Cross-Validation

Herwig Friedl   Erwin Stampfer

3rd May 2001

1 Introduction

Cross-validation is a resampling technique that is often used for the assessment of statistical models, as well as selection amongst competing model alternatives. Basically, it is a method to estimate the prediction error of statistical predictor functions. This technique can be very useful in data problems involving minimal distributional assumptions. It has found many applications ranging from linear regression, partial least squares, ridge regression, classification and discrimination, to smoothing and neural networks, in univariate as well as in multivariate settings.

Cross-validation is rooted in the well-known phenomenon that estimating prediction error on the same data used for model building tends to give downward-biased estimates. The reason for this is that the parameter estimates are optimized to reflect the peculiarities of the data-set. When new data arrive, the model usually performs worse than expected on the grounds of assessment measurements on the training set. A reliable method to estimate the prediction error is especially necessary for very flexible models, e.g. neural networks and tree-based classifiers, where overfitting is ubiquitous.

2 The Classical Concept

It was Stone [34] who gave one of the first clearly stated accounts on this topic. In the ideal situation a model is validated by applying it to new data deriving from the same source, i.e. having the same probability distribution. This will lead to a trustworthy evaluation of the prediction error. But in many cases one is forced to use the same data for model building and its assessment.

The idea of cross-validation is to partition the dataset into a training set and a test set. The test set can now be used for testing the model which was built on the training set. This is done repeatedly for a (possibly large) number of divisions of the data. The results of the test procedures are suitably combined and used for subsequent model selection and assessment. Omission of data items is at the heart of cross-validation, a point that has lead to confusion with the technically quite different jackknife method.

We exemplify this for the leave-one-out case and write $S = \{(x_i, y_i)|i = 1, \ldots, n\}$ for the sample, where x_i and y_i may be of quite general form. Let $\hat{y}(x; \alpha, S)$ denote a prediction rule, where α controls some aspect of the model, e.g. the complexity in variable selection problems or the smoothing parameter as in the constructed example below. To allow for data driven selections of α we write $\alpha = \alpha(S)$.

EXAMPLE 1. Figure 1(a) contains $n = 100$ data points (x_i, y_i) simulated from the model $y_i = m(x_i) + \epsilon_i$, where $m(x) = 2\sin(2x)$ for $x \in [-\pi/2, \pi/2]$ and $m(x) = 0$ elsewhere. We took x_i uniform on $[-\pi, \pi]$, $\epsilon_i \overset{iid}{\sim} N(0, 0.3^2)$ and fitted a local quadratic regression model (LOESS) to the data giving a smooth regression curve estimate $\hat{y}(x; \alpha, S)$. The parameter α specifies the (constant) number $n \cdot \alpha$ of neighboring observations of x_i used to calculate the local fits. Thus, the bandwidth α determines the degree of smoothness of the fitted curve. Choosing α too small leads to undersmoothing, i.e. the curve tends to picture idiosyncracies of the current dataset, whereas selecting α too large may result in a oversmoothed curve that ignores local features of the data. Figure 1(b) shows $\hat{y}(x; \alpha, S)$ for two values of α. Two questions arise here, namely how should we choose α and how should we assess the prediction ability of the resulting fit.

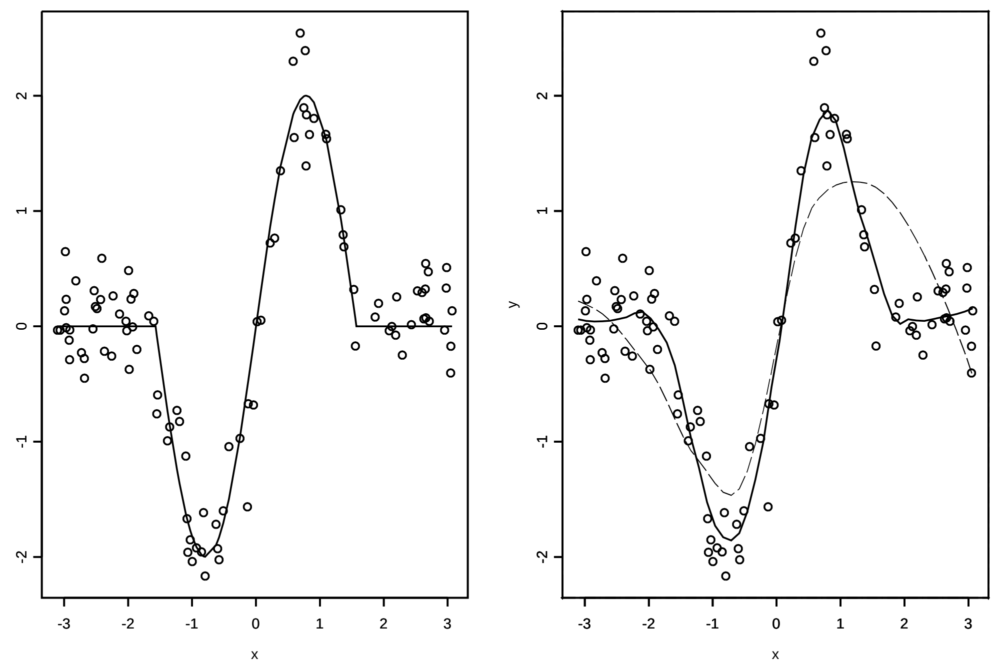

Figure 1: (a) simulated data with underlying modell $m(x)$; (b) LOESS fits with $\alpha = 1/3, 2/3$

Denote by $L(y, \hat{y})$ a loss function measuring the difference between the observed and the predicted value of y. The quadratic loss $L(y, \hat{y}) = (y - \hat{y})^2$ is widely used in regression-type models. The expected prediction error rate is defined as $\Delta(\alpha) = \mathrm{E}\left\{L[y_0, \hat{y}(x_0; \alpha, S)]\right\}$, where the expectation has to be taken with respect to new data (x_0, y_0) (from the same source) and all possible samples S. Let us consider the problem of selection and assessment of a model from a theoretical point of view. A straightforward approach is to select α equal to the value α_0 that minimizes $\Delta(\alpha)$. The resulting predictor can be assessed through $\Delta = \Delta(\alpha_0)$. However, the distribution of the data is usually unknown, so we have to find reliable estimates of $\Delta(\alpha)$ to implement this selection and assessment strategy.

A naive approach is to use the so-called apparent error. The naive selection is then given by the value $\alpha_A(S)$ that minimizes the apparent error $\hat{\Delta}_A(\alpha) = n^{-1}\sum_i L[y_i, \hat{y}(x_i; \alpha, S)]$. This selection may be assessed by $\hat{\Delta}_A[\alpha_A(S)]$, which typically underestimates the prediction error. In the above example this would select α as small as possible and will give a fitted curve that almost interpolates the data. We will see that cross-validation gives more sensible results in this case.

Let us first discuss cross-validatory assessment, i.e. assume that an arbitrary selection $\alpha(S)$ is available. We assess its performance using the leave-one-out cross-validated prediction error estimate given by

$$\hat{\Delta}_{CV} = \frac{1}{n}\sum_{i=1}^{n} L\left\{y_i, \hat{y}[x_i; \alpha(S_{-i}), S_{-i}]\right\} \tag{1}$$

where S_{-i} denotes the sample with the ith item omitted and $\alpha(S_{-i})$ is the choice for α based on the reduced sample S_{-i}. We expect this to be a more realistic estimate for Δ, because each summand compares y_i to $\hat{y}_i$, where the ith data point has not been used in the fitting procedure, i.e. is new data. Note that (1) can be rather cumbersome to evaluate, since it involves n fits to the model.

The cross-validatory selection of α is somewhat more complex. It is defined as the value $\alpha_{CV}(S)$ that minimizes the function

$$\hat{\Delta}_{CV}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{y}(x_i; \alpha, S_{-i})] \tag{2}$$

In some cases analytic expressions for $\alpha_{CV}(S)$ can be found, but very often one resorts to numerical methods or simulation techniques requiring a large amount of computation. Formulas (1) and (2) cover the main part of applications of cross-validation. Note the difference between them. The former allows assessment of any data dependent choice of α, not necessarily cross-validatory. In contrast the latter implements a specific selection strategy by using cross-validation.

Example 1 (continued). We consider three selections for α, $\alpha_1(S) = 1/3$ and $\alpha_2(S) = 2/3$, both not depending on the data at all, and the cross-validatory choice $\alpha_{CV}(S)$, which is implemented numerically by calculating (2) with quadratic loss $L(y, \hat{y}) = (y - \hat{y})^2$ for a grid of α's and selecting the minimizer of this function. These selections are subsequently assessed by the apparent error and by the cross-validated error $\hat{\Delta}_{CV}$. Figure 2(b) shows $\hat{\Delta}_{CV}(\alpha)$ for various values of α and suggests to use $\alpha_{CV}(S) = 0.15$. Besides, it indicates the problem of multiple local minima, a drawback which sometimes occurs with this technique. Figure 2(a) contains the fits $\hat{y}(x; \alpha, S)$ for $\alpha \in \{1/3, 2/3\}$ (dashed lines) and for the cross-validated choice 0.15 (solid line). The solid line gives a much better picture of $m(x)$ than the dashed do. However, it seems to undersmooth, the curve 'wiggling' too much in the margins. Table 1 summarizes the results, containing the selected values of α and their respective assessment statistics. Clearly, cross-validation yields fewer optimistic prediction error estimates. Note that $\hat{\Delta}_{CV}$ leads to the so-called 'two deep' analysis for the cross-validated selection $\alpha_{CV}(S)$. Applying (1) to $\alpha_{CV}(S_{-i})$ gives $\hat{\Delta}_{CV} = n^{-1} \sum_i L\{y_i, \hat{y}[x_i; \alpha_{CV}(S_{-i}), S_{-i}]\}$, where $\alpha_{CV}(S_{-i})$ is defined to be the minimizer of $(n-1)^{-1} \sum_{j \neq i} L[y_j, \hat{y}(x_j; \alpha, S_{-ij})]$, S_{-ij} being the sample without the ith and jth items. This value is not identical with the minimum of $\hat{\Delta}_{CV}(\alpha)$, which is 0.107.

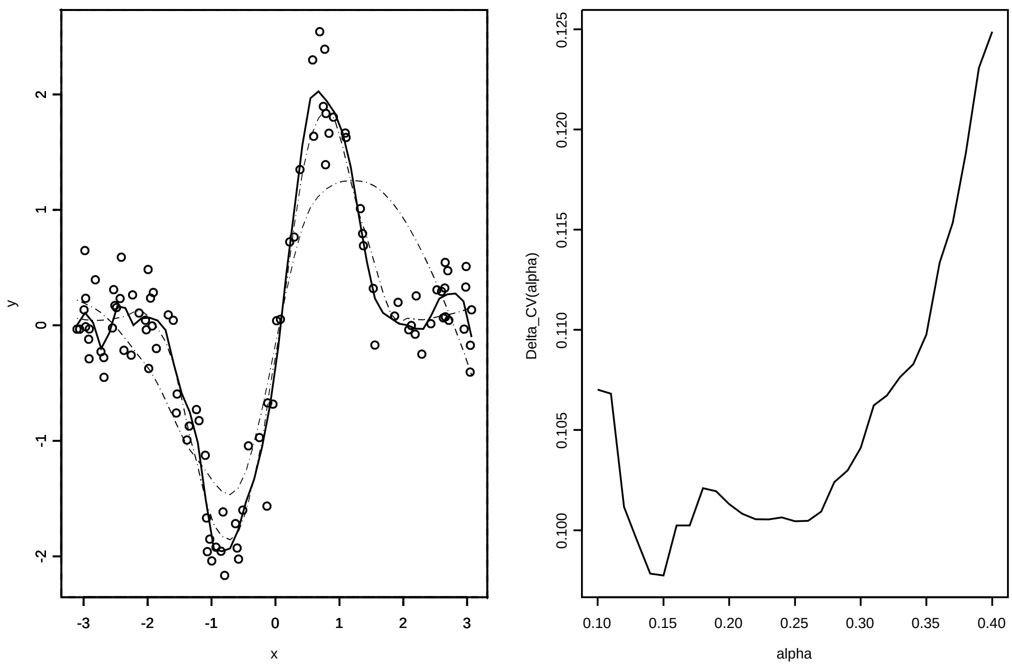

Figure 2: (a) LOESS fits based on $\alpha \in \{1/3, 2/3\}$ (dashed) and on cross-validated $\alpha_{CV}(S) = 0.15$ (solid); (b) $\hat{\Delta}_{CV}(\alpha)$ for $\alpha = 0.1, 0.11, \ldots, 0.4$

3 Regression Curve Fitting

Let us first state the concept of cross-validation for simple linear least squares regression models with $\mathrm{E}(y_i) = x_i^\top \beta$ and $\mathrm{var}(y_i) = \sigma^2$. In matrix notation we write $y = X\beta + \epsilon$, where $y = (y_1, \ldots, y_n)^\top$ contains the responses, $X = (x_1, \ldots, x_n)^\top$ is an $n \times p$ matrix including p explanatory covariates, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ denotes

Table 1: Assessment statistics for various α selections

Selection	$\hat{\Delta}_A$	$\hat{\Delta}_{CV}$
$\alpha_1(S) = 1/3$	0.094	0.115
$\alpha_2(S) = 2/3$	0.317	0.339
$\alpha_{CV}(S) = 0.15$	0.061	0.114

the vector of errors, assumed to be independent. With $\hat{\beta} = (X^\top X)^{-1} X^\top y$, the least squares predictor $\hat{y}(x; S)$ is $x^\top \hat{\beta}$. Note that here is no α in the predictor at all. For a cross-validatory assessment of the models' predictive performance, we further need the least squares predictor $\hat{y}(x_i; S_{-i}) = x_i^\top \hat{\beta}_{-i}$, which is based on the sample with (x_i, y_i) omitted. Specifying a quadratic loss in (1) provides the cross-validatory assessment statistic

$$\begin{aligned} \hat{\Delta}_{CV} &= \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{y}(x_i, S_{-i})]^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - x_i^\top \hat{\beta})^2}{(1 - h_i)^2} \end{aligned} \tag{3}$$

where h_i, the ith diagonal element of the $n \times n$ matrix $H = X(X^\top X)^{-1} X^\top$, is the leverage for the ith data point. Hence, cross-validation supplies an estimator of the models mean squared error of prediction (MSEP). Alternative MSEP estimates often require the estimation of the error variance whereas the cross-validation statistic does not. However, in general cross-validation overestimates the MSEP for the linear regression case (Bunke & Droge [4]), which is a disadvantage compared to other estimates like C_p (Mallows [28]) or those derived by the bootstrap (Efron [12]). Efron [14] investigated the relations between cross-validation, jackknife and some bootstrap estimates of the prediction error. He remarked that cross-validation gives reasonable estimates for smooth predictors. Bunke & Droge [4] showed that cross-validation does not give reliable results, especially for larger model dimensions. Most of these poor properties of the simple cross-validation criterion also hold for a more general case of regression curve fitting (see Efron [15]).

To see how cross-validation works for linear model selection, we additionally specifiy α as a subset of $\{1, \ldots, p\}$. Now we write $y = X(\alpha)\beta(\alpha) + \epsilon$ for the model M_α that only includes β_j if $j \in \alpha$. Given data we want to choose the best model M_{α^*} out of the $2^p - 1$ possible ones, meaning that covariates are included if their respective parameter is nonzero. We call models M_α good models if $\alpha \supset \alpha^*$. Now formula (2) with quadratic loss can be applied to this problem of variable selection, which amounts to choosing α equal to $\alpha_{CV}(S)$, that minimizes

$$\begin{aligned}\hat{\Delta}_{CV}(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}(x_i;\alpha,S_{-i})]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{[y_i - x_i^\top(\alpha)\hat{\beta}(\alpha)]^2}{[1-h_i(\alpha)]^2} \qquad (4)\end{aligned}$$

$h_i(\alpha)$ being the leverage of the ith observation under model M_α. Note that the above function corresponds to the PRESS criterion for model selection in general regression (see e.g., Cook & Weisberg [8]).

But this procedure has some deficiencies. Shao [32] pointed out that in linear models, leave-one-out cross-validation is asymptotically equivalent to the Akaike information criterion (AIC) (Akaike [1]), the C_p (Mallows [28]), the jackknife, and the bootstrap (Efron [14]). But these do not provide consistent model selection, meaning that they do not provide the best predictive model with probability 1, as $n \to \infty$. It has been found by Stone [35] that the probability of choosing a good model converges to one, but this is not true for choosing the best one. These results suggest that leave-one-out cross-validation will overfit in many situations, i.e. choose a too complex model.

More generally, we consider models of the form $y_i = m(x_i) + \epsilon_i$, where the regression function $m(\cdot)$ is estimated by a class of estimators $m(\cdot,\alpha)$. There exists a variety of techniques where those prediction estimators are also linear in the observations, i.e. $\hat{y}(x;\alpha,S) = \hat{m}(x,\alpha) = H(\alpha)y$. This includes variable selection in linear and in additive models, ridge regression, smoothing splines and kernel type estimates, amongst others (cf. Droge [11]). The projection matrix $H(\cdot)$ maps the observations onto the vector of predicted values. In all these models, α can be interpreted as a smoothing parameter. For many linear predictors the leave-one-out assessment criterion (1) for α, which is originally defined as the mean sum of squared leave-one-out residuals, can be rewritten in terms of weighted ordinary residuals, i.e.

$$\hat{\Delta}_{CV} = \frac{1}{n}\sum_{i=1}^{n}\frac{[y_i - \hat{y}(x_i;\alpha,S)]^2}{[1-h_i(\alpha)]^2} \qquad (5)$$

where $h_i(\alpha)$ corresponds to $H(\alpha)$. But there are again several drawbacks with the statistic (5). It generally overestimates the MSEP and fails for some nonlinear regression functions (see Droge [11] for an example). To account for these disadvantages modifications of the criterion were introduced. One is the generalized cross-validation (GCV) criterion by Craven & Wahba [9], which is based on replacing the weights $[1-h_i(\alpha)]^{-2}$ in (5) by their average value $[1-\mathrm{tr}(H(\alpha))/n]^{-2}$. Instead of calculating each h_i seperately as in (5) we still have to compute the trace of the matrix $H(\alpha)$. This calculation can be a burdensome task. Silverman [33] therefore proposed the asymptotic generalized cross-validation (AGCV) score which is much easier to calculate. Simulations showed that the AGCV criterion performs sometimes even better than the respective GCV score.

In order to get information on treatment contrasts whilst avoiding bias due to other environmental factors, Green [19] considered generalized least squares and its equivalent smoothing formulation to model yields in an agricultural field experiment. An extra parameter α was introduced in the variance function accounting for the spatial context, and its value was chosen by cross-validation and GCV. Both criteria showed a similar performance for this case.

Nychka [29] constructed confidence intervals for the value of α, which minimizes the average squared error, by using its GCV choice. Kent & Mohammadzadeh [25] focused on algorithmic aspects for finding the global minimum of the GVC criterion that sometimes can have multiple local minima.

The full cross-validation (FCV) criterion by Bunke at al. [5] derives from replacing the observed value y_i by its least squares prediction $\hat{y}(x_i; \cdot)$ instead of leaving it out. Hence, this idea also adjusts for the sample size n and results in $[1 + h_i(\alpha)]^2$ for the residual weights in (5). It also provides a valid assessment of the prediction performance even when ordinary cross-validation or GCV fails. A combined version gives the generalized full cross-validation (GFCV) criterion, where the residuals in (5) are weighted by $[1 + \mathrm{tr}(H(\alpha))/n]^2$. The properties of all four estimates were compared in Droge [11] where it was shown that both full versions outperform their traditional counterparts with respect to their bias and variance.

In the context of kernel regression estimation the optimal kernel bandwidth is chosen to minimize the prediction error. Its resubstitution (plug-in) estimate reflects an optimistic bias yielding a bandwidth h which is too small and, hence, often undersmooths the curve. Härdle & Marron [22] pinpointed that minimizing the respective cross-validation score is asymptotic equivalent to the minimization of the integrated mean squared error distance. They also showed that this desired property does not hold for the plug-in estimate. This problem can be handled by multiplying the score by a correction factor which depends on $n^{-1}h^{-1}$. An example for such a correction is the GCV score

$$\frac{1}{n}\sum_{i=1}^{n}\frac{[y_i - \hat{y}(x_i; h, S)]^2}{[1 - n^{-1}h^{-1}K(0)]^2}$$

where $K(0)$ is the kernel function evaluated at 0. As indicated in Härdle et al. [23] the ordinary cross-validation score can be (asymptoticaly) considered as the $n^{-1}h^{-1}$ corrected version of the plug-in estimate, hence showing up the desired result of asymptotic equivalence.

4 Generalizations

Leave-one-out cross-validation has been the main focus of research in the past decades. It has been observed that the prediction error estimate based on omitting only one data point tends to be quite variable, especially when the fitted values do not depend smoothly on S. Beyond it leave-one-out does not yield consistent model selection. Similar to the jackknife, some of these drawbacks can be remedied by using leave-n_a-out cross-validation, where one sets aside assessment data of size n_a instead of size 1 (Geisser [18]). An interesting discussion of those multifold cross-validation methods is given in Zhang [36].

A widespread version is the grouped or v-fold cross-validation (Breiman et al. [3]), where the dataset is randomly divided into v groups $S_{a,k}$, $k = 1, \ldots, v$, of roughly the same size $n_a = n/v$. Each group constitutes a test set to assess the results based on the remaining $n - n_a$ training cases. Leave-one-out cross-validation is applied to these groups leading to the v-fold cross-validation estimate

$$\hat{\Delta}_{vCV}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} L[y_i, \hat{y}(x_i; \alpha, S_{-k(i)})] \tag{6}$$

where $S_{-k(i)}$ are the data without $S_{a,k(i)}$ containing y_i. It has been observed that (6) reduces variance, but of course at the cost of a larger bias (Bunke & Droge [4]). To account for that, Burman [6] introduced a correction term and showed that the modified method works well for model selection purposes. Davison & Hinkley [10] recommended use of $v = \min\{n^{1/2}, 10\}$. Additionally, a repeated learning-testing method is discussed in Burman [6], where the data are repeatedly split randomly in two parts (thus it is essentially a bootstrap method). As for the v-fold method, a bias correction term has also to be used here. Both modifications can be utilized, if the leave-one-out version is computationally very expensive.

A more elaborate and expensive version of cross-validation involves leaving out all possible subsets of size n_a. Note that there is the (huge) number of $R = \binom{n}{n_a}$ possibilities to form assessment sets $S_{a,r}$, $r = 1, \ldots, R$, of size n_a and training sets $S_{t,r}$ of size $n - n_a$. Formulas (1) and (2) essentially remain the same, with the distinction that one takes the average prediction error over all R data partitions. In what follows we focus on selection problems, i.e. on

$$\hat{\Delta}_{CV,n_a}(\alpha) = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{n_a} \sum_{j \in S_{a,r}} L[y_j, \hat{y}(x_j; \alpha, S_{t,r})] \tag{7}$$

For linear regression the probability of selecting the best model based on (7) converges to one (for $n \to \infty$) provided that $n - n_a \to \infty$ and $n_a/n \to 1$ (Shao [32]).

Normally it is not feasible to calculate R model fits, so versions of cross-validation have been devised, that are computationally more efficient. We are going to describe some of them. Let $\mathcal{D}$ be a set of subsets of $\{1, \ldots, n\}$, all of size n_a. We impose the following balance conditions on $\mathcal{D}$: (a) every i, $1 \leq i \leq n$, appears in the same number of subsets in $\mathcal{D}$; (b) every pair i, j, $1 \leq i < j \leq n$, appears in the same number of subsets in $\mathcal{D}$. This method is called balanced incomplete cross-validation (BICV). With cardinality of $\mathcal{D}$ equal to $R_{\mathcal{D}}$ we have

$$\hat{\Delta}_{BICV,n_a}(\alpha) = \frac{1}{R_{\mathcal{D}}} \sum_{r=1}^{R_{\mathcal{D}}} \frac{1}{n_a} \sum_{j \in S_{a,r}} L[y_j, \hat{y}(x_j; \alpha, S_{t,r})] \tag{8}$$

Another possibility is to use a Monte-Carlo approximation to (7). It consists of randomly selecting R test sets $S_{a,r}$ of size n_a from the whole dataset and doing the calculations in (7) only for these ones. We denote this estimate by $\hat{\Delta}_{MCCV}(\alpha)$. For the linear regression model, we write (y_{-r}, x_{-r}) for the data matrix without the assessment set $S_{a,r}$. It can be shown that

$$\| y_{-r} - \hat{y}(x_{-r}; \alpha, S_{t,r}) \|^2 = \| [I_{n_a} - H_{-r}(\alpha)]^{-1}[y_{-r} - X_{-r}(\alpha)\hat{\beta}(\alpha)] \|^2 \tag{9}$$

where $\| a \| = a^\top a$ and $H_{-r}(\alpha) = X_{-r}(\alpha)(X^\top(\alpha)X(\alpha))^{-1}X_{-r}^\top(\alpha)$. This gives

$$\hat{\Delta}_{CV,n_a}(\alpha) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_a} \| [I_{n_a} - H_{-r}(\alpha)]^{-1}[y_{-r} - X_{-r}(\alpha)\hat{\beta}(\alpha)] \|^2 \tag{10}$$

and similar formulas for $\hat{\Delta}_{BICV,n_a}(\alpha)$ and $\hat{\Delta}_{MCCV,n_a}(\alpha)$. An analytic first-order approximation of $\hat{\Delta}_{BICV}(\alpha)$ is given by Shao [32] who has shown that (linear) model selection based on $\hat{\Delta}_{BICV}(\alpha)$, its first-order approximation, as well as $\hat{\Delta}_{MCCV}(\alpha)$, is consistent.

So far we have discussed situations where the test set is independent of the training set. Burman et al. [7] extended cross-validation to observations $y_1, \ldots, y_N$ from a general stationary sequence. A least squares fit of the p parameter model

$$y_i = \theta_1 + \theta_2 y_{i-1} + \cdots + \theta_p y_{i-p+1} \tag{11}$$

gives estimates $\hat{\theta}_1, \ldots, \hat{\theta}_p$. Thus, the cases here are the (overlapping) sequences $y_i, \ldots, y_{i-p+1}$, $i = p, \ldots, N$, and the expected prediction error is

$$\Delta = E[(\tilde{y}_p - \hat{\theta}_1 - \hat{\theta}_2\tilde{y}_{p-1} - \cdots - \hat{\theta}_p\tilde{y}_1)^2] \tag{12}$$

where $\tilde{y}_1, \ldots, \tilde{y}_N$ is another process with the same distribution as $y_1, \ldots, y_N$ but independent of it. One may apply ordinary cross-validation resulting in

$$\hat{\Delta}_{CV} = \frac{1}{n}\sum_{i=p}^{N}(y_i - \hat{\theta}_{1,-i} - \hat{\theta}_{2,-i}y_{i-1} - \cdots - \hat{\theta}_{p,-i}y_{i-p+1})^2 \tag{13}$$

where $\hat{\theta}_{j,-i}$, $j = 1, \ldots, p$, is the least squares estimate of θ_j with the ith case omitted. However, this estimate for the prediction error may be very poor because of the dependence of the observations. For calculation of the ith least squares estimate Burman et al. [7] proposed to delete a block of h cases on either side of the ith case. The remaining $N-p-2h$ observations are used to fit the model which is assessed on the ith observation. Note that the $2h$ observations are not used at all in one cycle of this procedure, ensuring almost independent training and test sets. So this strategy is leave-$(2h+1)$-out with respect to omission and leave-one-out cross-validation with respect to assessment. The choice of h depends on the underlying structure of the process which is usually unknown. Burman et al. proposed to use a fixed fraction p, $0 \leq p \leq 1/2$, of $N-p+1$ for omission and to correct the underuse of the sample by a simple term. In a simulation study he found that corrected h-block cross-validation works well for p in a range of $1/5$ to $1/2$.

5 Classification Error Estimates

In classification problems one likes to predict the membership of an observation to one out of a number of classes. A widely used loss function is $L(y, \hat{y}) = 1$ for $y \neq \hat{y}$, and 0 otherwise, amongst others differentially weighting the misclassification cases. The expectation $\Delta = \mathrm{E}\,[L(y, \hat{y})]$ is called the error rate and can be estimated by the apparent error $n^{-1} \sum_i L[y_i, \hat{y}(x_i; S)]$, i.e. the proportion of observed errors made by $\hat{y}(x; S)$ on its training set. As in regression problems the apparent error is positively biased for the error rate. Cross-validation circumvents this problem, by leaving out observations, calculating the classification function on the training set and applying it to the test set. The cross-validated estimate of the error rate again counts misclassified observations, but these where not used in the training phase.

Among the first authors proposing cross-validation to obtain more trustworhty estimates of the error rate was Lachenbruch & Mickey [26], recommending its usage especially when normality of the data is questionable. Further useful material may be found in Hand [21] and Ripley [31].

Very insightful work on this topic is provided by Efron ([13] [14]), and Efron & Gong [16], especially stressing the connection of cross-validation to the bootstrap and jackknife. It is noted there, that leave-one-out cross-validation reduces the bias of the error rate estimate from $O(1/n)$ to $O(1/n^2)$, but leads to possibly unacceptable high variability for small n. This again may be remedied by using leave-n_a-out cross-validation.

Example 2. The magnesium and creatinine contents in the urine of $n = 90$ children from two different parts in Styria, Austria, were monitored. One question concerned the possibility of discriminating between individuals from the control or from the industrial region. Based on a heteroskedastic (quadratic) discrimination model we aimed to estimating the classifiction error. The apparent error rate (plug-in estimate) and different variants of cross-validated estimates (leave-one-out, leave-n_a-out, v-fold) are given in Table 2. Compared to the cross-validated estimates the apparent error seems to be optimistically biased. We utilize a Monte Carlo approximation ($R = 1000$) to the leave-61-out cross-validation estimate, where $n_a = 61$ results from the recommendation in Shao [32] to use test sets of size $n^{3/4} = 29$. For $v = 5$ the error estimate seems to be too large, being considerably smaller for $v = n^{1/2} = 9$.

6 Some Environmental Applications

In a comparative study of different kriging methods to model the deposition of acid through rainfall, Haas [20] analyzed their respective cross-validation residuals. Laslett [27] empirically compared kriging and GCV spline methods. Diagnostic plots of the set of leave-one-out cross-validation distances were used in Bradley & Haslett [2] in order to detect some spatial anomalies in the data points.

Table 2: Classification error estimates for the urine data

Method	Estimate
Apparent error	0.122
Leave-one-out	0.133
Leave-61-out	0.166
Five-fold	0.222
Nine-fold	0.133

Hjorth & Holmqvist [24] adopted cross-validation for variable selection in multivariate autoregression for weather data. They combined the idea of stepwise regression with cross-validation. Since they were dealing with dependent cyclic (five-year) data, they simply left out all observations of one entire year at a time.

Indirect GVC (Oehlert [30]) is another idea appropriate for dependent observations, and was introduced for time series models on data from a network of monitoring stations. The covariance matrix Σ weights the residual vectors in the GCV criterion; that is, for a class of linear predictors $H(\alpha)y$, the projection matrix is chosen to minimize the criterion $n[y - H(\alpha)y]^\top \Sigma^{-1} [y - H(\alpha)y][\mathrm{tr}(I - H(\alpha))]^{-2}$.

7 Software

Modern statistical software makes use of cross-validation in many methodological fields. SAS allows for cross-validated density estimates, prediction error estimates in discriminance analysis, as well as GCV in smooth curve fitting. Amongst the most important applications of cross-validation is choosing the number of factors in partial least squares regression. There are a few SAS macros for cross-validation, offered by independent authors, e.g. `CVLR` (C. Moore, `http://www.arches.uga.edu/~ctmoore/prof/index.htm`) for cross-validation in logistic regression, or `SRS` (F. Carinci, `http://statbone.cmns.mnegri.it/software/misc/srs.html`) designed for simple random sampling for v-fold cross-validation.

S-PLUS provides a number of procedures that are enhanced by cross-validation options, especially bandwidth selection for density estimation, prediction error estimation for discriminance analysis, as well choosing the penalty coefficient for smoothing splines. A more general-purpose function for cross-validation of least squares regression is contained in the package of Efron & Tibshirani [17]. Davison & Hinkley [10] provide a function to cross-validate the prediction error of generalized linear models.

References

[1] Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

[2] Bradley, R. and Haslett, J. (1992). High-interaction diagnostics for geostatistical models of spatially referenced data. *The Statistician* **41**, 371-380.

[3] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth and Brooks, Belmont.

[4] Bunke, O. and Droge, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models. *The Annals of Statistics* **12**, 1400-1424.

[5] Bunke, O., Droge, B. and Polzehl, J. (1993). Model selection and variable transformations in nonlinear regression. *CORE Discussion Paper*, 9327 CORE, UCL, Belgium.

[6] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503-514.

[7] Burman, P., Chow, E. and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika* **81**, 351-358.

[8] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, London.

[9] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics* **31**, 377-403.

[10] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications.* Cambridge University Press, Cambridge.

[11] Droge, B. (1996). Some comments on cross-validation. In: *Statistical Theory and Computational Aspects of Smoothing* (Härdle, W., Schimek, M. G. eds.), 178-199, Physica-Verlag, Heidelberg.

[12] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1-26.

[13] Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans.* SIAM, Philadelphia.

[14] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316-331.

[15] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461-470.

[16] Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37**, 36-48.

[17] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

[18] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Society* **70**, 320-328.

[19] Green, P. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika* **72**, 527-537.

[20] Haas, T. C. (1990). Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association* **85**, 950-963.

[21] Hand, D. J. (1997). *Construction and Assessment of Classification Rules.* Wiley, New York.

[22] Härdle, W. and Marron, J. S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72**, 481-484.

[23] Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association* **83**, 86-95.

[24] Hjorth, U. and Holmqvist, L. (1981). On model selection based on validation with applications to pressure and temperature prognosis. *Applied Statistics* **30**, 264-274.

[25] Kent, J. T. and Mohammadzadeh, M. (2000). Global optimization of the generalized cross-validation criterion. *Statistics and Computing* **10**, 231-236.

[26] Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1-10.

[27] Laslett, G. M. (1994). Kriging and splines: an empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association* **89**, 391-400.

[28] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[29] Nychka, D. (1991). Choosing a range for the amount of smoothing in nonparametric regression. *Journal of the American Statistical Association* **86**, 653-664.

[30] Oehlert, G. W. (1993). Regional trends in sulfate wet deposition. *Journal of the American Statistical Association* **88**, 390-393.

[31] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

[32] Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486-494.

[33] Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B* **47**, 1-52.

[34] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 111-147.

[35] Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B* **41**, 276-278.

[36] Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics* **21**, 299-313.

Herwig Friedl, Erwin Stampfer
Institute of Statistics
Technical University Graz
Steyrergasse 17
A-8010 Graz, Austria.