2_p.$$

Furthermore, if $\theta = \binom{\theta^{(1)}}{\theta^{(2)}}$ where $\theta^{(1)}$ and $\theta^{(2)}$ are $p-r$ and r vectors respectively, let $\widehat{\theta}_0$ denote the maximum likelihood estimate of θ under the constraint that $\theta^{(1)} = \theta_0^{(1)}$, then

$$R_n(\widehat{\theta}_0) := n\underline{\psi}(\widehat{\theta}_0)^t I^{-1}(\widehat{\theta}_0)\underline{\psi}_n(\widehat{\theta}_0) \longrightarrow_{(d)} \chi^2_{p-r}.$$

Here $p-r$ is the number of parameters specified by the null hypothesis.

The test that rejects H_0 when $R_n > k_{d.f.,\alpha}$, where d.f. denotes the degrees of freedom, is known as the *Rao score test*.

4.5 Gaussian Linear Model and Wald Test

Consider a Gaussian linear model $Y = X\beta + \epsilon$, with n observations. If $\frac{1}{n}(X^tX) \overset{n\to+\infty}{\longrightarrow} \Sigma$ for some deterministic value Σ, then it is relatively straightforward to show that $\widehat{\beta}$ is consistent; $\widehat{\beta} \overset{n\to+\infty}{\longrightarrow}_{(P)} \beta$.

Furthermore, if $\frac{1}{n}(X^tX) \overset{n\to+\infty}{\longrightarrow} \Sigma$, then under the relaxed assumption that errors are mean zero, each with variance σ^2 and independent, a central limit result can be proved, that

$$\sqrt{n}(\widehat{\beta} - \beta) \overset{n\to+\infty}{\longrightarrow}_{(d)} N(0, \Sigma^{-1}).$$

The Wald Test The Wald test for $H_0 : R\beta = q$ assumes asymptotic normality as $n \to +\infty$ and, in this case,

$$\frac{(\widehat{\beta} - \beta)^t R^t (R(X^tX)^{-1}R^t)^{-1} R(\widehat{\beta} - \beta)}{S^2} \overset{n\to+\infty}{\longrightarrow} \chi^2_J.$$

4.6 Criteria for Model Selection

The most common model selection methods are:

1. AIC (Akaike Information Criterion), AICc (corrected AIC) and related methods such as Mallow's C_p criterion;
2. Cross Validation;
3. BIC (Bayesian Information Criterion) and related methods such as MDL (minimum description length) and Bayesian Model Selection.

It is important to distinguish between two different, but related, goals for model selection:

1. Finding the model that gives the best prediction (without any assumptions that any of the proposed models are correct);

2. Assume that one of the proposed models is the 'true' model and find it.

Generally speaking, AIC (and the strongly related C_p criterion) and cross validation are used for 1., while BIC is used for 2..

4.7 The Akaike Information Criterion (AIC)

Consider a selection of k possible models, $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where each model is a set of densities:

$$\mathcal{M}_j = \{p(y; \theta_j) : \theta_j \in \Theta_j\}$$

and there is data $Y_1, \ldots, Y_n$ drawn from a density f.

Note: For AIC, no assumption is made that f is in any of the models.

For example, for a regression problem, think of $\epsilon_j = Y_j - \sum_{k=0}^{p} \beta_k x_k$, where $\theta_j = (\beta_0, \ldots, \beta_p)$ is the unknown parameter vector and $1, x_1, \ldots, x_p$ are known; $\epsilon_1, \ldots, \epsilon_n$ is the observed random sample from f (hypothesised to be $N(0, \sigma^2)$ when normal errors are assumed).

Let $\widehat{\theta}_j$ denote the MLE of model j. An estimate of f, based on model j is $\widehat{p}_j$ where $\widehat{p}_j(y) = p(y; \widehat{\theta}_j)$. The quality of $\widehat{p}_j$ as an estimator of f may be measured by the *Kullback Leibler Divergence*

$$D_{KL}(f; g) = \int f(x) \log \frac{f(x)}{g(x)} dx.$$

(Note that $D_{KL}(f; g) \geq 0$ for any densities f and g - this is an easy consequence of Jensen's inequality - and $D_{KL}(f; f) = 0$). Here:

$$D_{KL}(f; \widehat{p}_j) = \int f(y) \log \left(\frac{f(y)}{\widehat{p}_j(y)} \right) dy = \int f(y) \log f(y) dy - \int f(y) \log \widehat{p}_j(y) dy.$$

The first term does not depend on j; therefore minimising the Kullback-Leibler divergence is the same as maximising

$$K_j := \int f(y) \log p(y; \widehat{\theta}_j) dy.$$

The aim of the Akaike Information Criterion is to obtain, at least approximately, the model j and Maximum Likelihood Estimates $\widehat{\theta}_j$ which minimise $D_{KL}(f; p(\widehat{\theta}_j))$ from the candidate models.

The quantity K_j has to be estimated. Let us drop the subscript j. The aim is therefore to estimate

$$K = \int f(y) \log p(y; \widehat{\theta}) dy.$$

If θ were fixed, then $\int f(y) \log p(y;\theta) dy = \mathbb{E}_f[\log p(Y;\theta)]$. Therefore, the first thing one would think of for estimating K would be $\overline{K} = \frac{1}{n}\sum_{j=1}^{n} \log p(Y_j, \widehat{\theta})$. Unfortunately, a bias is introduced with this estimate, since the estimator $\widehat{\theta}$ is based on the random sample $Y_1, \ldots, Y_n$. The following lemma shows the size of the bias and is the basis of the AIC criterion.

Lemma 4.5. *Let*

$$\overline{K} = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i, \widehat{\theta})$$

in other words, $\overline{K}$ *is the sample average of* $\log p(Y, \widehat{\theta})$*. Then*

$$n(\overline{K} - K) \overset{n\to+\infty}{\longrightarrow} d$$

where d *is the dimension of the parameter vector* θ*.*

Proof At the end of the lecture. □

This leads to the definition of the Akaike Information Criterion:

Definition 4.6 (Akaike Information Criterion)**.** *The AIC for a model with* d *parameters, based on* n *observations is defined as:*

$$AIC = 2n\widehat{K} = 2nl(\widehat{\theta}) - 2d.$$

where $\widehat{K}$ *is the estimate of* K*,* l *is the log likelihood function,* $\widehat{\theta}$ *is the maximum likelihood estimate for the model parameter vector* θ *and* d *is the dimension of the parameter space.*

It is clear from the proof of the lemma that there are rather many approximations and assumptions used here; the AIC is a very crude tool.

The AICc (Akaike Information Criterion corrected) The AIC is an asymptotic result and the AIC often tends to overestimate the number of parameters. The AICc is a modification to accommodate this;

$$AICc = 2nl(\widehat{\theta}) - 2d - \frac{2d(d+1)}{n-d-1}.$$

Under the assumption that the model is a Gaussian linear model, it can be shown that this is equivalent to minimising the Kullback Leibler distance. This is left as an exercise.

4.8 Mallow's C_p Statistic

For the specific problem of linear regression, consider a model:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

where the errors are mean zero, each with the same variance $\mathrm{Var}(\epsilon) = \sigma^2$. The problem is to select a subset of p regressor variables, where $p \leq k$. If p regressors are selected from k, the C_p tatistic is:

$$C_p = \frac{\mathrm{SSE}_p}{S^2} - n + 2p$$

where n is the number of observations, SSE_p is the error sum of squares for the model with p covariates and S^2 is the error sum of squares for the full model (with all k covariates). This is derived from the following consideration: for a model $Y = X\beta + \epsilon$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_p; \beta_{p+1}, \ldots, \beta_{p+q})^t$ and $\epsilon \sim N(0, \sigma^2 I)$, suppose we include the first p parameters. Then:

$$\widehat{Y} = X_p (X_p^t X_p)^{-1} X_p^t Y$$

where X_p denotes the design matrix with p parameters included. Then:

$$SSE = (Y - \widehat{Y})^t (Y - \widehat{Y}) = Y^t (I - X_p (X_p^t X_p)^{-1} X_p^t) Y$$

It turns out that (this is left as an exercise, the computations are similar to those in examples from 'Statistics'):

$$\mathbb{E}[SSE] = |\mathbb{E}[\widehat{\beta}] - \beta|^2 + \sigma^2 (n - p)$$

where β denotes the parameter vector with the first p parameters. S^2 is an approximation for σ^2. If we have the true model and sufficient data so that $S^2 \simeq \sigma^2$, then

$$\mathbb{E}[C_p] \simeq p.$$

4.9 Cross Validation

Cross validation turns out to be a more reliable tool than AIC and the Mallow C_p statistic, although it has, in general, far larger computational cost. There are various flavours of cross validation. The data are split into a training set and a test set. The models are fitted to the training set and are used to predict the test set. Usually, many such splits are used and the results are averaged over the splits.

To keep things simple, consider a single split.

Consider models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. Assume that there are $2n$ data points. Split the data randomly into two halves, denoted $D = (Y_1, \ldots, Y_n)$ and $T = (Y_1^*, \ldots, Y_n^*)$. Use D to find the MLEs $\widehat{\theta}_j$. define:

$$\widehat{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i^*, \widehat{\theta}_j).$$

Note that $\mathbb{E}[\widehat{K}_j] = K_j$; there is no bias, since $\widehat{\theta}_j$ is independent of Y_i^*. Assume that $|\log p(y;\theta)| \leq B < +\infty$. Hoeffding's inequality tates that if $X_1, \ldots, X_n$ are independent random variables, X_i strictly bounded by the interval $[a_i, b_i]$ and $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, then:

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \geq t) \leq \exp\left\{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

and

$$\mathbb{P}\left(\left|\overline{X} - \mathbb{E}[\overline{X}]\right| \geq t\right) \leq 2\exp\left\{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

By Hoeffding's inequality,

$$\mathbb{P}\left(\max_j \left|\widehat{K}_j - K_j\right| > \epsilon\right) \leq 2ke^{-2n\epsilon^2/(2B^2)}$$

Let

$$\epsilon_n = \sqrt{\frac{2B^2\log(2k/\alpha)}{n}}$$

then

$$\mathbb{P}\left(\max_j \left|\widehat{K}_j - K_j\right| > \epsilon_n\right) \leq \alpha.$$

If we choose $\widehat{j} = \text{argmax}_j \widehat{K}_j$, then, with probability at least $1-\alpha$,

$$K_{\widehat{j}} \geq \max_j K_j - 2\sqrt{\frac{2B^2\log(2k/\alpha)}{n}} = \max_j K_j - O\left(\frac{\log k}{n}\right).$$

The best model is chosen with high probability. The loss function is different for different problems. For ordinary least squares regression, for example, the loss function is $\mathbb{E}[|Y - \mu(X)|^2]$ and the cross validation score function is:

$$\frac{1}{n}\sum_{j=1}^{n}(Y_i^* - \mu(X_i^*))^2.$$

4.10 Bayesian Information Criterion (BIC)

The BIC for a model is defined as

$$BIC = l(\widehat{\theta}) - \frac{d}{2}\log n$$

where n is the number of data points, $\widehat{\theta}$ is the MLE of the parameter vector θ, the parameter space is d dimensional. It is virtually the same as the Minimum Description Length (MDL) criterion. The penalty is harsher than the AIC and hence the BIC chooses models with fewer parameters. A sketch of the derivation is given here:

Put a prior $\pi_j(\theta_j)$ over parameter vector θ_j, given that model j is correct. Put a prior p_j on model j. Then, by Bayes rule,

$$\mathbb{P}(\mathcal{M}_j|Y_1,\ldots,Y_n) \propto p(Y_1,\ldots,Y_n|\mathcal{M}_j)p_j.$$

Also,

$$p(Y_1,\ldots,Y_n|\mathcal{M}_j) = \int p(Y_1,\ldots,Y_n|\mathcal{M}_j,\theta_j)\pi_j(\theta_j)d\theta_j = \int L(\theta_j)\pi_j(\theta_j)d\theta_j.$$

The model j is chosen to maximise $\mathbb{P}(\mathcal{M}_j|Y_1,\ldots,Y_n)$. This is equivalent to choosing j to maximise

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j.$$

For a log-likelihood l, based on n independent observations $Y_1,\ldots,Y_n$ and a parameter vector $\theta = (\theta_1,\ldots,\theta_d)^t$, we have a Taylor expansion:

$$l(\theta) \simeq l(\widehat{\theta}_{ML}) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{k=1}^{n}\log p(X_k,\widehat{\theta}_{ML})\right)(\theta_i-\widehat{\theta}_i)(\theta_j-\widehat{\theta}_j)$$

By the law of large numbers, $\lim_{n\to+\infty}\frac{1}{n}\sum_{k=1}^{n}\log p(Y_k,\theta_0) = \mathbb{E}[\log p(Y,\theta_0)]$. Similarly,

$$\lim_{n\to+\infty} -\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(Y_k,\theta_0) \to I_{ij}(\theta_0)$$

where I is the Fisher information. The posterior distribution will be asymptotically $N(\widehat{\theta},\frac{1}{n}I(\widehat{\theta})^{-1})$. From this,

$$\begin{aligned}
&\log\int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j\\
&\quad\simeq \log\int e^{l(\widehat{\theta}_j)-\frac{n}{2}(\theta-\widehat{\theta}_j)^t I(\widehat{\theta}_j)(\theta-\widehat{\theta}_j)}\pi_j(\theta)d\theta + \log p_j\\
&\quad\simeq l_j(\widehat{\theta}_j) - \frac{d_j}{2}\log n + \frac{1}{2}\log|I(\widehat{\theta})| + \frac{d_j}{2}\log(2\pi) + \log p_j\\
&\quad\simeq l_j(\widehat{\theta}_j) - \frac{d_j}{2}\log n = BIC_j.
\end{aligned}$$

where, to go from second last to last line, the terms of order 1 have been removed. (the terms involving the prior are of a lower order, so they do not appear in the final formula).

BIC behaves quite differently from AIC and cross validation, since it is based on an entirely different set of principles. BIC assumes that, among the collection of models, there is a 'correct' model and that the aim is to find the model most likely to be true in the Bayesian sense.

AIC and cross-validation are trying to find the model from the collection which will give the best prediction, and make no assumptions about existence of a 'correct' model in the class being considered.

Model Averaging with BIC Suppose we want to predict a new observation Y. Let $D = (Y_1, \ldots, Y_n)$ be the observed data. Then

$$p(y|D) = \sum_j p(y|D, \mathcal{M}_j)\mathbb{P}(\mathcal{M}_j|D)$$

where

$$\mathbb{P}(\mathcal{M}_j|D) = \frac{\int L(\theta_j)\pi_j(\theta_j)d\theta_j}{\sum_s \int L(\theta_s)\pi_s(\theta_s)d\theta_s} \simeq \frac{e^{BIC_j}}{\sum_s e^{BIC_s}}.$$

4.11 Example: Normal Observations

Suppose $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, 1)$ and we want to compare two models:

$$M_0 : N(0, 1) \qquad M_1 : N(\mu, 1).$$

We consider three approaches: hypothesis testing, AIC and BIC.

Hypothesis Testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. The test statistic is:

$$Z = \sqrt{n}\overline{Y}$$

which has $N(0, 1)$ distribution if H_0 is true. H_0 is rejected if $|Z| \geq z_{\alpha/2}$.

AIC The likelihood is proportional to:

$$L(\mu) = \prod_{i=1}^{n} e^{-(y_i-\mu)^2/2} = e^{-n(\overline{y}-\mu)^2/2}e^{-nS^2/2}$$

where $S^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. It follows that:

$$l(\mu) = -\frac{n}{2}(\overline{Y} - \mu)^2 - \frac{nS^2}{2}.$$

The AIC scores are:

$$\begin{cases} \frac{1}{2}AIC_0 = l(0) - 0 = -\frac{n}{2}\overline{Y}^2 - \frac{nS^2}{2} \\ \frac{1}{2}AIC_1 = l(\widehat{\mu}) - 1 = -\frac{nS^2}{2} - 1 \end{cases}$$

since $\widehat{\mu} = \overline{Y}$. It follows that Model 1 is chosen if and only if $AIC_1 > AIC_0$; in other words, if and only if

$$|\overline{Y}| \geq \frac{\sqrt{2}}{\sqrt{n}}.$$

If the hypothesis test were carried out at $\alpha = 0.05$, then it would give $|\overline{Y}| \geq \frac{1.96}{\sqrt{n}}$.

BIC Here,

$$\begin{cases} BIC_0 = l(0) = -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2} \\ BIC_1 = l(\widehat{\mu}) - \frac{1}{2}\log n = -\frac{nS^2}{2} - \frac{1}{2}\log n \end{cases}$$

which leads to model 1 being chosen if

$$|\overline{Y}| > \sqrt{\frac{\log n}{n}}.$$

Proof of Lemma 4.5

Suppose that θ_0 minimises $D_{KL}(f, p(.,\theta))$, so that (under Kullback-Leibler divergence) $p(.,\theta_0)$ is the closest density to f in the model to the true density. Let

$$l(y,\theta) = \log p(y,\theta)$$

and let

$$s_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} \nabla_\theta \log p(Y_j,\theta)$$

be the score; $\widehat{\theta}$ satisfies $s(\widehat{\theta}) = 0$. Let $H(\theta)$ the matrix defined by: $H_{jk}(\theta) = \frac{\partial^2}{\partial\theta_j\partial\theta_k}\mathbb{E}_f[\log p(Y,\theta)]$ and let $\widehat{H}_{jk}(\theta) = \frac{\partial^2}{\partial\theta_j\partial\theta_k}\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i,\theta)$, so that $\widehat{H}(\theta) \to H(\theta)$ in probability (by the law of large numbers).

Let $Z_n = \sqrt{n}(\widehat{\theta} - \theta_0)$ and let

$$J_{jk}(\theta) = \mathbb{E}_f[s_j(Y,\theta)s_k(Y,\theta)].$$

Since $\widehat{\theta}$ is the value such that $s_n(\widehat{\theta}_n) = 0$, it follows by the central limit theorem that if $\widehat{\theta} \overset{n\to+\infty}{\longrightarrow} \theta_0$, then $\sqrt{n}s_n(\widehat{\theta}) \overset{n\to+\infty}{\longrightarrow} N(0,J)$.

By Taylor's expansion theorem,

$$s_j(\theta_0) = s_j(\widehat{\theta}) + \sum_{k=1}^{d}\widehat{H}_{jk}(\theta^*)(\widehat{\theta}_k - \theta_{0k})$$

where $|\theta^* - \theta_0| \leq |\widehat{\theta} - \theta_0|$. Since $s_j(\widehat{\theta}) = 0$ (by definition of $\widehat{\theta}$), it follows that, provided $\widehat{\theta} \overset{n\to\infty}{\longrightarrow} \theta_0$ in probability,

$$Z_n \overset{n\to+\infty}{\longrightarrow} N(0, J^{-1}HJ^{-1}).$$

Now let $M(y,\theta)$ denote the matrix such that $M_{jk}(y,\theta) = \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(y,\theta)$, so that $\int f(y)M(y,\theta)dy = H(\theta)$. Then Taylor's expansion theorem gives (ignoring third and higher order terms):

$$\begin{aligned} K &\simeq \int f(y)\left(\log p(y;\theta_0) - (\widehat{\theta}-\theta_0)^t s(y,\theta_0) + \frac{1}{2}(\widehat{\theta}-\theta_0)^t M(y,\theta_0)(\widehat{\theta}-\theta_0)\right)dy \\ &= K_0 - \frac{1}{2n}Z_n^t H Z_n \end{aligned}$$

where

$$K_0 = \int f(y)\log p(y,\theta_0)dy.$$

The second term drops out because it has mean 0. A Taylor expansion of $\overline{K}$ gives:

$$\begin{aligned} \overline{K} &\simeq \frac{1}{n}\sum_{i=1}^n\left(\log p(Y_i,\theta_0) + (\widehat{\theta}-\theta_0)^t s(Y_i,\theta_0) + \frac{1}{2}(\widehat{\theta}-\theta_0)^t M(Y_i,\theta_0)(\widehat{\theta}-\theta_0)\right) \\ &= K_0 + A_n + (\widehat{\theta}-\theta_0)^t S_n - \frac{1}{2n}Z_n^t\widehat{H}_n Z_t \\ &\simeq K_0 + A_n + \frac{1}{\sqrt{n}}Z_n^t S_n - \frac{1}{2n}Z_n^t H Z_n \end{aligned}$$

where

$$\widehat{H}_n = \frac{1}{n}\sum_{i=1}^n M(Y_i,\theta_0) \xrightarrow{P} H$$

and

$$A_n = \frac{1}{n}\sum_{i=1}^n (l(Y_i,\theta_0) - K_0)$$

so that

$$n(\overline{K} - K) \simeq nA_n + \sqrt{n}Z_n^t S_n \simeq nA_n + Z_n^t J Z_n.$$

For any random vector ϵ with $\mathbb{E}[\epsilon] = \mu$ and $\mathrm{Var}(\epsilon) = \Sigma$ and a symmetric matrix A,

$$\mathbb{E}[\epsilon^t A\epsilon] = \mathrm{trace}(A\Sigma) + \mu^t A\mu.$$

Therefore:

$$n\mathbb{E}[\overline{K} - K] \simeq n\mathbb{E}[A_n] + \mathbb{E}\left[Z_n^t H Z_n\right] = 0 + \text{trace}(HJ^{-1}HJ^{-1}) = \text{trace}((HJ^{-1})^2)$$

from which:

$$n(K - \overline{K}) = -\text{trace}((HJ^{-1})^2).$$

If the model is correct, then $H = J$ so that $\text{trace}((HJ^{-1})^2) = \text{trace}(I) = d$ and hence

$$n(K - \overline{K}) = -d.$$

□