Jackknife Resampling

Herwig Friedl Erwin Stampfer

3rd May 2001

The jackknife is a particular **Resampling method** that aims primarily at the calculation of the bias and the variance of estimates, without making very restrictive distributional assumptions. It is nonparametric and easy to apply in quite general settings. As with other resampling techniques (like **Bootstrap resampling** or **Cross-Validation**), the jackknife can be computationally very intensive. In the simplest case the jackknife resamples are generated by sequentially deleting single cases from the original sample (delete-one jackknife). A more generalized jackknife technique uses resamples that are based on multiple case deletion (delete-d jackknife). The methodology was introduced by Quenouille in [24].

The main part of this article focuses on independently and (in the introductory part also) identically sampled data. Then we discuss the basics of the jackknife and its applications to nonidentically distributed observations that can be described by linear, generalized linear, nonlinear, and Cox's regression models. Finally, we give an account of jackknife estimates for dependent data, e.g. in the context of **generalized estimating equations** or special **time series** models.

To describe the jackknife method we first start with a sample $Y = (Y_1, \ldots, Y_n)$ and assume that the Y_is are drawn independently and identically (iid) from a distribution F. Our interest lies on some quantity θ, possibly vector valued, that can be an unknown parameter β or any function of it, say $g(\beta)$. We estimate θ by the statistic $\hat{\theta} = \hat{\beta}(Y)$ or $\hat{\theta} = g(\hat{\beta}(Y))$ and we are interested in distributional characteristics of this statistic. Denote the sample without the ith datum by $Y_{(i)}$. Now the ith deletion value of $\hat{\theta}$ is defined as $\hat{\theta}_{(i)} = \hat{\beta}(Y_{(i)})$ or $\hat{\theta}_{(i)} = g(\hat{\beta}(Y_{(i)}))$, respectively. Let $\hat{\theta}_{(\cdot)} = n^{-1} \sum_i \hat{\theta}_{(i)}$ denote the mean over all deletion values. As defined by Quenouille [25] the jackknife bias, a nonparametric estimate of $\mathrm{E}_F(\hat{\theta} - \theta)$, equals

$$\widehat{\mathrm{bias}}_J(\hat{\theta}) = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) \tag{1}$$

which results in the bias-reduced jackknife estimate of θ

$$\begin{aligned} \tilde{\theta} &= \hat{\theta} - \widehat{\mathrm{bias}}_J(\hat{\theta}) \\ &= n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)} \end{aligned} \tag{2}$$

A simple justification for (2) relies on the assumption that $\mathrm{bias}(\hat{\theta}) = an^{-1} + O(n^{-2})$, where a is an unknown term not depending on n. Straightforward calculation shows that $\text{bias}(\tilde{\theta}) = O(n^{-2})$. For example, if $\hat{\theta} = \overline{Y}$, (1) results in $\widehat{\text{bias}}_J(\overline{Y}) = 0$, which is the true bias. Often, $\tilde{\theta}$ is still biased, but the magnitude of its bias is often reduced from $O(n^{-1})$ to $O(n^{-2})$. For a quadratic functional $\hat{\theta}$, (1) is unbiased for estimating its true bias (cf., Efron [7]).

Another application of jackknife resampling is estimating the variance of $\hat{\theta}$. This was first proposed by Tukey in [29], introducing so-called jackknife pseudo-values

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}. \tag{3}$$

Their mean value corresponds to the ordinary (bias-reduced) jackknife point estimator of θ given in (2). Note that in the special case of $\hat{\theta} = \overline{Y}$, the respective pseudo-values are $\tilde{\theta}_{(i)} = Y_i$, which are iid by assumption. Speculating that the pseudo-values (3) are iid variates in general, the jackknife estimate of the variance-covariance matrix of $\hat{\theta}$ is defined as their empirical second moment, i.e.

$$\begin{aligned}
\widehat{\text{var}}_J(\hat{\theta}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_{(i)} - \frac{1}{n}\sum_{j=1}^{n} \tilde{\theta}_{(j)} \right) \left(\tilde{\theta}_{(i)} - \frac{1}{n}\sum_{j=1}^{n} \tilde{\theta}_{(j)} \right)^{\top} \\
&= \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right) \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^{\top} \qquad (4)
\end{aligned}$$

Equivalently, (4) may also be interpreted as the variation of the $\hat{\theta}_{(i)}$ around their mean $\hat{\theta}_{(\cdot)}$ corrected by $(n-1)/n$ to account for the fact that only $n-1$ observations are used to construct $\hat{\theta}_{(i)}$. Again, for the mean of iid variates, (4) gives the standard estimate $\widehat{\text{var}}_J(\overline{Y}) = n^{-1}S^2$, where S^2 is the empirical variance of the Y_is. If we consider the smooth statistic $\overline{Y}^2$ as an estimate of $\mathrm{E}^2(Y) = \mu^2$, then (1) results in $\widehat{\text{bias}}_J(\hat{\theta}) = (n(n-1))^{-1}\sum_i (Y_i - \overline{Y})^2$, i.e. the jackknife yields the right answer and the corresponding $\tilde{\theta}$ in (2) is known to be unbiased.

For two independent samples $Y^1 = (Y_1^1, \ldots, Y_{n_1}^1)$ and $Y^2 = (Y_1^2, \ldots, Y_{n_2}^2)$, where the qantity of interest is a function of their population parameters, i.e. $\theta = g(\theta^1, \theta^2)$, the two-sample jackknife introduced by Arvesen [1] is appropriate. This includes the difference of means or the ratio of variances as customary examples. Let $\hat{\theta} = g(\hat{\theta}^1, \hat{\theta}^2)$ be the estimate of θ based on all values in both samples and define $\hat{\theta}^1_{(i)} = g(\hat{\theta}^1_{(i)}, \hat{\theta}^2)$, $i = 1, \ldots, n_1$, and $\hat{\theta}^2_{(j)} = g(\hat{\theta}^1, \hat{\theta}^2_{(j)})$, $j = 1, \ldots, n_2$. Using the two sample-specific sets of pseudo-values $\tilde{\theta}^1_{(i)} = n_1\hat{\theta} - (n_1 - 1)\hat{\theta}^1_{(i)}$ and $\tilde{\theta}^2_{(j)} = n_2\hat{\theta} - (n_2 - 1)\hat{\theta}^2_{(j)}$, from (2) the bias-reduced two-sample jackknife estimate of θ is their joint mean

$$\tilde{\theta} = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} \tilde{\theta}^1_{(i)} + \sum_{j=1}^{n_2} \tilde{\theta}^2_{(j)} \right) \tag{5}$$

On the other hand, the sum of both jackknife variances serves as the two-sample jackknife variance

$$\widehat{\text{var}}_J(\hat{\theta}) = \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \left(\tilde{\theta}^1_{(i)} - \frac{1}{n_1}\sum_{k=1}^{n_1} \tilde{\theta}^1_{(k)} \right)^2 + \frac{1}{n_2(n_2-1)} \sum_{j=1}^{n_2} \left(\tilde{\theta}^2_{(j)} - \frac{1}{n_2}\sum_{l=1}^{n_2} \tilde{\theta}^2_{(l)} \right)^2 \tag{6}$$

In many relevant cases the bias of $\widehat{\text{var}}_J$ in (4) is positive (see, Efron [7]). In other words (4) is conservative in the sense that it always overestimates the variability of $\hat{\theta}$. Some authors noted that $\widehat{\text{var}}_J$ tends to give more reliable estimates than does $\widehat{\text{bias}}_J$. But the sample median or extremes are striking examples where the jackknife variance estimator (4) is not even consistent and therefore totally fails. This unfavorable behavior always occurs when the statistic $\hat{\theta}$ is not smooth enough and, hence, cannot be approximated linearly. Sample quantiles are prominent examples resulting in inconsistent jackknife variance estimates. In contrast, the bootstrap resampling technique always gives consistent variance estimates. Another deficiency of the jackknife is that it does not directly provide distribution estimators. Both disadvantages can be avoided when we do not base the jackknife on eliminating only single variates but leave out a group of d variates. This resampling plan is called the delete-d jackknife, a technique that has been proposed in Wu [30].

Let s_* be a random subset with d elements from $\{1, \ldots, n\}$ and denote $\hat{\theta}_{(s_*)} = \hat{\theta}(Y_{(s_*)})$ the statistic based on the remaining $n-d$ data points that are obtained by leaving out all $\{Y_i, i \in s\}$. In total there are $N = \binom{n}{d}$ different subsets s_j, $j = 1, \ldots, N$. With the respective deletion values $\hat{\theta}_{(s_j)}$ we obtain the delete-d jackknife variance estimator

$$\widehat{\text{var}}_{J-d}(\hat{\theta}) = \frac{n-d}{dN} \sum_{j=1}^{N} \left(\hat{\theta}_{(s_j)} - \frac{1}{N} \sum_{k=1}^{N} \hat{\theta}_{(s_k)} \right) \left(\hat{\theta}_{(s_j)} - \frac{1}{N} \sum_{k=1}^{N} \hat{\theta}_{(s_k)} \right)^{\top} \tag{7}$$

by using the same arguments as before. For (7) to be consistent, the fraction d of n (as well as n) needs to be large. It can be considered as a term correcting for the nonsmoothness of the statistic $\hat{\theta}$. For the special choice $d = 1$, the delete-d jackknife equals the ordinary delete-one jackknife method.

Another important advantage of the delete-d jackknife method is that it also allows construction of a consistent estimate of the sampling distribution of $\hat{\theta}$, i.e. $\Pr(\sqrt{n}(\hat{\theta} - \theta) \leq y)$ can be estimated by the so-called cumulative jackknife histogram

$$\Pr_* \left\{ \sqrt{\frac{n(n-d)}{d}} \left(\hat{\theta}_{(s_*)} - \hat{\theta} \right) \leq y \right\} = \frac{1}{N} \sum_{j=1}^{N} I \left\{ \sqrt{\frac{n(n-d)}{d}} \left(\hat{\theta}_{(s_j)} - \hat{\theta} \right) \leq y \right\} \tag{8}$$

Here, $\Pr_*$ indicates that $Y_{(s)}$ is considered as a simple random (re)sample of size $n-d$ without replacement from Y, i.e. $\Pr_*(Y_{(s_j)}) = N^{-1}$, and $I(\cdot)$ is an indicator with value 1 if its argument holds. Note that the jackknife delete-d variance estimator (7) for $\hat{\theta}$ is exactly the variance of the respective jackknife histogram, because

$$\begin{aligned} \text{var}_* \left(\sqrt{\frac{n-d}{d}} \left(\hat{\theta}_{(s)} - \hat{\theta} \right) \right) &= \frac{n-d}{Nd} \sum_{j=1}^{N} \left(\hat{\theta}_{(s_j)} - \hat{\theta} - \frac{1}{N} \sum_{k=1}^{N} (\hat{\theta}_{(s_k)} - \hat{\theta}) \right)^2 \\ &= \widehat{\text{var}}_{J-d}(\hat{\theta}) \end{aligned} \tag{9}$$

Wu [31] showed that the jackknife distribution estimate has a convergence rate of $O(n^{-1/2})$. For smooth functions of the mean Booth & Hall [2] improved this rate to

$O(n^{-5/6})$ by mixing the jackknife estimator with the limiting normal distribution. Note that this is still worse than the respective result for the bootstrap estimate which is $O(n^{-1})$. This is due to the fact that the jackknife procedure is based on resampling without replacement whereas the original sample was drawn with replacement from a population considered to be of infinite size. This attitude is exactly reflected by the bootstrap which is based on resampling with replacement.

When $\hat{\theta}$ is difficult to evaluate or n is very large, then it might already become a very burdensome task to apply the delete-one jackknife. So various simplifications have been proposed. An obvious extension results when the n observations are partitioned in g subsets of (approximately) the same cardinality h, i.e. $n = gh$. The deletion estimates $\hat{\theta}_{(i)}$, $i = 1, \ldots, g$, are then calculated on the original observations, with the ith subset deleted. This is sometimes called the grouped jackknife variance estimate

$$\widehat{\mathrm{var}}_{Jg}(\hat{\theta}) = \frac{g-1}{g} \sum_{i=1}^{g} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right) \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^{\top} \tag{10}$$

which therefore only requires g evaluations of the statistic. If there is no apparent way of forming groups, this should be done randomly. Another possibility is to draw a random subsample of size h without replacement and to calculate h copies of delete-one jackknife values by repeatedly leaving out one data point from this subsample (Shao & Tu [28]). Balanced and random subsampling have been proposed for the delete-d jackknife variance estimator, where we draw a random sample of subsets instead of evaluating $\hat{\theta}$ on all subsets of size $n-d$. If the calculation of $\hat{\theta}$ needs a large number of iterations, e.g. when $\hat{\theta}$ is defined as the root of an estimating equation (EE), then approximations to the respective pseudo-values can sometimes be used. These are based on the outcome of a single iteration step starting in $\hat{\theta}$ instead of fully iterating until convergence. In fact, this one-step approximation behaves well for Newton-type methods with quadratic convergence rates. Jorgensen [15] established a correction factor that makes the one-step jackknife variance estimator a better approximation of its full version.

We now describe in more detail the use of the jackknife within the context of regression models for independent responses y_i, $i = 1, \ldots, n$. In what follows we always suppose the units, left out in the jackknife procedure, to be iid. Let the conditional model for the mean be of the simple linear form $\mathrm{E}(y_i|x_i) = \mu_i = x_i^{\top}\beta$ with homoskedastic variability $\mathrm{var}(y_i|x_i) = \sigma^2$. Using matrix notation this is equivalent to $y = X\beta + \epsilon$, where $y = (y_1, \ldots, y_n)^{\top}$ contains the responses, $X = (x_1, \ldots, x_n)^{\top}$ is a matrix of dimension $n \times p$ including intercept and all explanatory variables considered fixed in the conditional model for μ_i, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\top}$ denotes the vector of iid error terms of zero mean and identical variance σ^2. Then the **least squares** estimator $\hat{\beta} = (X^{\top}X)^{-1}X^{\top}y$ has variance-covariance matrix $\mathrm{var}(\hat{\beta}) = \sigma^2(X^{\top}X)^{-1}$ and can be estimated by means of leaving out sequentially the ith complete component (y_i, x_i) resulting in $\hat{\beta}_{(i)}$, say. These jackknife values can be easily computed as

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{y_i - x_i^\top \hat{\beta}}{1 - h_i} (X^\top X)^{-1} x_i \tag{11}$$

where h_i denotes the ith diagonal element of the $n \times n$ hat matrix $X(X^\top X)^{-1} X^\top$. Like $\hat{\beta}$ their average, $\hat{\beta}_{(\cdot)} = n^{-1} \sum_i \hat{\beta}_{(i)}$, is also unbiased but equals $\hat{\beta}$ for identical h_is only. Let $\theta = g(\beta)$ denote a function of the parameter β estimated by $\hat{\theta} = g(\hat{\beta})$ with respective jackknife values $\hat{\theta}_{(i)} = g(\hat{\beta}_{(i)})$. Miller [23] has discussed the ordinary jackknife variance

$$\widehat{\mathrm{var}}_J(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_{(i)} - \tilde{\theta}_{(\cdot)} \right) \left(\tilde{\theta}_{(i)} - \tilde{\theta}_{(\cdot)} \right)^\top \tag{12}$$

with pseudo-values as defined in (3), or equivalently

$$\tilde{\theta}_{(i)} = \hat{\theta} - (n-1) \left(\hat{\theta}_{(i)} - \hat{\theta} \right) \tag{13}$$

Although (12) is biased for estimating $\mathrm{var}(\hat{\theta})$, it is a consistent estimate. The latter is generally not true for the respective jackknife bias estimate

$$\widehat{\mathrm{bias}}_J(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\theta} - \tilde{\theta}_{(i)} \right) \tag{14}$$

This inconsistency of the jackknife bias is mainly due to the imbalance of the considered model. Because the responses y_i differ in their mean value $\mathrm{E}(y_i) = x_i^\top \beta$, they are not identically distributed. A suitable measure for these differences are the values of h_i. Various weighting schemes have been proposed to correct for the imbalance in the data. The weigthed jackknife estimates are all consistent and based on some weigthed versions of the pseudo-values. Hinkley [14] defined

$$\tilde{\theta}_{(i)} = \hat{\theta} - n(1 - h_i)(\hat{\theta}_{(i)} - \hat{\theta}) \tag{15}$$

and used them in the degrees-of-freedom-corrected version of (4),

$$\widehat{\mathrm{var}}_{JH}(\hat{\theta}) = \frac{1}{n(n-p)} \sum_{i=1}^{n} \left(\tilde{\theta}_{(i)} - \tilde{\theta}_{(\cdot)} \right) \left(\tilde{\theta}_{(i)} - \tilde{\theta}_{(\cdot)} \right)^\top \tag{16}$$

and in

$$\widehat{\mathrm{bias}}_{JH}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\theta} - \tilde{\theta}_{(i)} \right) \tag{17}$$

For $\theta = \beta$, we get $\tilde{\beta} = \hat{\beta}$ and, hence, the jackknife bias is exactly zero and correctly reflects the unbiasedness of the least squares estimate $\hat{\beta}$. In this case, (16) reduces to

$$\widehat{\mathrm{var}}_{JH}(\hat{\beta}) = \frac{n}{n-p} \sum_{i=1}^{n} (1 - h_i)^2 \left(\hat{\beta}_{(i)} - \hat{\beta} \right) \left(\hat{\beta}_{(i)} - \hat{\beta} \right)^\top \tag{18}$$

If in addition the model only includes an intercept (iid case), then $p = 1$ and $h_i = n^{-1}$, which gives the ordinary pseudo-values in (3). It has been shown by Wu [30] that (16) is biased but robust against variance heterogeneity.

Alternative jackknife estimates, proposed by Wu [30], are

$$\widehat{\mathrm{var}}_{JW}(\hat{\theta}) = \sum_{i=1}^{n}(1-h_i)\left(\hat{\theta}_{(i)} - \hat{\theta}\right)\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^{\top} \tag{19}$$

and

$$\widehat{\mathrm{bias}}_{JW}(\hat{\theta}) = \sum_{i=1}^{n}(1-h_i)\left(\hat{\theta}_{(i)} - \hat{\theta}\right) \tag{20}$$

The jackknife bias estimates (17) and (20) are generally equal. It can be further shown that the jackknife variance estimates (16) and (19) are equivalent for D-optimal designs in which $h_i = p/n$. If the h_is are not at all provided by statistics software, they also can be replaced by their average value p/n. Similarly, there is also evidence to replace the factor $(n-1)$ by $(n-p)$ in (12). In linear models with heteroskedastic error terms, where we have to handle the unequal variances σ_i^2, the variance-covariance matrix of $\hat{\theta}$ can be estimated robustly by one of the proposed jackknife variances.

If the responses follow a **nonlinear regression** model, $y = f(x, \beta) + \epsilon$, the least squares estimate $\hat{\beta}$ minimizes the criterion $\sum_i [y_i - f(x_i, \beta)]^2$ which, typically, has to be done iteratively. Let $\hat{\beta}_{(i)}$ denote the estimate of β obtained by deleting the ith pair (y_i, x_i) in the sample. Then the ordinary jackknife variance estimator of $\hat{\theta} = g(\hat{\beta})$ is again

$$\widehat{\mathrm{var}}_{J}(\hat{\theta}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \tilde{\theta}_{(\cdot)}\right)\left(\tilde{\theta}_{(i)} - \tilde{\theta}_{(\cdot)}\right)^{\top} \tag{21}$$

where $\hat{\theta}_{(i)} = g(\hat{\beta}_{(i)})$. However, the main problem here is the number of iterations that are possibly required to calculate the jackknife values $\hat{\beta}_{(i)}$. Instead of iterating each time until convergence, the one-step jackknife estimator can be an appealing and attractive alternative. Let the iteration start in the original estimate $\hat{\beta}$ and do only one iteration based on the reduced sample to get $\hat{\theta}^1_{(i)} = g(\hat{\beta}^1_{(i)})$. The calculation of the respective jackknife variance

$$\widehat{\mathrm{var}}^1_{J}(\hat{\theta}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}^1_{(i)} - \tilde{\theta}^1_{(\cdot)}\right)\left(\tilde{\theta}^1_{(i)} - \tilde{\theta}^1_{(\cdot)}\right)^{\top} \tag{22}$$

only needs n single iterations in total. Since we always have to evaluate the second derivative of the criteria function whenever a jackknife value is calculated, Fox et al. [11] have further utilized an approximation to these matrices (linear jackknife estimates). Both jackknife estimates perform quite well and they are again robust against misspecifications of $\mathrm{var}(\epsilon)$.

The same (ordinary or one-step) jackknife techniques are also applied to generalized linear models, Cox's regression models (with randomly censored data), as well as for other complex models, where the maximum (partial) likelihood estimate of the parameter is a solution of the (nonlinear) score function or defined by any EE. While in survival models we also have to adjust the risk set for the deletion of the ith observation, the simple leaving-one-out or the one-step procedure are applied in a straightforward manner for **generalized linear models**. However, if non-canonical link models are considered, the iteratively reweighted least squares procedure is not of a Newton type with quadratic convergence and one-step estimates might fail (see Jorgensen [16] for a better approximation).

A main application of statistical models is the prediction of a random variate y_0 at a vector $x = x_0$ of explaining variables. For example, a fitted linear regression $x^\top \hat\beta$ may be used for predicting a new response y_0 at $x = x_0$ by $x_0^\top \hat\beta$, or a fitted discrimination function $\hat d(x)$, designed to discriminate between a number of classes, may be used to predict class membership for $x = x_0$ by $\hat d(x_0)$. To cover a wide range of cases we let $\eta_Z(x)$ denote a generic prediction rule, constructed on a dataset Z, for example, consisting of the pairs $z_i = (y_i, x_i)$, $i = 1, \ldots, n$, in the regression case.

A prediction rule $\eta_Z(x)$ may become quite complex, so it is of eminent importance to construct measures for its ability to predict the response at (arbitrary) values of x. It is useful to think of z_i as iid random vectors coming from a $(p+1)$-dimensional distribution F. In the classical regression case one may take F to equal the product of $F_X(x)$ and $F_{Y|X}(y|x)$, $F_X(x)$ being arbitrary and $F_{Y|X}(y|x)$ having mean $x^T\beta$ and variance σ^2. We further assume that $\eta_Z(x)$ is of functional form, i.e. there exists $\eta(x, F)$ not depending on n such that $\eta_Z(x) = \eta(x, \hat F)$, with $\hat F$ being the empirical distribution function of Z. We may think of $\eta(x, F)$ as the Bayes rule for discriminating two normal populations with expectation μ_1, μ_2 respectively, and variance Σ and $\eta(x, \hat F)$ its usual plug-in estimate. Now we define a loss function $Q(y, \eta)$, measuring the error between the observation y and its prediction η. $Q(y, \eta)$ may be squared loss $(y - \eta)^2$ for regression problems or $I(y \neq \eta)$ for discrimination problems. We are interested in certain expectations of the loss function.

Define $\theta_0 = \theta(F)$ to equal $\mathrm{E}_{0F}\left[Q(y_0, \eta(x_0, F))\right]$, where E_{0F} denotes the expectation with respect to a future observation $(y_0, x_0) \sim F$, giving the prediction error when F is known. For discriminant analysis this is the probability of an erroneous classification (error rate). The obvious estimate of θ_0 is the apparent error

$$\begin{aligned}\hat\theta = \theta(\hat F) &= \mathrm{E}_{0\hat F}\left[Q(y_0, \eta(x_0, \hat F))\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \eta(x_i, \hat F)) \qquad (23)\end{aligned}$$

It is well known that $\hat\theta$ is (optimistically) biased for θ_0 so we may use the jackknife to improve on it, leading to the bias estimate

$$\widehat{\mathrm{bias}}_J(\hat\theta) = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\hat\theta_{(i)} - \hat\theta\right) \qquad (24)$$

with $\hat{\theta}_{(i)} = \theta(\hat{F}_{(i)}) = (n-1)^{-1}\sum_{j\neq i} Q[y_j, \eta(x_j, \hat{F}_{(i)})]$. Remember that $\eta(x_j, \hat{F}_{(i)})$ is based on all the data with the exception of the ith pair, so recalculation of, for example n discriminant functions is necessary here. In the case of a Bayes classification procedure this bias-reduced estimate may be seen as a lower bound for the achieveable error rate (cf., Ripley [26]).

Usually one is more interested in estimating the so-called true error $\theta_t = \theta(F, \hat{F}) = \mathrm{E}_{0F}\left[Q(y_0, \eta(x_0, \hat{F}))\right]$, which measures the ability of a given rule $\eta_Z(x) = \eta(x, \hat{F})$ to predict future responses. Of course θ_t cannot be evaluated because it depends on F. An obvious estimate is $\theta(\hat{F}, \hat{F})$ which equals the apparent error $\hat{\theta}$ given in (23). It is not evident how to use the jackknife to improve on $\hat{\theta}$ as an estimate of θ_t. One approach considered by Efron & Gong [8] and Gong [13] is the following. Let $\omega(F) = \mathrm{E}[\theta(F, \hat{F}) - \theta(\hat{F}, \hat{F})] = \mathrm{E}(\theta_t - \hat{\theta})$ denote the expected excess error, where the expectation is taken over the sample Z. Note that if we can find an unbiased estimator for $\omega(F)$, we could correct $\hat{\theta}$ such that it has the same expectation as θ_t. The obvious estimate of $\omega(F)$ is $\hat{\omega} = \theta(\hat{F}, \hat{F}) - \theta(\hat{F}, \hat{F}) = 0$, which is apparently biased. The jackknife bias is

$$\widehat{\mathrm{bias}}_J(\hat{\omega}) = (n-1)\hat{\omega}_{(\cdot)} \tag{25}$$

where

$$\begin{aligned}
\hat{\omega}_{(\cdot)} &= \frac{1}{n}\sum_{i=1}^{n} \hat{\omega}_{(i)} = \frac{1}{n}\sum_{i=1}^{n} [\theta(\hat{F}, \hat{F}_{(i)}) - \theta(\hat{F}_{(i)}, \hat{F}_{(i)})] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n} Q(y_j, \eta(x_j, \hat{F}_{(i)})) - \frac{1}{n-1}\sum_{j\neq i} Q(y_j, \eta(x_j, \hat{F}_{(i)}))\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[Q(y_i, \eta(x_i, \hat{F}_{(i)})) - \hat{\theta}_{(i)}\right] \qquad (26)
\end{aligned}$$

The bias-reduced estimate $\hat{\theta} + (n-1)\hat{\omega}_{(\cdot)}$ is supposed to have approximately the same expectation as θ_t (cf., Ripley [26]).

So far we have reviewed jackknife techniques that are appropriate for independent observations. However, sometimes we have to model data with a cluster structure. This means that responses from different clusters are assumed to be independent but those within the same cluster are correlated. Let y_{ij} denote the jth response, $j = 1, \ldots, n_i$, in the ith cluster, $i = 1, \ldots, n$, with respective explanatory vectors x_{ij}, each of dimension $p \times 1$. As before we define the response vector and the design matrix in the ith cluster as $y_i = (y_{i1}, \ldots, y_{in_i})^\top$ and $X_i = (x_{i1}, \ldots, x_{in_i})^\top$. In addition to the mean model $\mathrm{E}(y_{ij}|x_{ij}) = \mu_{ij} = h(X_i\beta)$ and the variance structure $\mathrm{var}(y_{ij}|x_{ij}) = V(\mu_{ij})$, described by the diagonal matrix $V_i = \mathrm{diag}(V(\mu_{i1}), \ldots, V(\mu_{in_i}))$, we further assume that $\mathrm{corr}(y_i|X_i) = R(\alpha)$ holds for each cluster, where α denotes the parameter (vector) in the working correlation matrix. Liang & Zeger [20] introduced the concept of **generalized estimating equations** (GEEs) by utilizing the resulting variance-covariance structure

$C_i = C_i(\alpha, \beta) = V_i^{1/2} R(\alpha) V_i^{1/2}$. The **quasi-likelihood** estimate $\hat{\beta}$ solves the GEE

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta}^{\top} C_i^{-1}(y_i - \mu_i) = 0 \tag{27}$$

a multivariate extension of the quasi-score function in generalized linear models. If the number of clusters n is large enough then the jackknife can be applied by leaving out single clusters instead of single observations in order to keep the dependence structure in the data. No matter how we have specified the working correlation structure $R(\alpha)$, the resulting jackknife estimate is again a consistent estimator.

The employment of the jackknife technique in more complex dependency models, as in general **time series** models, is very limited so far. However, if the stationary sequence of random variables is m-dependent, as in moving-average models, there is a remaining chance to work with groups of independent observations. We partition the entire dataset $\{Y_1, \dots, Y_n\}$ into g groups each of h observations, i.e. $\{Y_{(j-1)h+1}, Y_{(j-1)h+2}, \dots, Y_{jh}\}$, $j = 1, \dots, g$. For large values of h these groups are considered as being approximately independent. The jackknife is now applied on each of these groups. In contrast to the grouped jackknife used in (10), here the observations are grouped according to their order. Carlstein [4] has discussed the grouped jackknife variance estimator of the overall sample mean $\overline{Y}$ which is

$$\widehat{\mathrm{var}}_{Jg}(\overline{Y}) = \frac{1}{g(g-1)} \sum_{i=1}^{g} \left(Z_i - \overline{Y}\right)^2 \tag{28}$$

with group-specific means $Z_i = h^{-1} \sum_{j=1}^{h} Y_{(i-1)h+j}$. Whilst none of the groups overlap, the moving block jackknife in Künsch [17] is based on $n - h + 1$ overlapping blocks $B_i = \{Y_i, Y_{i+1}, \dots, Y_{i+h-1}\}$, $i = 1, \dots, n-h+1$, again each of size $h < n$. The moving block jackknife variance estimator of a statistic $\hat{\theta} = \hat{\theta}(Y)$ is then

$$\widehat{\mathrm{var}}_{Jb}(\hat{\theta}) = \frac{1}{n(n-h+1)h} \sum_{i=1}^{n-h+1} \left(\tilde{\theta}_{(i)} - \frac{1}{n-h+1} \sum_{j=1}^{n-h+1} \tilde{\theta}_{(j)}\right)^2 \tag{29}$$

with pseudo-values $\tilde{\theta}_{(i)} = n\hat{\theta} - (n-h)\hat{\theta}_{(i)}$ where $\hat{\theta}_{(i)}$ is the statistic based on all observations not from block B_i.

Another modified jackknife variance estimator for m-dependent data has been provided by Shao [27]. He studied

$$\widehat{\mathrm{var}}_{JS}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} \sum_{|j-i| \le m} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right) \left(\hat{\theta}_{(j)} - \hat{\theta}_{(\cdot)}\right) \tag{30}$$

where $\hat{\theta}_{(\cdot)} = n^{-1} \sum_i \hat{\theta}_{(i)}$ and $\hat{\theta}_{(i)}$ denotes the statistic from leaving out the ith datum only. The same idea was applied in Lele [19] for non-iid random variables $Y = (Y_1, \dots, Y_n)$, where $\sum_i g_i(Y, \theta) = 0$ is a linear EE for θ, e.g. the score function for nonidentically distributed Y_is, conditional least squares EEs for a Markov process or a maximum pseudo-likelihood EE for spatial processes. If the EEs are not uncorrelated, e.g. $\mathrm{E}(g_i g_j) \neq 0$ for $j \in N(i)$, where $N(i)$ includes all neighbors of i, then $\widehat{\mathrm{var}}_{JS}(\hat{\theta})$ can be applied but $|j - i| \leq m$ has to be replaced by $j \in N(i)$.

There are several other applications of the jackknife method. Kuo [18] employed this resampling technique to evaluate the variance and the mean squared error of composite estimators. A jackknife-based estimate of the population size in capture-recapture models allowing for heterogenious capture probabilities is utilized in Burnham & Overton [3]. Gladen [12] applied the jackknife to estimate the proportion of affected fetuses in teratological experiments. A more general application is the derivation of jackknife-based adjustments to goodness-of-fit statistics introduced in Fay [10] for complex sample designs.

The jackknife has been proposed in many circumstances for estimating the bias and the variance of an estimator. A review is given by Miller [22]. However, multifunctional software is rather scarce. An **S-PLUS** macro `jackknife`, described in the book by Efron & Tibshirani [9], is available. This macro allows use of the jackknife to a rather arbitrary statistic in a generic manner by using the ordinary or the grouped resampling strategy. Another source is the book by Davison & Hinkley [5] providing the S-PLUS macro `jack.after.boot` to produce jackknife-after-bootstrap plots by calculating the jackknife influence values from a bootstrap output object. A general-purpose algorithm for the leave-one-out and the grouped jackknife techniques for an arbitrary statistic (defined by a routine supplied by the user) is decribed in Matloff [21]. Jackknife analyses for simple random samples can also be perfomed with SAS. The `jack` macro can be used to compute approximate standard errors (SEs), bias-corrected estimates, and confidence intervals assuming a normal sampling distribution. This macro and its description is available at `http://ewe3.sas.com/techsup/download/stat/jackboot.sas`.

Example: The purpose of the following study is to illustrate the jackknife methodology and its usefulness for many data problems. Amongst others, the creatinine and magnesium content in urines of $n = 90$ children from two regions (site 1: Kapfenberg, $n_1 = 38$, and site 2: Knittelfeld $n_2 = 52$) in Styria, Austria, was monitored in order to analyse the heavy-metal loadings in human beings.

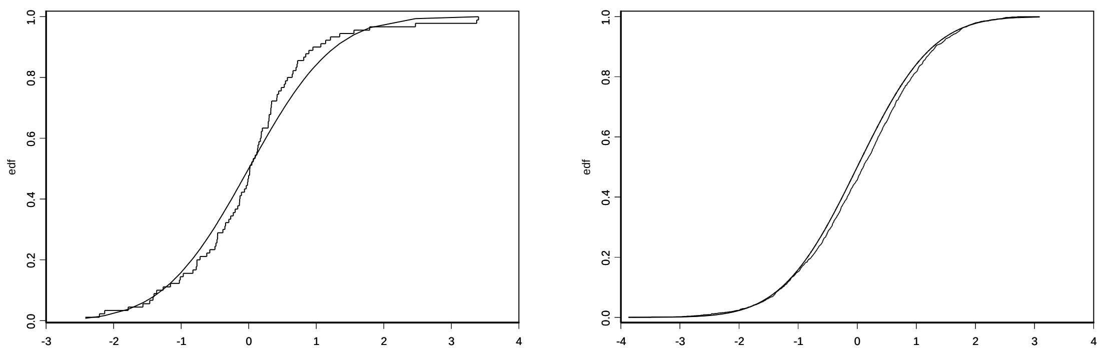
First we treat our sample as iid from a bivariate distribution with correlation ρ. For the entire sample the correlation coefficient between both chemicals is $r = 0.409$ with estimated asymptotic SE 0.0887 (based on assuming a bivariate normal distribution) or with 0.0732 (by applying the nonparametric delta method). Table 1 summarizes the results for different jackknife methods. The jackknife estimates for the bias are quite similar for all resampling methods and can be neglected with respect to the magnitude of r. Both, the delete-one and the delete-45 jackknife yield SEs which are comparable to the result of the delta method. The grouped jackknife shows quite high variability with varying group size h. One may recommend its use only for situations where heavy computations would be necessary otherwise. The studentized jackknife histograms in Figure 1 are based on (8) and compared to the

Table 1: Jackknife bias and SE for Pearson's correlation coefficient r

Jackknife	Bias	SE
Delete-one	0.00254	0.0757
Delete-45	0.00498	0.0778
Grouped ($h = 5$)	0.00197	0.0664
Grouped ($h = 10$)	0.00179	0.0961

standard normal distribution function. More precisely, the figure shows the empirical distribution functions of $\sqrt{(n-d)/d}(r_{(s_j)} - r)/\widehat{\mathrm{SE}}(r)$, where we used $\widehat{\mathrm{SE}}(r) = 0.0732$, its nonparametric delta method estimate. We choose $d = 1$ and $d = 45$, respectively, where $N = 2000$ subsets are generated randomly to reflect the exact behaviour of the delete-d jackknife procedure. In contrary to the delete-one jackknife histogram, the delete-45 histogram conforms very well with the limiting normal distribution. Hence, standard confidence intervals should therefore be based on the delete-d jackknife, where d is chosen sufficiently large.

Figure 1: Jackknife histograms ((a) delete-one; (b) delete-45) for the bias-corrected standardized Pearson correlation coefficient



In the following the jackknife is used in linear and generalized linear regression models to show its applicability for non-iid situations. To compare the local heavy-metal concentrations in the urine of different children, such elements are often standardized with respect to the creatinine content. Hence, in the next step we consider a linear regression model for magnesium (Mg) as the dependent variable and with site and creatinine as explanatory variates. The resulting least squares estimates (SEs) are the site-specific intercepts $\hat{\beta}_1^1 = 16.737$ (34.256) and $\hat{\beta}_1^2 = 7.866$ (26.483), with the respective slopes $\hat{\beta}_2^1 = 219.264$ (29.897) and $\hat{\beta}_2^2 = 99.104$ (18.137). Because the intercepts do not differ significantly from zero, the ratio of slopes β_2^1/β_2^2 is an absolute measure of the location effect. Its plug-in estimate $\hat{\beta}_2^1/\hat{\beta}_2^2 = 2.212$

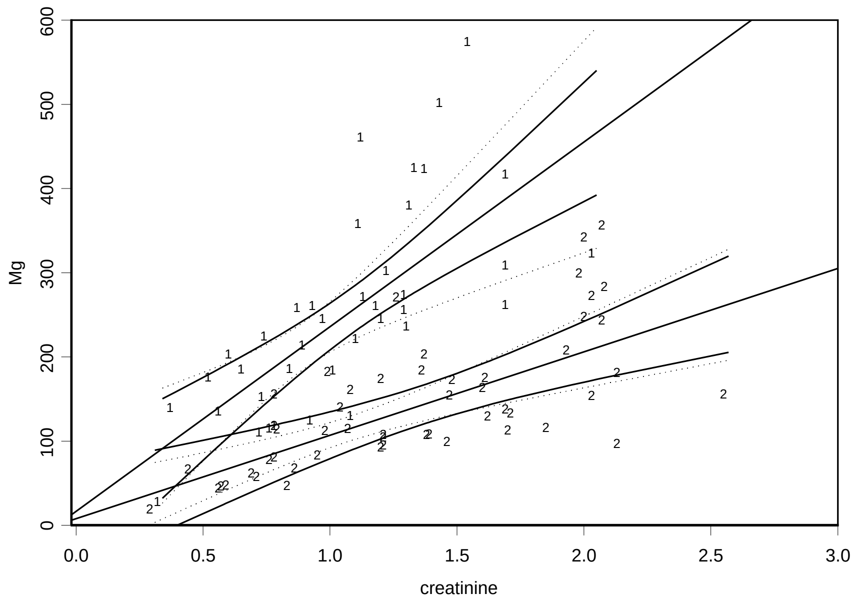


Figure 2: Magnesium/creatinine scatterplot with least squares estimates, pointwise 99% confidence region for the regression line μ (least squares results: solid lines; jackknife delete-one results: dotted lines)

reflects the general tendency that the mean magnesium contents at the industrial region (first site) are more than twice those of the second site (the control region). Applying the ordinary and weighted jackknife gives the results in Table 2. Note that for the ratio of slopes we utilized the two-sample jackknife as the appropriate delete-one version and thus used (5) and (6) to calculate bias and SE, respectively. As mentioned before, only Hinkley's and Wu's jackknife procedures yield the correct bias for the least squares estimates. Generally, (17) equals (20) and is comparable to the two-sample bias from (5) for the parameter ratio. The SEs are based on (12) or (6), (16) or (18), (19), and do not differ that much; however, the delete-one results are slightly larger. Figure 2 displays the results of a jackknife analysis of the fitted mean $\hat{\mu} = x^\top \hat{\beta}$, which is a linear function in the parameters. Hence, for a jackknife version of the standard pointwise $1 - \alpha$ confidence region $\hat{\mu} \pm t_{n-p;1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu})}$, drawn as solid lines, we simply replace $\widehat{\mathrm{var}}(\hat{\mu})$ by $\widehat{\mathrm{var}}_J(\hat{\mu}) = x^\top \widehat{\mathrm{var}}_J(\hat{\beta}) x$, shown as dotted lines. Instead of $\widehat{\mathrm{var}}_J(\hat{\beta})$, one might also use $\widehat{\mathrm{var}}_{JH}(\hat{\beta})$ or $\widehat{\mathrm{var}}_{JW}(\hat{\beta})$.

To assess the capability of the model to predict the response of a future observation, the apparent error 4754.9 is much too optimistic. Application of (24) results in a jackknife bias of -291.8 correcting the apparent error to 5046.7 as an estimate of the optimal prediction error. The jackknife bias (25) of the expected excess error is 609.6, thus the corrected estimate of the true error is 5364.5, which is quite close

Table 2: Jackknife bias and SE estimates for the parameters in a linear regression model

	Slope site 1, β_2^1		Slope site 2, β_2^2		β_2^1/β_2^2	
Jackknife	Bias	SE	Bias	SE	Bias	SE
Delete-one	5.482	47.053	0.448	18.075	0.059	0.616
Hinkley	0	42.271	0	17.288	0.068	0.573
Wu	0	44.162	0	17.521	0.068	0.590

Table 3: Jackknife bias and SE estimates for the parameters in the heteroskedastic linear model

	Slope site 1, β_2^1		Slope site 2, β_2^2		β_2^1/β_2^2	
Jackknife	Bias	SE	Bias	SE	Bias	SE
Delete-one	2.081	41.501	-1.021	14.435	0.045	0.496
One-step	-7.463	40.448	0.835	13.420	-0.053	0.478

to the cross-validated estimate 5368.1 for the true error.

Since the data show increasing variability we also apply a heteroskedastic GLM with identity link and variance proportional to the mean. Under this model the site effect parameters are even closer to zero, i.e. $\hat{\beta}_1^1 = 1.922$ (29.636) and $\hat{\beta}_1^2 = -0.345$ (16.237), whereas the slopes $\hat{\beta}_2^1 = 232.981$ (29.671) and $\hat{\beta}_2^2 = 105.154$ (13.284) are somewhat larger than in the homoskedastic model. The plug-in estimate of the slope ratio is $\hat{\beta}_2^1/\hat{\beta}_2^2 = 2.216$, being almost equal to the previous result. The ordinary jackknife technique, its two-sample version for the ratio, and its approximation by the one-step procedure (22) leads to the results in Table 3. In comparison to the one-step method the delete-one as the two-sample jackknife seems to give more dependable results, but of course at computationally higher costs.

References

[1] Arvesen, J. N. (1969). Jackknifing U-statistics. *Annals of Mathematical Statistics* **40**, 2076-2100.

[2] Booth, J. G. and Hall, P. (1993). An improvement of the jackknife distribution function estimator. *The Annals of Statistics* **21**, 1467-1485.

[3] Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**, 625-633.

[4] Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics* **14**, 1171-1179.

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

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

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

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

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

[10] Fay, R. E. (1985). A jackknife chi-square test for complex samples. *Journal of the American Statistical Association* **80**, 148-157.

[11] Fox, T., Hinkley, D. V., and Larntz, K. (1980). Jackknifing in nonlinear regression. *Technometrics* **22**, 123-129.

[12] Gladen, B. (1979). The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects. *Journal of the American Statistical Association* **74**, 278-283.

[13] Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap. Excess error estimation in forward logistic regression. *Journal of the American Statistical Association* **81**, 108-113.

[14] Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19**, 285-292.

[15] Jorgensen, M. A. (1987). Jackknifing fixed points of iterations. *Biometrika* **74**, 207-211.

[16] Jorgensen, M. A. (1993). Influence functions for iteratively defined statistics. *Biometrika* **80**, 253-265.

[17] Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17**, 1217-1241.

[18] Kuo, L. (1989). Composite estimation of totals for livestock surveys. *Journal of the American Statistical Association* **84**, 421-429.

[19] Lele, S. (1991). Jackknifing linear estimating equations: Asymptotic theory and applications in stochastic processes. *Journal of the Royal Statistical Society, Series B* **52**, 253-267.

[20] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[21] Matloff, N. S. (1980). Algorithm AS 148: The jackknife. *Applied Statistics* **29**, 115-117.

[22] Miller, R. G. (1974). The jackknife–a review. *Biometrika* **61**, 1-15.

[23] Miller, R. G. (1974). An unbalanced jackknife. *The Annals of Statistics* **2**, 880-891.

[24] Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B* **11**, 68-84.

[25] Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353-360.

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

[27] Shao, J. (1992). Jackknife estimator for an m-dependent stationary process. *Acta Mathematicae Applicatae Sinica* **8**, 115-123.

[28] Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap.* Springer, New York.

[29] Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics* **29**, 614.

[30] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *The Annals of Statistics* **14**, 1261-1350.

[31] Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *The Annals of Statistics* **18**, 1438-1452.

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