



Research Report
No. 86

Linear Approximations for Functional Statistics in Large-Sample Applications

Tim C. Hesterberg and Stephen J. Ellis

Revision 1, Revision Date: October 14, 1999

Acknowledgments: This work was supported by NSF Phase I SBIR Award No. DMI-9861360.

MathSoft, Inc.
1700 Westlake Ave. N, Suite 500
Seattle, WA 98109-9891, USA
Tel: (206) 283-8802
FAX: (206) 283-6310

E-mail: timh@statsci.com
sellis@statsci.com
Web: www.statsci.com/Hesterberg/tilting

Linear Approximations for Functional Statistics in Large-Sample Applications

T. C. Hesterberg and S. J. Ellis

Abstract

We discuss methods for obtaining linear approximations to a functional statistic, with particular application to bootstrapping medium to large datasets. Existing methods use analytical approximations, finite-difference derivatives, or linear regression using bootstrap results. Finite-difference methods require an additional n evaluations of a functional statistic (where n is the number of observations in the data set), and regression methods require that the number of bootstrap samples B is substantially larger than n . We develop regression-type methods that allow B to be much smaller, and that require no dedicated bootstrap samples. The method uses a prespecified or adaptively chosen design matrix.

Key Words: Bootstrap tilting, concomitants of order statistics, importance sampling, jackknife, stratified sampling, variance reduction.

1 Introduction

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 Efron and Tibshirani (1993). The original data are $\mathcal{X} = (X_1, X_2, \dots, X_n)$, a sample from an unknown distribution (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 used 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 \bar{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 with an estimate \hat{F} . We focus on the nonparametric bootstrap, for which \hat{F} is the empirical distribution. Let $\mathcal{X}^* = (X_1^*, X_2^*, \dots, X_n^*)$ be 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 bootstrap samples, sample \mathcal{X}_b^* for $b = 1, \dots, B$ with replacement from \mathcal{X} , then let

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

The focus of this report is on computationally-efficient methods for obtaining (generalized) linear approximations for functional statistics. Such approximations are used for a number of applications — standard errors, the acceleration constant for the bootstrap BC-a interval (Efron (1987)), importance sampling in bootstrap applications (Johns (1988); Davison and Hinkley (1988)), concomitants of order statistics for bootstrap variance reduction (Efron (1990); Do and Hall (1992)), control variates and post-stratification (Hesterberg (1995); Hesterberg (1996)), bootstrap tilting inferences (Efron (1981); DiCiccio and Romano (1990); Hesterberg (1997); Hesterberg (1998)), and bootstrap tilting diagnostics (Hesterberg (1997); Hesterberg (1998)).

A “generalized linear approximation” to $\hat{\theta}^*$ is determined by a vector \mathbf{L} of length n , with elements L_j corresponding to each of the original observations x_j , such that

$$\psi(\theta(\mathcal{X}^*)) \doteq L^* \equiv \sum_{j=1}^n L_j P_j^* \quad (3)$$

for some smooth monotone increasing function ψ , where $P_j^* = M_j/n$ and M_j is the number of times x_j is included in \mathcal{X}^* . The special case where $\psi(T^*) = T^* - T(\mathcal{X})$ is a standard linear approximation.

For example, Figure 1 shows a generalized linear approximation for bootstrapping the sample standard deviation $(n^{-1} \sum (x_i^* - \bar{x}^*)^2)^{1/2}$. (The divisor is n rather than $n-1$ so that the statistic is functional.) The curvature could be removed in this case by the transformation $\psi(\theta) = \theta^2$.

In Section 2 we discuss “knife” methods — the jackknife and related methods — for obtaining linear approximations. In Section 3 we discuss regression methods, including the new “design-based” regression method in Section 3.1.

2 Knife Methods

In this section we review a number of methods based on functional derivatives.

We restrict consideration to distributions with support on the observed data. Then we may describe a distribution in terms of the probabilities $\mathbf{p} = (p_1, \dots, p_n)$ assigned to the original observations; \hat{F} corresponds to $\mathbf{p}_0 = (1/n, \dots, 1/n)$. Let $\theta(\mathbf{p})$ be the corresponding parameter estimate (which depends implicitly on \mathcal{X}).

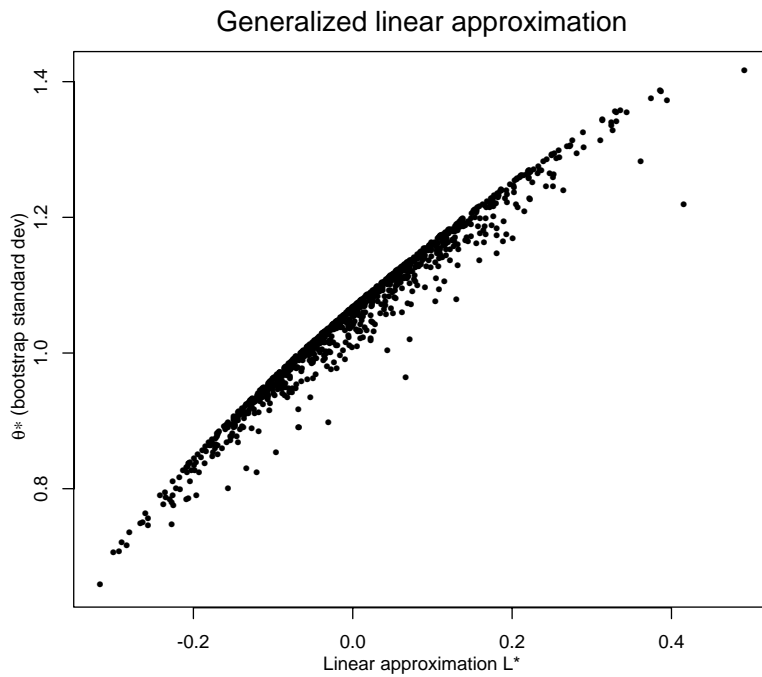


Figure 1: Generalized Linear Approximation for bootstrapping the sample standard deviation. The data are a random sample of size $n = 40$ from a standard normal distribution, $B = 1000$, and the linear approximation is obtained by the infinitesimal jackknife (empirical influence function).

The “knife” approximations in this section are of the form

$$L_j = \frac{\theta(\mathbf{P}_0 + \epsilon(\delta_j - \mathbf{P}_c)) - \theta(\mathbf{P}_0)}{\epsilon} \quad (4)$$

for some ϵ . These approximations are Taylor-series or finite-difference approximations to the gradient of the function $\theta(\mathbf{P})$. Four choices of ϵ are noteworthy:

$$\begin{aligned} \text{negative jackknife : } & \epsilon = -1/(n - 1) \\ \text{influence function : } & \epsilon \rightarrow 0 \\ \text{positive jackknife : } & \epsilon = 1/(n + 1) \\ \text{butcher knife : } & \epsilon = n^{-1/2} \end{aligned} \quad (5)$$

The first three are the negative jackknife, influence function (or infinitesimal jackknife), and positive jackknife approximations of (Efron (1982)), the fourth is the butcher knife of (Hesterberg (1995)).

The infinitesimal jackknife (influence function) requires analytical calculations, or numerical approximation by a small value of ϵ . Using a numerical approximation, or using any of the other methods, requires an additional n function evaluations. It is this expense that the new methods described below are intended to avoid.

The two jackknives can be calculated using software that does not explicitly support weights, by deleting each observation in turn, or repeating an observation twice. The butcher knife can also be approximated in this manner, by repeating an each observation in turn k times, with $k = \text{round}(1 + \frac{\sqrt{n}}{1-1/\sqrt{n}})$, this corresponds to $\epsilon = (k - 1)/(n + k)$.

The butcher knife can be used for some non-smooth statistics such as the median for which the other methods fail.

3 Regression Methods

We turn now to regression methods, which may be used to obtain linear approximations for any statistic, even one not defined for weighted samples. They also do not require n extra function evaluations; however, depending on the method, they may require that B be substantially larger than it would otherwise be.

Regression methods utilize existing bootstrap samples to obtain linear approximations. Let $M_{b,j}$ be the number of times original observation x_j is included in the b 'th bootstrap sample and let $P_{b,j}^* = M_{b,j}/n$. A linear regression without an intercept of the form

$$\hat{\theta}_b^* = \sum_{j=1}^n \hat{\beta}_j P_{b,j}^* + \text{residual}_b, \quad (6)$$

yields coefficients which are centered to obtain the linear approximation

$$\mathbf{L}_j = \hat{\beta}_j - \bar{\beta} \quad (7)$$

where $\bar{\beta} = (1/n) \sum_{i=1}^n \hat{\beta}_i$. The intercept must be omitted because otherwise the regression would be singular, because $\sum_{j=1}^n P_j^* = 1$. This linear approximation was obtained by (Efron (1990)).

Hesterberg (1995) generalizes this procedure by obtaining the regression approximation as above, calculating the corresponding linear approximation L^* (right side of 3), smoothing L^* (as the response variable) against $\hat{\theta}^*$ to estimate a smooth nonlinear transformation $\hat{\psi}$, and then performing another regression using $\hat{\psi}(\hat{\theta}^*)$ in place of $\hat{\theta}^*$:

$$\hat{\psi}(\hat{\theta}_b^*) = \sum_{j=1}^n \hat{\beta}_j P_{b,j}^* + \text{residual}_b, \quad (8)$$

The procedure is motivated by the ACE algorithm (Breiman and Friedman (1985)). This gives more accurate coefficients in some problems — using the linear transformation reduces the residual standard deviation, and hence provides linear regression coefficients with smaller variance for a given sample size B .

If the bootstrap samples were obtained using importance sampling, then (6) and (7) are replaced by weighted regressions.

Both the regression and ACE procedures utilize B observations to estimate n regression coefficients. To do this accurately requires that B be substantially larger than n . This makes these procedures impractical in many situations, involving large or even moderate samples. For example, B could be as small as 60 when using bootstrap tilting to obtain confidence intervals (Hesterberg (1997)).

3.1 Regression against a design matrix

We describe in this section a procedure using regression on fewer degrees of freedom.

To motivate the procedure, consider the case where there are duplicate values among the original n data points, e.g. if the underlying distribution is discrete. Then the corresponding values of L_j should also be duplicated, and fewer than n unique regression coefficients would be needed. Or, suppose that observations are not exactly duplicated, but are similar; then the corresponding regression coefficients should be similar; this knowledge could be used to reduce the Monte Carlo variability in those observations.

We implement those thoughts using a “design-based” method for obtaining linear approximations. Let h be a “design transformation,” such that $h(x_i)$ is a p -dimensional vector, usually with $p \ll n$, and let $\bar{h}_b^* = n^{-1} \sum_{i=1}^n h(x_{b,i}^*) = \sum_{i=1}^n P_{b,i}^* h(x_i)$ be the vector containing the average of the design transformations for all observations in a bootstrap sample b . A regression of the form

$$\hat{\theta}_b^* = \sum_{j=1}^p \beta_j \bar{h}_{b,j}^* + \text{residual}_b \quad (9)$$

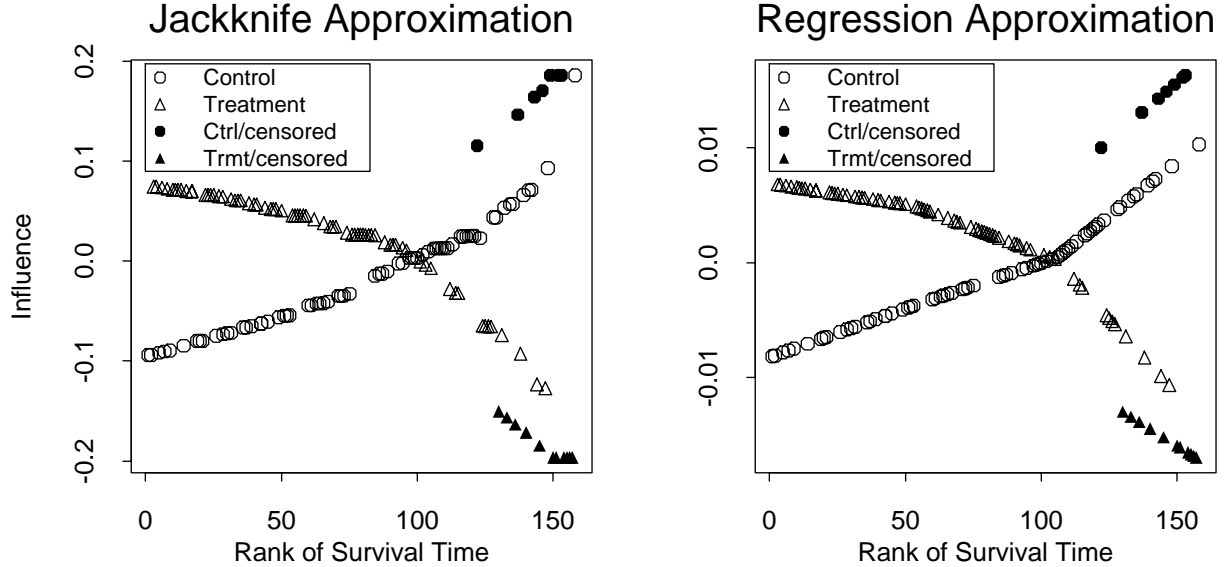


Figure 2: Approximations to influence function values, based on the positive jackknife (left panel) and a linear regression with low degrees of freedom.

yields regression coefficients β_j , $j = 1, \dots, p$. The first element of $h(x_j)$ would typically be identically 1, in which case β_1 is an intercept term. The vector \mathbf{L} is then determined by:

$$L_i = \sum_{j=1}^p \beta_j h(x_i)_j. \quad (10)$$

Note that this vector must in general be linearly transformed before it can be used in a (non-generalized) linear transformation.

Optionally, $\hat{\theta}^*$ can be replaced with $\psi(\hat{\theta}^*)$ in (9), thus combining the ACE algorithm idea with this design matrix method.

An example is shown in Figure 2. 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 θ is the treatment coefficient in a Cox proportional hazards regression model. The left panel for the figure uses the positive jackknife, while the right panel uses a regression against a design transformation with with $p = 12$ terms (including the intercept).

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

3.2 Choice of design matrix

The design transformation should be chosen so that $\hat{\theta}^* \doteq \sum_{j=1}^p \beta_j \bar{h}_j^*$, for some unknown coefficients β_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 $\sum_{j=1}^n \hat{L}_j p_j^*$ —but saves 158 function evaluations. Adding additional terms results in higher correlation — correlation .9923 with $p = 20$ and .9928 with $p = 26$ (the additional terms were added by increasing the number of knots used for b-splines; the knot placements were not optimized).

Choosing the design transformation is an art, similar to that of variable selection in ordinary linear regression. Many of the same techniques can be utilized, such as t - and F -statistics for determining whether the addition of terms results in a substantial reduction in residual variance, and stepwise regression. Techniques borrowed from Multivariate Adaptive Regression Splines (Friedman (1991)) should be particularly suitable.

There is less need to obtain a parsimonious model here than in most linear regression applications, because interpretability of results is not necessary, and because the p coefficients are not used directly, but only indirectly after a linear transformation to the vector \mathbf{L} . As long as B is much larger than p , adding additional terms causes little harm. Simulation results, in Tables 1–5, support the general rule that it is critical to include certain terms (which vary by problem), and that adding additional terms does not hurt. Those simulations are based on the correlation of the linear approximations L^* with $\hat{\psi}(\hat{\theta}^*)$; additional simulations should be done that focus on the variability of the elements of \mathbf{L} .

Our rule of thumb is to require that $B \geq 50 + 3p$, but more work should be done to quantify the effect of different values of B and p ; we suspect that say $B \geq 50 + p$ may be adequate. If $p = n$ and all columns of the design matrix are linearly independent, the procedure gives the same results as the earlier regression procedure.

It should be straightforward to create a “tail-specific” version of the design-based regression procedure, based on the tail-specific regression procedure of (Hesterberg (1995)), but we have not done so.

Summary

The key contribution of this report is the development of a “design-based linear approximation” method for obtaining linear approximations in bootstrap situations cheaply. The procedure does not require additional function evaluations, in contrast to “knife” methods,

Table 1: Average Adjusted R^2 of Transformed Replicates $\hat{\psi}(\hat{\theta}^*)$ with Linear Approximation L^* (Normal Data, Statistic: Two-sample Correlation)

N	B	Linear Approximation Method								
		JACK	REG	ACE	DM-1	DM-2	DM-3	DMA-1	DMA-2	DMA-3
10	50	0.77	0.77	0.78	0.30	0.78		0.26	0.78	
	100	0.74	0.76	0.77	0.28	0.76		0.26	0.77	
	200	0.73	0.75	0.76	0.27	0.76		0.26	0.76	
	400	0.74	0.76	0.76	0.27	0.76		0.26	0.76	
80	100	0.97	0.96	0.96	0.07	0.97	0.97	0.04	0.97	0.97
	200	0.97	0.97	0.97	0.06	0.97	0.97	0.05	0.97	0.97
	400	0.97	0.97	0.97	0.05	0.97	0.97	0.05	0.97	0.97

The methods used are positive jackknife, regression, ACE, design matrix, and design matrix with ACE. The data (\mathbf{x}, \mathbf{y}) are jointly normal with $\rho = 0.5$, with sample size 10 or 80. For each sample size, 100 random data sets are generated; from each data set, four sets of bootstrap samples are generated, with sizes $B = 50, 100, 200$, and 400 (the $n = 80, B = 50$ case was omitted). The linear approximation methods are applied to the bootstrap samples, and corresponding L^* computed. Then the best-fit $\hat{\psi}$ is found for each method, using smoothing splines with 4 degrees of freedom, and the squared correlation (R^2) with L^* is recorded. The R^2 values are adjusted according to degrees of freedom (DF) as $R_a^2 = 1 - (1 - R^2) \cdot (B - 1)/(B - p)$, where p is 1 for jackknife, n for regression, $n + 3$ for ACE (3 is the nonlinear DF of the smoothing), k for design matrix, and $k + 3$ for design matrix with ACE (k is the number of columns of the design matrix, including the intercept). Each of the cells of the table is an average of 100 R_a^2 values.

Each of the design matrices has an intercept term. In addition, the design matrix for DM-1 and DMA-1 has (\mathbf{x}, \mathbf{y}) , for DM-2 and DMA-2 has $(\mathbf{x}, \mathbf{y}, \mathbf{x}^2, \mathbf{xy}, \mathbf{y}^2)$, and for DM-3 and DMA-3 has $(\mathbf{x}, \mathbf{y}, \mathbf{x}^2, \mathbf{xy}, \mathbf{y}^2, \mathbf{x}^3, \mathbf{x}^2\mathbf{y}, \mathbf{xy}^2, \mathbf{y}^3)$. Including the intercept, this last design matrix has 10 columns, and thus gives an identical fit to that of the regression method when $n = 10$. Thus, these redundant results are omitted from the table. Since the correlation coefficient can be written in linear and quadratic terms of \mathbf{x} and \mathbf{y} , *a priori* we expected the second design matrix to give the best fit, which indeed happened. We also expected that the first design matrix would give poor results due to underfitting, and that the third design matrix would not improve on the second but would also not do (much) worse; the results match these expectations.

Table 2: Average Adjusted R^2 of Transformed Replicates $\hat{\psi}(\hat{\theta}^*)$ with Linear Approximation L^* (Normal Data, Statistic: One-sample Variance)

N	B	Linear Approximation Method								
		JACK	REG	ACE	DM-1	DM-2	DM-3	DMA-1	DMA-2	DMA-3
10	50	0.87	0.87	0.86	0.23	0.87	0.87	0.18	0.86	0.86
	100	0.86	0.86	0.85	0.20	0.86	0.86	0.18	0.86	0.86
	200	0.85	0.85	0.85	0.19	0.85	0.85	0.18	0.85	0.85
	400	0.85	0.85	0.85	0.19	0.85	0.85	0.18	0.85	0.85
80	100	0.99	0.99	0.99	0.06	0.99	0.99	0.04	0.99	0.99
	200	0.99	0.99	0.99	0.06	0.99	0.99	0.04	0.99	0.99
	400	0.99	0.99	0.99	0.04	0.99	0.99	0.04	0.99	0.99

The data (\mathbf{x}) are univariate standard normal. In addition to the intercept term, the design matrix for DM-1 and DMA-1 has (\mathbf{x}), for DM-2 and DMA-2 has (\mathbf{x}, \mathbf{x}^2), and for DM-3 and DMA-3 has ($\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3$). Like the correlation coefficient, the sample variance is quadratic, so *a priori* we expected the second design matrix to give the best fit, which indeed happened. Similarly, the first design matrix gives poor results due to underfitting and the third design matrix does not improve on the second. For other details on this simulation, see Table 1.

Table 3: Average Adjusted R^2 of Transformed Replicates $\hat{\psi}(\hat{\theta}^*)$ with Linear Approximation L^* (Exponential Data, Statistic: One-sample Variance)

N	B	Linear Approximation Method								
		JACK	REG	ACE	DM-1	DM-2	DM-3	DMA-1	DMA-2	DMA-3
10	50	0.898	0.894	0.890	0.511	0.899	0.898	0.479	0.893	0.893
	100	0.893	0.892	0.892	0.490	0.894	0.894	0.474	0.892	0.893
	200	0.899	0.898	0.899	0.493	0.899	0.899	0.485	0.899	0.899
	400	0.897	0.896	0.897	0.490	0.897	0.897	0.486	0.897	0.898
80	100	0.996	0.994	0.994	0.561	0.996	0.996	0.547	0.996	0.996
	200	0.996	0.995	0.995	0.550	0.996	0.996	0.543	0.996	0.996
	400	0.995	0.995	0.995	0.554	0.995	0.995	0.551	0.995	0.995

The data are exponential(1). See Table 2 for further details.

Table 4: Average Adjusted R^2 of Transformed Replicates $\hat{\psi}(\hat{\theta}^*)$ with Linear Approximation L^* (Normal Data, Statistic: Ratio of Means)

N	B	Linear Approximation Method						
		JACK	REG	ACE	DM-1	DM-2	DMA-1	DMA-2
10	50	0.9976	0.9958	0.9981	0.9986	0.9971	0.9989	0.9985
	100	0.9974	0.9965	0.9982	0.9985	0.9972	0.9987	0.9984
	200	0.9972	0.9972	0.9983	0.9984	0.9974	0.9986	0.9984
	400	0.9972	0.9974	0.9984	0.9984	0.9976	0.9985	0.9984
80	100	0.9998	0.9980	0.9982	0.9998	0.9997	0.9998	0.9998
	200	0.9998	0.9989	0.9994	0.9998	0.9997	0.9998	0.9998
	400	0.9998	0.9993	0.9997	0.9998	0.9998	0.9998	0.9998

The data (\mathbf{x}, \mathbf{y}) are independently normal with mean vector $(3, 9)$; each has unit variance. In addition to the intercept term, the design matrix for DM-1 and DMA-1 has (\mathbf{x}, \mathbf{y}) and for DM-2 and DMA-2 has $(\mathbf{x}, \mathbf{y}, \mathbf{x}^2, \mathbf{xy}, \mathbf{y}^2)$. We expected the quadratic design matrix not to improve on the first, and that was the result. For other details on this simulation, see Table 1.

Table 5: Average Adjusted R^2 of Transformed Replicates $\hat{\psi}(\hat{\theta}^*)$ with Linear Approximation L^* (Exponential Data, Statistic: Ratio of Means)

N	B	Linear Approximation Method						
		JACK	REG	ACE	DM-1	DM-2	DMA-1	DMA-2
10	50	0.977	0.960	0.980	0.984	0.971	0.988	0.984
	100	0.974	0.966	0.981	0.983	0.972	0.985	0.983
	200	0.971	0.970	0.981	0.982	0.974	0.983	0.982
	400	0.970	0.972	0.980	0.980	0.973	0.982	0.981
80	100	0.998	0.983	0.985	0.998	0.997	0.998	0.998
	200	0.998	0.990	0.995	0.998	0.998	0.998	0.998
	400	0.998	0.994	0.997	0.998	0.998	0.998	0.998

The data (\mathbf{x}, \mathbf{y}) are independent exponentials plus a constant vector $(0, 2)$. See Table 4 for further details.

which require n function evaluations, which is expensive if n is large and/or θ is expensive to compute. It does not require that B be much larger than n . It is suitable for non-smooth functions, such as the sample median, unlike most knife methods. It does not require analytical calculations by the user, and can be implemented in general-purpose bootstrap software.

The new method does require that the user specify a design matrix, or that an automated procedure such as a variation of stepwise regression be used to select the design matrix.

The method produces accurate linear approximations in a variety of test problems.

We have written an S-PLUS function `resamp.get.L` that takes as input a `bootstrap` object and uses any of the methods described above to compute L ; for the design matrix method the user must also supply the design matrix.

Further study is needed to quantify the effect of choosing the design matrix adaptively, to quantify how large B should be in order to obtain desired levels of accuracy, to study the variability of individual elements of \mathbf{L} as a function of degrees of freedom in the design matrix, and to obtain a “tail-specific” version of the method.

References

- Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80:580–619.
- Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint Approximation in Resampling Methods. *Biometrika*, 75:417 – 431.
- DiCiccio, T. J. and Romano, J. P. (1990). Nonparametric Confidence Limits by Resampling methods and Least Favorable Families. *International Statistical Review*, 58(1):59–76.
- Do, K. A. and Hall, P. (1992). Distribution estimation using concomitants of order statistics, with application to Monte Carlo simulations for the bootstrap. *Journal of the Royal Statistical Society, Series B*, 54(2):595–607.
- Efron, B. (1981). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, 9:139 – 172.
- Efron, B. (1982). *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.
- Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, 82:171 – 200.

- Efron, B. (1990). More Efficient Bootstrap Computations. *Journal of the American Statistical Association*, 85(409):79 – 89.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–67.
- Hesterberg, T. C. (1995). Tail-Specific Linear Approximations for Efficient Bootstrap Simulations. *Journal of Computational and Graphical Statistics*, 4(2):113–133.
- Hesterberg, T. C. (1996). Control Variates and Importance Sampling for Efficient Bootstrap Simulations. *Statistics and Computing*, 6(2):147–157.
- Hesterberg, T. C. (1997). The bootstrap and empirical likelihood. In *Proceedings of the Statistical Computing Section*, pages 34–36. American Statistical Association.
- Hesterberg, T. C. (1998). Bootstrap Tilting Inference and Large Datasets. Grant application to N.S.F.
- Johns, M. V. (1988). Importance Sampling for Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 83(403):701–714.