**Bootstrap Tilting Inference and Large Data Sets**
Proposal to NSF SBIR Program

Tim C. Hesterberg
MathSoft, Inc., 1700 Westlake Ave. N., Suite 500
Seattle, WA 98109-3044, U.S.A.
TimH@statsci.com
June 11, 1998

This Small Business Innovation Research Phase I project is for research on confidence intervals and hypothesis tests using fast bootstrap methods, and ways to make bootstrapping feasible for large data sets.

Classical inference (intervals and tests) methods are known to be inaccurate when the underlying assumptions are violated, the usual case in practice. For example, skewness causes the usual $t$-test to be in error. The new methods would be an order of magnitude (power of $\sqrt{(}n)$, where $n$ is the sample size) more accurate in general than classical inferences.

Bootstrap methods are a promising alternative to classical inferences, and can handle complex statistics including modern robust statistics, but are slow and have been little used in practice. The methods proposed are 17 times faster than other bootstrap methods.

The methods are fast enough to be seamlessly incorporated into standard software, alongside or instead of classical inferences. This provides statistical practitioners a realistic alternative to easy but inaccurate classical inferences and non-robust methods. The competitive advantage to the firm that does this first is a major opportunity. Furthermore, the large sample methods would be attractive in the thriving data mining market.

Key words: bootstrap, resampling, tilting, importance sampling, least-favorable family, data mining

## Contents

## 1 Identification and Significance of the Opportunity

The confidence intervals and hypothesis tests used most often in statistical practice are based on normal approximations and theoretical derivations based on assumptions about the underlying distributions. Unfortunately, these classical methods are commonly used even when the assumptions are violated, causing substantial errors. For example, the errors caused by skewness when performing a $t$-test for the mean are $O(n^{-1/2})$ ($n$ is the sample size), an order of magnitude larger than $O(n^{-1})$ difference between using Students-$t$ and normal quantiles. The actual Type I error probability can easily be double the desired value. Similar situations exist throughout statistical practice. There exists an opportunity to change that.

The bootstrap is a powerful tool for statistical inference that substitutes raw computing power for theoretical analysis. It approximates the distribution of a statistic using only the observed data, without resorting to asymptotic and other approximations simply for mathematical and computational tractability. Resampling methods (including the bootstrap) "replace 'theory from a book', typified by t-tables and F-tables, by 'theory from scratch', generated anew by the computer for each new data analysis problem" [10]. Bootstrap methods can often be applied in more complex real applications than competing methods, without requiring the user to perform analytical calculations. The interest in bootstrap methods in statistical research has been enormous; a search of the Current Index to Statistics yielded over 1500 articles published through 1996 on the bootstrap. A number of existing bootstrap procedures are "second order correct" under general conditions, an order of magnitude more accurate than classical methods. But the impact on statistical practice has not been as great, due in large part to the slowness of bootstrapping.

We propose to develop bootstrap methods that are Fast, fast enough to be used routinely and automatically alongside classical inferences. Whenever a statistician requests a $t$-interval or hypothesis test—for one or two problems, linear regression, or a wide variety of other procedures—the software could give the bootstrap tilting answers as well, and warn when the classical answers may be inaccurate.

The new methods are based on bootstrap tilting, proposed not long after the invention of the bootstrap [11] but nearly overlooked since then, with the notable exception of theoretical work by [9], who show that bootstrap tilting intervals are second order correct. With the right implementation the method can be must faster than other bootstrap methods, e.g. requiring only 60 bootstrap replications instead of 1000 for comparable accuracy. This is fast enough for routine use, for software to provide by default without annoying users (depending on the size of the data and speed of the statistic). Furthermore, some tilting methods should be more accurate than even other existing bootstrap procedures.

In addition, we propose to make bootstrapping feasible in much larger problems without analytical calculations. Tilting and many existing bootstrap methods require evaluating a statistic say 60 or 1000 times for the actual bootstrapping, plus an additional $n$ times. This is impractical for large data sets where $n$ is ten thousand or more. We propose ways to avoid the additional effort. The methods are not limited to simple statistics, but also handle robust and other modern statistical methods.

The proposed research, if successful, would offer a wide range of scientists and engineers much better methods of inference than they currently use. The combination of speed, accuracy, and ability to handle complex statistics and large data sets, can steer practitioners away from easy but inaccurate classical inferences and non-robust methods.

The firm that first seamlessly provides these bootstrapping capabilities would enjoy a major competitive advantage. Providing the methods for routine use inside a wide range of statistical testing and modeling functions would justify a new release of the MathSoft product line and a major marketing push, worth millions of dollars.

## 2 Background and Technical Approach

We begin with a short introduction to the bootstrap, then discuss new methods in subsequent sections; for a more complete introduction to the bootstrap see [16]. We conclude this background

section with a discussion of S-Plus and current bootstrap software.

The original data is $\mathcal{X} = (x_1, x_2, \ldots, x_n)$, a sample from an unknown distribution $F$, which may be multivariate. Let $\theta = \theta(F)$ be a real-valued functional parameter of the distribution, such as its mean, interquartile range, or slope of a regression line, and $\hat{\theta} = \theta(\hat{F})$ the value estimated from the data. The sampling distribution of $\hat{\theta}$

$$G(a) = P_F(\hat{\theta} \leq a) \tag{1}$$

is needed for statistical inference. In simple problems the sampling distribution can be approximated using methods such as the central limit theorem and the substitution of sample moments such as $\overline{x}$ and $s$ into formulas obtained by probability theory. This may not be sufficiently accurate or even possible in many real, complex situations.

The bootstrap principle is to estimate some aspect of $G$, such as its standard deviation, by replacing $F$ by an estimate $\hat{F}$. In this proposal we consider nonparametric problems for which $\hat{F}$ is the empirical distribution. Let $\mathcal{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ be a "resample" (a bootstrap sample) of size $n$ from $\hat{F}$, denote the corresponding empirical distribution $\hat{F}^*$, and write $\hat{\theta}^* = \theta(\hat{F}^*)$. In simple problems the bootstrap distribution $P_{\hat{F}}(\hat{\theta}^* \leq a)$ can be calculated or approximated analytically, but it is usually approximated by Monte Carlo simulation—for some number $B$ of resamples, sample $\mathcal{X}_b^*$ for $b = 1, \ldots, B$ with replacement from $\mathcal{X}$, then let

$$\hat{G}(a) = B^{-1} \sum_{b=1}^{B} I(\hat{\theta}_b^* \leq a). \tag{2}$$

There are two levels of approximation here—approximating (1) by $P_{\hat{F}}(\hat{\theta} \leq a)$, and estimating the latter by Monte Carlo simulation. We consider both levels in this proposal.

Similarly the sampling distribution of a (possibly pivotal) statistic $T = T(\hat{F}, F)$

$$J(a) = P_F(T \leq a) \tag{3}$$

can be approximated by $P_{\hat{F}}(T^* \leq a)$ where $T^* = T(\hat{F}^*, \hat{F})$, and implemented by Monte Carlo sampling

$$\hat{J}(a) = B^{-1} \sum_{b=1}^{B} I(T_b^* \leq a). \tag{4}$$

For example, the bias of $\hat{\theta}$ is the mean of the sampling distribution of $T = \hat{\theta} - \theta$, and can be estimated by the mean of $T^*$. Another example is the $t$-statistic used for bootstrap-$t$ confidence intervals [11], $T = (\hat{\theta} - \theta)/s(\hat{F})$ where $s(\hat{F})$ is an estimate of the standard deviation of $\hat{\theta}$.

We restrict consideration to distributions with support on the observed data; methods described below could be extended to parametric situations or smoothed bootstrapping, but that is beyond the scope of Phase I of this proposal. Then we may describe a distribution in terms of the probabilities $\mathbf{p} = (p_1, \ldots p_n)$ assigned to the original observations; $\hat{F}$ corresponds to $\mathbf{p}_0 = (1/n, \ldots, 1/n)$. Let $\theta(\mathbf{p})$ be the corresponding parameter estimate (which depends implicitly on $\mathcal{X}$). Also write $\mathbf{p}^* = (M_1^*/n, \ldots, M_n^*/n)$ for the vector corresponding to resample $\mathcal{X}^*$, where $M_i^*$ is the number of times $x_i$ is included in $\mathcal{X}^*$. For later use, let

$$U_i(\mathbf{p}) = \lim_{\epsilon \to 0} \epsilon^{-1}(\theta(\mathbf{p} + \epsilon(\delta_i - \mathbf{p})) - \theta(\mathbf{p})) \tag{5}$$

where $\delta_i$ is the vector with 1 in position $i$ and 0 elsewhere. When evaluated at $\mathbf{p}_0$ these derivatives are known as the empirical influence function, or infinitesimal jackknife.

A fundamental assumption in the application of the bootstrap is that the theoretical bootstrap distributions $P_{\hat{F}}(\hat{\theta}^* \leq a)$ and $P_{\hat{F}}(T^* \leq a)$ accurately approximate (1) and (3), respectively; in

other words that $\hat{F}$ can substitute for the unknown $F$. Theoretical treatments of some aspects of the assumption are summarized in [20], using Edgeworth expansions, and [41], using functional analysis. We weaken the assumption by using the sampling distributions of $\hat{\theta}^*$ and $T^*$ under certain distributions other than $\hat{F}$ which belong to "least-favorable" families (described below). These families play a major role in other bootstrap procedures [11, 13, 8].

In Section 2.1 we discuss bootstrap tilting inferences, in which confidence intervals and hypothesis tests are obtained using least-favorable families. We discuss using tilting to improve bootstrap-$t$ inferences in Section 2.2, and implementation issues for large samples in Section 2.3.

### 2.1 Bootstrap Tilting Inference

In this section we discuss bootstrap tilting hypothesis tests, which might prove to be both more accurate and computationally more efficient than currently popular bootstrap inference methods. We propose research dealing with implementation details that affect both asymptotic and finite-sample accuracy and computational efficiency.

Consider testing $H_0$: $\theta = \theta_0$. In a one-parameter parametric problem one would compare the observed $\hat{\theta}$ with a critical value of its null distribution, obtained by sampling from the parametric distribution $F_{\theta_0}$ rather than $F_{\hat{\theta}}$. In a more general parametric setting, with one parameter $\theta$ of interest and a number of nuisance parameters, one might find the maximum likelihood estimate of the parameters under the null hypothesis, then compare the observed value of some statistic (a pivotal statistic, likelihood ratio, or $\hat{\theta}$) with its estimated null distribution. Again, sampling is from a distribution consistent with the null hypothesis.

Similarly, bootstrap sampling for a hypothesis test should be from a distribution consistent with the null distribution. This seems to conflict with the usual bootstrap practice of sampling from the observed distribution, but in fact the bootstrap principle is to sample from the best estimate of the underlying distribution, given the information available, which may include the constraint implied by the null hypothesis. For instance [39, 40] sample in this way, for testing independence, rotational invariance, symmetry, and similar problems. Others (e.g. [4]) sample in various ways consistent with the null hypothesis in two-sample and multi-sample problems. Bootstrap tilting hypothesis tests also sample this way, and were used by [45] for a one-sample mean and suggested by [32] for comparing two means.

The maximum likelihood estimate of the distribution, consistent with $H_0$ and with support on the observed data, maximizes $\prod p_i$ subject to $p_i \geq 0$, $\sum p_i = 1$, and $\theta(\mathbf{p}) = \theta_0$. In the case of a mean, $\theta(\mathbf{p}) = \sum p_i x_i$, $U_i(\mathbf{p}) = x_i - \overline{x}$, and the solution can be written in the form

$$p_i = c(1 - \tau(x_i - \overline{x}))^{-1}, \tag{6}$$

where $\tau$ is a "tilting" parameter and $c$ normalizes the probabilities to sum to 1. The value of $\tau$ that satisfies the last constraint is found numerically. These probabilities are a special case of what we call "maximum likelihood tilting" (ML tilting), and are shown in Figure 1. Here the unweighted sample mean is less than the null hypothesis value, so tilting places higher probabilities on the larger values of $x$ to make the weighted mean match $\theta_0$.

In bootstrap tilting hypothesis testing, the null distribution of $\hat{\theta}$ is estimated by resampling from the weighted empirical distribution, and $H_0$ is rejected in favor of $H_a$: $\theta > \theta_0$ if the estimated $p$-value is less than $\alpha$,

$$P_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) < \alpha, \tag{7}$$

where $F_\tau$ is the weighted empirical distribution induced by tilting with parameter $\tau$.

The procedure can be generalized to nonlinear statistics, and by substituting another single-parameter family for the maximum likelihood tilting family. The chosen family should be least-favorable, i.e. inference within a family is not easier, asymptotically, than in the full $(n-1)$-dimensional family. We consider four families in this proposal,

$$\mathcal{F}_1 : p_i \quad = \quad c\exp(\tau U_i(\mathbf{p}_0))$$

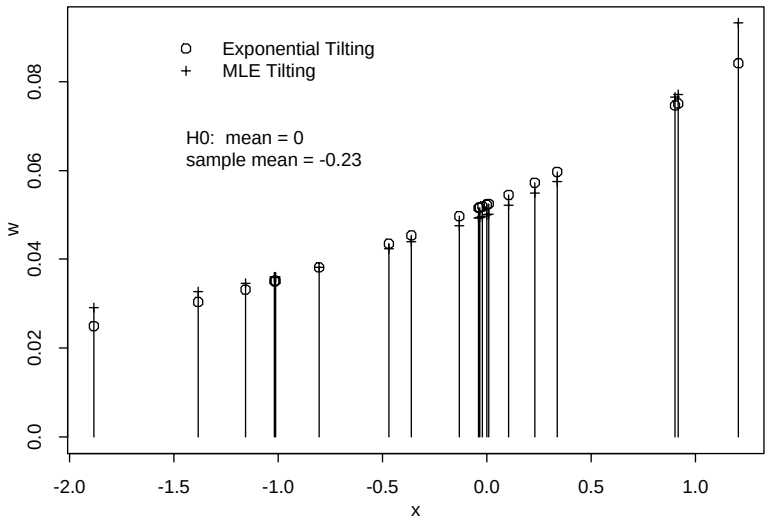


Figure 1: Exponential and Maximum Likelihood Tilting for a mean.

$$\begin{aligned} \mathcal{F}_2 : p_i &= c\exp(\tau U_i(\mathbf{p})) \\ \mathcal{F}_3 : p_i &= c(1-\tau U_i(\mathbf{p}_0))^{-1} \\ \mathcal{F}_4 : p_i &= c(1-\tau U_i(\mathbf{p}))^{-1}, \end{aligned} \tag{8}$$

each indexed by a tilting parameter $\tau$, where each $c$ normalizes the corresponding vector to add to 1. $\mathcal{F}_1$ and $\mathcal{F}_2$ are well-known as "exponential tilting", and coincide if $\theta$ is a mean; these weights are also shown in Figure 1. Similarly $\mathcal{F}_3$ and $\mathcal{F}_4$ are ML tilting and are the same as (6) for a mean. $\mathcal{F}_4$ gives the maximum likelihood solution for nonlinear statistics. In the sequel we write $\mathbf{p}_\tau$ and $F_\tau$ for the corresponding probability vector and weighted empirical distribution, respectively. Note that $\tau = 0$ corresponds to $\mathbf{p}_0$ and $\hat{F}$.

For any family, $\tau$ is found numerically to solve

$$\theta(\mathbf{p}_\tau) = \theta_0 \tag{9}$$

and the decision to reject is based on the estimated $p$-value under weighted bootstrap sampling (7).

### 2.1.1 Bootstrap tilting intervals

Bootstrap tilting hypothesis tests are consistent with the bootstrap tilting confidence intervals defined by [11], in that the test rejects $H_0$ iff the confidence interval excludes $\theta_0$. After choosing a least-favorable family, the lower limit of a one-sided $(1-\alpha)$ interval is found by solving

$$P_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) = \alpha \tag{10}$$

in $\tau$, then defining the lower limit as

$$\theta_\alpha = \theta(F_\tau).$$

Upper limits are found similarly. [9] show that bootstrap tilting intervals are second-order correct under general assumptions, i.e. that the one-sided coverage errors are $O(n^{-1})$ (they consider only $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_4$). This is the same rate as for better-known procedures such as the bootstrap-$t$ [11] and BC-$a$ [13] intervals.

Bootstrap tilting corresponds to an exact method in single-parameter parametric problems, where the lower limit of the confidence interval is defined to be that value $\theta_\alpha$ for which $P_{\theta_\alpha}(\hat{\theta}^* > \hat{\theta})$, where $\hat{\theta}$ is the estimate from the observed data and $\hat{\theta}^*$ is the random estimate obtained from a new sample. Here, by restricting to a least-favorable family, the problem is reduced to a single-parameter parametric family.

### 2.1.2 Choice of least-favorable family

There are two important implementation decisions for either confidence intervals or hypothesis tests: which least-favorable family to use, and how (10) is solved or (7) is evaluated. Investigation of these details is the heart of our proposed contributions to bootstrap tilting inference.

The bootstrap literature contains little discussion of the merits of the different least-favorable families, but simulations have tended to use $\mathcal{F}_1$ because it offers some computational advantages. $\mathcal{F}_4$ corresponds to maximum likelihood estimation subject to a null hypothesis, and is the family used in empirical likelihood (EL) inference [36, 37, 21]; both limit support to the observed values and find the restricted maximum likelihood vector of probabilities. But where EL inference is based on asymptotic approximations, in bootstrap tilting all probabilities are estimated by sampling. [7] study bootstrap likelihood and EL, and discuss relative advantages of EL and bootstrap methods, and [30] discusses connections between the bootstrap and EL.

We propose to compare the families, in terms of accuracy and computational efficiency. We suggest that $\mathcal{F}_4$ should give the most accurate inferences in finite-sample problems—the actual type I error and coverage rates should most closely match the nominal values. First, using derivatives (5) evaluated at $\mathbf{p}_\tau$ rather than $\mathbf{p}_0$, e.g. using $\mathcal{F}_4$ rather than $\mathcal{F}_3$, results in more conservative inferences in nonlinear problems—wider confidence intervals and smaller type I errors. Since in practice most bootstrap inferences tend to be anti-conservative with finite samples (see simulation results collected in [41]), these more conservative inferences should be more accurate.

Second, ML tilting should be more accurate than exponential tilting. Taylor-series expansions of the families in (8) in terms of $\tau$ about 0 agree to the first two terms, but the quadratic term for ML tilting is double that of exponential tilting. The result is apparent in Figure 1, where the ML tilting probabilities are larger than exponential tilting probabilities at *both* extremes of the distribution; they are smaller in the middle because the probabilities are normalized. When sampling from weighted bootstrap distributions, using ML tilting gives $\hat{\theta}^*$ a larger variance, so that confidence intervals are wider and hypothesis tests are less likely to reject $H_0$. Again, these more conservative inferences should be more accurate. Furthermore, a result by [27] implies that when $\theta$ is the mean, $H_0$ is true, and the weights are obtained by ML tilting so that $\sum_i p_i x_i = \theta_0$, then the weighted variance $\sum_i p_i (x_i - \theta_0)^2$ has bias of order $O(n^{-2})$, so that the bootstrap estimate of the variance of the sample mean is biased by a factor $O(n^{-2})$. In contrast the usual bootstrap estimate of variance is biased by a factor $n^{-1}$, as is the bias obtained using exponential tilting. Similar results should hold for nonlinear statistics. The relatively small bias for ML tilting should result in more accurate inferences.

However, using derivatives that implicitly depend on $\tau$ can be expensive. $\mathcal{F}_2$ and $\mathcal{F}_4$ can be found by minimizing the backward and forward Kullback-Leibler distances between $\mathbf{p}$ and $\mathbf{p}_0$, respectively, which requires constrained numerical optimization in $(n-1)$ dimensions. In contrast, $\mathcal{F}_1$ and $\mathcal{F}_3$ require only solving univariate equations in $\tau$. We propose using a two-step approximation to $\mathcal{F}_2$ or $\mathcal{F}_4$: first tilt using $U_i(\mathbf{p}_0)$ to find $\mathbf{p}_\tau^{(1)}$, then calculate $U_i(\mathbf{p}_\tau^{(1)})$ and tilt again to find an updated $\mathbf{p}_\tau^{(2)}$. Similar updating was used in another bootstrap context by [26], and in empirical likelihood by [44].

### 2.1.3 Numerical solution for tilting—Importance Sampling Reweighting

The next major implementation detail is the numerical solution of (10). This involves finding the value of $\tau$ for which resampling from $F_\tau$ yields a tail probability of $\alpha$.

One approach is to sample from the weighted empirical distribution $F_\tau$ for different values of $\tau$, estimate the tail probabilities for each $\tau$, smooth the estimated probabilities, and numerically find the $\tau$ for which the value of the smooth curve is $\alpha$. Because tail probabilities are relatively difficult to estimate using Monte Carlo simulation, this requires a large number of resamples (typically 1000 [13]) for each candidate value of $\tau$. This can be expensive. [8] suggest one alternative, the "automatic percentile method", which requires bootstrap sampling only from one candidate $F_\tau$ (in each tail for two-sided intervals) in addition to sampling from $\hat{F}$; this would typically require 3000 resamples. The automatic percentile method may also be used as an iterative process, whose fixed point is the bootstrap tilting endpoint; iterating more than once should give more accurate

endpoints, but requires more resamples.

A much more efficient approach [11] uses importance sampling reweighting (ISR), a non-traditional application of importance sampling. We review this method here before turning to its application in bootstrap tilting inference and later in bootstrap diagnostics. Variations have appeared under other names, e.g. likelihood ratio sensitivity analysis, likelihood ratio gradient estimation, the score function method, polysampling, likelihood ratio reweighting, importance sampling sensitivity analysis, and importance reweighting [2, 38, 43, 24, 29, 6].

Importance sampling is traditionally used to obtain more accurate answers in Monte Carlo simulation by concentrating effort on important regions in the sample space. In order to estimate an integral $\int Y(\mathcal{X})f(\mathcal{X})d\mathcal{X}$ one could generate $B$ observations from density $f$ and compute the average observed value of $Y$, $B^{-1}\sum_{b=1}^{B} Y_b$. Alternately, by rewriting the integral as $\int (Y(\mathcal{X})f(\mathcal{X})/g(\mathcal{X}))g(\mathcal{X})d\mathcal{X}$, where $g$ dominates $f$, one could generate observations from $g$, and report the average observed value of $(Yf/g)$. If $g$ is well chosen, so that $g$ is larger than $f$ in "important" regions where $Y$ is relatively large, then $(Yf/g)$ has smaller variance (under $g$) than does $Y$ (under $f$) [23].

The name "importance sampling" and the association with estimating integrals obscure the more general utility of the procedure. The procedure utilizes samples from a "design distribution" $g$ in order to estimate the distribution for $Y$ that would be obtained under sampling from the "target distribution" $f$. It need not be the case that $f$ is fixed and $g$ is chosen for variance reduction; in bootstrap tilting $g$ is chosen for convenience, and a single set of observations (resamples) from $g$ is used for estimation under an infinite number of target distributions.

[11] lets the design distribution be $\hat{F}$, and generates a single set of $B$ resamples by simple bootstrap sampling (with equal probabilities). Let $M^*_{b,i}$ be the number of times $x_i$ is included in $\mathcal{X}^*_b$. Then for any target distribution be $F_\tau$, with probabilities $\mathbf{p}_\tau$ on the observed data, the likelihood ratio $W = f/g$ for $\mathcal{X}^*_b$ is

$$W_b = \prod_{i=1}^{n} (np_i)^{M^*_{b,i}}. \tag{11}$$

For any $\tau$, an estimate of the left side of (10) is

$$\hat{P}_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) = B^{-1} \sum_{b=1}^{B} W_b I(\hat{\theta}^* \geq \hat{\theta}). \tag{12}$$

This procedure has a number of advantages. Sampling is simpler because no weights are involved, and a single set of resamples is used for both sides in a two-sided confidence interval. The estimated tail probabilities are a smooth monotone function of $\tau$, simplifying root-finding and eliminating the need for smoothing. Finally, by a fortunate coincidence, the unweighted empirical distribution is a well-known, nearly optimal, design distribution for the traditional role of importance sampling as a variance reduction technique, at least for the mean and exponential tilting. The advantage relative to simple Monte Carlo sampling is by a factor of about 17 for estimating a tail probability that is about 0.025. Thus, where $B = 1000$ replications are required for sufficient accuracy for other bootstrap confidence intervals based on percentiles [13] 60 might suffice here. This is a major computational savings, that appears not to be mentioned in the literature except in the forthcoming [30].

The computational advantage is even greater relative to the bootstrap BC-$a$ interval [13], the most common second-order-correct bootstrap interval (because $z_0$ is estimated from bootstrap results). The results of small simulation comparing the accuracy of bootstrap tilting confidence intervals with $B = 100$ replications to BC-$a$ intervals with $B = 2000$ are shown in Table 1; the tilting intervals are more accurate.

Figure 2 shows the bootstrap distribution for the treatment coefficient in a Cox proportional hazards regression (the center curve), together with two bootstrap distributions obtained by tilting (using ISR) such that the probabilities of falling above the original $\hat{\theta}$ (shown by a vertical line) are

Table 1: Simulation Variability

| Method | $p = .025$ | $p = 0.5$ | $p = .95$ | $p = .975$ |
|---|---|---|---|---|
| Var of BC-$a$, $B = 2000$ | 1.60 | 1.44 | 4.65 | 5.11 |
| Var of Tilting, $B = 100$ | 1.16 | 1.43 | 2.86 | 2.48 |
| Relative Efficiency | 28 | 20 | 33 | 41 |

Variance of BC-$a$ and Bootstrap (Exponential) tilting confidence intervals for the mean, $n = 11$, data from [18]. $B$ is the number of bootstrap samples used. The relative efficiency is ratio of variances, corrected for the difference in sample size; this gives the relative number of bootstrap samples required for comparable accuracy.

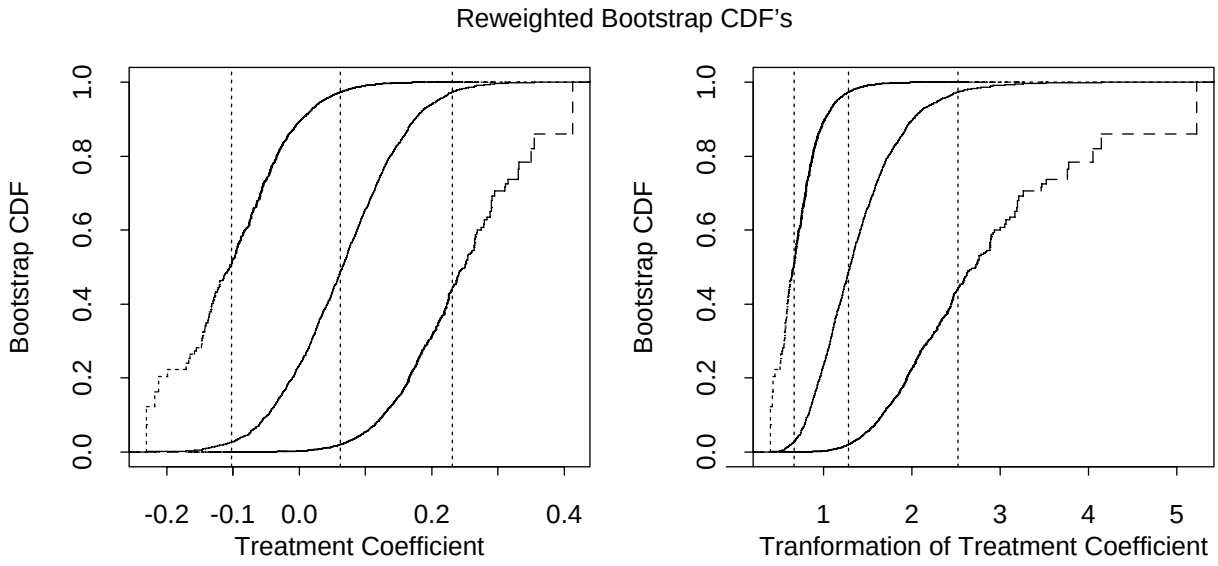


Figure 2: Importance Sampling Reweighting. The three curves are obtained from the same resamples, but reweighted to tilt the distribution left, not reweighted, and right. Vertical lines are at $\hat{\theta}$ and the two values of $\theta(F_\tau)$. In the left panel $\theta$ is the treatment coefficient for a Cox model for the head and neck data, in the right panel $\theta$ is a nonlinear transformation of the coefficient.

about 0.025 and 0.975 for the left and right curves, respectively. All three distribution estimates make their vertical jumps at the same locations, but the sizes of the jumps for the two outer curves depend on the weights $W_b$. The leftmost curve takes large jumps at the left and is inaccurate there, but is very accurate near $\hat{\theta}$, so the probability in (10) is accurately estimated.

The data used here are provided by Dr. Michael LeBlanc of the Fred Hutchinson Cancer Research Center, consisting of survival times of 158 patients in a head and neck cancer study; 18 of the observations were right-censored. The control group received surgery and radiotherapy, while the treatment group also received chemotherapy. The statistic $\theta$ is the treatment coefficient in a Cox proportional hazards regression model. The coefficient is the log of the estimated ratio of the hazards rates between groups; some users may be interested in bootstrapping the hazard ratio $\exp(\theta)$ directly, a nonlinear transformation of $\theta$. The right panel is for such a transformation (actually $\exp(4\theta)$, for greater nonlinearity for presentation purposes), and illustrates that bootstrap tilting is invariant under transformations—the endpoints of a confidence interval for this transformed coefficient are the same as the transformation of the endpoints for the untransformed coefficient.

However, there are a number of factors that may increase the computational burden. First, if the derivatives (5) are estimated numerically (e.g. using the jackknife), an additional $n$ resamples

are needed, and $n$ may be much larger than 60; we propose a way to mitigate this in Section 2.3.

Second, estimates can be unstable if $\theta$ is nonlinear. In the head and neck example $\theta$ is nearly linear (correlation 0.993 between $\theta^*$ and a linear approximation), so all of the large jumps in the leftmost curve in Figure 2 occur on the left side of that curve. But for nonlinear situations large jumps could occur on the right, where accuracy matters. We have observed this when bootstrapping the correlation coefficient for the law school data [12]. We propose to use a defensive mixture distribution [27] that produces estimates that are more robust against nonlinearity. This involves using a small number of resamples from $F_\tau$ in addition to the resamples from $\hat{F}$; if $\lambda B$ of the $B$ resamples are from $F_\tau$, then the jump size $W_b/B$ is bounded above by $(\lambda B)^{-1}$. Using a defensive mixture has another advantage. The weighted cumulative distribution function with weights $W_b/B$ has the range $[0, B^{-1}\sum_b W_b]$ rather than $[0,1]$, and the upper limit can be very different from 1, e.g. over 10% off for one of the curves in Figure 2. The upper limit is much less variable when defensive mixtures are used. The curves actually plotted in that figure were simply normalized to the range $[0,1]$, but this reduces the accuracy for estimating tail probabilities; with defensive mixtures we may use more accurate normalization methods [27].

Third, $\hat{F}$ is a nearly optimal design distribution (for nearly linear problems) for exponential tilting, but not for ML tilting; a more accurate design would involve exponentially tilting the ML tilting probabilities back to the center, e.g. $p_i' = p_i c' \exp(\tau' U_i)$ where $p_i$ is obtained by ML tilting, $c'$ is a normalizing constant, and $\tau'$ tilts the distribution back toward the center so that $\theta(\mathbf{P}^*) = \hat{\theta}$. This design places slightly higher probabilities on the more extreme observations (large and small values of $U_i$) than does $\hat{F}$. We propose to investigate this design, and an alternative that uses exponential tilting but with inferences adjusted to approximate ML tilting; exponential tilting has an additional advantage, that that the computation of (11) is particularly convenient [11], $W_b = (nc)^n \exp(\sum M_{b,i}^* U_i)$.

ISR can also be used to estimate the $p$-value for a bootstrap tilting hypothesis test.

In summary, bootstrap tilting confidence intervals and hypothesis tests are potentially very accurate and computationally efficient. They are second-order accurate, and may have smaller errors of order $O(n^{-1})$ than do other second-order accurate procedures, particularly when family $\mathcal{F}_4$ is used. The use of ISR makes their implementation computationally efficient, perhaps requiring only 60 resamples rather than 1000. However, further work is needed before these procedures are ready for widespread use with complex statistics. Our plans are described in Section 3.

### 2.2 Bootstrap-$t$

In this section we describe the bootstrap-$t$ interval, two problems with it, and possible improvements based on tilting.

Let $T = (\hat{\theta} - \theta)/s(\hat{F})$ where $s(\hat{F})$ is an estimate of the standard deviation of $\hat{\theta}$; the lower endpoint of a bootstrap-$t$ $(1-\alpha)$ confidence interval is

$$L_1 = \hat{\theta} - s(\hat{F})\hat{J}^{-1}(1-\alpha) = \hat{\theta} - s(\hat{F})T^*_{(B(1-\alpha))} \tag{13}$$

The procedure is second-order correct under general circumstances [19], but has exhibited poor finite-sample performance, e.g. it "fails spectacularly" [20] when applied to the correlation coefficient.

**Transformation invariance** The problem in [20] is that the interval is not transformation-invariant. Let $\psi(\theta)$ be a smooth increasing transformation, and let $T' = (\psi(\hat{\theta}) - \psi(\theta))/s_\psi(\hat{F})$ where $s_\psi(\hat{F})$ is an estimate of the standard deviation of $\psi(\hat{\theta})$; the bootstrap-$t$ endpoint for $\psi(\theta)$ is not in general equal to $\psi$ of the endpoint for $\theta$. It is generally recognized that a variance-stabilizing transformation should often be used first. Let $v(\theta) = \mathrm{Var}(\hat{\theta}|\theta)$ denote the variance of $\hat{\theta}$ as a function of $\theta$. Then an approximate variance-stabilizing transform $\psi$ can be defined as

$$\psi(\theta) = \int (\hat{v}(\theta))^{-1/2} d\theta; \tag{14}$$

the indefinite integral is evaluated numerically.

[42] estimates a variance-stabilizing transformation from the data, using the double bootstrap. For some number $B_1$ of first-level resamples (say 100), he generates $B_2$ (say 25) second-level resamples, lets $\hat{\omega}_b$ be the sample variance of the values of $\hat{\theta}^{**}$ from the second level resamples from $\mathcal{X}_b^*$, performs a scatterplot of $\hat{\omega}_b$ against $\hat{\theta}_b^*$ for $b = 1, \ldots, B_1$, smoothes the scatterplot to obtain $\hat{v}$, then uses (14).

As an aside, we note that [42] does not actually use bootstrap-$t$ intervals on the transformed scale, but rather a basic bootstrap interval (defined below) on that scale.

We propose an alternate procedure, using bootstrap tilting. This uses only a single set of bootstrap observations. For any $\tau$, compute probabilities $\mathbf{p}_\tau$ (8), the corresponding weighted statistic $\theta\tau = \theta(F_\tau)$, the bootstrap weights (11), and let

$$\hat{v}(\theta_\tau) = \mathrm{Var}_{F_\tau}(\hat{\theta}^*)$$

be the variance of the weighted bootstrap distribution. This defines a relationship between variance and $\theta$, implicitly in terms of $\tau$. For example, in Figure 2 we see the cumulative distribution functions for the weighted bootstrap distributions for three values of $\tau$ ($\tau_0$, and two values chosen so that $\theta_{\tau_1}$ and $\theta_{\tau_2}$ are at approximately 0.025 and 0.975 confidence limits for $\theta$) for each of two statistics. In the right panel the variance of the weighted bootstrap distribution is strongly dependent on $\theta$, in the left panel nearly independent.

In Figure 3 we see the functional relationships between variance and $\theta$ estimated by the different procedures. The top left panel is for the data of [18], (9.6, 10.4, 13.0, 15.0, 16.6, 17.2, 17.3, 21.8, 24.0, 26.9, 33.8). The statistic is the mean. The positive slope of all curves in the middle is due to the positive skewness of the data; the corresponding variance-stabilizing transformation would be concave. Both tilting procedures produce somewhat higher estimates of variance than the double bootstrap for values of $\theta$ farther from $\hat{\theta}$. We investigate this further in the remaining three panels in the figure. Here the data are artificial, 20 samples of size 16, each formed by reflecting 8 standard normal variates about the mean, then standardizing to variance 1. This is a situation in which the true variance is constant, not dependent on $\theta$. The use of symmetric data ensures that the estimated variance relationships have slope 0 at $\theta = 0$ (except for simulation error in iterated bootstrapping), making it easier to view the second derivatives of the curves. In this example we use $B_1 = 500$ and $B_2 = \infty$ (analytical calculations replace the second level of bootstrapping); even with these relatively large values for $B_1$ and $B_2$ the iterated bootstrapping curves exhibit considerable variability. The tilting curves are more stable.

The exponential tilting curves tend to have negative curvature, and the ML tilting curves positive curvature (near the center); this suggests that a family intermediate between exponential and ML tilting might be preferred for variance stabilization, because the ideal curve in this artificial situation is known to be flat.

In summary, it appears that the tilting methods give more stable results in estimating variance-stabilizing transformations, with far less computational effort, and that a family intermediate between the exponential and ML tilting families may give the least-biased results.

**Bootstrap-$t$ intervals too long** The second criticism of bootstrap-$t$ intervals is that they tend to be too long [13, 34, 41]. This is generally attributed to instability in the estimates of standard deviation. *Our diagnostics suggest another explanation*—that the bootstrap standard deviations are too small, when $\hat{\theta}^*$ is not close to $\hat{\theta}$. See e.g. the smoothed double bootstrap curve in Figure 3.

In retrospect, this is not surprising. Consider a simple example. Recall that $\bar{X}$ and $s$ are independent when sampling from normal populations. But when taking *resamples* from normal samples, $\bar{X}^*$ and $s^*$ are not independent—$s^*$ tends to be smaller when $(\bar{X}^* - \bar{X})$ is large; see the second panel of Figure 3. When computing a bootstrap-$t$ statistic, the standard error in the denominator tends to be small when the numerator is large, causing the distribution of $T^*$ to have long tails.

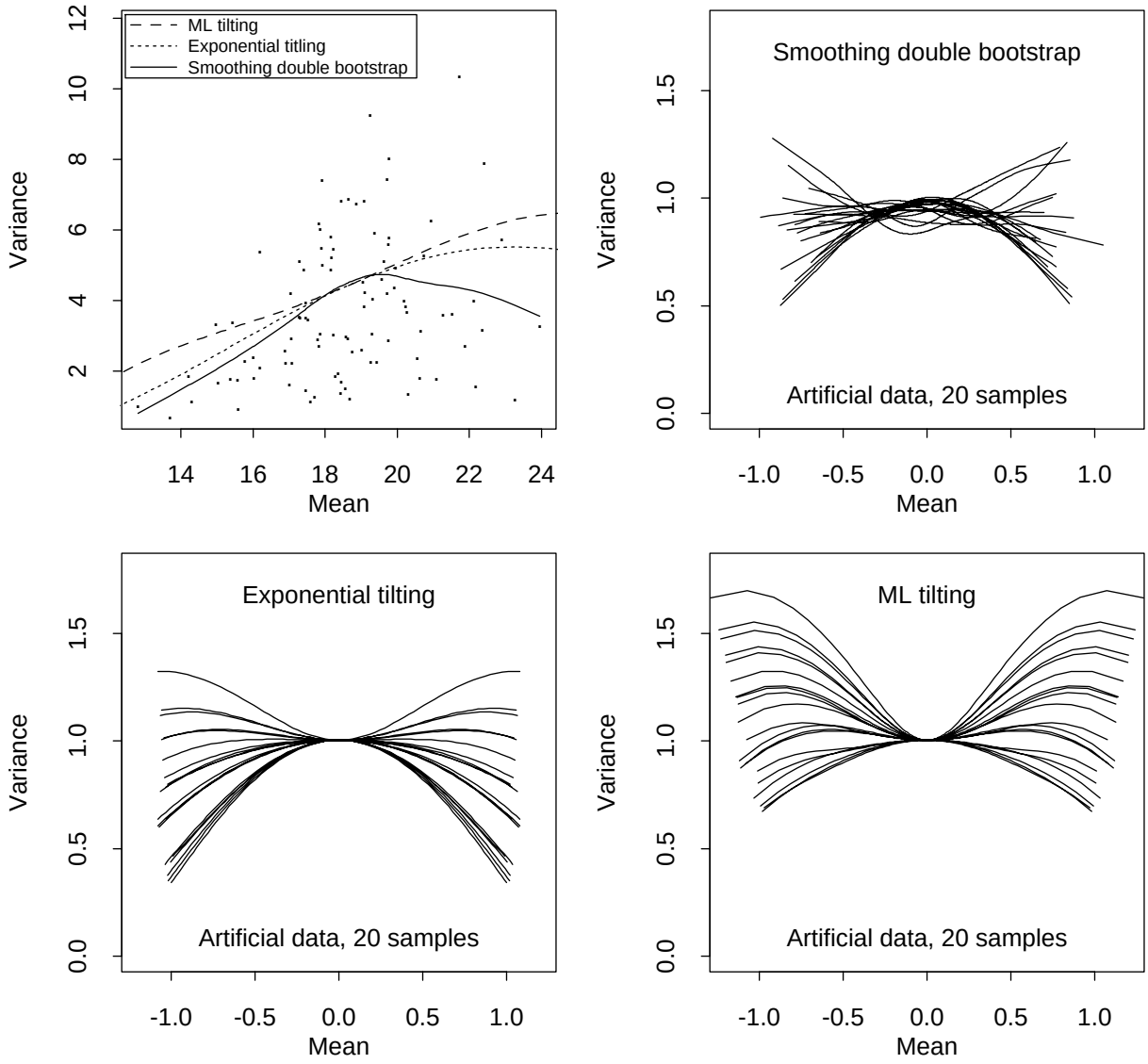


Figure 3: Estimating Variance as a function of $\theta$. The top left panel shows the estimated functional relationship obtained by iterated bootstrapping, exponential and ML tilting. The plotted points are the sample variances of 25 second-level resamples against $\theta$ for the corresponding first-level resample. A scatterplot smooth produces the iterated bootstrap estimate. The other three panels show the curves produced by the same three procedures for 20 sets of artificial symmetric approximately normal data.

We propose to adjust the denominator based on the ratio of the exponential and ML tilting estimates of $\hat{v}(\theta^*)$.

**Implementation by ISR** Both variance-stabilization and denominator adjustment can be implemented by ISR, but in Figure 2 the two outer curves are inaccurate over half of their ranges. This is due to the ISR method used to create this plot, in which the design distribution was $\hat{F}$. An alternative is to actually sample from the corresponding distributions $F_{\tau_1}$ and $F_{\tau_2}$ (here $\tau_1$ is the negative value that gives the left curve, and $\tau_2$ the positive value that gives the right curve).

We propose to develop a more accurate procedure that shares resamples across the three distributions, using ISR with a mixture design distribution using roughly equal numbers of resamples from $\hat{F}$ and each of the two tilted distributions. The three distributions $F_{\tau_1}$, $\hat{F}$, and $F_{\tau_2}$ can effectively share resamples, except on the outside of the two outer distributions. This would make the center curve extremely accurate in both tails, and the two outer curves extremely accurate in one tail each and would not hurt their accuracy in the other tail. Thus the number of resamples required is $3B$ (and $B$ could be reduced). Distribution function curves for intermediate values of $\tau$ could also be estimated using ISR without further resampling.

**Bootstrap Tilting-$t$ Interval** A common use for iterated bootstrapping is to calibrate bootstrap confidence interval procedures [34, 3]; giving an increase of one order of accuracy for every level of bootstrap iteration under fairly general circumstances [22], but this is computationall expensive. Bootstrap tilting might serve the same role, at least for one level. Indeed, it already has—we show here that bootstrap tilting confidence intervals can be interpreted in this way. Let $T = T(\hat{F}, F) = \hat{\theta} - \theta$, and $T^* = \hat{\theta}^* - \hat{\theta}$, and let $\hat{J}^{-1}(q)$ denote the $q$ percentile of $\hat{J}$. Treating $T$ as a pivot yields the approximation $P(\hat{\theta} - \theta < \hat{J}^{-1}(q)) = q$, which can be inverted to yield confidence intervals. The lower endpoint of a one-sided $(1-\alpha)$ confidence interval and its Monte Carlo approximation are

$$L_2 = \hat{\theta} - \hat{J}^{-1}(1-\alpha) = \hat{\theta} - (\hat{\theta}^*_{(B(1-\alpha))} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}^*_{(B(1-\alpha))} \tag{15}$$

where $\hat{\theta}^*_{(k)}$ is the $k$'th order statistic of the bootstrap distribution. This interval is relatively common—[41] (page 141) indicate that "this method is used in practice more frequently than any other bootstrap method, especially when the problem under consideration is complex", but often appears in the bootstrap literature with no name. We follow [6] in calling it the "basic bootstrap".

Now suppose that $\hat{J}$ is estimated by sampling not from $\hat{F}$, but from $F_\tau$, where $\theta(F_\tau)$ equals the yet-to-be-determined endpoint $L_3$. In principle, this should be more accurate; for example, this yields exact endpoints in one-parameter parametric problems. Now $\hat{J}_{F_\tau}$ becomes the distribution of $\hat{\theta}^* - L_3$ when sampling from $F_\tau$, and the $\alpha$ quantile of this distribution is the $\alpha$ quantile of $\hat{\theta}^*$, minus $L_3$. In place of (15), we solve

$$\begin{aligned} L_3 &= \hat{\theta} - \hat{J}^{-1}_{F_\tau}(1-\alpha) \\ &= \hat{\theta} - (\hat{G}^{-1}_{F_\tau}(1-\alpha) - L_3). \end{aligned}$$

Simplification yields the bootstrap tilting equation (10)!

In other words, bootstrap tilting inference is equivalent to using $T = T(\hat{F}, F) = \hat{\theta} - \theta$ as a pivotal statistic, calibrated by estimating the distribution of $T$ using bootstrap tilting calibration (BTC) at the endpoint. This improves the accuracy from first order, with one-sided coverage errors of $O(n^{-1/2})$, to second order.

We propose applying BTC to the bootstrap-$t$ interval (13) to create the "bootstrap-tilting-$t$" interval; [33] use a similar procedure in a parametric context. The calibrated version would solve

$$L_4 = \theta(F_\tau) = \hat{\theta} - s(F_\tau)\hat{J}^{-1}_{F_\tau}(1-\alpha).$$

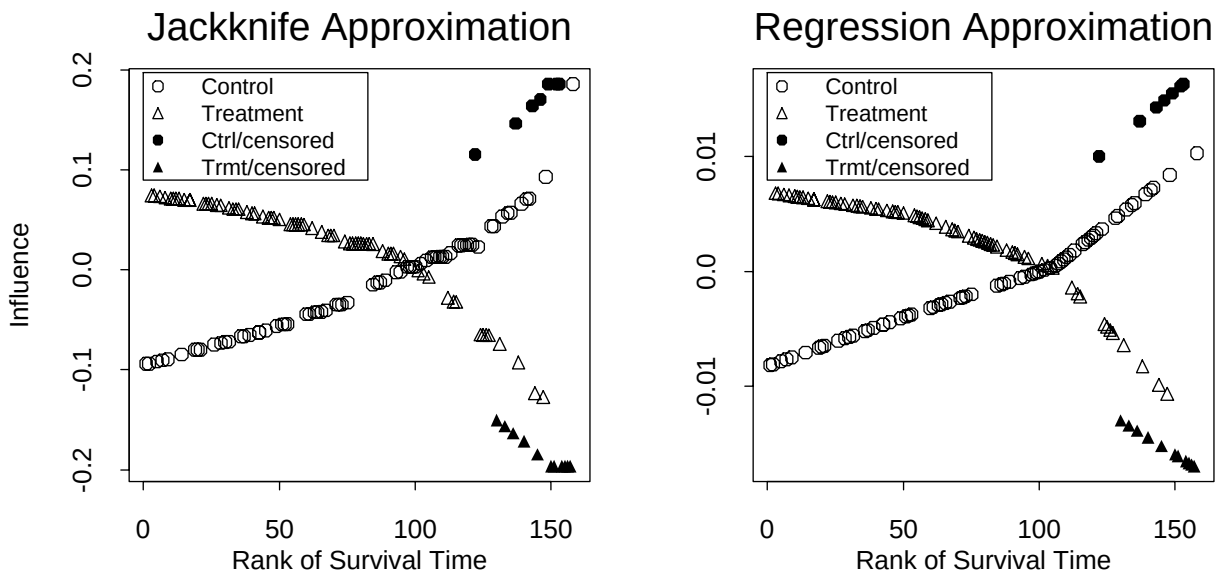


Figure 4: Approximations to influence function values, based on the positive jackknife (left panel) and a linear regression with low degrees of freedom.

The bootstrap-$t$ interval is already second-order correct under general circumstances (e.g. [20]); the calibrated version might be third-order correct. Even if it is not, it should result in a more accurate finite-sample procedure, in two ways. First, it should reduce the finite-sample inaccuracy that results from the lack of transformation-invariance of the bootstrap-$t$. Second, its denominator (standard error) would not be too small when its numerator is large. In this case the numerator is $\hat{\theta}^* - \theta(F_\tau)$ (rather than $\hat{\theta}^* - \hat{\theta}$), and a "large" numerator means that $\hat{\theta}^*$ is far from $\theta(F_\tau)$ but *close* to $\hat{\theta}$ (close on the side of $\hat{J}_{F_\tau}$ that determines the critical value used for confidence intervals), so denominators are not deflated.

In summary, bootstrap tilting offers ways to overcome two known problems with bootstrap-$t$ intervals—lack of transformation invariance, and too-long intervals—in a computationally efficient way. Furthermore, combining tilting and bootstrap-$t$ ideas yields a fast new bootstrap-tilt-$t$ interval which may be more accurate in finite samples than other available intervals.

### 2.3 Large-sample linear approximations

In this section we consider two issues that arise in practice when bootstrapping large data sets—estimating the values of $U_i$ (5) cheaply, and obtaining standard errors for use in bootstrap-$t$ intervals.

We are interested in methods that users may apply without doing analytical calculations, such as those indicated in (5). The derivatives can be approximated using finite differences, such as jackknife and positive jackknife [12] and butcher knife [26] approximations. The positive jackknife approximations are shown in the left panel of Figure 4, for the treatment coefficient for the head and neck data. Calculating these required an additional $n = 158$ function evaluations, in addition to the $B$ evaluations required for the bootstrap. This is expensive for large $n$ and complex $\theta$.

An alternate approximation to (5) was proposed by [14], involving linear regression of the form

$$\hat{\theta}_b^* = \sum_{j=1}^{n} \beta_j p_{b,j}^* + \epsilon.$$

[26] found improved performance by replacing $\hat{\theta}_b^*$ with $\psi(\hat{\theta}_b^*)$ for a linearizing transformation $\psi$, estimated from the data. These methods do not require $n$ extra function evaluations, but do use linear regression with $n$ coefficients, which requires $B$ to be very large for accurate estimation if $n$ is large.

We propose here a regression method on fewer degrees of freedom. Let $h$ be a "design transformation," such that $h(x_j)$ is a $p$-dimensional vector with $p \ll n$, and let $\overline{h}^*_b = n^{-1}\sum_{i=1}^n h(x^*_{b,j}) = \sum_{i=1}^n p^*_{b,j} h(x_j)$ be the vector containing the average of the design transformations for all observations in a resample $b$. A regression of the form

$$\hat{\theta}^*_b = \sum_{j=1}^p \beta_j \overline{h}^*_{b,j} + \epsilon_b$$

yields regression coefficients. Optionally, $\hat{\theta}^*$ could be replaced with $\psi(\hat{\theta}^*)$ in the regression. Let $L_i = \sum_{j=1}^p \beta_j h(x_i)_j$, then $(L_1, \ldots, L_n)$ approximates $(U_1, \ldots, U_n)$, modulo a linear transformation. An example for the head and neck data, based on a linear regression with $p = 12$ terms (11 degrees of freedom), is shown in the right panel of Figure 4.

The design transformation should be chosen so that $\hat{\theta}^* \doteq \sum_{j=1}^p \beta_j \overline{h}^*_j$, for some unknown coefficients $\beta_j$. It should include an intercept, dummy variables (for discrete components of $x_j$), continuous variables and/or polynomial, b-spline, or other nonlinear transformations of the continuous variables, and possibly interaction terms. In this example we split the data into four groups based on treatment and censoring status, used separate intercepts for each group, used separate slopes for the two censored groups, and used linear b-splines with two interior knots for the two non-censored groups, for 12 total degrees of freedom. The result is a slightly less accurate—the correlation between $\hat{\theta}^*$ and the regression approximation $\sum_{j=1}^n \hat{L}_j p^*_j$ is 0.989, while it is 0.993 for the jackknife linear approximation $\sum_{j=1}^n \hat{U}_j p^*_j$—but saves 158 function evaluations. Choosing the design transformation is an art. It might be (partially) automated using stepwise regression or multivariate adaptive regression splines [17]. Diagnostics to guide analysts would be helpful.

An alternative procedure, based on clustering the data and regression against the cluster proportions, did not work as well. The estimates of $U_i$ are constant within each cluster, whereas the linear regression procedure allows for linear (or quadratic, etc.) relationships within clusters.

### 2.3.1 Standard errors for the bootstrap-$t$

The bootstrap-$t$ procedure requires an estimate $s(\hat{F})$ of the standard error of $\hat{\theta}$, and the bootstrap analog $s(\hat{F}^*)$. Where no easier estimate is available, a standard estimate is

$$s(\hat{F}) = \sqrt{n^{-2}\sum_i U_i^2(\mathbf{p}_0)}, \tag{16}$$

with bootstrap analog

$$s(\hat{F}^*) = \sqrt{n^{-2}\sum_i U_i^2(\mathbf{p}^*)}. \tag{17}$$

When $U_i$ is approximated by finite-difference methods such as the jackknife, this requires $n$ additional function evaluations for the original sample, and $nB$ total additional evaluations for the $B$ resamples. This is very expensive for large $n$ and $B$ and complex $\theta$.

We propose to eliminate the $nB$ additional samples required to calculate all of the $U_i(\mathbf{p}^*_b)$ by re-using the values of $U_i$ from the first-level sample. Consider the linear approximation

$$\theta(\mathbf{p}) \doteq \theta(\mathbf{p}_0) + \sum_i U_i(\mathbf{p}_0) p_i. \tag{18}$$

Suppose that the approximation is accurate for both first and second-level resamples, i.e. if either $\mathbf{p}^*$ or $\mathbf{p}^{**}$ is substituted for $\mathbf{p}$. Using the known covariance structure of $\mathbf{p}^*$, this approximation leads to (16) (except for a factor of $n/(n-1)$), and also yields an approximation to (17)

$$\hat{s}(\hat{F}^*) = \sqrt{n^{-2}\sum_i M^*_i (U_i(\mathbf{p}_0) - \overline{U}^*)^2} \tag{19}$$

where $M_i^*$ is the number of times $x_i$ is included in $\mathcal{X}^*$ and $\overline{U}^* = \sum p_i^* U_i(\mathbf{p}_0)$. In contrast, a variation of (18) with $\mathbf{p}_0$ replaced with $\mathbf{p}^*$ yields (17).

Note that the use of (19) requires the $n$ additional function evaluates for evaluating $U_i(\mathbf{p}_0)$, but not the additional $nB$ evaluations required for evaluating each $U_i(\mathbf{p}_b^*)$. This would give major computational savings, making the bootstrap-$t$ interval more practical.

However, this procedure may work poorly where (18) is inaccurate, particularly for second-level resamples. A remedy is to work $\psi(\theta)$ rather than $\theta$ directly, where $\psi$ is a *linearizing* transformation such that

$$\psi(\theta(\mathbf{p})) \doteq \psi(\theta(\mathbf{p}_0)) + \sum_i U_i'(\mathbf{p}_0) p_i$$

where $U'$ is like (5) but evaluated for $\psi(\theta)$ rather than $\theta$. [26] obtains linearizing transformations based on a scatterplot smooth of $\hat{\theta}_b^*$ against $\sum_i U_i(\mathbf{p}_0) p_{b,i}^*$. This does not require additional functional evaluations.

In summary, we propose two methods for reducing the computational cost of bootstrap tilting and bootstrap-$t$ intervals, for problems with large $n$ and complex statistics for which analytical derivatives are unavailable.

### 2.4 S-Plus and bootstrap software

The software that would be developed under this proposal would become part of S-Plus, is an extremely powerful and flexible data analysis environment, built on the S language originally developed at Bell Labs [5], which includes some 2000 built-in functions covering exploratory data analysis, data management, high-level programming, etc.

S-Plus is extensible, using functions written in the S-Plus object-oriented language, C and Fortran. There is an enthusiastic user community; users have posted 245 packages to statlib (see http://lib.stat.cmu.edu/S), most containing multiple functions. Many new statistical procedures are made available for general use in this way.

The design of S-Plus is uniquely suitable for bootstrapping. S-Plus is a high-level programming environment, not just a statistical package. Efron, inventor of the bootstrap, noted [15] that "my bootstrapping has increased considerably with the switch to S, a modern interactive computing language. ... My guess is that the bootstrap (and other computer-intensive methods) will really come into its own only as more statisticians are freed from the constraints of batch-mentality processing systems like SAS." [35] adds "The S language may continue to provide the simplest bootstrap programming in the future."

That prediction has come true. S-Plus now includes an easy-to-use bootstrapping function; the first two lines here perform bootstrap sampling, save the results in an object "BootstrapResults," and create a histogram of the bootstrap distribution with overlaid density curve:

```
BootstrapResults <- bootstrap(lung.survival, mean, args.stat=list(trim=.25))
plot( BootstrapResults, main="Trimmed mean survival time")
summary( BootstrapResults )
```

The `summary` command prints a number of results including the standard deviation and percentiles of the bootstrap distribution, and the bootstrap BC-$a$ [13] confidence intervals. This `bootstrap` function can be used with virtually any statistic, including those defined by users, and is accessible through a graphical user interface.

In addition, both [16] and [6] use S-Plus in their books, and provide sets of bootstrap functions written in S-Plus to accompany their books.

There are no competitors who can provide the nearly the same level of bootstrap capability. The design of most packages effectively precludes this.

Further bootstrap software is being developed at MathSoft, under an NIH SBIR 2R44CA67734-02 project "Statistical Software for Resampling Methods". This software will provide a greater variety of bootstrap methods, variance reduction techniques to make the bootstrapping faster, training materials, etc., with a particular emphasis on biostatistical applications. That project

includes the implementation of the bootstrap tilting confidence interval as a post-processing step on the output of the `bootstrap` function, using a certain one-step approximation to $\mathcal{F}_4$, implemented by importance sampling using design distribution $\hat{F}$ (18 lines in Section 4.1.2 of that proposal). That work will be performed prior to work under this proposal, and is not included in the Specific Objectives or Research Plan below.

That project, and the software that accompanies [6], provide a sound foundation for the current proposal: (1) Some variance reduction techniques already developed could be used with the new methods to further reduce the necessary number of bootstrap replications even below 60. The notable exception is importance sampling, because ISR for tilting and bootstrap-tilting-$t$ is inherently more efficient than the importance sampling that can be done for other bootstrap methods. (2) Some of the new work can use data structures already developed. (3) The new general-purpose function could use the tilting function mentioned in the previous paragraph as a model. (4) Material on the new methods could be added to existing documentation and training materials.

We propose to go beyond these foundations, in ways we describe next.

## 3 Phase I Research Objectives

Most of this proposal deals with new or forgotten bootstrap inference methods. There is always resistance to new methods, particularly if they require time and energy to learn and use. In order for the proposed work to be a technical and commercial success,

- the new methods must be substantially better than other available alternative, in terms of speed and/or accuracy,
- the methods must be available in easy-to-use software, preferably provided automatically,
- the wide statistical community must learn about and accept the methods.

In Phase I we plan to address the first point, demonstrating technical feasibility by showing that the new methods are better, and to begin to address the second and third points, will be further addressed in Phase II. These lay the groundwork for the last point.

**Specific Objectives for Phase I** We offer the following specific objectives:

- Perform initial simulation studies to compare the four tilting families (8) with respect to statistical accuracy and speed. Investigate implementation methods, including different design distributions for importance sampling, nonlinear transformations of the tilting parameter $\tau$ to make the numerical solution of (9) better conditioned, and hybrids of the exponential and ML tilting families, to try to capture the speed of exponential tilting and the accuracy of ML tilting.
- Perform initial studies to investigate variance-stabilizing transformations by tilting and importance sampling reweighting, comparing exponential and ML tilting, and importance sampling design distributions. Investigate using the difference between exponential and ML tilting estimates of variance given $\theta$ in order to adjust the denominator of bootstrap-$t$ statistics.
- Perform initial studies for the bootstrap-tilting-$t$ confidence interval and hypothesis method.
- Perform initial studies to investigate the accuracy of large-sample linear approximations.
- Investigate approximating standard errors for bootstrap samples using the influence values from the original sample, with and without linearizing transformations.
- Summarize the results of the above investigations in one or more technical reports.
- Perform a final simulation study, carefully comparing the methods found to be best in the initial studies to extant methods, including normal-based inferences, bootstrap BC-$a$ methods, bootstrap-$t$, empirical likelihood, and bootstrap ABC methods. Test problems would include the sample mean, median and other robust alternatives, least-squares and robust linear regression, generalized linear and generalized additive models, and correlation. Prepare a technical report, and a report for submission to a peer-reviewed statistical journal. This report would focus on statistical accuracy and give results for computational efficiency, but not describe implementation methods in detail.

- Prepare software for use by beta-testers that implements the best methods. This software would likely be in the form of one general-purpose S-Plus function that implements the methods for arbitrary statistics, and one example with the methods built into an existing function such as the `t.test` function for inference for a mean.
- Obtain feedback from at least 15 beta testers and statistical researchers, on their experience with the beta software and impressions of the reports.
- Present results at one or more conferences.

This list is ambitious. However, the bulk of the work would be done in S-Plus, an efficient language for prototyping new ideas. The lowest priorities are improvements on the usual (non-tilting) bootstrap-$t$ (variance stabilization, denominator adjustment, and approximate standard errors), because the bootstrap-$t$ does not offer the inherent computational advantage of tilting.

**Phase II** Further work, to be performed in Phase II, includes:

- Implement the best methods more carefully in a general-purpose S-Plus function. This coding would be done primarily in the S-Plus language, to provide flexibility both in terms of what statistics may be handled, and to allow the researchers to experiment with and improve our methods.
- Build the best methods into functions for specific purposes, such as functions that perform $t$-tests, linear and nonlinear regression, and robust location and regression methods, for seamless use by practitioners who already use these functions; by default the new inferences would be provided in addition to existing inferences, with warnings whenever the new inferences indicate that the other inferences may be inaccurate.

  These implementations will take advantage of analytic influence functions evaluations and other tricks to make these special-purpose implementations faster. Much of this coding would be done in C or Fortran for speed.
- Add the best methods to selected statistical functions in MathSoft's MathCad and Axum products.
- Modify some existing functions in S-Plus to allow user-supplied weights, which is necessary for the tilting methods to work.
- Investigate ways to choose the design transformation for large-sample linear approximations automatically, or let the user specify it using the formula language in S-Plus.
- Incorporate the software within a graphical user interface.
- Prepare documentation aimed at the general statistical audience.
- Develop extensions to handle stratified-sample problems, including two-sample tests. This should be relatively straightforward, making use of multi-sample influence functions or approximations.
- Develop extensions to multiple-parameter problems such as analysis of variance for categorical explanatory variables and $\chi^2$ tests of independence in contingency tables. This may not be straightforward. It should be possible to solve the multi-parameter analog of (9), by letting $\tau$ have dimension equal to the number of parameters to be tested. However, there is not a simple analog to the $p$-value (7) in multi-parameter problems. Using empirical likelihood to determine shapes of regions may provide an answer.
- Extensions to handle the smoothed bootstrap. This should be straightforward, at least in simple problems. Exponential tilting may be more accurate than ML tilting when combined with smoothing.
- Extensions to handle time series and other dependent data. This is not straightforward.
- Extensions to parametric problems.

## 4 Phase I Research Plan

The work would be carried out by Dr. Hesterberg and a programmer. The anticipated timetable for this work is shown in Table 2.

Table 2: Time line for Phase I work

| Task | Month | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Tilting | H | PH | | | | |
| Boot-$t$ improvements | | H | PH | | | |
| Boot-tilt-$t$ | H | HP | PH | | | |
| SE for boot-$t$ | | H | PH | | | |
| Large-sample approx | | | | | H | PH |
| Study best candidates | | | PH | PH | | |
| Write reports | | | | H | H | |
| Beta software | | | HP | PH | PH | |
| Feedback | | | | | H | H |

"H" denote Hesterberg and "P" denotes programmer, with the person spending most time listed first.

Initial investigations will be begun by Hesterberg, later with the assistance of the program. The programmer has the primary responsibility for the large-scale simulations of the best candidates from the initial investigations.

Investigation of large-sample approximations occurs late because that work is not a prerequisite for other work. This work may be moved up so the methods can be included in the reports and beta software.

## 5 Commercial Potential

### 5.1 Mission and Main Products

MathSoft Data Analysis Products Division's primary mission is to develop, market, and support cutting edge scientific computing software environments for high-interaction graphical analysis of multivariate data, modern statistical methods (e.g., robust and nonparametric methods), data clustering and classification and mathematical computing.

One of MathSoft's main products is the S-Plus interactive computing environment for graphics, data analysis, statistics and mathematical computing. S-Plus is a super-set of the S object-oriented language and system developed at AT&T Bell Laboratories [1]. MathSoft's customer base represents almost every major industry, with particular strength in high-tech manufacturing, biotechnology, engineering and finance. S-Plus is available in both UNIX and Windows versions.

While S-Plus has traditionally held the higher end of the statistical market, MathSoft is reaching out the a broader market, with a new easy-to-use graphical user interface (GUI), broader marketing, the creation of lower-cost "student" and "standard" versions, and other initiatives. There are currently about 20,000 users for S-Plus, and this number is growing rapidly. MathSoft is also adding statistical capabilities to its other main products—MathCad, a mathematical analysis package with 1,100,000 licences sold, and Axum, a technical graphics package, with 25,000 users.

The company has well-established teams for software development, quality assurance, marketing, sales, and teaching short courses.

### 5.2 Commercialization of Technology

MathSoft has an outstanding record in the commercialization of advanced data analysis technology. Our core product, S-Plus, is a commercial version of the S language developed in the research environment of AT&T Bell Laboratories. In fact, MathSoft DAPD would not exist if it not for our ability to commercialize data analysis software. MathSoft has an established record of commercializing advanced data analysis software developed partially using Government funds under the SBIR program and the NASA EOCAP program. Partially supported by these awards, MathSoft has commercialized and shipped six products: S+Wavelets, S+DOX, S-Plus for ARC/INFO, S+ArcView GIS, S+SpatialStats, S+Garch, and S+SDK, and incorporated other capabilities into the core S-Plus product. New methods developed here would be included in the core

product.

### 5.3 Commercialization of fast bootstrap inference methods

The new inference methods would be deployed initially in S-Plus, and subsequently in Axum and MathCad.

A key part of our strategy is to include new methods within common standard functions, such as those to compute $t$-tests and linear regression, and to provide warnings when the new methods indicate that the normal-based inferences may be inaccurate. The intent is to give the new methods credibility, and to discredit normal-based inferences. While the inaccuracy of normal-based methods is well known, they are commonly used in the absence of realistic alternatives. By providing the side-by-side alternative, we hope to make software users reluctant to just use normal-based inferences.

We plan to spread the news about the new capabilities through a combination of word of mouth, journal articles initially aimed at the statistical research community and subsequently at the wider population of practicing statisticians, presentations by MathSoft employees and consultants at conferences, courses offered by MathSoft, marketing and sales efforts by MathSoft, other publicity—e.g. press from a high-profile court case or FDA decision, if the results from inaccurate normal-based methods and newer methods disagree, particularly if they fall on either side of the magic 5% level, and outreach to the statistical education community. Hesterberg is a former teacher and maintains close ties to leading statistical educators. These efforts should gradually increase demand and result in new sales.

The new capabilities would also be added to some less-common but important functions, such as those providing robust alternatives to the sample mean and least-squares regression, promoting the use of these functions, and generating sales among people who wish to use these alternatives but have not because of the lack of easy inferences. Some statistical educators are particularly eager for this, which would expose students to S-Plus and increase future demand.

If the proposed research meets expectations, the new capabilities would be adequate justification for new releases of S-Plus, Axum, and MathCad, with a major marketing push, worth millions of dollars to MathSoft.

### 5.4 Commercialization of large sample methods

The market for statistical analysis for large data sets ("data mining") is large and growing, estimated at $8 billion per year and growing by 40% per year by META Group, an industry research firm. S-Plus is currently a leading player in this market, offering attractive methods such as classification and regression trees, clustering, factor analysis, Trellis graphics, linear, nonlinear and logistic regression, and predictive models. The methods proposed here for making bootstrapping feasible with larger data sets would allow allow analysts to not only obtain estimates, but obtain confidence intervals to indicate how accurate those estimates are. This capability would let MathSoft to increase its penetration in this market.

## 6 References


[1] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The New S Language*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1988.

[2] R. J. Beckman and M. D. McKay. Monte Carlo estimation under different distributions using the same simulation. *Technometrics*, 29:153–160, 1987.

[3] R. Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83:687–697, 1988.

[4] D. D. Boos, P. Janssen, and N. Veraverbeke. Resampling from centered data in the two sample problem. *J. Statist. Plan. Inferencee*, 21:327–345, 1989.

[5] J.M. Chambers and T.J. Hastie. *Statistical Models in S.* Wadsworth, California, 1992.

[6] A. Davison and D. Hinkley. *Bootstrap Methods and their Applications.* Cambridge University Press, 1997.

[7] A. C. Davison, D. V. Hinkley, and B. J. Worton. Bootstrap likelihoods. *Biometrika*, 79(1):113–130, 1992.

[8] T. J. DiCiccio and J. P. Romano. The automatic percentile method: accurate confidence limits in parametric models. *The Canadian Journal of Statistics*, 17(2):155–169, 1989.

[9] T. J. DiCiccio and J. P. Romano. Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review*, 58(1):59–76, 1990.

[10] B. Efron. Computer intensive methods in statistics. In *200th Anniversary Volume*. Lisbon Academy of Sciences, 1981.

[11] B. Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9:139 – 172, 1981.

[12] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. National Science Foundation – Conference Board of the Mathematical Sciences Monograph 38. Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[13] B. Efron. Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, 82:171 – 200, 1987.

[14] B. Efron. More efficient bootstrap computations. *Journal of the American Statistical Association*, 85(409):79 – 89, 1990.

[15] B. Efron. discussion of "Bootstrap: More than a stab in the dark?". *Statistical Science*, 9:396–398, 1994.

[16] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[17] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–67, 1991.

[18] R. L. Graham, D. V. Hinkley, P. W. M. John, and S. Shi. Balanced design of bootstrap simulations. *Journal of the Royal Statistical Society, Series B*, 52:185–202, 1990.

[19] P. Hall. On the bootstrap and confidence intervals. *Annals of Statistics*, 14(4):1431–1452, 1986.

[20] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1992.

[21] P. Hall and B. La Scala. Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58(2):109–127, 1990.

[22] Peter Hall and Michael A. Martin. On bootstrap resampling and iteration. *Biometrika*, 75(4):661–671, 1988.

[23] J. M. Hammersley and D. C. Hanscomb. *Monte Carlo Methods*. Methuen, London, 1964.

[24] Tim C. Hesterberg. *Advances in Importance Sampling*. PhD thesis, Statistics Department, Stanford University, 1988.

[25] Tim C. Hesterberg. Tail-specific linear approximations for efficient bootstrap simulations. *Journal of Computational and Graphical Statistics*, 4(2):113–133, June 1995.

[26] Tim C. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[27] Tim C. Hesterberg. Estimates and confidence intervals for importance sampling sensitivity analysis. *Mathematical and Computer Modeling*, 23(8/9):79–86, 1996.

[28] Tim C. Hesterberg. The bootstrap and empirical likelihood. In *Proceedings of the Statistical Computing Section*. American Statistical Association, 1997. in press, available at http://www.statsci.com/Hesterberg/index.html.

[29] D. V. Hinkley. Bootstrap significance tests. *Bulletin of the Intenational Statistical Institute*, pages 65–74, 1989.

[30] Paul Kabaila. Some properties of profile bootstrap confidence intervals. *Australian Journal of Statistics*, 35(2):205–214, 1993.

[31] W. Loh. Calibrating confidence coefficients. *Journal of the American Statistical Association*, 82:155–162, 1987.

[32] M. P. Meredith and J. G. Morel. discussion of "Bootstrap: More than a stab in the dark?". *Statistical Science*, 9:404–406, 1994.

[33] Art Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.

[34] Art Owen. Empirical likelihood confidence regions. *Annals of Statistics*, 18:90–120, 1990.

[35] M. I. Reiman and A. Weiss. Sensitivity analysis via likelihood ratios. In *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289, 1986.

[36] J. P. Romano. A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708, 1988.

[37] Joseph P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, 17:141–159, 1989.

[38] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995.

[39] R. Tibshirani. Variance stabilization and the bootstrap. *Biometrika*, 75:433 – 444, 1988.

[40] J. W. Tukey. Configural polysampling. *SIAM REVIEW*, 29:1–20, 1987.

[41] A. T. A. Wood, K. A. Do, and B. M. Broom. Sequential linearization of empirical likelihood constraints with application to u- statistics. *Journal of Computational and Graphical Statistics*, 5(4):365–385, 1996.

[42] G. A. Young. Resampling tests of statistical hypotheses. In D. Edwards and N. E. Raun, editors, *Compstat: Proceedings in Computational Statistics*, pages 233–238, Heidelberg, 1988. Physica-Verlag.