Chapter 7

Model Assessment and Selection

We have defined various smoothers and nonparametric estimation techniques. In classical statistical theory we usually assume that the underlying model generating the data is in the family of models we are considering. In this case bias is not an issue and efficiency (low variance) is all that matters. Much of the theory in classical statistics is geared towards finding efficient estimators.

In this course we try not make the above assumptions. Furthermore, for the techiniques we have shown (and will show) asymptotic and finite sample bias and variance estimates are not always easy (many times impossible) to find in closed form. In this Chapter we discuss monte carlos simulation, in sample approximation, and resampling methods that are commonly used to get estimates of prediction error.

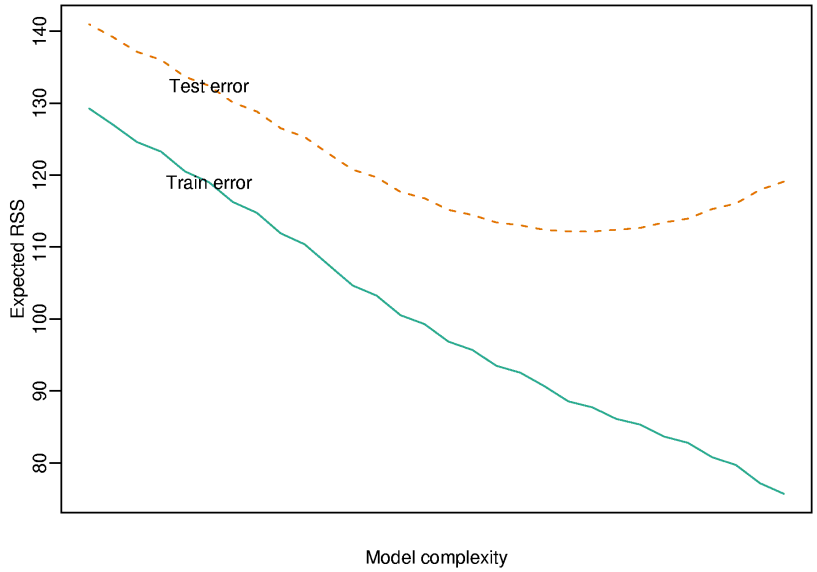

Figure 7.1: Expected test and train errors. Notice the expected train error is the EPE.

Remember that the main difficulty with model assessment and selection is that the observed prediction error for training data becomes smaller with model complexity regardless of the prediction ability on the test data. See figure 7.1.

In this Chapter we will look at a specific example: choosing smoothing parameters. For the methods defined in this class, the complexity of the model being considered is controled by the smoothing parameter. Remember how most of the smoothers we have defined have some parameter that controls the smoothness of the curve estimate. For kernel smoothers we defined the scale parameter, for local regression we defined the span or bandwidth, and for penalized least squared

probelsm we have the penalty term. We will call all of these *the smoothing parameter* and denote it with λ. It should be clear from the context which of the specific smoothing parameters we are referring to.

7.1 Introduction

Typically there are two parts to solving a prediction problem: model selection and model assessment. In model selection we estimate the performance of various competing models with the hope of choosing the best one. Having chosen the final model, we assess the model by estimating the prediction error on new data.

Remember that the best model is defined as the one with the lowest EPE:

$$\text{EPE}(\lambda) = E[L\{Y - \hat{f}_\lambda(X)\}]$$

Where Y and X are drawn at random from the population and the expectation averages anything that is random.

Typical loss function are squared error, $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$, and absolute error, $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$.

We define training error is the observed average loss

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

With squared errro loss this is the residual sum of squares divide by N, which we will call the Average Squared Error (ASE).

For categorical data, using square loss doesn't make much sense. Typical loss functions are 0-1 loss, $L(G, \hat{G}(X)) = 0$ if $G = \hat{G}(X)$, 0 otherwise, and the log-likelihood: $L(G, \hat{G}(X)) = -2 \sum_{k=1}^{K} I(G = k) \log \hat{p}_k(X) = -2 \log \hat{p}_G(X)$. The latter is also called *cross-entropy*. Notice the -2 is used so that for normal error it is becomes equivalent to the loss function.

The training errors are obtained as in the continous example. For 0-1 loss it is simple the percentage of times we are wrong in the training data. For the likelihood loss we simply use the observerd log-likelihood times -2/N:

$$-\frac{2}{N} \sum_{i=1}^{N} \log \hat{p}_{g_i}(x_i)$$

As we have discussed various times the training error underestimates the test error or EPE. In today's lectures we describe ways of getting better estimates of EPE.

7.2 Split Samples

When the amount of data and computation time permits it, there is no method better than data splitting. The idea is simple: Divide the data in three parts: train,

validation, and test. We use the train and validation data to select the best model and the test data to assess the chosen model.

The recipe is the following:

1. In the first part, model selection, the validation model is treated as the test data. We train all competing model on the train data and define the best model as the one that predicts best in the validation set. We could replit the train/validation data, do this many times, and select the method that, on average, best performs.

2. Because we chose the best model among many competitors, the observed performance will be a bit biased. Therefore, to appropriately assess performance on independent data we look at the performance on the test set.

3. Finally, we can resplit everything many times and obtain average results from steps 1) and 2).

There is no obvious choice on how to split the data. It depends on the signal to noise ratio which we, of course, do not know. A common choice is 1/2, 1/4, and 1/4.

There are two common problems:

When the amout of data is limited, the results from fitting a model to 1/2 the data can be substantially different to fitting to all the data. An extreme example: We have 12 data points and want to consider a regression model with 7 parameters.

Model fitting might have high computational requirements.

In this Chapter we describe some *in-sample* methods for model selection as well as less biased split sample methods. We also describe monte-carlo simulations which we can use to find, in theory, the best model without even collecting data.

7.3 Bias-Variance tradeoff

We want to estimate f and assume our data comes from the following model:

$$Y_i = f(X_i) + \epsilon_i$$

with the ϵ IID, independent of X, and variance σ^2.

Suppose we are using loess and want to decide what is the best span λ.

To quantify "best", we say it is the λ that minimizes the expected prediction error:

$$\text{EPE}(\lambda) = E[\{Y - \hat{f}_\lambda(X)\}^2] \tag{7.1}$$

Where, as mentioned, Y and X are drawn at random from the population and the expectation averages anything that is random.

The above is better understood in the following way. Let $\hat{f}_\lambda$ be the estimate obtained with the training data. Now, imagine that we get a completely independent data point. Let's simplify by assuming $X = x^*$ is fixed. So what we are looking to minimize is simply

$$\mathrm{E}[Y^* - \hat{f}_\lambda(x^*)]$$

This can be broken up into the following pieces.

$$\sigma^2 + \{\mathrm{E}[\hat{f}_\lambda(x^*)] - f(x^*)\}^2 + \mathrm{var}[\hat{f}_\lambda]$$

The first term is due to unpredictable measurement error. There is nothing we can do about it. The second term is bias of the estimatore (squared) and the last term is the estimator's variance.

Notice that the above calculation can be done because the Y_i^*s are independent of the estimates $\hat{f}_\lambda(x_i)$s, the same can't be said about the Y_is.

In general, we want to pick a λ that performs well for all x. If instead of just one new points we obtain N then we would have

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}[Y_i^* - \hat{f}_\lambda(x_i^*)] = \sigma^2 + \{\mathrm{E}[\hat{f}_\lambda(x_0)] - f(x_0)\}^2 + \mathrm{var}[\hat{f}_\lambda]$$

If we instead assume X is random we can use expectations instead of averages and we are back to our original equation (7.1).

7.3.1 Bias-variance trade-off for linear smoothers

Define $\mathbf{S}_\lambda$ as the hat matrix for a particular smoother when the smoothing parameter λ is used. The "smooth" will be written as $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$.

Define

$$\mathbf{v}_\lambda = \mathbf{f} - \mathrm{E}(\mathbf{S}_\lambda \mathbf{y})$$

as the *bias* vector.

Define $\mathrm{ave}(\mathbf{x}^2) = n^{-1} \sum_{i=1}^n x_i^2$ for any vector $\mathbf{x}$. We can derive the following formulas:

$$\begin{aligned} \mathrm{MSE}(\lambda) &= n^{-1} \sum_{i=1}^n \mathrm{var}\{\hat{f}_\lambda(x_i)\} + \mathrm{ave}(\mathbf{v}_\lambda^2) \\ &= n^{-1} \mathrm{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda') \sigma^2 + n^{-1} \mathbf{v}_\lambda' \mathbf{v}_\lambda \\ \mathrm{EPE}(\lambda) &= \{1 + n^{-1} \mathrm{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda')\} \sigma^2 + n^{-1} \mathbf{v}_\lambda' \mathbf{v}_\lambda. \end{aligned}$$

Notice for least-squares regression $\mathbf{S}_\lambda$ is idempotent so that $\mathrm{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda') = \mathrm{tr}(\mathbf{S}_\lambda) = \mathrm{rank}(\mathbf{S}_\lambda)$ which is usually the number of parameters in the model. This is why we will sometimes refer to $\mathrm{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda')$ as the *equivalent number of parameters* or degrees of freedom of our smoother.

7.4 Monte-Carlo Simulations

We will demonstrate using the example above.

With Monte Carlo simulation we try to create data ourselves (with a computer) using a random model that we hope is similar to reality.

In the example above, we need to decide what is f, the points x, or the distribution of X, and the distribution of ϵ. Notice the possibilities are endless.

So for our loess example, say we think the process we are interested in describing produces and f similar to $2\sin(1/x)$. The rest of the model is specified by (7.2):

$$y_i = 2\sin(1/x) + \epsilon_i, i = 1, \ldots, n \tag{7.2}$$

with the ϵ_i IID $N(0, 1)$.

Figure 7.2 is one instace of our data:

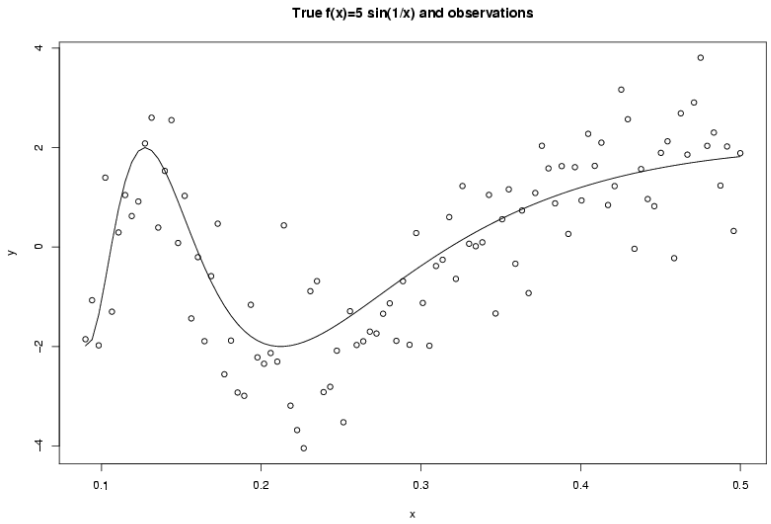

Figure 7.2: Outcomes of model with $f(x) = 2\sin(1/x)$ and IID normal errors with $\sigma^2 = 1$

Here is the R code I used:

```
B <- 1000
sigma <- 1
lambdas <- seq(0.15,0.6,len=30) ##lambdas to try
trainerror <- vector(‘‘numeric’’,length=length(lambdas))
testerror <- vector(‘‘numeric’’,length=length(lambdas))
for(i in 1:B){ ## we want the same y for all competitors
  x <- sort(runif(N,.09,.5))
  f <- 2*sin(1/x)
  y <- f + rnorm(N,0,sigma)
  testy <- f + rnorm(N,0,sigma)
  for(j in seq(along=lambdas)){
    yhat <- loess(y~x,span=lambdas[j])$fitted
    trainerror[j] <- trainerror[j] + sum((y-yhat)^2)/B
    testerror[j] <- testerror[j] + sum((testy-yhat)^2)/B
  }
}
plot(-lambdas,trainerror,ylim=range(c(trainerror,testerror)),xlab=’
complexity’’,ylab=’’Expected RSS’’,type=’’l’’,col=1,lty=1,xaxt=’’n’
lines(-lambdas,testerror,col=2,lty=2)
```

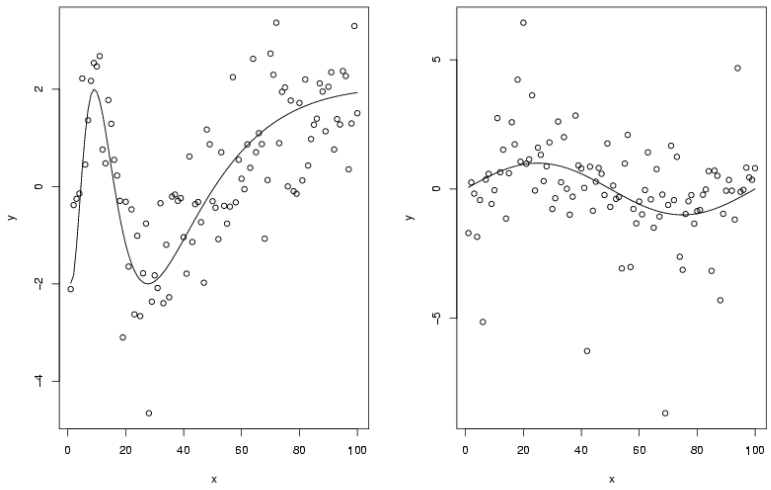

Figure 7.3: Outcomes of model (7.2)

7.5 Cross Validation: Choosing smoothness parameters

In the section, and the rest of the class, we will denote with $\hat{f}_\lambda$ the estimate obtained using smoothing parameter λ. Notice that usually what we really have is the smooth $\hat{\mathbf{f}}_\lambda$.

We will use the model defined by (7.2). Figure 7.3 shows one outcome of this model with normal and t-distributed errors.

In practice it is not common to have a new set of data $y_i^*, i = 1, \ldots, n$. Cross-validation tries to imitate this by leaving out points (x_i, y_i) one at a time and estimating the smooth at x_i based on the remaining $n - 1$ points. The cross-validation sum of squares is

$$\text{CV}(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2$$

where $\hat{f}_\lambda^{-i}(x_i)$ indicates the fit at x_i computed by leaving out the $i - th$ point.

We can now use CV to choose λ by considering a wide span of values of λ, computing CV(λ) for each one, and choosing the λ that minimizes it. Plots of CV(λ) vs. λ may be useful.

Why do we think this is good? First notice that

$$\begin{aligned} \text{E}\{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2 &= \text{E}\{y_i - f(x_i) + f(x_i) - \hat{f}_\lambda^{-i}(x_i)\}^2 \\ &= \sigma^2 + \text{E}\{\hat{f}_\lambda^{-i}(x_i) - f(x_i)\}^2. \end{aligned}$$

Using the assumption that $\hat{f}_\lambda^{-i}(x_i) \approx \hat{f}_\lambda(x_i)$ we see that

$$\text{E}\{\text{CV}(\lambda)\} \approx \text{EPE}(\lambda)$$

However, what we really want is

$$\min_\lambda \text{E}\{\text{CV}(\lambda)\} \approx \min_\lambda \text{EPE}(\lambda)$$

but the law of large numbers says the above will do.

Why not simply use the averaged squared residuals

$$\text{ASR}(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{f}_\lambda(x_i)\}^2?$$

It turns out this under-estimates the EPE. Notice in particular that the estimate $\hat{f}(x_i) = y_i$ always has ASR equal to 0! But we know the EPE will not be small.

Later we will learn of a couple of ways we can adjust the ASR to form "good" estimates of the MSE.

7.5.1 CV for linear smoothers

Now we will see some of the practical advantages of linear smoothers.

For linear smoothers in general it is not obvious what is meant by $\hat{f}_\lambda^{-i}(x_i)$. Let's give a definition...

Notice that any reasonable smoother will smooth constants into constants, i.e. $\mathbf{S1} = \mathbf{1}$. If we think of the rows $\mathbf{S}_{i\cdot}$ of $\mathbf{S}$ as weights of a kernels, this condition is requiring that all the n weights in each of the n kernels add up to 1. We can define $\hat{f}_\lambda^{-i}(x_i)$ as the "weighted average"

$$\mathbf{S}_{i\cdot}\mathbf{y} = \sum_{j=1}^{n} S_{ij} y_j$$

but giving zero weight to the ith entry, i.e.

$$\hat{f}_\lambda^{-i}(x_i) = \frac{1}{1 - S_{ii}} \sum_{j \neq i} S_{ij} y_j.$$

From this definition we can find CV without actually making all the computations again. Lets see how:

Notice that

$$\hat{f}_\lambda^{-i}(x_i) = \sum_{j \neq i} S_{ij} y_j + S_{ii} \hat{f}_\lambda^{-i}(x_i).$$

The quantities we add up to obtain CV are the squares of

$$y_i - \hat{f}_\lambda^{-i}(x_i) = y_i - \sum_{j \neq i} S_{ij} y_j - S_{ii} \hat{f}_\lambda^{-i}(x_i).$$

Adding and subtracting $S_{ii} y_i$ we get

$$y_i - \hat{f}_\lambda^{-i}(x_i) = y_i - \hat{f}_\lambda(x_i) + S_{ii}(y_i - \hat{f}_\lambda^{-i}(x_i))$$

which implies

$$y_i - \hat{f}_\lambda^{-i}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}}$$

and we can write

$$\text{CV}(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}} \right\}^2$$

so we don't have to compute $\hat{f}_\lambda^{-i}(x_i)$!

Lets see how this definition of CV may be useful in finding the MSE.

Notice that the above defined CV is similar to the ASR except for the division by $1 - S_{ii}$. To see what this is doing we notice that in many situations $S_{ii} \approx [\mathbf{S}_\lambda \mathbf{S}_\lambda]_{ii}$ and $1/(1 - S_{ii})^2 \approx 1 + 2S_{ii}$ which implies

$$\text{E}[\text{CV}(\lambda)] \approx \text{EPE}(\lambda) + 2\text{ave}[\text{diag}(\mathbf{S}_\lambda)\mathbf{v}^2].$$

Thus CV adjusts ASR so that in expectation the variance term is correct but in doing so induces an error of $2S_{ii}$ into each of the bias components.

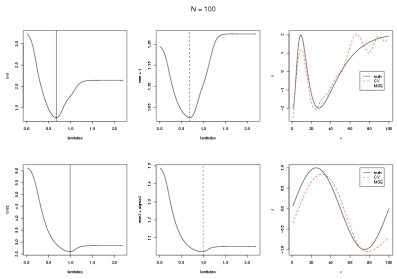

Figure 7.4: CV, MSE, and fits obtained for the normal and t models.

In Figure 7.4 we see the CV and MSE for $n = 100$ and $n = 500$ observatios

7.6 Mallow's C_p

The following three sections describe the related ways of choising the best model using only the training data. These are sometimes call in-sample methods. They were originally developed in the context of parametric models. For example, Mallow's C_p was developed for chosing the number of covariates in a regression model (Mallows 1973).

The basic idea is to start with to try estimate the expected difference between ASR and EPE. Remember ASR is a random quantity and EPE is not!

The larger the model, the more ASR understimates EPE. For a linear model with p covariates, Mallow's C_p estimates this bias with $2 * d/N *^2$. A problem here

is that we need to estimate $\hat{\sigma}^2$. Which model do we use? Typically, a big model (small bias) is used. Below I include some notes on the calculations as presented by the Mallow.

The C_p statistic is defined as a criteria to assess fits when models with different numbers of parameters are being compared. It is given by

$$C_p = \frac{\text{RSS}(p)}{\sigma^2} - N + 2p \tag{7.3}$$

If model(p) is correct then C_p will tend to be close to or smaller than p. Therefore a simple plot of C_p versus p can be used to decide amongst models.

In the case of ordinary linear regression, Mallow's method is based on estimating the mean squared error (MSE) of the estimator $\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{Y}$,

$$\text{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2$$

via a quantity based on the residual sum of squares (RSS)

$$\begin{aligned} \text{RSS}(p) &= \sum_{n=1}^{N}(y_n - \mathbf{x}_n\hat{\boldsymbol{\beta}}_p)^2 \\ &= (\mathbf{Y} - \mathbf{X}_p\hat{\boldsymbol{\beta}}_p)'(\mathbf{Y} - \mathbf{X}_p\hat{\boldsymbol{\beta}}_p) \\ &= \mathbf{Y}'(\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p')\mathbf{Y} \end{aligned}$$

Here $\mathbf{I}_N$ is an $N \times N$ identity matrix. By using a result for quadratic forms, presented for example as Theorem 1.17 in Seber's book, page 13, namely

$$\text{E}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = \text{E}[\mathbf{Y}']\mathbf{A}\text{E}[\mathbf{Y}] + \text{tr}[\boldsymbol{\Sigma}\mathbf{A}]$$

$\mathbf{\Sigma}$ being the variance matrix of $\mathbf{Y}$, we find that

$$\begin{aligned}
\mathrm{E}[\mathrm{RSS}(p)] &= \mathrm{E}[\mathbf{Y}'(\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p')\mathbf{Y}] \\
&= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \mathrm{tr}\left[\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\right]\sigma^2 \\
&= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \sigma^2\left(N - \mathrm{tr}\left[(\mathbf{X}_p'\mathbf{X}_p)(\mathbf{X}_p'\mathbf{X}_p)^{-1}\right]\right) \\
&= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \sigma^2(N-p)
\end{aligned}$$

where N is the number of observations and p is the number of parameters. Notice that when the true model has p parameters $\mathrm{E}[\mathrm{C}_p] = p$. This shows why, if model(p) is correct, C_p will tend to be close to p.

One problem with the C_p criterion is that we have to find an appropriate estimate of σ^2 to use for all values of p.

C_p for smoothers

A more direct way of constructing an estimate of EPE is to correct the ASR. It is easy to show that

$$\mathrm{E}\{\mathrm{ASR}(\lambda)\} = \left\{1 - n^{-1}\mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda')\right\}\sigma^2 + n^{-1}\mathbf{v}_\lambda'\mathbf{v}_\lambda$$

notice that

$$\mathrm{EPE}(\lambda) - \mathrm{E}\{\mathrm{ASR}(\lambda)\} = n^{-1}2\mathrm{tr}(\mathbf{S}_\lambda)\sigma^2$$

This means that if we knew σ^2 we could find a "corrected" ASR

$$\mathrm{ASR}(\lambda) + 2\mathrm{tr}(\mathbf{S}_\lambda)\sigma^2$$

with the right expected value.

For linear regression tr($\mathbf{S}_\lambda$) is the number of parameters so we could think of $2\text{tr}(\mathbf{S}_\lambda)\sigma^2$ as a penalty for large number of parameters or for un-smooth estimates.

How do we obtain an estimate for σ^2? If we had a λ^* for which the bias is 0, then the usual unbiased estimate is

$$\frac{\sum_{i=1}^n \{y_i - f_{\lambda^*}(x_i)\}^2}{n - \text{tr}(2\mathbf{S}_{\lambda^*} - \mathbf{S}_{\lambda^*}\mathbf{S}'_{\lambda^*})}$$

The usual trick is to chose one a λ^* that does little smoothing and consider the above estimate. Another estimate that has been proposed it the first order difference estimate

$$\frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2$$

Once we have an estimate $\hat{\sigma}^2$ then we can define

$$C_p = \text{ASR}(\lambda) + n^{-1} 2\text{tr}(S_\lambda)\hat{\sigma}^2$$

Notice that the p usually means number of parameters so it should be C_λ.

Notice this motivates a definition for degrees of freedoms.

7.7 AIC

Akaike (1977) developed a correction for more general situations, i.e. not just the squared error case. The AIC derives a correction for the training error with the more general likelihood loss. To do this

The AIC is simply:

$$AIC = -\frac{2}{N}\text{loglik} + 2d/N$$

This reduces to Mallow's C_p in the case of Gaussian likelihood. Below is the derivation as shown by Akaike (1977).

Remember that the number of parameters can be defined by smoothers too!

Suppose we observe a realization of a random variable Y, with distribution defined by a parameter $\boldsymbol{\beta}$

$$\prod_{\mathbf{x}_i \in N_0} f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \equiv f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \tag{7.4}$$

where $\mathbf{y}$ is the observed response associated with the covariates $\mathbf{X}$ and $\boldsymbol{\beta} \in \mathbb{R}^P$ is a $P \times 1$ parameter vector.

We are interested in estimating $\boldsymbol{\beta}$. Suppose that before doing so, we need to choose from amongst P competing models, generated by simply restricting the general parameter space R^P in which $\boldsymbol{\beta}$ lies.

In terms of the parameters, we represent *the full model* with P parameters as:

$$\text{Model(P): } f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_P), \boldsymbol{\beta}_P = (\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_P)'.$$

We denote the "true value" of the parameter vector $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^*$.

Akaike (1977) formulates the problem of statistical model identification as one of selecting a model $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_p)$ based on the observations from that distribution, where the particular restricted model is defined by the constraint $\beta_{p+1} = \beta_{p+2} = \ldots = \beta_P = 0$, so that

$$\text{Model(p): } f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_p), \boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p, 0, \ldots, 0)' \tag{7.5}$$

We will refer to p as the *number of parameters* and to Ω_p as the sub-space of $\mathbb{R}^P$ defined by restriction (7.5). For each $p = 1, \ldots, P$, we may assume model(p) to estimate the non-zero components of the vector $\boldsymbol{\beta}^*$. We are interested in a criterion that helps us chose amongst these P competing estimates.

Akaike's original work is for IID data, however it is extended to a regression type setting in a straight forward way. Suppose that the conditional distribution of Y given $\mathbf{x}$ is know except for a P-dimensional parameter $\boldsymbol{\beta}$. In this case, the probability density function of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ can be written as

$$f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \tag{7.6}$$

with $\mathbf{X}$ the design matrix with rows $\mathbf{x}_i$.

Assume that there exists a true parameter vector $\boldsymbol{\beta}^*$ defining a true probability density denoted by $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*)$. Given these assumptions, we wish to select $\boldsymbol{\beta}$, from one of the models defined as in (7.5), "nearest" to the true parameter $\boldsymbol{\beta}^*$ based on the observed data $\mathbf{y}$. The principle behind Akaike's criterion is to define

"nearest" as the model that minimizes the Kullback-Leibler Information Quantity

$$\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) = \int \{\log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) - \log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})\} \, f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \, d\mathbf{y}. \tag{7.7}$$

The analytical properties of the Kullback-Leibler Information Quantity are discussed in detail by Kullback (1959) . Two important properties for Akaike's criterion are

1. $\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) > 0$ if $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \neq f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$
2. $\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) = 0$ if and only if $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) = f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$

almost everywhere on the range of $\mathbf{Y}$. The properties mentioned suggest that finding the model that minimizes the Kullback-Leibler Information Quantity is an appropriate way to choose the "nearest" model.

Since the first term on the right hand side of (7.7) is constant over all models we consider, we may instead maximize

$$\begin{aligned} H(\boldsymbol{\beta}) &= \int \log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \, d\mathbf{y} \\ &= \sum_{i=1}^{n} \int \log f(y_i; \mathbf{X}, \boldsymbol{\beta}) \, f(y_i; \mathbf{x}_i, \boldsymbol{\beta}^*) \, dy_i. \end{aligned} \tag{7.8}$$

Let $\hat{\boldsymbol{\beta}}_p$ be the maximum likelihood estimate under Model(p). Akaike's procedure for model selection is based on choosing the model which produces the estimate

that maximizes $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}}_p)\right]$ amongst all competing models. Akaike then derives a criterion by constructing an asymptotically unbiased estimate of $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}}_p)\right]$ based on the observed data.

Notice that $H(\hat{\boldsymbol{\beta}}_p)$ is a function, defined by (7.8), of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_p$, which is a random variable obtained from the observed data. A natural estimator of its expected value (under the true distribution of the data) is obtained by substituting the empirical distribution of the data into (7.8) resulting in the log likelihood equation evaluated at the maximum likelihood estimate under model(p)

$$l(\hat{\boldsymbol{\beta}}_p) = \sum_{i=1}^{n} \log f(y_i; \mathbf{x}_i, \hat{\boldsymbol{\beta}}_p).$$

Akaike noticed that in general $l(\hat{\boldsymbol{\beta}}_p)$ will overestimate $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}})\right]$. In particular Akaike found that under some regularity conditions

$$E_{\boldsymbol{\beta}^*}\left[l(\hat{\boldsymbol{\beta}}_p) - H(\hat{\boldsymbol{\beta}}_p)\right] \approx p.$$

This suggests that larger values of p will result in smaller values of $l(\hat{\boldsymbol{\beta}}_p)$, which may be incorrectly interpreted as a "better" fit, regardless of the true model. We need to "penalize" for larger values of p in order to obtain an unbiased estimate of the "closeness" of the model. This fact leads to the Akaike Information Criteria which is a bias-corrected estimate given by

$$\mathrm{AIC}(p) = -2l(\hat{\boldsymbol{\beta}}_p) + 2p. \tag{7.9}$$

See, for example, Akaike (1973) and Bozdogan (1987) for the details.

7.8 BIC

Objections have been raised that minimizing Akaike's criterion does not produce asymptotically consistent estimates of the correct model Notice that if we consider Model(p^*) as the correct model then we have for any $p > p*$

$$\Pr\left[AIC(p) < AIC(p^*)\right] = \Pr\left[2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\} > 2(p - p^*)\right]. \qquad (7.10)$$

Notice that, in this case, the random variable $2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\}$ is the logarithm of the likelihood ratio of two competing models which, under certain regularity conditions, is known to converge in distribution to $\chi^2_{p-p^*}$, and thus it follows that the probability in Equation (7.10) is not 0 asymptotically. Some have suggested multiplying the penalty term in the AIC by some increasing function of n, say $a(n)$, that makes the probability

$$\Pr\left[2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\} > 2a(n)(p - p^*)\right]$$

asymptotically equal to 0. There are many choices of $a(n)$ that would work in this context. However, some of the choices made in the literature seem arbitrary.

Schwarz (1978) and Kashyap (1982) suggest using a Bayesian approach to the problem of model selection which, in the IID case, results in a criterion that is similar to AIC in that it is based on a penalized log-likelihood function evaluated at the maximum likelihood estimate for the model in question. The penalty term in the Bayesian Information Criteria (BIC) obtained by Schwarz (1978) is the AIC penalty term p multiplied by the function $a(n) = \frac{1}{2}\log(N)$.

The Bayesian approach to model selection is based on maximizing the posterior probabilities of the alternative models, given the observations. To do this we must

define a strictly positive prior probability $\pi_p = \Pr[\text{Model}(p)]$ for each model and a conditional prior $d\mu_p(\boldsymbol{\beta})$ for the parameter given it is in Ω_p, the subspace defined by Model(p). Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the response variable and define the distribution given $\boldsymbol{\beta}$ following (7.6)

$$f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \boldsymbol{\beta})$$

The posterior probability that we look to maximize is

$$\Pr\left[\text{Model}(p)|\mathbf{Y} = \mathbf{y}\right] = \frac{\int_{\Omega_p} \pi_p f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) d\mu_p(\boldsymbol{\beta})}{\sum_{q=1}^{P} \int_{\Omega_q} \pi_q f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) d\mu_q(\boldsymbol{\beta})}$$

Notice that the denominator depends neither on the model nor the data, so we need only to maximize the numerator when choosing models.

Schwarz (1978) and Kashyap (1982) suggest criteria derived by taking a Taylor expansion of the log posterior probabilities of the alternative models. Schwarz (1978) presents the following approximation for the IID case

$$\log \int_{\Omega_p} \pi_p f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) d\mu_p(\boldsymbol{\beta}) \approx l(\hat{\boldsymbol{\beta}}_p) - \frac{1}{2} p \log n$$

with $\hat{\boldsymbol{\beta}}_p$ the maximum likelihood estimate obtained under Model(p).

This fact leads to the Bayesian Information Criteria (BIC) which is

$$\text{BIC}(p) = -2l(\hat{\boldsymbol{\beta}}_p) + p \log n \tag{7.11}$$

Kyphosis Example

The AIC and BIC obtained for the gam are:

```
AIC(Age) =  83                BIC(Age) = 90
AIC(Age,Start) = 64           BIC(Age,Start) = 78
AIC(Age,Number) = 73          BIC(Age,Number) = 86
AIC(Age,Start,Number) = 60    BIC(Age,Start,Number) = 81
```

Bibliography

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in Petrov, B. and Csaki, B., editors, *Second International Symposium on Information Theory*, pp. 267–281, Budapest: Academiai Kiado.

[2] Bozdogan, H. (1987), "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrica*, 52, 345–370.

[3] Bozdogan, H. (1994), "Mixture-model cluster analysis using a new informational complexity and model selection criteria," in Bozdogan, H., editor, *Multivariate Statistical Modeling*, volume 2, pp. 69–113, The Netherlands: Dordrecht.

[4] Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley & Sons.

[5] Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

[6] Shibata, R. (1989), "Statistical aspects of model selection," in Williems, J. C., editor, *From Data to Model*, pp. 215–240, New York: Springer-Verlag.

[7] Efron B. and Tibshirani, R.J (1993), *An Introduction of the Bootstrap*. Chapman and Hall/CRC: New York.

[8] Efron, B. (1979). Bootstrap Methods: Another Look At the Jackknife, *The Annals of Statistics* 7, 1–26.