Model Selection – Optimization versus Simplicity

There are two opposing philosophies on how to draw inference from empirical data:

- Find the model that fits the data best.
- Keep it simple. Don't choose a complicated model over a simpler model if the simpler model suffice (Occam's razor).

Ideas like maximum-likelihood and empirical loss minimization work by the first principle.

Model selection/test theory work by the second principle to compensate for the fact that an optimization procedure generally overfits to the given data set.

Model or Predictor Assessment

Another problem of importance is

Having fitted a final predictor $\hat{f}$, how will it actually perform?

The *training error*

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

generally *underestimates* $\text{EPE}(\hat{f})$. The expected *generalization* or *test* error

$$\text{Err} = E(\text{EPE}(\hat{f}))$$

is the expected EPE.

$\text{EPE}(\hat{f})$ can only really be estimated if we have an independent *test data set*.

In the book at this point (page 220) they introduce the test error – conditioning on the training data. This is the same as the expected prediction error for the estimated predictor $\hat{f}$.

Figure 7.1 – The Bias-Variance Tradeoff

Realizations of training error $\bar{\text{err}}$ and expected prediction error $\text{EPE}(\hat{f})$ estimated on an independent test dataset as functions of model complexity. Also the estimates of the expectation of $\bar{\text{err}}$ and $\text{EPE}(\hat{f})$ are shown.

The Bias-Variance Tradeoff for nearest neighbors

Consider the case with $Y = f(X) + \varepsilon$ where X and ε are independent, $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$.

With $\hat{f}_k$ the k-nearest neighbor regressor the expected *test error*, conditionally on the training data $X_1 = x_1, \ldots, X_N = x_N$, in x_0 is

$$\begin{aligned} E((Y - \hat{f}_k(x_0))^2 \mid X = x_0) &= E((Y - f(x_0))^2 \mid X = x_0) \\ &\quad + E((f(x_0) - E(\hat{f}_k(x_0)))^2 \mid X = x_0) \\ &\quad + E((\hat{f}_k(x_0) - E(\hat{f}_k(x_0)))^2 \mid X = x_0) \\ &= \sigma^2 + \underbrace{\left[f(x_0) - \frac{1}{k} \sum_{l \in N_k(x_0)} f(x_l) \right]^2}_{\text{Squared bias}} + \underbrace{\frac{\sigma^2}{k}}_{variance} \end{aligned}$$

Small choices of k (complex model) will give a large variance and generally a smaller bias, and vice versa for large choices of k (simple model).

Figure 2.4 – Bias-Variance Tradeoff for k-nearest neighbors

The test error here is an estimate of $\text{EPE}(\hat{f})$ for the estimated predictor based on an *independent* dataset.

The Train-Validate-Test Idea

In a data rich situation we split the data *before doing anything else* into three subsets.

- On the *training* data we estimate all parameters besides tuning parameters (model complexity parameters).
- On the *validation* data we estimate prediction error for the estimated predictors and optimize over tuning parameters and models.
- On the *test* data we estimate the expected prediction error for the chosen predictor – no model selection here, please.

Problem: We are almost never in a data rich situation.

Can we justify to throw away data that can be used for estimation, and thus reduction of variance, for the purpose of estimating parameters of secondary importance?

Figure 7.2 – Space of Models

Digesting Figure 7.2 provides a core understanding of the bias-variance tradeoff between complex and simple models. This understanding should be obtained in close connection with reading about the bias-variance decomposition in Section 7.3. One should note that the nice additive decomposition into a squared bias term and a variance term of the expectation of the prediction error in x_0 is a consequence of the choice of the loss function being the squared error loss. For the 0-1 loss often used in classification things work out differently, see Exercise 7.2.

Setup

In the following discussion $(X_1, Y_1), \ldots, (X_N, Y_N)$ denote N i.i.d. random variables, with X_i a p-dimensional vector.

A concrete realization is denoted $(x_1, y_1), \ldots, (x_N, y_N)$ and we use boldface, e.g. $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ and $\mathbf{y} = (y_1, \ldots, y_N)^T$ to denote vectors.

We can not distinguish in notation between $\mathbf{X}$ – the matrix of random variables $X_1, \ldots, X_N$ – and $\mathbf{X}$ – the matrix of a concrete realization $x_1, \ldots, x_N$.

Mallows' C_p

With $\hat{\mathbf{f}} = P\mathbf{y}$ where P is a projection onto a d-dimensional subspace define

$$\bar{\text{err}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mathbf{f}}_i)^2 = \frac{1}{N}||\mathbf{y} - \hat{\mathbf{f}}||^2.$$

By a standard decomposition

$$\frac{1}{N}E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X}) = \frac{1}{N}E(||\mathbf{Y} - \hat{\mathbf{f}}||^2|\mathbf{X}) + \frac{2d}{N}\sigma^2$$

The expected *in-sample error*

$$\text{Err}_{\text{in}} = \frac{1}{N}E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X})$$

can thus be estimated by

$$C_p = \hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2d}{N}\hat{\sigma}^2.$$

Note that $\mathbf{Y}^{\text{new}} - P\mathbf{Y}$ and $\mathbf{Y} - P\mathbf{Y}^{\text{new}}$ have the same conditional distributions given $\mathbf{X}$ and note that $\mathbf{Y} - P\mathbf{Y}^{\text{new}} = (I-P)\mathbf{Y} + P(\mathbf{Y} - \mathbf{Y}^{\text{new}})$ where the two terms are orthogonal. This implies that

$$\begin{aligned} E(||\mathbf{Y}^{\text{new}} - P\mathbf{Y}||^2|\mathbf{X}) &= E(||\mathbf{Y} - P\mathbf{Y}^{\text{new}}||^2|\mathbf{X}) \\ &= E(||(I-P)\mathbf{Y}||^2|\mathbf{X}) + E(||P(\mathbf{Y} - \mathbf{Y}^{\text{new}})||^2|\mathbf{X}) \end{aligned}$$

Since the vector $\mathbf{Y} - \mathbf{Y}^{\text{new}}$ has (conditional) mean 0 and (conditional) covariance matrix $2\sigma^2 I$ the second expectation above equals $2\sigma^2 d$. One derivation relies on trace computations

as follows:

$$\begin{aligned} E(||P(\mathbf{Y}-\mathbf{Y}^{\text{new}})||^2|\mathbf{X}) &= E((\mathbf{Y}-\mathbf{Y}^{\text{new}})^T P^T P(\mathbf{Y}-\mathbf{Y}^{\text{new}}))|\mathbf{X}) \\ &= E(\text{tr}(P(\mathbf{Y}-\mathbf{Y}^{\text{new}})(\mathbf{Y}-\mathbf{Y}^{\text{new}})^T)|\mathbf{X}) \\ &= \text{tr}(PE((\mathbf{Y}-\mathbf{Y}^{\text{new}})(\mathbf{Y}-\mathbf{Y}^{\text{new}})^T|\mathbf{X})) \\ &= \text{tr}(P2\sigma^2 I) \\ &= 2\sigma^2\text{tr}(P) = 2\sigma^2 d, \end{aligned}$$

where we have used the trace formula $\text{tr}(AB) = \text{tr}(BA)$, the projection formula $P^T P = P^2 = P$, the linearity of the trace, and the fact that $\text{tr}(P) = d$ because P has precisely d eigenvectors with eigenvalue 1 and $N - d$ eigenvectors with eigenvalue 0.

Mallows' C_p

$$\hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2d}{N}\hat{\sigma}^2.$$

is an equivalent of *Mallows' C_p statistic* – with $\hat{\sigma}^2$ estimated from a "low-bias" model with p degrees of freedom;

$$\hat{\sigma}^2 = \frac{1}{N-p}||\mathbf{y} - Q\mathbf{y}||^2$$

where Q is a projection on a p-dimensional space.

If $\mathbf{S}$ is a *linear smoother*, that is, $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$, one can generalize C_p as

$$\hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2\text{trace}(\mathbf{S})}{N}\hat{\sigma}^2$$

with $\hat{\sigma}^2$ estimated from a "low-bias" model, e.g. as

$$\hat{\sigma}^2 = \frac{1}{N - \text{trace}(2\mathbf{S}_0 - \mathbf{S}_0^2)}||\mathbf{y} - \mathbf{S}_0\mathbf{y}||^2.$$

for a "low-bias" smoother $\mathbf{S}_0$.

The complete justification of the above generalization of Mallows' C_p to general smoothers is sketched in the book and will be treated in theoretical exercises.

Using C_p

The classical use of C_p is when $\mathbf{X}$ is $N \times p$ of rank p and $Q = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

For any choice of d columns we compute C_p and select the model with the smallest value of C_p.

This is equivalent to best subset selection for each d followed by choosing d that minimizes

$$NC_p = \text{RSS}(d) + \frac{2d}{N-p}\text{RSS}(p)$$

As a function of d the classical definition of C_p,

$$\tilde{C}_p = \frac{NC_p(N-p)}{\mathrm{RSS}(p)} - N = \frac{(N-p)\mathrm{RSS}(d)}{\mathrm{RSS}(p)} + 2d - N,$$

is a monotonely increasing function of our C_p.

Minimizing C_p or $\tilde{C}_p$ is equivalent.

The historically correct definition of Mallows' C_p is as $\tilde{C}_p$ above in the framework of multiple linear regression, see e.g. Wikipedia. When used as a model selection tool in this framework we can just as well consider C_p as we have defined. They select the same model. $\tilde{C}_p$ looks related to the F-test statistic of testing a d-dimensional model against the larger p-dimensional alternative. Indeed,

$$\frac{1}{p-d}\tilde{C}_p = F_{d,p} + 2$$

where $F_{d,p}$ is the F-test statistic. Our C_p is, however, easier to generalize and compare to other methods.

Generalization Error

Instead of the expected in-sample error we can consider the expected *generalization* or *test error*

$$\mathrm{Err} = E(L(Y, \hat{f}(X))) = E(E(L(Y, \hat{f}(X))|\mathbf{X}, \mathbf{Y})) = E(\mathrm{EPE}(\hat{f}))$$

Here (X, Y) is independent of $(X_1, Y_1), \ldots, (X_N, Y_N)$ that enter through $\hat{f}$.

Err is the *expectation* over the dataset of the *expected prediction error* for the estimated predictor $\hat{f}$.

A small value of Err tells us that *the estimation methodology is good* and will on average result in estimators with a small EPE. It *does not* guarantee that a concrete realization, $\hat{f}$, has a small EPE!

Likelihood Loss

The *generalized* decision theoretic setup has sample spaces E and F, *action space* $\mathcal{A}$, *decision rule* $f : E \to \mathcal{A}$ and loss functions $L : F \times \mathcal{A} \to [0, \infty)$. If h_a for $a \in \mathcal{A}$ denotes a collection of densities on F we define the *minus-log-likelihood* loss function as

$$L(y, a) = -\log h_a(y)$$

The empirical loss for $(x_1, y_1), \ldots, (x_N, y_N)$ when using decision rule f is

$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = -\frac{1}{N}\log\prod_{i=1}^{N} h_{f(x_i)}(y_i)$$

Expected prediction error equals the expectation of (conditional) cross entropies.

$$\mathrm{EPE}(f) = \int \underbrace{\int -\log h_{f(x)}(y) g(y|x) \mathrm{d}y}_{\text{cross entropy}} g_1(x) \mathrm{d}x$$

The standard example in this context of the need for the general setup as compared to the setup where $\mathcal{A} = F$ and f is simply the predictor is when F is discrete. For instance, if $F = \{0, 1\}$ we might want "the action space" to be the set of probability measures on F – represented as $\mathcal{A} = [0, 1]$ and $p \in [0, 1]$ is the probability of $Y = 1$. A "decision" can then be the computation of $f(x) = \Pr(Y = 1 \mid X = x)$ – the conditional probability that $Y = 1$ given $X = x$. What is perhaps more common in this case is that $\mathcal{A} = \mathbb{R}$ and a "decision" is the computation of the logit of $\Pr(Y = 1|X = x)$, that is,

$$f(x) = \mathrm{logit}(\Pr(Y = 1|X = x)) = \log \frac{\Pr(Y = 1|X = x)}{\Pr(Y = 0|X = x)}$$

the log-odds of $Y = 1$ conditionally on $X = x$.

Akaike's Information Criteria – AIC

We take $\mathcal{A} = \{f_\theta(x, \cdot)\}_{\theta \in \Theta, x \in E}$ with Θ being d-dimensional and $f_\theta : E \times F \to [0, \infty)$ such that $f_\theta(x, \cdot)$ is a probability density on F. Let $\hat{\theta}_N$ denote the MLE.

With likelihood loss we define the equivalent of the expected in-sample error

$$\mathrm{Err}_{\mathrm{loglik,in}} = -\frac{1}{N} \sum_{i=1}^{N} E(\log f_{\hat{\theta}_N}(x_i, Y_i^{\mathrm{new}})|\mathbf{X})$$

Then one derives (difficult) the approximation

$$\mathrm{Err}_{\mathrm{loglik,in}} \simeq \frac{1}{N} E(l_N(\hat{\theta}_N)) + \frac{d}{N}$$

where the minus-log-likelihood function in $\hat{\theta}_N$

$$l_N(\hat{\theta}_N) = -\frac{1}{N} \sum_{i=1}^{N} \log f_{\hat{\theta}_N}(x_i, y_i)$$

is the equivalent of $\bar{\mathrm{err}}$ when using likelihood loss.

We let $Y_1, \ldots, Y_N$ and $Y_1^{\mathrm{new}}, \ldots, Y_N^{\mathrm{new}}$ be conditionally independent with the same distribution given $X_1 = x_1, \ldots, X_N = x_N$. The minus-log-likelihood

$$l_N(\theta) = -\sum_{i=1}^{N} \log f_\theta(x_i, Y_i)$$

and the minus-log-likelihood for the new data

$$l_N^*(\theta) = -\sum_{i=1}^{N} \log f_\theta(x_i, Y_i^{\mathrm{new}}).$$

Letting $\hat{\theta}_N$ denote the MLE for the original dataset and $\tilde{\theta}_N$ the MLE for the new dataset then a Taylor expansion of l_N^* around $\tilde{\theta}_N$ yields

$$l_N^*(\hat{\theta}_N) = l_N^*(\tilde{\theta}_N) + \frac{1}{2}(\hat{\theta}_N - \tilde{\theta}_N)^T D^2 l_N(\tilde{\theta}_N)(\hat{\theta}_N - \tilde{\theta}_N) + \text{remainder}_N.$$

Under suitable regularity assumptions there is a θ_0 such that

$$\frac{1}{\sqrt{N}} D_\theta l_N(\theta_0)^T \xrightarrow{\mathcal{D}} N(0, K(\theta_0))$$

and

$$\frac{1}{N} D_\theta^2 l_N(\theta_0) \xrightarrow{P} I(\theta_0)$$

and the two esimators are independent and asymptotically $N(\theta_0, \frac{1}{N} I(\theta_0)^{-1} K(\theta_0) I(\theta_0)^{-1})$-distributed. Consequently

$$\sqrt{N}(\hat{\theta}_N - \tilde{\theta}_N) \xrightarrow{\mathcal{D}} N(0, 2I(\theta_0)^{-1} K(\theta_0) I(\theta_0)^{-1})$$

$$\begin{aligned} E\left(\frac{1}{N} l_N^*(\hat{\theta}_N)|\mathbf{X}\right) &\simeq E\left(\frac{1}{N} l_N^*(\tilde{\theta}_N)|\mathbf{X}\right) + \frac{1}{N}\text{trace}(E\left(N(\hat{\theta}_N - \tilde{\theta}_N)(\hat{\theta}_N - \tilde{\theta}_N)^T|\mathbf{X}\right) I(\theta_0)) \\ &\simeq E\left(\frac{1}{N} l_N(\hat{\theta}_N)|\mathbf{X}\right) + \frac{1}{N}\text{trace}(I(\theta_0)^{-1} K(\theta_0)) \end{aligned}$$

$E\left(\frac{1}{N} l_N^*(\hat{\theta}_N)|\mathbf{X}\right)$ is for the likelihood loss the equivalent of the expected in-sample error for quadratic loss. If $I(\theta_0) = K(\theta_0)$ the trace simplifies to the trace of the $d \times d$ identity matrix and is thus equal to d. This always happens if Θ contains the true parameter. To make likelihood loss for the Gaussian model (with known variance) equivalent to squared error loss we usually multiply everything by 2 and define the esimator of twice this expected in-sample error as

$$\text{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d}{N}.$$

For a more general quantity that does not rely on the model being true we need to replace d by $\text{trace}(I(\theta_0)^{-1} K(\theta_0))$ with the latter quantity having the obvious draw-back that it depends upon the unknown matrices $I(\theta_0)$ and $K(\theta_0)$, which have to be estimated also. Simple estimators are

$$\hat{K} = \frac{1}{N} \sum_{i=1}^{N} D_\theta \log f_{\hat{\theta}_N}(x_i, y_i)^T D_\theta \log f_{\hat{\theta}_N}(x_i, y_i) \quad \text{and} \quad \hat{I} = \frac{1}{N} D_\theta^2 l_N(\hat{\theta}_N)$$

which gives

$$\text{NIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2}{N}\text{trace}(\hat{I}^{-1}\hat{K}).$$

If the distribution of Y given $X = x$ is $N(f(x), \sigma^2)$ for an unknown mean value function $f(x)$, if we take $\Theta = \mathbb{R}^p$, if we assume that σ^2 is fixed, and if we let $f_\theta = X^T\theta$ then

$$2\sigma^2 l_N(\hat{\theta}_N) = ||\mathbf{y} - \mathbf{X}\hat{\theta}_N||^2$$

and we see in this case that $\sigma^2 \text{AIC} = C_p$. We derived C_p and thus AIC as a valid model selection quantity even if the model, as in this general case, is wrong. It is no problem to

show explicitly (and perhaps surprisingly) in this case that the identity $I(\theta_0) = K(\theta_0)$ in fact holds. Here θ_0 is the θ that minimizes $E((f(X) - X^T\theta)^2)$. If we consider logistic regression instead this result does not hold. For logistic regression let $p(x) = \Pr(Y = 1|X = x)$ denote the true conditional probability and let W denote the $N \times N$ diagonal matrix with $p(x_i)(1-p(x_i))$ in the diagonal. Then it is straight forward to show that $I(\beta_0) = \mathbf{X}^T W(\beta_0)\mathbf{X}$ but $K(\beta_0) = \mathbf{X}^T W\mathbf{X}$ – which does not depend upone β_0 – hence

$$\text{trace}(I(\theta_0)^{-1}K(\theta_0)) = \text{trace}((\mathbf{X}^T W(\beta_0)\mathbf{X})^{-1}\mathbf{X}^T W\mathbf{X}).$$

With

$$p_\beta(x) = \frac{\exp((1, x^t)\beta}{1 + \exp((1, x^t)\beta)}$$

the β_0 is the minimizer of

$$E(-p(X)\log p_\beta(X) - (1-p(X))\log(1-p_\beta(X))) = E(-p(X)(1, X^T)\beta + \log(1+\exp((1, X^T)\beta))).$$

One good starting point for a more theoretical treatment of AIC and other aspects of statistical decision theory and model selection is *Pattern Recognition and Neural Networks* by Brian D. Ripley. The more recent book *Model selection and model averaging* by Gerda Claeskens and Nils Lid Hjort is also recommended.

AIC

$$\text{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d}{N}$$

We use AIC for model selection by choosing the model among several possible that *minimizes* AIC.

Assumptions and extensions:

- The models considered *must be true*. If they are *not*, d must in general be replaced by a more complicated quantity d^* leading to the model selection criteria

$$\text{NIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d^*}{N}.$$

- For linear regression with Gaussian errors and fixed variance $d^* = d$ even when the model is wrong, but this does not hold in general, e.g. logistic regression.
- The estimator $\hat{\theta}_N$ must be the MLE. Extensions to non-MLE and non-likelihood loss setups are possible with d replaced again by a more complicated d^*.

Practical BIC

With the same framework as for AIC

$$\text{BIC} = 2l_N(\hat{\theta}_N) + d\log(N)$$

We choose among several models the one with the smallest BIC.

Up to the scaling by $1/N$, BIC is from a practical point of view AIC with 2 replaced by $\log(N)$. The theoretical derivation is, however, completely different.

For $N > e^2 \simeq 7.4$, BIC penalizes complex models more than simple models compared to AIC.

All the preceeding computations with AIC and BIC have been done in the framework of the *conditional* distribution of Y given X. This framework with the likelihood loss has the strongest resemblence to the pure prediction-loss statistical decision theoretic framework, though we have to allow for a more general "action space" to accomodate all situations of practical interest. We can also consider AIC and BIC in the framework of the joint distribution of (X, Y).

Cross-Validation

Let $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$ and denote by $\hat{f}^{-k}$ for $k = 1, \ldots, K$ the estimator of f based on the data (x_i, y_i) with $\kappa(i) \neq k$.

The (x_i, y_i) with $\kappa(i) = k$ work as a test dataset for $\hat{f}^{-k}$ and

$$\hat{\mathrm{EPE}}(\hat{f}^{-k}) = \frac{1}{N_k} \sum_{i:\kappa(i)=k} L(y_i, \hat{f}^{-k}(x_i))$$

with $N_k = |\{i|\kappa(i) = k\}|$

The K-fold κ-cross-validation estimator of Err is the weighted average

$$\begin{aligned} CV_\kappa &= \sum_{k=1}^{K} \frac{N_k}{N} \hat{\mathrm{EPE}}(\hat{f}^{-k}) \\ &= \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i)) \end{aligned}$$

Figure 7.8 – Err as a Function of N

We should write $\mathrm{Err} = \mathrm{Err}(N)$ as a function of the sample size. If $\hat{f}_N$, $f \in \mathcal{F}$ and f minimizes EPE then $\mathrm{EPE}(\hat{f}_N) \geq \mathrm{EPE}(f)$ and

$$\mathrm{Err}(N) = E(L(Y, \hat{f}_N(X))) \geq \mathrm{EPE}(f)$$

If we have a *consistent* estimator; $\hat{f}_N \to f$, then

$$\mathrm{Err}(N) \to \mathrm{EPE}(f).$$

Cross-Validation

Among several models we will choose the model with smallest CV_κ. How to choose κ? How to choose K?

We aim for $N_1 = \ldots = N_K$ in which case

$$E(CV_\kappa) = \mathrm{Err}(N - N_1).$$

With a *steep learning curve* at N we need N_1 to be small or we *underestimate* Err.

Extreme case; N-fold or *leave-one-out* cross-validation with $\kappa(i) = i$ leads to an almost unbiased estimator of $\mathrm{Err}(N)$, but the strong correlation of the $\hat{\mathrm{EPE}}(\hat{f}^{-i})$'s works in the direction of given a larger variance. Recommendations are that 5- or 10-fold CV is a good compromise between bias and variance.

The choice of κ is also of some interest. For N-fold cross validation there is just one choice.

It may be recommended that κ is chosen as a random subdivision of the data into groups of prespecified sizes. If we divide the dataset into groups like $\{1, \ldots, N_1\}$, $\{N_1+1, \ldots, N_1+N_2\}$ we risk that there is some structure in the data that is related to their current ordering, which mess up the result. It could be that the data had be grouped somehow or sorted. But if κ is chosen randomly what makes one choice more appropriate than another? If we generate just a single, random κ it seems most appropriate to keep the same κ for all models considered, but we can also generate $\kappa_1, \ldots, \kappa_B$ and compute the estimator

$$CV = \frac{1}{B}\sum_{i=1}^{B} CV_{\kappa_i}$$

instead. This estimator removes the arbitrary fluctuations of CV_κ that are due to a specific choice of κ at the expense of doing a considerable amount of extra computations.

The Wrong and The Right Way to Cross-Validate

- Mess with the data to find variables/methods that seem to be useful.
- Estimate parameters using the selected variables/methods and use cross-validation to choose tuning parameters.

WRONG

Don't mess with the data before the cross-validation.

Cross-Validation must be out side of all modeling steps, including filtering or variable selection steps.

The only thing that one is allowed to is to do computations or selections based on the x-values alone. This could be to rule out x-values that show a very low variance, say, or different forms of transformations of the x-values.

Estimates of Expected Prediction Error

If $\hat{f}$ is estimated based on a data set, we can only get an estimate of $\text{EPE}(\hat{f})$ by an *independent* test set $(x_1, y_1), \ldots, (x_B, y_B)$ as

$$\hat{\text{EPE}}(\hat{f}) = \frac{1}{B}\sum_{b=1}^{B} L(y_b, \hat{f}(x_b)).$$

Cross-validation provide estimates $\hat{\text{Err}}$ of the expected generalization error.

- $\text{EPE}(\hat{f})$ is a random variable with mean Err.
- $\hat{\text{Err}}$ is a random variable with mean Err.

Can $\hat{\text{Err}}$ be regarded as an approximation/estimate of $\text{EPE}(\hat{f})$?

Figure 7.15 – The Relation Between $\hat{\text{Err}}$ and $\text{EPE}(\hat{f})$

The simulation study reveals that despite the fact that $\hat{\text{Err}}$ is computed by cross-validation on the same dataset and $\hat{f}$ is computed, $\hat{\text{Err}}$ and $\text{EPE}(\hat{f})$ show almost no relation, and if there is a relation it is even one with a negative correlation!

Classification and The Confusion Matrix

For a classifier with two groups we can decompose the errors:

	Predicted y	
Observed y	1	0
1	$\Pr(Y=1, f(X)=1)$	$\Pr(Y=1, f(X)=0)$
0	$\Pr(Y=0, f(X)=1)$	$\Pr(Y=0, f(X)=0)$

This is the *confusion matrix* and

$$\text{EPE}(f) = \Pr(Y=0, f(X)=1) + \Pr(Y=1, f(X)=0).$$

As with $\text{EPE}(\hat{f})$ the confusion matrix can only be estimated using an independent test dataset. "Estimates" based on e.g. cross-validation are estimates of $E(\Pr(Y=k, \hat{f}(X)=l))$.

EXTRA: Linear Smoother Bias-Variance Decomposition

Assumptions: $\hat{\mathbf{f}} = \mathbf{SY}$ and conditionally on $\mathbf{X}$ the Y_i's are uncorrelated with common variance σ^2.

Then with $\mathbf{f} = E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}^{\text{new}}|\mathbf{X})$ and $\mathbf{Y}^{\text{new}}$ independent of $\mathbf{Y}$

$$\begin{aligned}
E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X}) &= E(||\mathbf{Y}^{\text{new}} - \mathbf{SY}||^2|\mathbf{X}) \\
&= E(||\mathbf{Y}^{\text{new}} - \mathbf{f}||^2|\mathbf{X}) + ||\mathbf{f} - \mathbf{Sf}||^2 \\
&\quad + E(||\mathbf{S}(\mathbf{f} - \mathbf{Y})||^2|\mathbf{X}) \\
&= N\sigma^2 + \underbrace{||(I - \mathbf{S})\mathbf{f}||^2}_{\text{Bias}(\lambda)^2} + \sigma^2 \text{trace}(\mathbf{S}^2) \\
&= \sigma^2(N + \text{trace}(\mathbf{S}^2)) + \text{Bias}(\lambda)^2
\end{aligned}$$

where we use that $E(\hat{\mathbf{f}}|\mathbf{X}) = E(\mathbf{SY}|\mathbf{X}) = \mathbf{Sf}$.

To derive the above formula we use the following decomposition for any $\mathbf{Y}^{\text{new}}$:

$$\begin{aligned} ||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2 &= ||(\mathbf{Y}^{\text{new}} - \mathbf{f}) + (\mathbf{f} - \mathbf{Sf}) + (\mathbf{Sf} - \hat{\mathbf{f}})||^2 \\ &= ||\mathbf{Y}^{\text{new}} - \mathbf{f}||^2 + ||\mathbf{f} - \mathbf{Sf}||^2 + ||\mathbf{Sf} - \hat{\mathbf{f}})||^2 \\ &\quad + 2(\mathbf{Y}^{\text{new}} - \mathbf{f})^T(\mathbf{f} - \mathbf{Sf}) \\ &\quad + 2(\mathbf{Y}^{\text{new}} - \mathbf{f})^T(\mathbf{Sf} - \hat{\mathbf{f}}) \\ &\quad + 2(\mathbf{f} - \mathbf{Sf})^T(\mathbf{Sf} - \hat{\mathbf{f}}) \end{aligned}$$

The first and third cross-products have zero mean because $\mathbf{f}$ is the mean of $\mathbf{Y}^{\text{new}}$ and $\hat{\mathbf{f}}$ has mean $\mathbf{Sf}$. If $\mathbf{Y}^{\text{new}} \perp \mathbf{Y}$ the mean of the second cross-product factorizes and is also zero. It is also possible to give a formula for the squared bias, which depend on $\mathbf{f}$.

$$\text{Bias}(\lambda)^2 = \text{trace}((I - \mathbf{S})^2\mathbf{f}\mathbf{f}^T).$$

If we take $\mathbf{Y}^{\text{new}} = \mathbf{Y}$ instead in the decomposition above, the mean of the second cross-product becomes

$$\begin{aligned} 2E(\mathbf{Y} - \mathbf{f})^T(\mathbf{Sf} - \mathbf{SY})|\mathbf{X}) &= -2E(\mathbf{Y} - \mathbf{f})^T\mathbf{S}(\mathbf{Y} - \mathbf{f})|\mathbf{X}) \\ &= -2E(\text{trace}(\mathbf{S}(\mathbf{Y} - \mathbf{f})(\mathbf{Y} - \mathbf{f})^T)|\mathbf{X}) \\ &= -2\text{trace}(\mathbf{S}\underbrace{E((\mathbf{Y} - \mathbf{f})(\mathbf{Y} - \mathbf{f})^T|\mathbf{X})}_{=\sigma^2 I}) \\ &= -2\sigma^2\text{trace}(\mathbf{S}) \end{aligned}$$

EXTRA: Estimation of σ^2 using low bias estimates

Take

$$\text{RSS}(\hat{\mathbf{f}}) = \sum_{i=1}^{N}(y_i - \hat{\mathbf{f}}_i)^2$$

is a natural estimator of $E(||\mathbf{Y} - \hat{\mathbf{f}}||^2|\mathbf{X})$. Its mean is then

$$\sigma^2(N - (\text{trace}(2\mathbf{S} - \mathbf{S}^2)) + \text{Bias}(\lambda)^2.$$

Choosing a *low-bias* – that is trace$(2\mathbf{S} - \mathbf{S}^2$ is large – model, we expect $\text{Bias}(\lambda)^2$ to be negligible and we estimate σ^2 as

$$\hat{\sigma}^2 = \frac{1}{N - \text{trace}(2\mathbf{S} - \mathbf{S}^2)}\text{RSS}(\hat{\mathbf{f}}).$$

From this point of view

$$\text{trace}(2\mathbf{S} - \mathbf{S}^2)$$

can be justified as the *effective degrees of freedom.*

Note that for a projection P we have $P^2 = P$ and hence

$$\text{trace}(2P - P^2) = \text{trace}(P^2) = \text{trace}(P) = \dim(\text{image}(P)).$$

There exists a discussion in the literature on what the most suitable generalization of the degrees of freedom is. One reference is the book *Generalized Additive Models* by Hastie

and Tibshirani. In the context above trace($2\mathbf{S} - \mathbf{S}^2$) turned out to play the same role as the degrees of freedom does in the usual variance estimator in a regression setup. In other contexts we will see that trace($\mathbf{S}$) pops out as the relevant replacement of the degrees of freedom. Historically at least the computation of the trace of $\mathbf{S}$ was faster and therefore preferred. One should perhaps simply remember not to put too must interpretation into the value of the "effective degrees of freedom" but simply view the number as one-dimensional quantification of model complexity.