Resampling Methods in Estimator Assessment

the bootstrap, the jackknife and their relationship

Amir H. Kadivar

School of Computer Science, McGill University

Abstract

Historically the first method to explicitly seek a resampling based estimation of bias and variance of an arbitrary estimator was one put forward in the 1950s by Quenouille for which the term jackknife was coined a few years later by Tukey. Thereafter there has been extensive study of similar methods, the most well known among which is Efron's bootstrap. The ideas generalize readily (or easily) to other measures of error (e.g prediction error) and also to more complicated situations (e.g regression analysis and model selection) leading to a family of assessment methods arising and taking advantage from the age of cheap and fast computation power. Here we introduce the basic ideas behind the bootstrap, the jackknife, a few flavors of both, a neat relationship between them shown by Efron, along with some real examples of their behavior.

1 Introduction

We consider the problem of estimating some aspect t of an unknown probability distribution F from a set X of samples drawn i.i.d from the distribution:

$$x_1, x_2, \ldots, x_n \sim F$$

The observed probability distribution resulting from the samples is one that puts $\frac{1}{n}$ weight over any observed value, and in the literature is referred to by *empirical probability distribution* and is denoted by $\hat{F}$. $\hat{F}$ is obviously a discrete probability distribution defined by observed frequencies:

$$\hat{f}_k = \frac{\#(x_i = k)}{n}$$

We will refer to the estimator of $t(F)$ as $s(\hat{F})$. The first problem to consider right here would be that considering observations to be independent samples of F implies that we are discarding any temporal information lying in the observed samples. We confine ourselves to the case where we can assume so and hence no information is lost by building the estimator as a function that operates on $\hat{F}$. The most trivial choice for $s(\hat{F})$ would be $t(\hat{F})$ which in the literature is referred to by the *plug-in* estimator. Plug-in estimators although not often the best choice, have two favorable features. First obviously is their simplicity, and second is the fact that they usually have low bias compared to their variance [4] and thus a

low $\frac{\text{bias}}{\text{standard error}}$ a feature that makes the estimator more reliable. For most cases within this review we consider only plug-in estimators.

Once an estimator is defined, or more than one estimators are available and one wants to compare their performance, resampling methods address the problem of estimating different error measures of the estimator, the most common among which are bias and variance. They define a new hypothetical problem in which the input space is limited to those values of x observed through X and the distribution over the input space is assumed to be $\hat{F}$, which in the original problem is unknown. They then simulate the observation-estimation process for a number of iterations, where in each iteration a number of samples are drawn from $\hat{F}$, and a new estimation instance is generated. The (now computable) bias or variance (or any other error measure) of the estimator is then used as an estimate of the actual bias or variance of $s(\hat{F})$.

Although the bootstrap provides a more general platform to define and analyze other preceding methods, we will follow the historic order of events as in [3]. We first introduce the jackknife and the way it was regarded when it was first introduced in the 1950s. Then we consider the bootstrap, and look at the jackknife through the vantage point introduced by Efron.

2 Influence curves and linear expansion of statistal functionals

Influence curves were introduced in the context of robust statistics. As mentioned in [7] they can be regarded as first order derivatives of statistical functionals, and for any specific distribution result in a function over the same domain as that of the the probability distribution. We denote the space of all real functions over some domain $\mathcal{X}$ by $\mathbb{F}$, of which the space of all probability distributions is a subset. For any functional t defined over $\mathbb{F}$, the influence curve of t at $F \in \mathbb{F}$ is defined by the following:

$$IC_t : \mathbb{F} \times \mathbb{R}^p \to \mathbb{R}$$

$$IC_t(F, x) = \lim_{\epsilon \downarrow 0} \frac{t((1-\epsilon)F + \epsilon\delta_x) - t(F)}{\epsilon} \tag{1}$$

where δ_x is the degenerate distribution at x. Essentially $F_\epsilon = t((1-\epsilon)F + \epsilon\delta_x)$ is a new probability distribution resulted by *contamination* of the distribution F at x by the amount ϵ (it is easy to check the probability distribution properties of this new function). To see the first order derivative behavior of $IC(F, .)$ we note that F and t are defined as:

$$\begin{aligned} F &: \mathbb{R}^p \to \mathbb{R} \\ t &: \mathbb{F} \to \mathbb{R} \end{aligned}$$

and that in this sense $t(F)$ is a function defined over an infinite dimensional space $\mathbb{F}$, in which the generalization of the gradient vector would be a infinite dimensional vector with its entries being the rate of change over infinitesimally small variations in each dimension. Each dimension of $\mathbb{F}$ being the value any F assigns to any of samples $x \in \mathcal{X}$, it would be obvious that infinitesimal variations of F could be

built through $\epsilon\delta_x$. The rest of the mathematical manipulation of the contaminated expression resulting in $t((1-\epsilon)F + \epsilon\delta_x)$ is to ensure that the resulting function is itself a probability distribution.

Analysis of the influence curve and proving bounds such as *gross-error-sensitivity* [7]:

$$\gamma^* = \sup_x |IC(., x)|$$

or asymptotic variance of $t(.)$:

$$\mathbb{E}_F[\int IC^2(F, x)dx]$$

fall in the domain of statistic robustness. We here focus on the von Mises expansion of an estimator, which is crucial in understanding and derivation of resampling based estimates of estimator error measures.

2.1 von Mises expansion

We follow the derivation of [5] for a series expansion of $t(F)$. We first note that all these derivations for t could be translated to expansions of $s(\hat{F})$, since the two function are similar functions in nature. For two probability distributions $G_1, G_2 \in \mathbb{F}$ and for some functional t, we define the function:

$$A_{G_1,G_2,t} : \mathbb{R} \to \mathbb{R}$$

$$A_{G_1,G_2,t}(\kappa) = t(\kappa G_1 + (1-\kappa)G_2)$$

which, for simplicity, we will just refer to as $A(\kappa)$. Under regularity conditions on t one can ensure that A is analytical around zero, and has a Taylor expansion around zero at 1. Thus we will have:

$$A(1) \simeq A(0) + A^{(1)}|_{\kappa=0}$$

For deriving the first order derivative of A we have the following by definition:

$$A^{(1)}(\kappa) = \lim_{\epsilon \to 0} \frac{A(\kappa+\epsilon) - A(\kappa)}{\epsilon}$$

using a rearrangement for $A(\kappa)$:

$$A(\kappa) = t(\kappa G_1 + (1-\kappa)G_2) = t(G_2 + \kappa(G_1 - G_2))$$

the first order derivative would be:

$$\begin{aligned} A^{(1)}(\kappa)|_{\kappa=0} &= \lim_{\epsilon \to 0} \frac{A(\kappa+\epsilon) - A(\kappa)}{\epsilon} = \lim_{\epsilon \to 0} \left[\frac{t[G_2 + (\kappa+\epsilon)(G_1 - G_2)] - t[G_2 + \kappa(G_1 - G_2))]}{\epsilon} \right] \\ \textit{[setting } \kappa \textit{ to zero]} &= \lim_{\epsilon \to 0} \frac{t[G_2 + \epsilon(G_1 - G_2)] - t(G_2)}{\epsilon} = \lim_{\epsilon \to 0} \frac{t(\epsilon G_1 + (1-\epsilon)G_2) - t(G_2))}{\epsilon} \end{aligned}$$

We keep the result we have got here, and consider the case of applying the expansion to estimating $t(F)$ by $t(\hat{F})$ which is the case of using the plug-in principle. For $A(\kappa) = A_{\hat{F},F,t}$ since $A(1) = t(\hat{F})$ and $A(0) = t(F)$ we get:

$$t(\hat{F}) \simeq t(F) + A^{(1)}|_{\kappa=0}$$

Now we notice that for any *empirical probability distribution* $\hat{F}$ defined from frequencies of the observations $x_1, x_2, \ldots, x_n$ we have:

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

and hence the first order derivative derived above could be rewritten as:

$$A^{(1)}|_{\kappa_0} = \lim_{\epsilon\to 0}\frac{t((1-\epsilon)F + \epsilon\hat{F}) - t(F))}{\epsilon} = \lim_{\epsilon\to 0}\frac{t((1-\epsilon)F + \epsilon\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}) - t(F))}{\epsilon}$$

Comparing this with the influence curve $IC(F, x)$ definition we notice that the above formulation is essentially the same, expect for the fact that $\epsilon\delta_x$ is replaced by $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. We remember the way $IC(.,.)$ can be regarded as an infinite dimensional gradient vector of $t(F)$ over all dimensions of $\mathbb{F}$, since

$$t(.) : \mathbb{F} \to \mathbb{R}$$

$$IC(.,.) : \mathbb{F} \to \mathbb{F}$$

In that sense $A^{(1)}|_{\kappa_0}$ can be regarded as a directional derivative, where the *direction* is this member of $\mathbb{F}$:

$$u = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

Using the chain rule (again assuming regularity conditions allowing us to do so) we would get:

$$A^{(1)}|_{\kappa_0} = \lim_{\epsilon\to 0}\frac{t((1-\epsilon)F + u) - t(F))}{\epsilon} = \frac{\partial t(F)}{\partial u} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial t(F)}{\partial \delta_{x_i}} = \frac{1}{n}\sum_{i=1}^{n} IC(F, x_i) \tag{2}$$

And thus we have derived the linear von Mises expansion for the plug-in principle:

$$t(\hat{F}) \simeq t(F) + \frac{1}{n}\sum_{i=1}^{n} IC(F, x_i)$$

where x_is are the observed samples.

3 Derivation of the jackknife

The following bias-corrected estimator was first introduced by Quenouille in [9]:

3.1 The jackknife estimate of bias

From the linear expansion:

$$t(\hat{F}) \simeq t(F) + \frac{1}{n}\sum_{i=1}^{n} IC(F, x_i)$$

one can think of improving the plug-in estimator by adding the second term on the right hand side of the above formulation to the estimate $t(\hat{F})$. The way jackknife performs this is first to approximate the terms

$IC(F, x_i)$ by using $\hat{F}$ as an estimate of F and then estimating the derivative formulation of (1) by setting $\epsilon = \frac{-1}{n-1}$ instead of taking ϵ to its limit at zero. Thus for each term $IC(F, x_i)$ we will have:

$$IC_t(\hat{F}, x) \simeq \frac{t\left[(1+\frac{1}{n-1})\hat{F} - \frac{1}{n-1}\delta_x\right] - t(\hat{F})}{\frac{-1}{n-1}} = (n-1)\left[t(\hat{F}) - t\left[(1+\frac{1}{n-1})\hat{F} - \frac{1}{n-1}\delta_x\right]\right]$$

Investigation of the term $(1+\frac{1}{n-1})F - \frac{1}{n-1}\delta_x$ shows that it is the normalized frequency distribution of the sample set resulting from putting aside the i-th observation in the original sample set. In the literature this is referred to as the i-th *jackknife sample* and the resulting functional value $t[(1+\frac{1}{n-1})F + \frac{1}{n-1}\delta_x]$ is denoted by $t_{(i)}$. The resulting jackknife estimate of bias would be:

$$\mathbb{E}[t(\hat{F})] - t(F) \simeq \widehat{BIAS}_{jack} = (n-1)(\hat{t} - t_{(.)})$$

where $\hat{t} = t(\hat{F})$ and $t_{(.)} = \frac{1}{n}\sum_{i=1}^{n} t_{(i)}$. And the resulting bias-corrected jackknife estimator would be:

$$s_{jack}(\hat{F}) = n\hat{t} - (n-1)t_{(.)}$$

It is easy to see that the jackknife is reducing the leading term of the bias by a power of n, meaning that the bias of the above estimator will be of the following form [8]:

$$\mathbb{E}[s_{jack}] = t(F) + O_p(\frac{1}{n^2})$$

Although in order to derive this we used a couple of seemingly random approximations, [3] argues and proves that $\widehat{BIAS}_{jack}$ is an unbiased estimator of $\mathbb{E}[t(\hat{F})] - t(F)$.

3.2 The jackknife estimate of variance

In [11] Tukey suggested the use of the statistics $t_{(i)}$ introduced by Quenouille for estimating the variance of a functional:

$$\mathbb{E}_F[t(\hat{F}) - \mathbb{E}_F t(\hat{F})]]^2$$

resulting from multiple sampling of the original probability distribution, hence the subscript F in the expectation. What Tukey suggested was:

$$\mathbb{E}_F[t(\hat{F}) - \mathbb{E}_F t(\hat{F})]] \simeq \widehat{VAR}_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{t}_{(i)} - \hat{t}_{(.)})^2$$

This can be regarded as $n-1$ times the variance of sample variance of the jackknife samples. The need for such an *inflation factor* (by which [4] refers to the $n-1$ term) is that the jackknife samples are very similar to the original sample. Checking various functionals such as mean and variance suggests a term of such form being used, but one should remember that the choice of exactly $n-1$ is a "somewhat arbitrary convention in the literature" [4]. Later on, with the introduction of the bootstrap a theoretical justification of this variance estimate can be made, being a quadratic approximation of the bootstrap estimate of variance. But as mentioned in [8] and in chapter 3 of [3], there are various situations disverifying Tukey's

variance estimate.

Here we mention a derivation of a variation of the jackknife estimate of variance from [4]. We go back again to the linear expansion of our statistical functional we used to derive the jackknife estimate of bias:

$$t(\hat{F}) \simeq t(F) + \frac{1}{n}\sum_{i=1}^{n} IC(F, x_i)$$

It follows that:

$$\mathrm{var}_F t(\hat{F}) = \frac{1}{n}\mathrm{var}_F IC(F,x) = \frac{1}{n}\mathbb{E}_F[IC^2(F,x)] \tag{3}$$

where the last expression results from the fact that $\mathbb{E}_F[IC(F,x)] = 0$ [4]. Again using the same approximation of $IC(.,.)$ as the one we used for the bias estimate, we set $\epsilon = \frac{-1}{n-1}$ in (1) and we will get the following approximation:

$$\frac{1}{n}\mathbb{E}_F[IC^2(F,x)] \simeq \left(\frac{n-1}{n}\right)^2 \sum_{i=1}^{n} (\hat{t}_{(i)} - \hat{t}_{(.)})^2$$

This is not exactly the same as the the original jackknife estimate of Tukey, but is a very similar one and as mentioned before about the choice of the inflation factor (which here is $\left(\frac{n-1}{n}\right)(n-1)$ instead of just $n-1$) could be regarded as equally justifiable.

4 The bootstrap

We start with the bias estimate of the bootstrap to demonstrate its mathematical foundation. The bias of a symmetrically defined functional $s(\hat{F})$ (symmetric over $x_1, x_2, \ldots, x_n$) is defined by:

$$\mathbb{E}_F[s(\hat{F})] - t(F)$$

More precisely if we denote different empirical distributions that can arise from F by F^*, we can write:

$$\mathbb{E}_F[s(\hat{F})] = \int_{\mathbb{F}} p(F^*)s(F^*)dF^*$$

The basic idea behind the bootstrap is that since F is unknown, as an approximation to the above method, we plug in $\hat{F}$ instead of F in whatever approximation we are going to perform. First we derive F^* from the observed empirical distribution $\hat{F}$ (instead of the actual unknown distribution), and act like $\hat{F}$ is the actual underlying distribution. The independent samples taken i.i.d with replacement from the $\{x_1, x_2, \ldots, x_n\}$ giving rise to $\hat{F^*}$s are referred to in the literature by *bootstrap samples*. This being said we can rewrite the above as the following:

$$\mathbb{E}_F[s(\hat{F})] \simeq \int_{\hat{\mathbb{F}}} p(\hat{F^*})s(\hat{F^*})d\hat{F^*}$$

The above integral is then computed by a Monte Carlo simulation, by a finite number B of bootstrap samples.

If we refer to the original sample by $X = \{x_1, x_2, \cdots, x_n\}$, the bootstrap starts with generating B independent random samples, each consisting of n random observations selected at random with replacement from X, namely $X^{*^1}, X^{*^2}, \cdots, X^{*^B}$. Then the following approximations are used to estimate the bias:

$$t^{\hat{*}\cdot}_B = \frac{1}{B}\sum_{b=1}^{B} s(X^{*^b}) \simeq \int_{\hat{\mathbb{F}}} p(\hat{F^*}) s(\hat{F^*}) d\hat{F^*} \simeq \int_{\mathbb{F}} p(F^*) s(F^*) dF^* \tag{4}$$

$$s(\hat{F}) \simeq t(F)$$

and hence:

$$\widehat{BIAS}_{boot} = t^{\hat{*}\cdot}_B - s(\hat{F}) \simeq \mathbb{E}_F[s(\hat{F})] - t(F)$$

The rationale behind the bootstrap is pretty straightforward. If we put aside the facts that $\mathcal{X}$ probably has members not showing up in X and the fact that even on the ones in X, the probability distribution is not necessarily the one we assumed ($\hat{F}$), one can easily see why the bootstrap converges to the exact bias value for large enough B. In other words, assuming that $\mathcal{X}$ and F do not have any *hidden tricks up their sleeves* (namely information about F), the second $\simeq$ in (4) can be replaced by $=$, and the first one can be replaced by asymptotic equality:

$$t^{\hat{*}\cdot}_B = \frac{1}{B}\sum_{b=1}^{B} s(X^{*^b}) \approx \int_{\hat{\mathbb{F}}} p(\hat{F^*}) s(\hat{F^*}) d\hat{F^*} = \int_{\mathbb{F}} p(F^*) s(F^*) dF^*$$

or equivalently:

$$\lim_{B\to\infty} t^{\hat{*}\cdot}_B = \mathbb{E}_F[s(\hat{F})]$$

(notice the expectation being on F instead of $\hat{F}$, which is made possible through the above assumptions). To justify the above assumptions we notice that $\hat{F}$ is the nonparametric maximum likelihood estimator of F (and also asymptotically exact letting $n \to \infty$) and regarding the bootstrap as *the most available knowledge* from the observation X justifies the bootstrap as being the asymptotically optimum choice. So the remaining error caused by the mentioned assumptions is not something we could get rid of, as in any other estimation problem. This is essentially how we are going to look at the bootstrap as do [3] and [4]. In the next section when we introduce the viewpoint of resampling probability distributions, we introduce a faster converging bootstrap bias estimation according to [4] which only applies to plug-in estimators.

We follow the exact same scheme to derive the bootstrap variance. The same argument as made above can be made to justify the performance of the bootstrap in this case. The variance of the estimator $s(\hat{F})$ is $\mathrm{var}_F s(\hat{F})$ and is approximated by $\mathrm{var}_{\hat{F}} s(\hat{F^*})$ which with the same arguments as we used for the estimated bias can be performed by the following Monte Carlo estimation:

$$\mathrm{var}_F s(\hat{F}) \simeq \widehat{VAR}_{boot} = \frac{1}{B-1}\sum_{b=1}^{B} (t^{*^b} - t^{*\cdot}_B)^2$$

where $t^{*^b} = s(\hat{F^*}^b)$.

4.1 Immediate extension to more general problems

First we look at the application of the bootstrap to a parametric setting. Assuming, for instance, a normal distribution family for $\hat{F}$ we would build the maximum likelihood $\hat{F}$ in a trivial way as we did previously for our nonparametric maximum likelihood estimation. Now looking for example at the problem of bias estimation for this case, it can be seen that nothing essentially changes in 4 except for the fact that $\hat{F}$ is now the parametric maximum likelihood estimator of F and hence the bootstrap samples are derived from such distribution.
Efron in [3] argues that one can regard Fisher's method assessing a maximum likelihood estimator as a bootstrap method. A detailed discussion of the relationship of the two can be found in chapter 21 of [4], which would fall out of the concern of this paper.

This gets us to look at a fundamental feature of the bootstrap. We follow Efron's argument in [3] for this matter. The nature of $s(\hat{F})$ and the functional $t(F)$ it is estimating has nothing to do with the rationale of the bootstrap (the same holds for the jackknife). Furthermore, as we saw in the derivation of the variance estimate, there is nothing essentially different for the statistical aspect of $s(\hat{F})$ being the expectation or the variance or even any other aspect of the functional s. Under all these variations the bootstrap will follow the same flow of computation as we have demonstrated, while the same justifications hold for its viability (*this* does not hold for the jackknife). In any case the bootstrap is using $\hat{F}$ as the (parametric or nonparametric) maximum likelihood estimator of F and then computing the desired statistical aspect through a Monte Carlo approximation with finite B. We will get back to this feature later on when discussing prediction errors.

5 Another viewpoint: Resampling probability distribution

We now consider a simple representation of the bootstrap process. For every bootstrap iteration one can look at the frequencies of the observed x_i^* and build a probability distribution which we are going to refer to by $\mathbf{P}^{*^b}$, which is defined in essentially the same way we defined $\hat{F}$ in the first place. We denote $\hat{F}$ by $\mathbf{P}^0$ which is a uniform $\frac{1}{n}$ distribution. We notice that in this sense, we are fixing the originally observed x_is, and regard them as "all the possible" choices or equivalently "all the members of $\mathcal{X}$ that we have proof of existence". We now build all possible distributions (or some of them in the real world case were we can afford a finite number of bootstrap samples), and investigate the behavior of s according to those.
All the $\mathbf{P}^*$s lie on an n-dimensional simplex we would call φ_n, which is defined by:

$$\varphi_n = \{\mathbf{P}^* : P_i^* \geq 0, \sum_{i=1}^{n} P_i^* = 1\}$$

It is easy to check that since the $\mathbf{P}^*$s are just normalized frequency vectors, any statistical function will translate into a pretty similar function of $\mathbf{P}^*$. For example a functional that is linear in the sense that it does not make any use of cross terms of x_is and can be represented exactly using at most 1st order terms

of its von Mises expansion:

$$t(\hat{F}) = t(F) + \frac{1}{n}\sum_{i=1}^{n} IC(F, x_i)$$

would simply translate to a function over $\mathbf{P}^*$:

$$t(\mathbf{P}^*) = t(\mathbf{P}^0) + (\mathbf{P}^* - \mathbf{P}^0)\mathbf{U} \tag{5}$$

where the vector $\mathbf{U}$ is simply the vector consisting of $IC(F, x_i)$ s.
Similarly one can derive an equivalent form for the case where the functional is a quadratic function of F, which means it involves only double cross terms in its formulation:

$$t(\mathbf{P}^*) = t(\mathbf{P}^0) + (\mathbf{P}^* - \mathbf{P}^0)\mathbf{U} + \frac{1}{2}(\mathbf{P}^* - \mathbf{P}^0)^T\mathbf{V}(\mathbf{P}^* - \mathbf{P}^0) \tag{6}$$

To estimate the bias and variance of the estimator we would be interested in the probability distribution over φ_n. It can be easily seen that the unnormalized frequency vectors $n\mathbf{P}^*$ come from a multinomial distribution which will result in the following expression for the probability distribution over φ_n:

$$\mathbf{P}^* \underset{*}{\sim} (\mathbf{P}^0, \frac{\mathbf{I}}{n^2} - \frac{\mathbf{P}^{0^T}\mathbf{P}^0}{n})$$

In [3] it is insisted that we use a $*$ under the probability relation, to show that this distribution is produced by the statistician and not by uncertainty from nature.
From the above representation of the resampling scheme, one can see that the bootstrap tries to cover the whole simplex as $B \to \infty$ and the jackknife only checks the n mid-edge members of the simplex, that are all of the form:

$$\mathbf{P}_{(i)} = \left(\frac{1}{n-1}, \frac{1}{n-1}, \ldots, 0, \frac{1}{n-1}, \ldots, \frac{1}{n-1}\right)$$

Two facts can already been noticed. First that the jackknife is an approximation of the bootstrap by selecting only n specific members of φ_n to find a statistical aspect of the functional $s(\mathbf{P}^*)$ over φ_n. Second that the jackknife samples are on average closer to $\mathbf{P}^0$ than the average bootstrap samples. In the following section we provide a proof for the exact relationship of the bootstrap and the jackknife.

In [4] a faster converging bootstrap estimate is introduced. If we denote the respective probability distributions of the bootstrap samples by $\mathbf{P}^{*^1}, \mathbf{P}^{*^2}, \ldots, \mathbf{P}^{*^B}$, we define:

$$\overline{\mathbf{P}} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{P}^{*^b}$$

The only modification to the bootstrap bias estimate we derived earlier on, is to use $s(\overline{\mathbf{P}})$ instead of $s(\mathbf{P}^0) = s(\hat{F})$ as an approximation to $t(F)$:

$$\overline{BIAS}_{boot} = \hat{t}^{*\cdot}_B - s(\overline{\mathbf{P}})$$

Essentially both $\overline{BIAS}_{boot}$ and $\widehat{BIAS}_{boot}$ converge to the ideal bootstrap estimate, but [4] shows and argues why the former converges faster.

6 The relationship between the bootstrap and the jackknife

Now we have developed all the tools to compare the bootstrap and the jackknife and see how they are related. We first look again at the linear approximation of (2). The asymptotic bootstrap estimation ($B \to \infty$) of the bias directly estimates

$$\mathbb{E}_F t(\hat{F}) - t(F)$$

by

$$\mathbb{E}_{\hat{F}} t(\hat{F}^*) - t(\hat{F})$$

The jackknife estimate of bias on the other hand uses (2) and then approximates the terms $IC(F, x_i)$ by $IC(\hat{F}, x_i)$ and then approximating these terms again by setting $\epsilon = \frac{-1}{n-1}$ instead of taking the limit in (1). In this sense the jackknife is roughly a linear approximation of the bootstrap. This fact is further elaborated in the next sections using the bootstrap sample probability distributions.

For the case of variance of the plug-in estimator we go back to (3) and notice that the ideal (asymptotic) bootstrap estimates the left hand side directly, while the jackknife estimates the right hand side using the same approximation scheme it uses for estimating the bias. This fact is again further elaborated in the following section.

6.1 The jackknife approximation of the bootstrap variance estimate

Efron proves a neat relationship between the jackknife and the bootstrap estimations of variance, using the resampling distributions we discussed before. We remember that the bootstrap sampling induces a probability distribution which is n times multinomial distribution over the simplex φ_n.The functional t on the other hand builds a surface on the simplex, consisting of points $\langle \mathbf{P}^*, t(\mathbf{P}^*) \rangle$ for all $\mathbf{P}^* \in \varphi_n$.

The ideal bootstrap estimator of variance would compute the variance of $t(\mathbf{P}^*)$ according to this distribution over the surface described above, by a finite number of samples from the surface. The jackknife on the other hand estimates the variance by just looking at n values being $t(\mathbf{P}_{(i)}$. We can see that through the points $\langle \mathbf{P}_{(i)}, t(\mathbf{P}_{(i)}) \rangle$ passes a hyperplane which can be regarded as a linear estimate of the actual $\langle \mathbf{P}^*, t(\mathbf{P}^*) \rangle$ surface. Technically speaking, referring to the linear hyperplane defined by the jackknife samples by t_{LIN} surface, we can use (5) to formulate the values over t_{LIN} (since this would be a linear functional):

$$t(\mathbf{P}^*) \simeq t_{LIN}(\mathbf{P}^*) = c_0 + (\mathbf{P}^* - \mathbf{P}^0)\mathbf{U}$$

and hence:

$$\begin{aligned} \mathrm{var}_* t_{LIN}(\mathbf{P}^*) &= \mathrm{var}_*[c_0 + (\mathbf{P}^* - \mathbf{P}^0)\mathbf{U}] = \mathrm{var}_*[(\mathbf{P}^* - \mathbf{P}^0)\mathbf{U}] \\ &= \mathrm{var}_* \mathbf{P}^* \mathbf{U} = \mathbf{U}^T (\mathrm{var}_* \mathbf{P}^*)\mathbf{U} = \frac{1}{n^2} \mathbf{U}^T \mathbf{U} \end{aligned}$$

where the * as the subscript of variance means the variance over the simplex φ_n and over the distribution:

$$\mathbf{P}^* \underset{*}{\sim} \left(\mathbf{P}^0, \frac{\mathbf{I}}{n} - \frac{\mathbf{P}^{0^T} \mathbf{P}^0}{n^2} \right)$$

and the last expression in the computation above results from the covariance matrix above. Now we notice that since t_{LIN} is the hyperplane passing through jackknife samples, the values of the entries in $\mathbf{U}$ are the jackknife estimates of $IC(F, x_i)$ which are as we mentioned earlier:

$$(n-1)(\hat{t}_{(.)} - t_{(i)})$$

and hence we have proven that:

$$\text{var}_* t_{LIN}(\mathbf{P}^*) = \frac{n-1}{n}\widehat{VAR}_{jack}$$

which means that the jackknife estimate for variance is roughly (for the term $\frac{n-1}{n}$) equal to a linear approximation of the bootstrap estimate of variance.

6.2 The jackknife approximation of the bootstrap bias estimate

In the case of bias estimation, if we look at the same hyperplane we defined in the previous section both $\mathbb{E}_* t_{LIN}(\mathbf{P}^*) - t(\mathbf{P}^0)$ and the jackknife estimate of bias would be zero. In this case jackknife would be an exact estimate of the bootstrap. We now look at a more complicated surface which is approximately the surface defined by $\langle \mathbf{P}^*, t(\mathbf{P}^*)\rangle$. We look at the quadratic surface passing through the jackknife samples $\langle \mathbf{P}_{(i)}, t(\mathbf{P}_{(i)})\rangle$ and $\langle \mathbf{P}^0, t(\mathbf{P}^0)\rangle$. For this surface which we will refer to as t_{QUAD} one can write (6):

$$t_{QUAD}(\mathbf{P}^*) = c_0 + (\mathbf{P}^* - \mathbf{P}^0)\mathbf{U} + \frac{1}{2}(\mathbf{P}^* - \mathbf{P}^0)^T\mathbf{V}(\mathbf{P}^* - \mathbf{P}^0)$$

where the contents of $\mathbf{U}$ are the same as before: estimates of $IC(\hat{F}, x_i)$ and the contents of the matrix $\mathbf{V}$ would similarly be estimates of the second order influence curves at (F, x_i, x_j) which would be an obvious extension of the definition in (1):

$$\lim_{\epsilon \to 0} \frac{IC((1-\epsilon)F + \epsilon\delta_{x_j}), x_i) - IC(F, x_i)}{\epsilon}$$

We here shorten the flow of computation since it is quite similar to the one we did in the previous section. The quadratic approximation of the bootstrap estimate of bias would be:

$$\mathbb{E}_* t_{QUAD}(\mathbf{P}^*) - t_{QUAD}(\mathbf{P}^0) = \frac{1}{2}\text{tr}(\frac{\mathbf{I}}{n} - \frac{\mathbf{P}^{0^T}\mathbf{P}^0}{n^2}) = \frac{\text{tr}(\mathbf{V})}{2n^2}$$

On the other hand the jackknife estimate of bias is:

$$(n-1)(\hat{t}_{(.)} - \hat{t})$$

which can be found by $(n-1)$ times the average of:

$$\hat{t}_{(i)} - \hat{t} = \hat{t}(\mathbf{P}_{(i)}) - t(\mathbf{P}^0) = \hat{t}_{QUAD}(\mathbf{P}_{(i)}) - t_{QUAD}(\mathbf{P}^0)$$

which yields to

$$\sum_{i=1}^{n}(\mathbf{P}_{(i)} - \mathbf{P}^0)\mathbf{U} + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{P}_{(i)} - \mathbf{P}^0)^T\mathbf{V}(\mathbf{P}_{(i)} - \mathbf{P}^0)$$

which can easily be seen to cancel out to $\frac{\text{tr}(\mathbf{V})}{2n(n-1)}$. Thus we just proved that:

$$\mathbb{E}_* t_{QUAD}(\mathbf{P}^*) - t_{QUAD}(\mathbf{P}^0) = \left(\frac{n-1}{n}\right) \widehat{BIAS}_{jack}$$

which implies that the jackknife estimate of bias is roughly (for the factor of $\frac{n-1}{n}$) the bootstrap estimate of bias over a quadratic approximation of the surface of the functional.

7 Prediction error estimation

We consider the problem where based on a set of observations $X = x_1, \ldots, x_n$ of the form:

$$(t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)$$

one wants to assess the behavior of a predictor η_X, which for an unseen (*test*) sample t_0 predicts:

$$y_0 = \eta_X(t_0)$$

In chapter 5 of [3], Efron proposes a two-sample bootstrapping scheme for a regression analysis, which we do not mention here. Here, on the other hand, we are concerned with the estimation of statistical aspects of a random predictor η_X, for which we consider the observations x_i to be the concatenation of the actual observation (t_i) with the attached label (y_i) and perform the one-sample resampling scheme that we have been discussing.
For the case where y values are either 0 or 1, a trivial error function $Q[y_0, \eta(t_0)]$ can be defined. As before we would fix the observations X and limit our statistical analysis to different possible test cases. In this sense we would have:

$$x_0 = (t_0, y_0) \sim F$$

for the actual samples that would be given to the predictor and

$$x_0 = (t_0, y_0) \underset{*}{\sim} \hat{F}$$

for the bootstrapping process that we will perform to assess the behavior of the predictor.

We now go back to the general bootstrap formulation we defined before, and build the functional to be statistically assessed as follows:

$$R(X, F) = \mathbb{E}_F Q[y_0, \eta_X(t_0)] - \mathbb{E}_{\hat{F}} Q[y_0, \eta_X(t_0)]$$

This functional is referred to by the *expected excess error* which is the difference between the *apparent error* $\mathbb{E}_{\hat{F}} Q[y_0, \eta_X(t_0)]$ and the unknown *actual error* (the same expectation, but over F instead of $\hat{F}$).
Since $\hat{F}$ is a the probability distribution arising from a normalized frequencies summary, the second term in the above formulation of $R(X, F)$ would be:

$$\mathbb{E}_{\hat{F}} Q[y_0, \eta_X(t_0)] = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta_X(t_i)]$$

Using the "plug in $\hat{F}$ for F" motto we mentioned earlier, the bootstrap estimate of the expected excess error would be:

$$R^* = \mathbb{E}_{\hat{F}} Q[y_0, \eta_{X^*}(t_0)] - \mathbb{E}_{\hat{F}^*} Q[y_0, \eta_{X^*}(t_0)]$$

where the bootstrap samples are defined as before by random sampling with replacement from X resulting in

$$X^{*^b} = \{x_1^{*^b}, x_2^{*^b}, \ldots, x_n^{*^b}\}$$

notice that in the first term of the bootstrap approximation above, the predictor which is used is also η_{X^*} which means the predictor build from the bootstrap sample is assessed over all X and then averaged to get the pseudo-actual value of error. Denoting these different predictors by:

$$\eta_b^* = \eta_{X^{*b}}$$

and referring to the expected excess error by EEE we will get:

$$\widehat{EEE}_{boot} = \mathbb{E}_* \left[\sum_{i=1}^{n} (\mathbf{P}_i^0 - \mathbf{P}_i^*) Q[y_i, \eta^*(t_i)] \right]$$

where the expectation $\mathbb{E}_*$ is taken over the distribution:

$$\mathbf{P}^* \underset{*}{\sim} (\mathbf{P}^0, \frac{\mathbf{I}}{n^2} - \frac{\mathbf{P}^{0^T}\mathbf{P}^0}{n})$$

Similarly for the jackknife we could reduce the computations of the bootstrap by using:

$$\widehat{EEE}_{jack} = (n-1)R_{(.)} = \left(\frac{n-1}{n}\right) \sum_{i=1}^{n} R(\mathbf{P}_{(i)})$$

In [3] an expansion of the above jackknife estimate of expected excess error, and a comparison with the cross validation estimate is presented. A conjecture about the two being asymptotically equal is presented there as well, a proof of which was later provided in [6].

8 Experimental Results

We demonstrate the performance of the bootstrap and the jackknife in two situations. We have produced a sample of 100 observations, each a member of $\mathbb{R}^3$ in the following manner:

$$X = x_1, x_2, \cdots, x_100$$

$$x_{i1} \sim \mathcal{N}(10, 1)$$

$$x_{i2} \sim \mathcal{N}(2, 1)$$

$$x_{i3} \sim \mathcal{N}(5, 1)$$

and we compare the performance of the jackknife and the bootstrap for the plug-in estimator of two different functionals

$$t_1 = \frac{\overline{x_{i1}}^3}{\overline{x_{i2}}^2 + \overline{x_{i3}}^3}$$

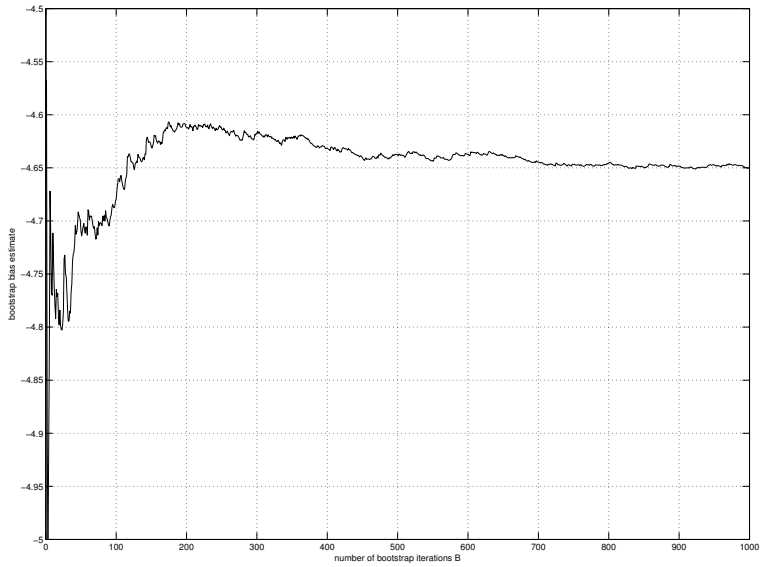

Figure 1: convergence to -4.65 of the bootstrap estimate for bias for t_1, the jackknife estimate of bias was 0.0067

and

$$t_2 = \frac{\overline{x_{i1}}}{\overline{x_{i2}}}$$

Obviously the first one is further from being a linear functional and we will see how the jackknife is less reliable in the first case. Figures 1 and 2 depict the convergence of the bootstrap for t_1 over 1000 iterations. It can be seen that something around 200 samples would have been enough to get a reasonable estimate in both cases. Figures 3 and 4 depict the convergence of the bootstrap for t_2.

9 Conclusion and some notes

We have introduced the foundations of bootstrap as a computationally intensive statistical method. The jackknife that predate the bootstrap can be regarded in the bootstrap framework as a lighter version which produces an approximation of the actual bootstrap estimate. As can be seen from the results of section 6, the reliability of the jackknife highly depends on the functional to be assessed being close to linear or not. In the latter case the jackknife might produce highly erronous results, the reasons of which have already been discussed.

Before the rise of the bootstrap different flavors of the jackknife were produced and analyzed, the most important of which are the infinitesimal jackknife introduced by Jaeckel which in a nutshell can be regarded as a more accurate jackknife. In the influence curve based derivation we used for the jackknife, the infinitesimal jackknife takes ϵ to 0 instead of using $\epsilon\frac{-1}{n-1}$ (as in the original jackknife) to approximate the influence curve value. In the context of the approximating surfaces discussed in section 6, the infinitesimal jackknife is the approximation of the functional on the tangent surface to the t surface at $\mathbf{P}^0$. Further

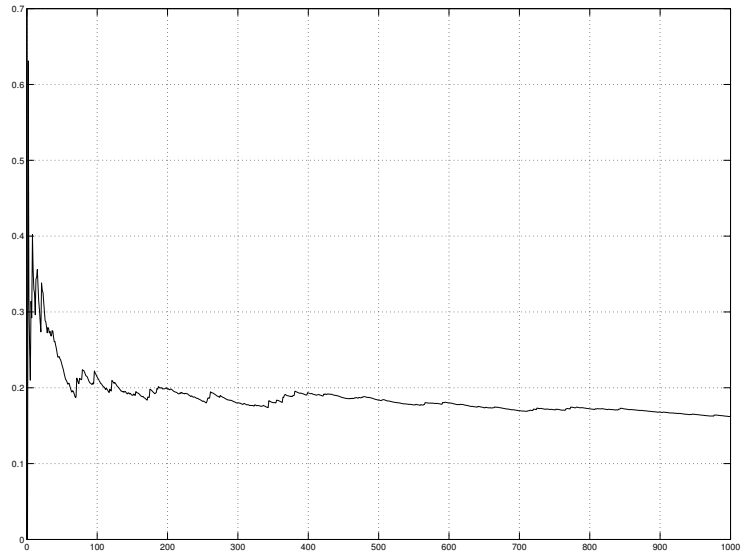

Figure 2: convergence to 0.17 of the bootstrap estimate for variance for t_1, the jackknife estimate of variance was 0.08

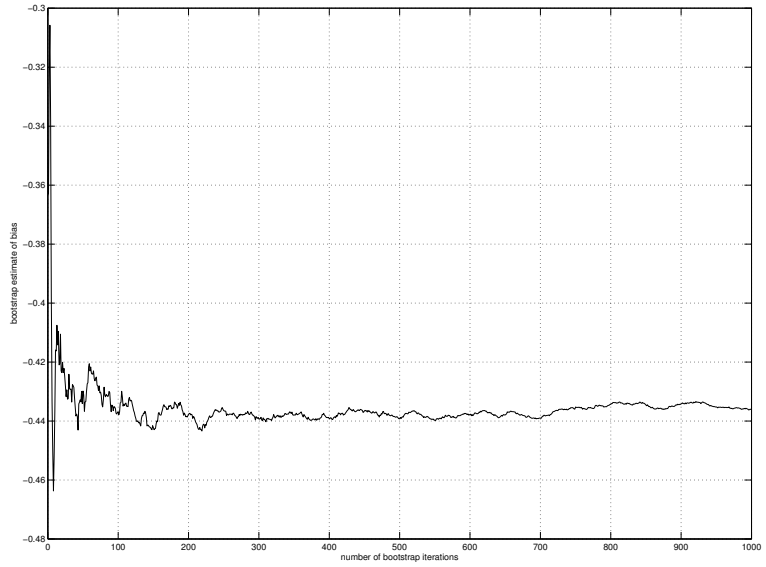

Figure 3: convergence to -0.44 of the bootstrap estimate for bias for t_2, the jackknife estimate of bias was 0.0004

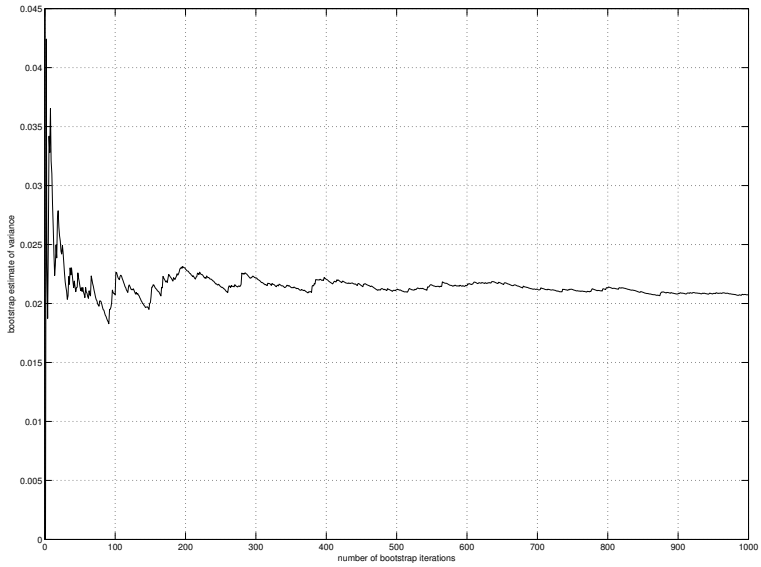

Figure 4: convergence to 0.02 of the bootstrap estimate for variance for t_2, the jackknife estimate of variance was 0.0007

discussions about the infinitesimal jackknife can be found in [8], [3], and [4].
Miller who has a rigorous study of the jackknife [8] introduced an amendment to the jackknife which he referred to as a *more trustworthy* jackknife, which essentially was a further correction of the bias estimation reducing the error order from $O_p(\frac{1}{n^2})$ to $O_p(\frac{1}{n^3})$. In 1971 [10] introduced a new method of bias correction which used two biased estimators s_1 and s_2 with biases b_1 and b_2 and introduced $\hat{s} = \frac{s_1 - Rs_2}{1-R}$ where $R = \frac{b_1}{b_2}$ as a bias corrected estimator, and argued that this outperforms Miller's more trustworthy jackknife.
[12] presents a study of the jackknife and the bootstrap in the regression analysis scenario, which we have completely skipped here. The basic ideas of the bootstrap were introduced in a couple of papers by Efron [2] and [1], which were almost entirely summarized in the two books [3] and [4], from which we have almost followed the choice of notation and proof flows.

References

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

[2] Bradley Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, December 1981.

[3] Bradley Efron. *The jackknife, the bootstrap, and other resampling plans*. Society for Industrial and Applied Mathematics, January 1987.

[4] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1994.

[5] Alfonso García-Pérez. Von mises approximation of the critical value of a test. *TEST*, 12:385–411, 2003. 10.1007/BF02595721.

[6] Gail Gong. Cross-validation, the jackknife, and the boostrap: Excess error estimation in forward logistic regression. Technical Report 80, Division of Biostatistics, Stanford University, Stanford, California, August 1982.

[7] Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, June 1974.

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

[9] M. H. Quenouille. Notes on bias estimation. *Biometrika*, 43(3/4):353–360, December 1956.

[10] W. R. Schucany, H. L. Gray, and D. B. Owen. On bias reduction in estimation. *Journal of the American Statistical Association*, 66(335):pp. 524–533, 1971.

[11] J. W. Tukey. Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, 29(2):614, 1958.

[12] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, December 1986.