Research Report
No. 84

Bootstrap Tilting Confidence Intervals

Tim Hesterberg

Revision 2, Revision Date: December 28, 1999

Acknowledgments: This work was supported by NSF 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
Web: www.statsci.com/Hesterberg

Bootstrap Tilting Confidence Intervals

Tim C. Hesterberg
December 28, 1999

Abstract

Bootstrap tilting confidence intervals could be the method of choice in many applications for reasons of both speed and accuracy. With the right implementation, tilting intervals are 37 times as fast as bootstrap BC-a limits, in terms of the number of bootstrap samples needed for comparable simulation accuracy. Thus 100 bootstrap samples might suffice instead of 3700.

Tilting limits have other desirable properties — second-order accuracy, transformation invariance, and certain variations of these limits offer better finite-sample coverage and/or shorter intervals on average than competing procedures.

Key Words: bootstrap, empirical likelihood, importance sampling, least favorable family.

Contents

1 Introduction

Bootstrap tilting confidence intervals and hypothesis tests could become the methods of choice in many inference problems for reasons of statistical accuracy and computational efficiency.

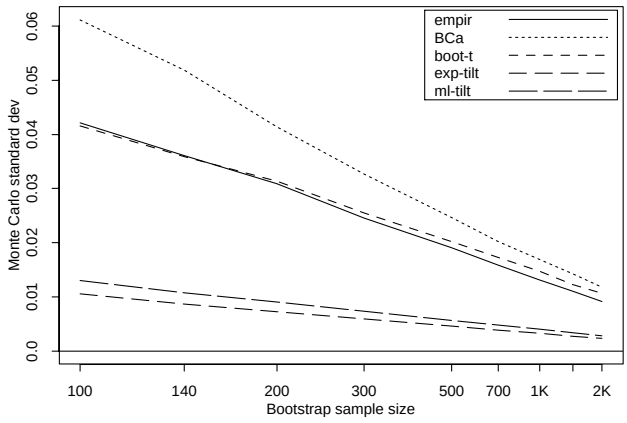

Figure 1: Estimated Monte Carlo variability (due to finite B) for one-sided 97.5% confidence intervals for the mean. There are 2000 datasets of normal data, $n = 40$; for each dataset and each value of B two sets of bootstrap samples are created and the sample variance of the two interval endpoints is calculated. Numbers shown are the square roots of the averages of the 2000 sample variances. *Comment*: tilting intervals have much smaller sampling variability.

Bootstrap tilting intervals were among the earliest bootstrap methods proposed (Efron 1981), and DiCiccio and Romano (1990) show that the methods are second-order accurate, comparable to the well-known BC-a and bootstrap-t methods, and an order of magnitude more accurate than normal-based or bootstrap percentile methods.

Efron (1981) also proposed the key idea in the efficient implementation of these methods, a creative use of importance sampling. The idea has largely escaped attention. However, using importance sampling and certain other implementation details, the intervals are very efficient computationally. Figure 1 is representative of the relative computational efficiency for various bootstrap methods in a variety of examples. Note that both tilting intervals suffer less variability with only 100 bootstrap samples than does the BC-a interval with 2000 samples. We show similar results for other examples later. In addition, tilting intervals offer better coverage and/or shorter intervals than competing intervals.

We begin with a short introduction to the bootstrap; for a more complete introduction to the bootstrap see (Efron and Tibshirani 1993).

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 or slope of a regression line, and $\hat{\theta} = \theta(\hat{F})$ the value estimated from the data. We require that θ be a functional statistic, i.e. it depends on the data only through the empirical distribution, with no dependence on sample size or order of the observations. 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 $\hat{F}$; then the sampling distribution can be estimated easily by Monte Carlo simulation (or by analytical approximations when available).

In this report 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 $\hat{\theta}^* = \theta(\hat{F}^*)$. For some number B of resamples (typically between 100 and 2000), 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 report.

The fundamental bootstrap assumption is that the theoretical bootstrap distribution accurately approximates the unknown sampling distribution (1), i.e. that $\hat{F}$ can substitute for the unknown F, at least for certain characteristics of the sampling distribution. Theoretical treatments of some aspects of the assumption are summarized in (Hall 1992), using Edgeworth expansions, and (Shao and Tu 1995), using functional analysis. We weaken the assumption by using the sampling distribution of $\hat{\theta}^*$ 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 (Efron 1981; Efron 1987; DiCiccio and Romano 1989).

In Section 2 we introduce tilting by way of hypothesis tests. We discuss tilting confidence intervals in Section 3, and the important topic of implementing tilting using importance sampling reweighting in Section 4. The choice of exponential or maximum likelihood tilting is discussed in Section 5, and whether to update derivatives in Section 6. Various adjustments to obtain better coverage properties are discussed in Section 7.

2 Hypothesis Tests

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_{θ_0} rather than $F_{\hat{\theta}}$. In a more general parametric setting, with one parameter θ 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 example Romano (1988); Romano (1989) sample in this way, for testing independence, rotational invariance, symmetry, and similar problems. Others (Boos et al. 1989) sample in various ways consistent with the null hypothesis in two-sample and multi-sample problems.

Bootstrap tilting hypothesis tests also involve sampling from distributions consistent with the null hypothesis, and were used by (Young 1988) for a one-sample mean, suggested by (Hinkley 1989) for comparing two means, and suggested by (Hall and Presnell 1999) for general one-sample problems.

In simple situations one may obtain a distribution consistent with the null hypothesis by modifying the observed values. For example, if θ is the mean of a univariate distribution, one could subtract $\theta_0 - \overline{x}$ from every observation.

Bootstrap tilting takes an alternate approach — the observed values are fixed, and unequal probabilities are placed on the observations. This allows greater generality — the distribution of x need not be univariate, and may include discrete or categorical variables — and can be implemented very efficiently; we return the latter point later.

In the sequel we restrict consideration to distributions with support on the observed data. Then we may describe a distribution in terms of the probabilities $\mathbf{p} = (p_1, \ldots p_n)$ 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 $\mathcal{X}^*$.

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, $\theta(\mathbf{p}) = \sum p_i x_i$, and the solution can be written in the form

$$p_i = c(1 - \tau(x_i - \overline{x}))^{-1}, \tag{3}$$

where τ is a "tilting" parameter and c normalizes the probabilities to sum to 1. The value of τ 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 2. 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 θ_0.

In bootstrap tilting hypothesis testing, the null distribution of $\hat{\theta}$ is estimated by resampling from the weighted empirical distribution, e.g. the p-value against the alternative H_a: $\theta > \theta_0$ is

$$P_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}), \tag{4}$$

where F_τ is the weighted empirical distribution induced by tilting with parameter τ.

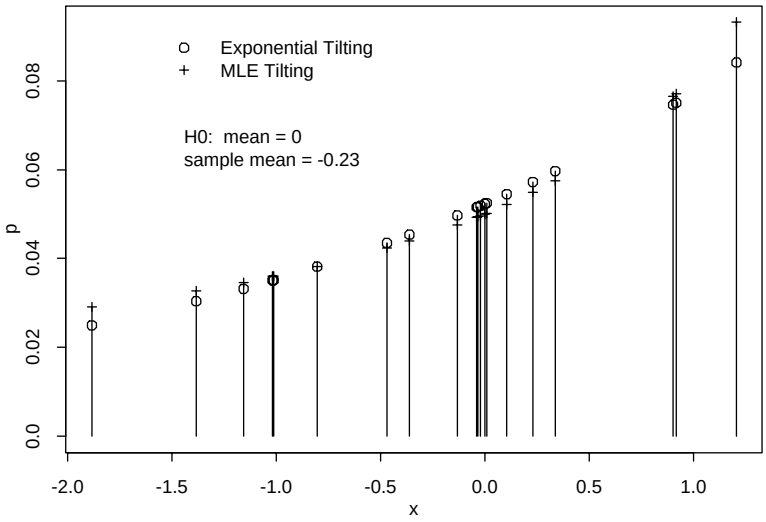

Figure 2: Exponential and Maximum Likelihood Tilting for a mean. *Comment*: ML tilting weights are higher on both ends, giving a distribution with larger variance.

The development to this point is similar to empirical likelihood (Owen 1988), except that in the latter p-values are obtained using a χ^2 approximation for the distribution of the log-likelihood statistic $-2\log(\prod p_i / \prod(1/n))$. The denominator is based on the unrestricted maximum likelihood distribution $\hat{F}$, which corresponds to probabilities $\mathbf{p}_0 = (1/n, \ldots, 1/n)$.

2.1 Nonlinear statistics

The procedure can be generalized to nonlinear statistics by substituting another single-parameter family for (3). The 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 report,

$$\begin{aligned}
\mathcal{F}_1 : p_i &= c\exp(\tau U_i(\mathbf{p}_0)) \\
\mathcal{F}_2 : p_i &= c\exp(\tau U_i(\mathbf{p})) \\
\mathcal{F}_3 : p_i &= c(1-\tau U_i(\mathbf{p}_0))^{-1} \\
\mathcal{F}_4 : p_i &= c(1-\tau U_i(\mathbf{p}))^{-1},
\end{aligned} \tag{5}$$

each indexed by a tilting parameter τ, where each c normalizes the corresponding vector to add to 1 and where

$$U_i(\mathbf{p}) = \lim_{\epsilon\to 0} \epsilon^{-1}(\theta(\mathbf{p}+\epsilon(\delta_i - \mathbf{p})) - \theta(\mathbf{p})) \tag{6}$$

where δ_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.

In the sequel we write $\mathbf{p}_\tau$ and F_τ for the corresponding probability vector and weighted empirical distribution for any of these families. Note that $\tau = 0$ corresponds to $\mathbf{p}_0$ and $\hat{F}$.

For any family, τ is found numerically to satisfy the null hypothesis,

$$\theta(\mathbf{p}_\tau) = \theta_0 \tag{7}$$

and the decision to reject is based on the estimated p-value (4).

$\mathcal{F}_1$ and $\mathcal{F}_2$ are well-known as "exponential tilting", and coincide if θ is a mean; these weights are also shown in Figure 2. Similarly $\mathcal{F}_3$ and $\mathcal{F}_4$ are ML tilting and coincide with (3) for a mean. $\mathcal{F}_3$ and $\mathcal{F}_4$ are not carefully defined in (5); they can be formally defined as distributions that respectively minimize the backward and forward Kullback-Leibler distances between $\mathbf{p}$ and $\mathbf{p}_0$, subject to (7). $\mathcal{F}_4$ gives the maximum likelihood solution for nonlinear statistics. We compare families in Sections 5 and 6.

3 Confidence Intervals

Bootstrap tilting hypothesis tests are consistent with the bootstrap tilting confidence intervals defined by (Efron 1981), in that the test rejects H_0 iff the confidence interval excludes θ_0. After choosing a least-favorable family, the lower limit of a one-sided $(1-\alpha)$ interval is found by solving

$$P_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) = \alpha \tag{8}$$

in τ, then defining the lower limit as

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

Upper limits are found similarly. DiCiccio and Romano (1990) 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 (Efron 1981) and BC-a (Efron 1987) 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 θ_α 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.

4 Implementation by Importance Sampling Reweighting

The most difficult step in implementing bootstrap tilting intervals is solving (8). This involves finding the value of τ for which resampling from F_τ yields a tail probability of α.

One approach is to sample from the weighted empirical distribution F_τ for different values of τ, estimate the tail probabilities for each τ, smooth the estimated probabilities, and numerically find the τ for which the value of the smooth curve is α. Because tail probabilities are relatively difficult to estimate using Monte Carlo simulation, this requires a large number of resamples (typically 1000) for each candidate value of τ. This can be expensive. Garthwaite and Buckland (1992) use the Robbins-Monroe algorithm, which would be similar in results and number of resamples required. DiCiccio and Romano (1989) suggest an alternative, the "automatic percentile method", which requires bootstrap sampling only from one candidate F_τ (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 (Efron 1981) 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. 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, importance reweighting, and recycling (Beckman and McKay 1987; Reiman and Weiss 1986; Tukey 1987; Hesterberg 1988; Hesterberg 1996; Davison and Hinkley 1997; Newton and Geyer 1994).

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) (Hammersley and Hanscomb 1964).

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.

Efron (1981) 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 F_τ, with probabilities $\mathbf{p}_\tau$ on the observed

data, the likelihood ratio $W = f/g$ for $\mathcal{X}_b^*$ is

$$W_b = \sum_{i=1}^{n} (np_i)^{M_{b,i}^*}. \tag{9}$$

Tail probability estimates

$$\begin{aligned} \hat{P}_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) &= B^{-1} \sum_{b=1}^{B} W_b I(\hat{\theta}^* \geq \hat{\theta}) \\ \hat{P}_{F_\tau}(\hat{\theta}^* \leq \hat{\theta}) &= B^{-1} \sum_{b=1}^{B} W_b I(\hat{\theta}^* \leq \hat{\theta}) \end{aligned} \tag{10}$$

are used for $\tau < 0$ and $\tau > 0$, respectively (the two estimates are not equivalent because $\sum_{b=1}^{B} W_b \neq B$).

This ISR procedure has a number of advantages. Sampling with equal probabilities is simpler, and a single set of resamples is used for both sides in a two-sided confidence interval, for every statistic if confidence intervals are required for multiple statistics (e.g. multiple regression coefficients), and for every α. The estimated tail probabilities are a smooth function of τ, 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, or for other statistics if sample sizes are large. The asymptotic relative efficiency compared to either sampling with probabilities $\mathbf{p}_\tau$ or to the bootstrap percentile interval is

$$\frac{\mathrm{Var}(L_{\mathrm{MC}})}{\mathrm{Var}(L_{\mathrm{IS}})} = \frac{\mathrm{Var}(L_{\mathrm{perc}})}{\mathrm{Var}(L_{\mathrm{IS}})} = \frac{\alpha(1-\alpha)}{(\exp(z_\alpha^2)\Phi(2z_\alpha) - \alpha^2)} \tag{11}$$

where Φ is the standard normal distribution function, $\Phi(z_\alpha) = \alpha$, and L_{MC}, L_{IS} and L_{perc} are the lower endpoints of one-sided $(1-\alpha)$ confidence intervals — L_{IS} is the exponential tilting interval using importance sampling, L_{MC} is the exponential tilting interval estimated using weighted Monte Carlo sampling, and L_{perc} is the bootstrap percentile interval estimated using simple Monte Carlo sampling. The variances are conditional on the observed data, the result is asymptotic as both $n \to \infty$ and $B \to \infty$, and depends on certain regularity conditions, that the statistic being bootstrapped is asymptotically linear and normal). The relative efficiency is about 17 for a two-sided 95% interval ($\alpha = 0.025$). Thus, if $B = 1000$ replications would give for sufficient accuracy for the bootstrap percentile interval (Efron 1987), 60 would suffice here.

The advantage is greater when α is smaller, e.g. when a Bonferroni procedure is used to set individual α values in a multiple-testing procedure. For $\alpha = .005$ the relative efficiency is approximately 68.

The computational advantage is even greater relative to the bootstrap BC-a interval (Efron 1987), probably the most common second-order-correct bootstrap interval, if z_0 is estimated from the data by the usual procedure $\hat{z}_0 = \Phi^{-1}(\hat{G}(\hat{\theta}))$. For general α, a, and z_0 the ratio of the Monte Carlo variance to that of the bootstrap percentile interval is

$$\frac{(p(1-p) + 2c(\min(p, p_0) - pp_0)) + c^2 p_0(1-p_0)}{(\alpha(1-\alpha))} \left(\frac{\min(\alpha, 1-\alpha)}{\min(p, 1-p)} \right)^2, \tag{12}$$

where $p_0 = \Phi(z_0)$, $z_\alpha = \Phi^{-1}(\alpha)$, $z_2 = z_0 + z_\alpha$, $z_3 = z_0 + z_2/(1 - az_2)$, $c = (1 + (1 - az_2)^{-2})\phi(z_3)/\phi(z_0)$, and $p = \Phi(z_3)$. Thus the Monte Carlo variability of the BC-a interval is greater than that of the bootstrap percentile interval by a factor of 2.18 (asymptotically, with $a = z_0 = 0$) for two-sided 95% intervals; this in turn implies an asymptotic relative efficiency of 37 for bootstrap tilting relative to the BC-a interval.

Figures 3–6 show the relative computational efficiency of bootstrap confidence interval procedures. These show the variability due to Monte Carlo sampling with a finite bootstrap sample-size B (there is an additional component of variability, due to differences between samples ($\mathcal{X}^*$), that is not included here). Note the remarkable similarity in relative computational efficiency for different statistics (mean, bivariate correlation, ratio of means, and variance), different sample sizes ($n = 10, 20, 80$), and different distributions (normal or skewed (exponential)). Both tilting intervals are markedly more efficient than the other intervals, especially when compared to the other second-order correct intervals (bootstrap-t and BC-a).

Tables 1 and 2 show the Monte Carlo efficiency for empirical likelihood tilting relative to the bootstrap percentile, BC-a, and bootstrap-t confidence intervals. Except for a few pathological cases, the relative efficiencies are around 17 compared to percentile intervals, and around 30 relative to BC-a intervals, for one-sided confidence levels of $\alpha = 0.025$ and $\alpha = 0.975$. The relative efficiencies are higher for $\alpha = 0.01$ and $\alpha = 0.99$. Results are for $B = 2000$ replications, and the same applications described earlier.

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

We defer discussion of coverage accuracy and non-Monte-Carlo variability (i.e. the variability that would still occur with $B = \infty$, due to random data) of competing methods to Section 5.

4.1 Mixture Design Distributions

Importance sampling using $\hat{F}$ as the design distribution is nearly ideal for exponential tilting when the statistic is the sample mean. In this case the weights (9) simplify to

$$W_b = (nc)^n \exp(\textstyle\sum M^*_{b,i} U_i) = (nc)^n \exp(\tau(\hat{\theta}^* - \hat{\theta}))$$

so that the weights are a function of $\hat{\theta}^*$, with exclusively small weights in the tail of interest, e.g. for those bootstrap samples that contribute to the tail of interest in (10).

Table 1: Relative Computational Efficiency of bootstrap exponential tilting and other confidence intervals, for 95% two-sided intervals. Numbers in the table are the Monte Carlo variance of the confidence interval listed, divided by the Monte Carlo variance of the exponential tilting interval. For example, the first table entry is 18, indicating that the bootstrap empirical interval requires 18 times as many bootstrap samples to reduce the simulation variance to the level of the exponential tilting interval.

θ	Distribution	n	α	Empirical	BC-a	Bootstrap-t
Mean	Normal	20	0.025	18	29	29
			0.975	18	30	31
		80	0.025	17	28	20
			0.975	17	27	17
	Exponential	20	0.025	21	26	29
			0.975	15	58	81
		80	0.025	20	27	18
			0.975	13	33	23
Cor	Normal	20	0.025	2	4	8
			0.975	7	11	35
		80	0.025	16	29	24
			0.975	17	26	25
Ratio of Means	Normal	20	0.025	19	32	28
			0.975	16	26	22
		80	0.025	17	29	19
			0.975	17	26	18
	Exponential	20	0.025	18	37	68
			0.975	32	32	22
		80	0.025	15	28	21
			0.975	20	26	18
Var	Normal	20	0.025	18	23	33
			0.975	0	3	5
		80	0.025	19	21	18
			0.975	12	48	29
	Exponential	20	0.025	13	25	68
			0.975	3	122	4874
		80	0.025	19	26	32
			0.975	11	234	199

Table 2: Relative Computational Efficiency of bootstrap exponential tilting and other confidence intervals, for 98% two-sided intervals. Numbers in the table are the Monte Carlo variance of the confidence interval listed, divided by the Monte Carlo variance of the exponential tilting interval.

θ	Distribution	n	α	Empirical	BC-a	Bootstrap-t
Mean	Normal	20	0.01	37	60	76
			0.99	40	60	80
		80	0.01	38	53	42
			0.99	36	49	40
	Exponential	20	0.01	48	44	77
			0.99	32	224	393
		80	0.01	43	44	42
			0.99	29	82	52
Cor	Normal	20	0.01	2	6	18
			0.99	7	9	84
		80	0.01	38	62	57
			0.99	38	46	62
Ratio of	Normal	20	0.01	41	61	76
Means			0.99	35	57	49
		80	0.01	38	54	42
			0.99	37	51	38
	Exponential	20	0.01	42	108	223
			0.99	77	55	53
		80	0.01	33	64	56
			0.99	52	46	40
Var	Normal	20	0.01	40	34	94
			0.99	0	6	23
		80	0.01	44	34	46
			0.99	24	142	62
	Exponential	20	0.01	40	52	446
			0.99	15	172	1447862
		80	0.01	43	40	85
			0.99	19	426	682

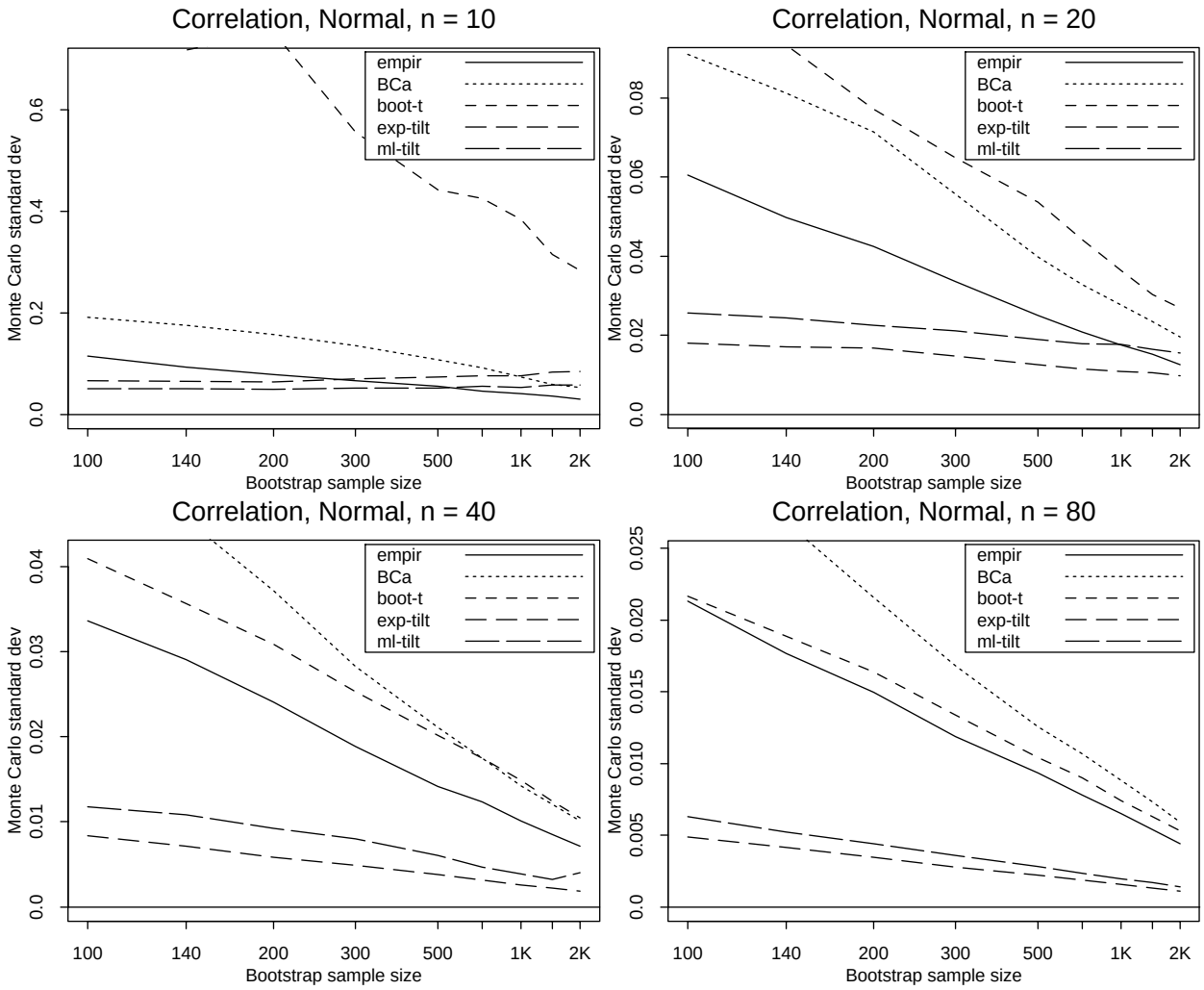

Figure 3: Estimated Monte Carlo variability (due to finite B) for one-sided confidence intervals ($\alpha = 0.025$) for the correlation coefficient, for bivariate normal data with correlation $(1/2)^{1/2}$. There are 2000 datasets of normal data, $n = 40$; for each dataset and each value of B two sets of bootstrap samples are created and the sample variance of the two interval endpoints is calculated. Numbers shown are the square roots of the averages of the 2000 sample variances. *Comment*: tilting intervals have much smaller sampling variability for large samples (for small samples the intervals should not be used for highly-nonlinear statistics like correlation without the optimization procedure described later).

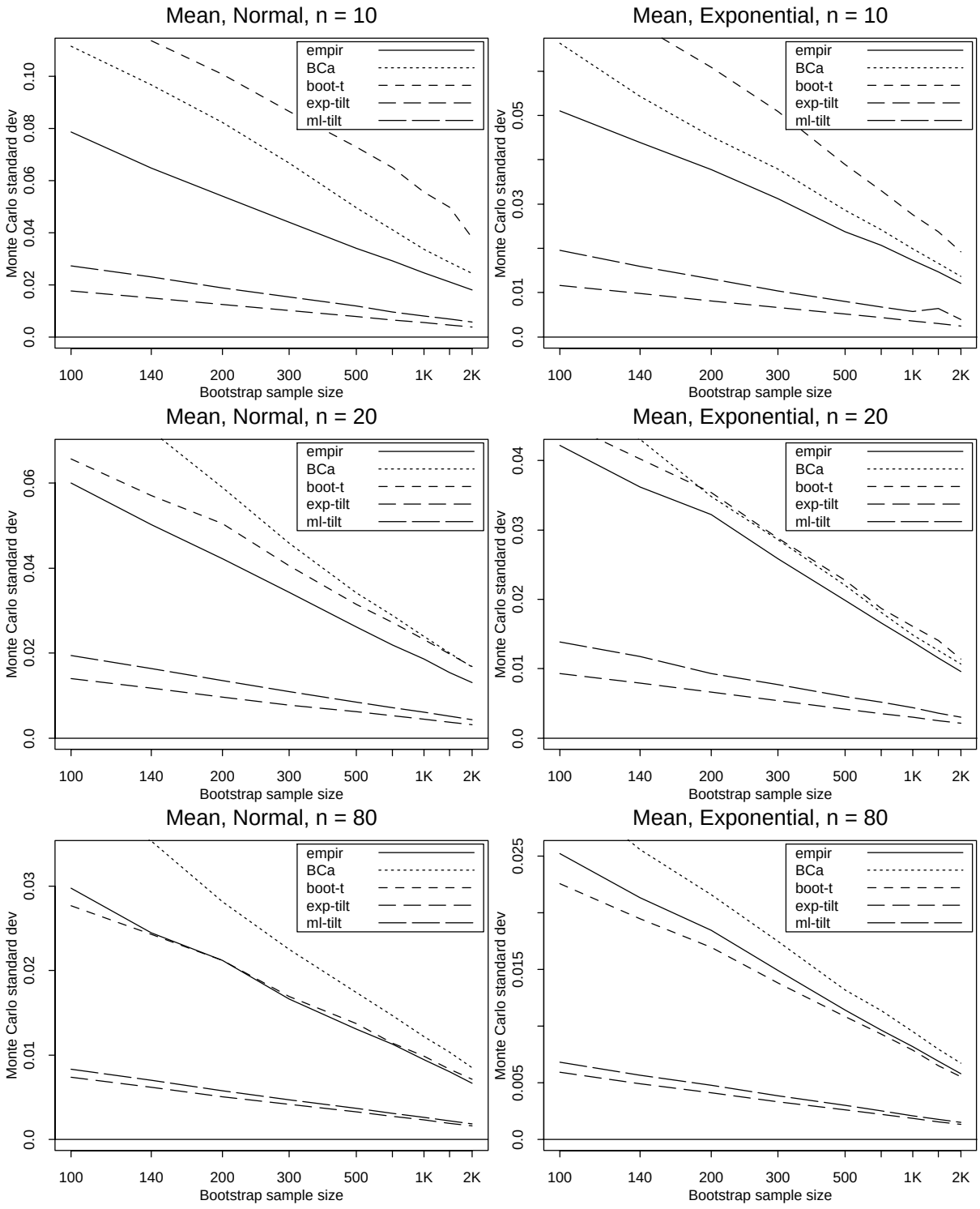

Figure 4: Estimated Monte Carlo variability for confidence intervals for the mean. Like Figure 3, but for the sample mean, for standard normal and exponential data.

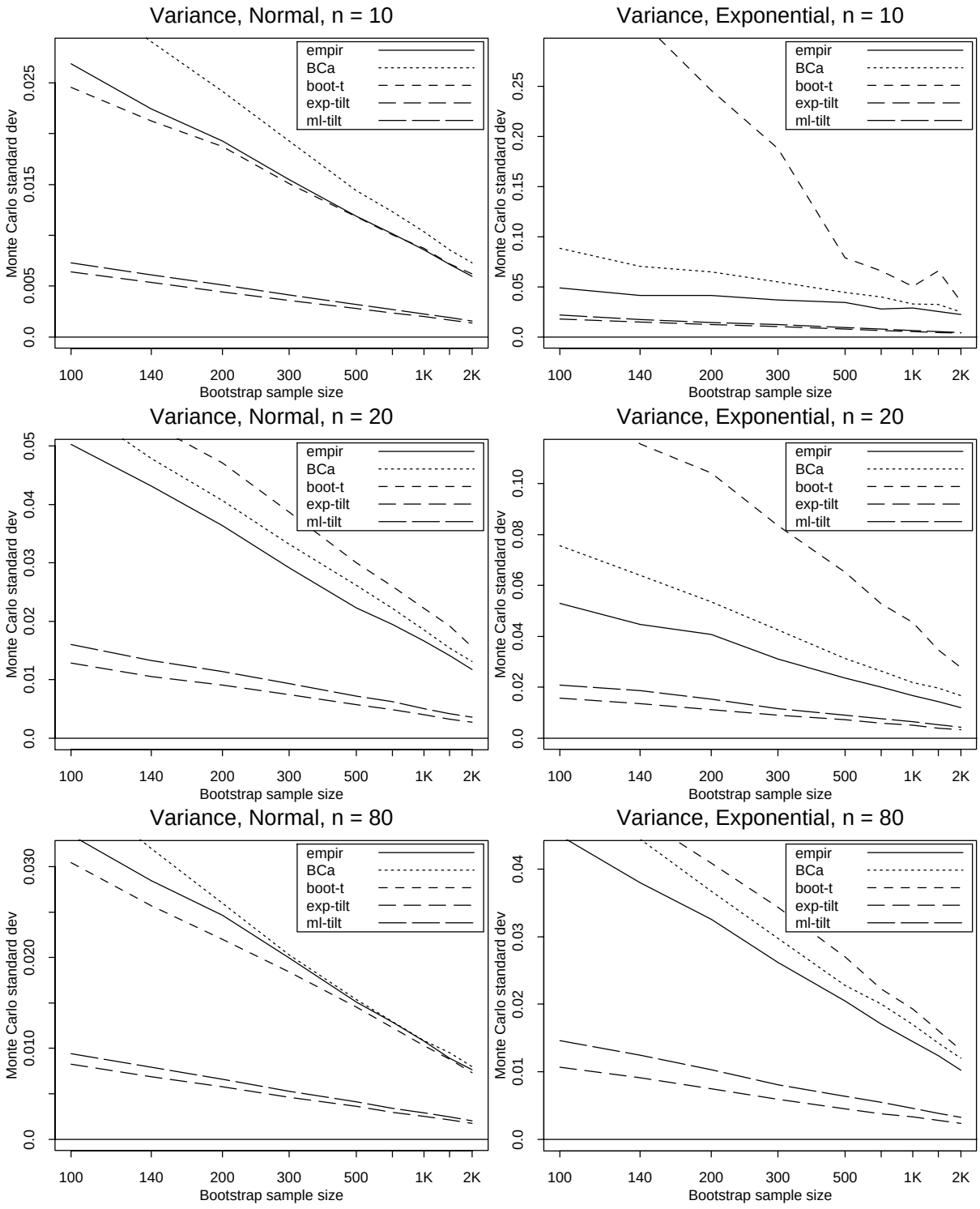

Figure 5: Estimated Monte Carlo variability for confidence intervals for the variance. Like Figure 3, but for the variance for standard normal and exponential data.

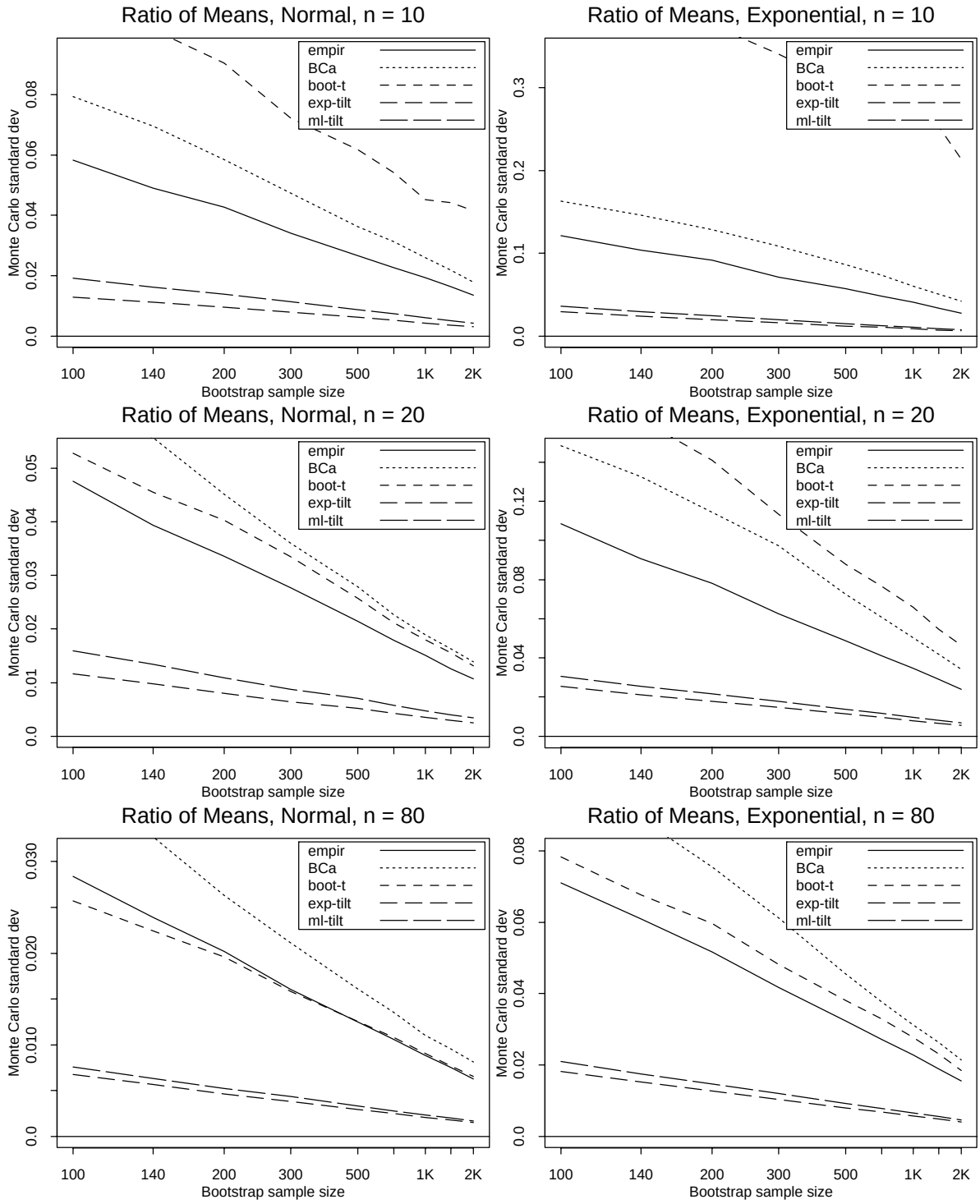

Figure 6: Estimated Monte Carlo variability for confidence intervals for ratios of means. Like Figure 3, but for the ratio of means, for bivariate normal data (uncorrelated, bivariate mean $(3, 9)$, variance 1) and exponential data (independent, minimum values for x and y are 0 and 2, respectively, standard scale).

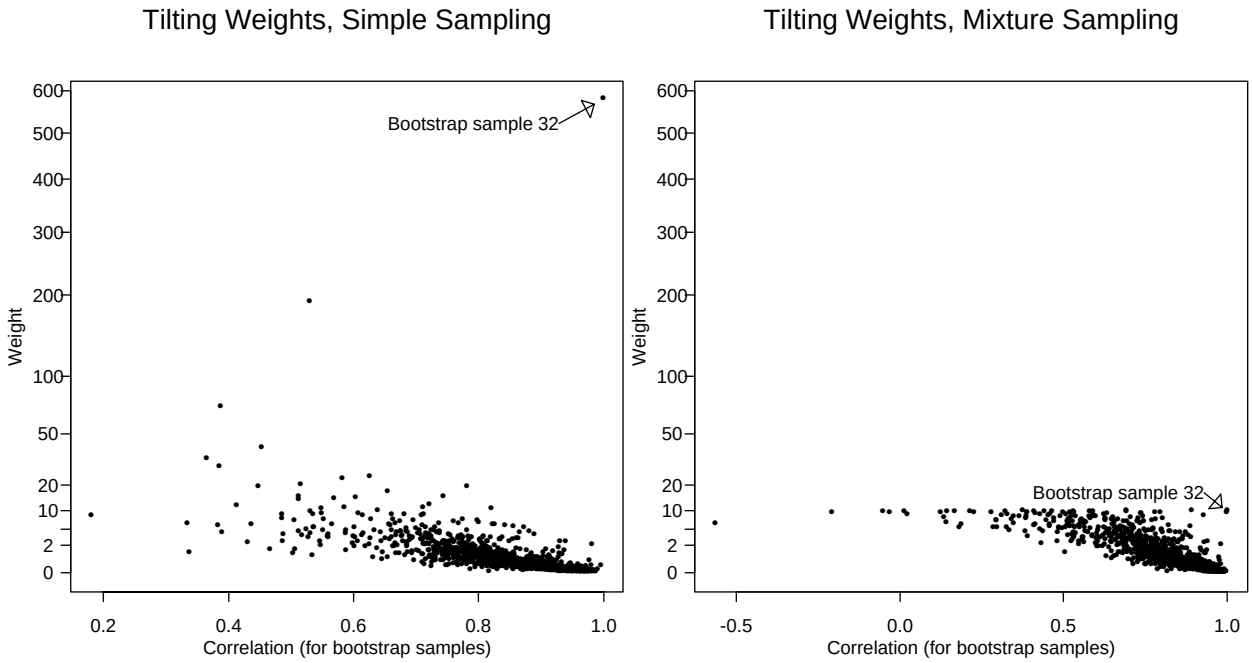

Figure 7: Simple random bootstrap sampling and defensive mixture sampling. Each point represents a bootstrap sample. All 2000 samples in the left plot were generated using simple Monte Carlo sampling. The right plot has the same initial 1600 samples, but the remaining 400 were generated using probabilities obtained by tilting to the left and right (200 samples each). With simple random sampling there are occasional outliers in tilting weights — large weights that occur on the wrong side — particularly with small samples and highly non-linear statistics like the correlation coefficient.

In general, we prefer to have small weights in the region of interest. The effective sample size for any region is roughly inversely proportional to the typical weight of an observation in that region (after normalizing by the sum of weights). For example, in the absence of importance sampling, all normalized weights would be exactly $1/B$. If importance sampling causes a particular region to be sampled about twice as often as it would be otherwise, the normalized weights in that region would be about $1/(2B)$

For nonlinear statistics the weights are no longer a function solely of $\hat{\theta}^*$, and for small n and/or highly nonlinear statistics it is possible for large weights to occur in the tail of interest — then the probability estimate (10) can be determined largely by the presence or absence of a small number of bootstrap samples with such large weights. For example, Figure 7 shows the exponential tilting weights for the correlation coefficient for a sample of size $n = 10$, with $B = 2000$. A single bootstrap sample has a weight of 600, or normalized weight of approximately $600/2000$.

A more robust design for nonlinear statistics is a (defensive) mixture design (Hesterberg

1995b), of the form

$$\sum_{k=1}^{K} \lambda_k \hat{F}_{\tau_k}. \tag{13}$$

Then for any τ, the importance sampling weights W_b are

$$W_b = \frac{\sum_{i=1}^{n} p_{\tau,i}^{M^*_{b,i}}}{\sum_k \left(\lambda_k \sum_{i=1}^{n} p_{\tau_k,i}^{M^*_{b,i}} \right)}. \tag{14}$$

For example, we may generate 80% of the resamples from $\hat{F}$ as before, calculate preliminary estimates τ_1 and τ_2 for a two-sided confidence interval, and generate an additional 10% from each of F_{τ_1} and F_{τ_2}, yielding a mixture design

$$.8\hat{F} + .1\hat{F}_{\tau_1} + .1\hat{F}_{\tau_2} \tag{15}$$

The use of even a relatively small fraction of resamples makes results robust, because now weights are bounded above by $W_b < 10$ (if $\tau = \tau_1$ or $\tau = \tau_2$). This is apparent in the right side of Figure 7.

Alternately, we may use design components $\hat{F}_{\tau_1/2}$ and $\hat{F}_{\tau_2/2}$; while this does not yield the same bound on the weights as does the use of $\hat{F}_{\tau_1}$ and $\hat{F}_{\tau_2}$, it does yield improved performance for mildly nonlinear statistics.

Stratification of mixture proportions Mixture designs may either use a true mixture distribution, in which the mixture component for each sample is chosen independently with probabilities $\lambda_1, \ldots, \lambda_K$, or a stratified mixture in which exactly $B\lambda_k$ samples are generated using distribution $\hat{F}_{\tau_k}$ (Hesterberg 1995b). The latter is more accurate.

Designs for ML tilting For ML tilting, mixture designs also improve robustness and performance, but there is one additional factor to consider. The nearly-optimal design for ML tilting for the sample mean is not $\hat{F}$, but rather of the form

$$p_i = c \frac{1}{(1 - \tau U_i)} \exp(\tau_r U_i) \tag{16}$$

where τ solves (8) and τ_r solves $\sum p_i x_i = \bar{x}$. Note that τ is the usual tilting parameter for ML tilting, that tilts the distribution to one side or the other — this is the distribution that would be used in the absence of the ISR implemetation. Then τ_r tilts the distribution back to match the original mean, $\tau_r \doteq -\tau$ — this is an example of the traditional use of importance sampling as a variance reduction technique.

The final set of probabilities give a weighted empirical distribution with the same weighted mean as the original $\bar{x}$, but with slightly larger variance than $\hat{F}$. This design places slightly

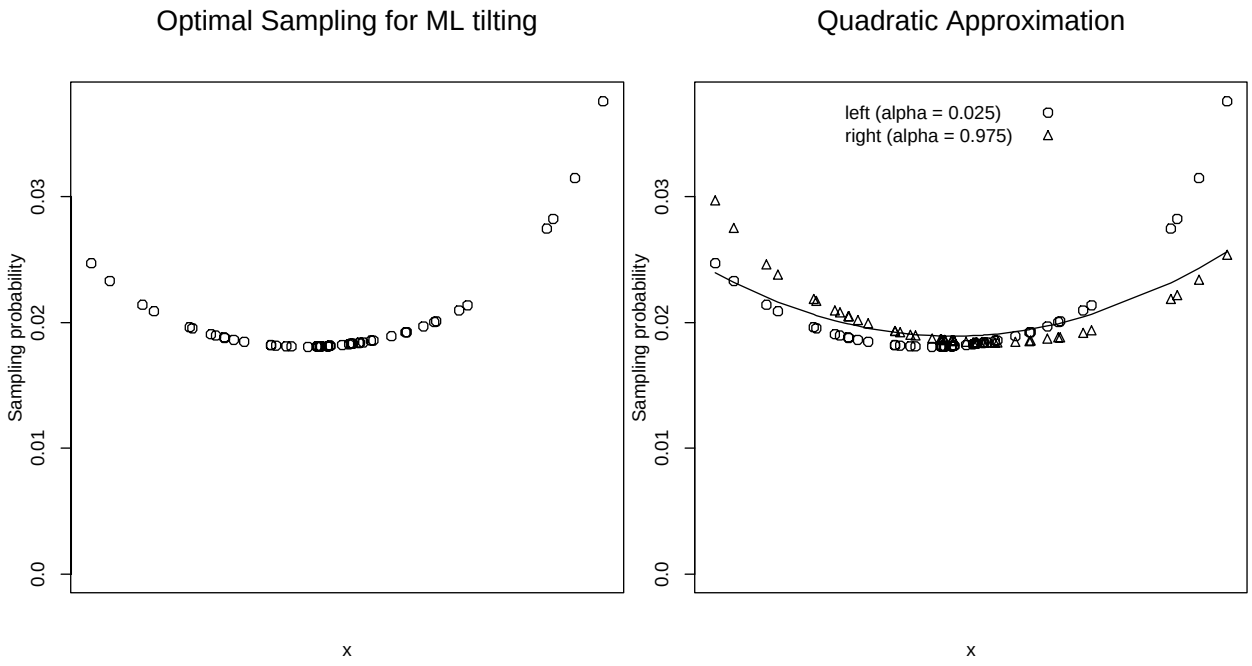

Figure 8: Optimal sampling weights for ML tilting. The left panel shows approximately optimal weights for the bootstrap sampling for the left side of a 95% two-sided confidence interval. The right panel shows approximately optimal weights for both sides, and a quadratic approximation. The data are a sample of size 40 from a normal distribution.

higher probabilities on the more extreme observations (large and small values of x_i, or more generally U_i) than does $\hat{F}$. One set of such probabilities is shown in the left panel of Figure 8.

This difference between $\hat{F}$ and the nearly-optimal design is one reason that exponential tilting performs better in Figures 3–6, which were all obtained using simple bootstrap sampling (another reason is that ML tilting intervals are wider see Section 5).

Curiously, the nearly-optimal probabilities are similar for both sides of a two-sided confidence interval, at least for reasonably large samples, so that little would lost by using a single set of sampling probabilities, either the average of the two probability vectors which are optimal for the two sides, or a quadratic approximation $p_i = c(1 + \tau^2 U_i^2)$

Figure 9 compares the results for four different sampling mechanisms:

1. Simple Monte Carlo ($\hat{F}$): nearly optimal for exponential tilting intervals, in this application (sample mean),

2. Mixture with 50% $\hat{F}$ and 25% each tilted to the left and right (using exponential tilting): this does not do as well for either exponential or ML tilting in this application, with a relative efficiency compared to Simple Monte Carlo of about 63% for exponential tilting intervals and 69% for ML tilting intervals (geometric mean of the ratios of Monte Carlo variances). However, for nonlinear statistics a mixture design like this would add needed robustness; the mixture components from the tails could be smaller than 50%.

3. 50% $\hat{F}$ and 25% each from the left and right nearly-optimal probabilities for ML tilting: this provides a modest improvement over simple Monte Carlo, with larger gains for small sample sizes — the relative efficiency compared to simple Monte Carlo is about 117% for $n = 10$ and 106% for $n = 80$.

4. 50% $\hat{F}$ and 50% from the average of the left and right nearly-optimal probabilities: results are practically equivalent to the previous, but this is slightly simpler in practice because only two mixture components are involved rather than three.

The latter two sampling mechanisms are used here only for ML tilting.

There are a number of disadvantages of using these nearly-optimal sampling probabilities for ML tilting (16), relative to sampling from $\hat{F}$. The first is that sampling with unequal probabilities is slower than sampling with equal probabilities. The second is that these probabilities depend on the desired confidence level for confidence intervals (in practice, a mixture of $\hat{F}$ and the nearly-optimal probabilities for highest confidence levels could be used). The third is that the sampling probabilities would differ for different dimensions of a multivariate statistic θ, and mixture components optimized for one dimension would be poor for another. The fourth is that sampling from $\hat{F}$ is simpler.

Because of these disadvantages, and because it does not appear that substantial variance reductions are obtained, we recommend not using nearly-optimal probabilities (16) for ML tilting.

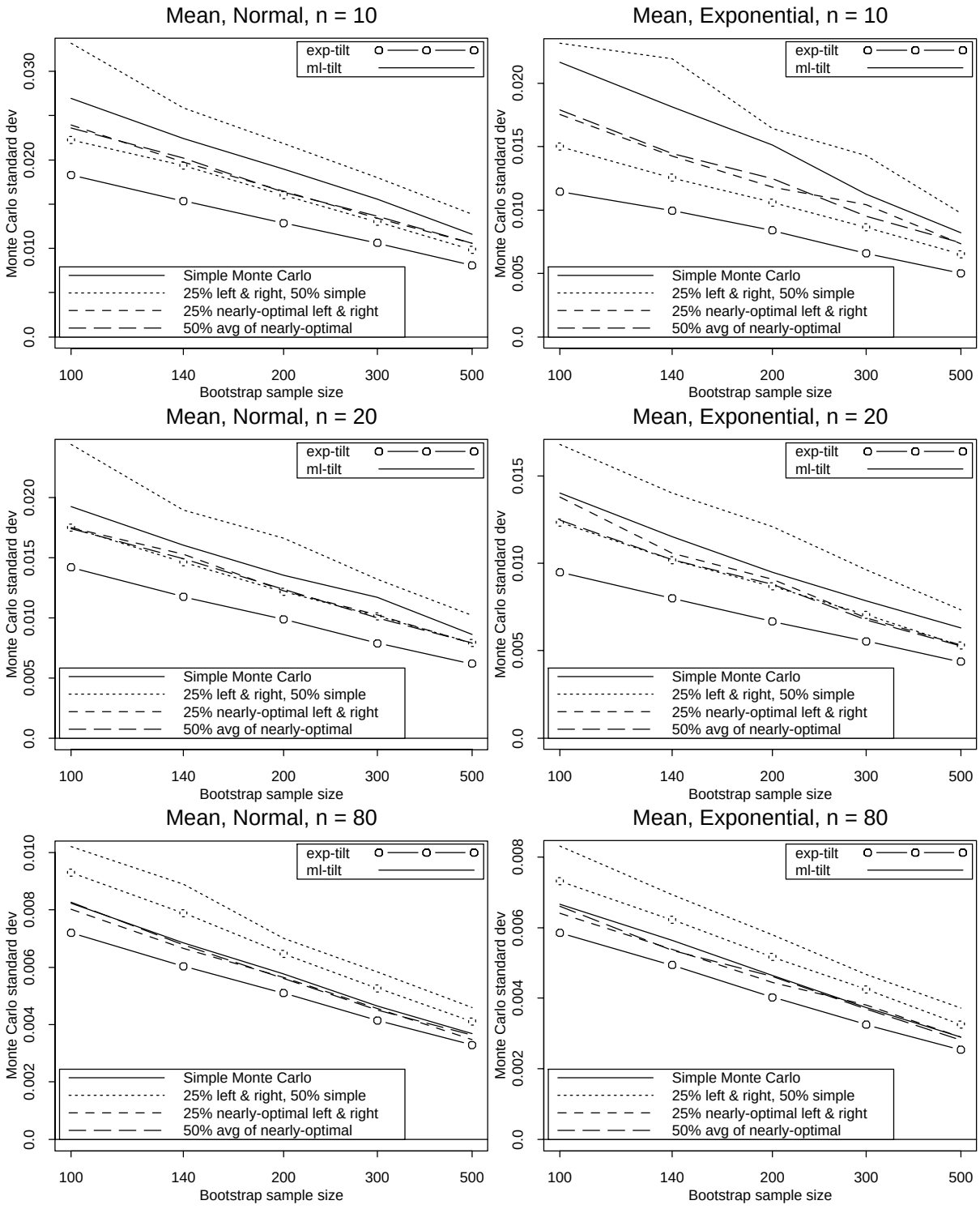

Figure 9: Estimated Monte Carlo variability under different sampling mechanisms. Like Figure 4, but comparing four sampling mechanisms.

5 Choice of Tilting Family

The four least-favorable (tilting) families $\mathcal{F}_1$–$\mathcal{F}_4$ have appeared in the bootstrap literature, but with little discussion of their relative merits. These families are distinguished by whether they use exponential or ML tilting, and whether derivatives are computed just once or are updated. We discuss updating derivatives in Section 6. Here we compare exponential and ML tilting, and other intervals, in terms of statistical accuracy — how closely the actual coverage probabilities match the nominal values.

In general ML tilting gives wider confidence intervals and higher coverage problems (and lower Type I error in hypothesis testing) than does exponential tilting. Taylor-series expansions of the families in (5) in terms of τ 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 2, 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.

Higher coverage is desirable because in practice most bootstrap and other confidence interval procedures tend to be anti-conservative with finite samples (see simulation results collected by (Shao and Tu 1995), or our examples in here.

Furthermore, a result by (Hesterberg 1995b) implies that when θ 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. The same result would hold for the linear part of nonlinear statistics. The relatively small bias for ML tilting should result in more accurate inferences.

The coverage accuracy of exponential and ML tilting, using families $\mathcal{F}_1$ and $\mathcal{F}_3$ (no updating of derivatives), and other bootstrap methods, is shown in Figures 10–15 for a variety of applications. The bootstrap tilting intervals are based on $B = 200$ bootstrap samples, and the other intervals on $B = 1999$ samples.

All intervals tend to under-cover, to produce intervals whose coverage probability is less than the desired value, particularly for small sample sizes. For example, in Figure 10, the intervals for the sample mean for normal data, most of the intervals excluded the true mean about 5% of the time on the low end when $n = 10$, and nearly 6% of the time on the upper end (see the top right panel), for nominal 97.5% one-sided intervals. The bootstrap-t intervals perform notably well in that example.

In asymmetric problems, such as the sample mean for exponential data (Figure 11), all of the intervals undercover badly on one side, while some slightly over-cover on the other.

The bootstrap percentile intervals do particularly poorly in this example. They are only

first-order correct, meaning that the coverage errors are $O(n^{-1/2})$, while the other procedures have errors of order $O(n^{-1})$. The usual Student's-t intervals for the mean are also only first-order correct. (In the special case of symmetric distributions the errors of the Student's-t and bootstrap percentile intervals are also $O(n^{-1})$.)

For the most part, ML tilting provides higher and more coverage than exponential tilting. However, ML tilting did poorly for the correlation and variance of small samples ($n = 10$). These are highly nonlinear problems with small sample sizes; indeed, $\theta(F_\tau)$ is a non-monotone function of τ for these statistics. ML tilting appears to be more sensitive to nonlinearity and non-monotonicity because it is based on distributions $\hat{F}_\tau$ which are farther from $\hat{F}$. Some form of updating should be used in such problems; see Section 6. Note that there is good diagnostic measure of the degree of linearity in a problem, based on the correlation of $\hat{\theta}^*_b$ and the linear approximation $L^*_b = n^{-1} \sum M^*_i U_i$.

The bootstrap-t interval provides the best coverage accuracy in our simulations, but is known to have two major flaws. First, it is not transformation invariant; (e.g. it does not produce equivalent intervals for an odds-ratio and the equivalent log-odds-ratio), and the high coverage comes at the price of very long confidence intervals. One aspect of the lack of transformation invariance is that bootstrap-t intervals can include impossible values, e.g. a bootstrap-t confidence interval for a correlation coefficient can include correlations greater than 1. Figures 16–21 show the average confidence interval lengths for the same applications as above. The bootstrap-t intervals are substantially longer than other intervals.

The next set of figures shows the relationship between coverage and confidence interval length. Figures 22–27 show the average confidence interval lengths for the same applications as above.

For the most part, all confidence interval methods trace out approximately the same trajectory in the coverage/length plane. This indicates that most of the difference in confidence interval length between various intervals is explained by the difference in coverage probabilities — the shorter intervals are shorter not because they are intrinsically better, but because they just do not include the true parameter often enough.

However, even after calibrating for coverage equivalent probability, the bootstrap-t intervals are substantially longer than other intervals, especially for the correlation coefficient.

The right panel of Figure 27 also shows that the ML tilting interval breaks down — as the nominal coverage probability increases the actual coverage barely changes and even decreases, for upper confidence limits for the variance of samples of size 10. We return to this problem in Section 6.

This comparison was based on 200 bootstrap replications for the tilting intervals and 1999 replications for the other intervals. The best interval in terms of a combination of statistical properties and computational expense is the ML tilting interval.

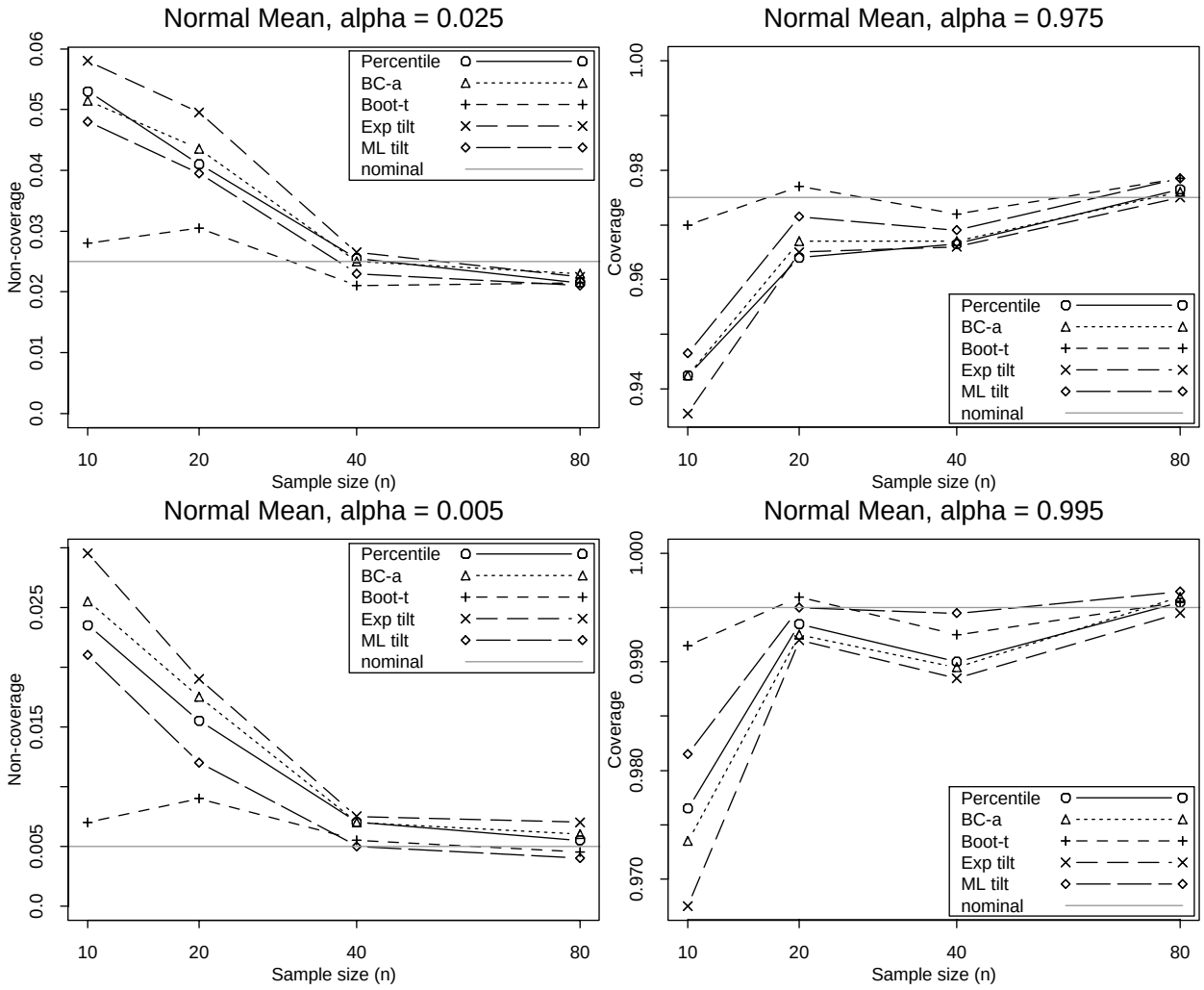

Figure 10: Coverage Accuracy. One-sided confidence intervals for the sample mean of normal data. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated and one of each kind of bootstrap confidence interval was generated, using $B = 200$ bootstrap samples for the tilting intervals and $B = 1999$ bootstrap samples for the bootstrap percentile, bootstrap-t, and BC-a intervals. The standard errors are approximately $(.025 \times .975/2000) = .0035$ for intervals with nominal coverage of 0.025 or 0.975. *Comment*: all under-cover for small n — the bootstrap-t does the best and exponential tilting the worst.

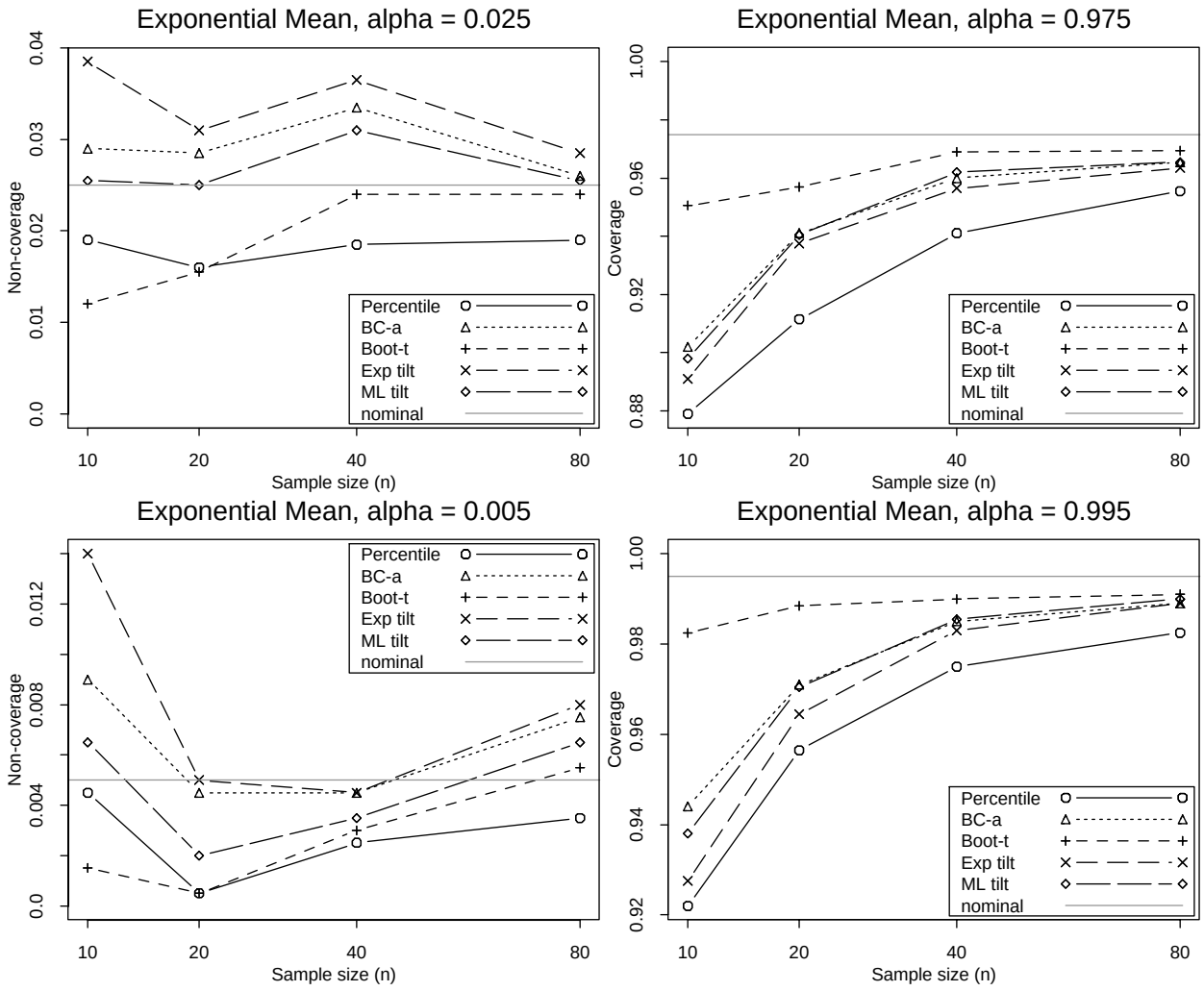

Figure 11: Coverage Accuracy. One-sided confidence intervals for the sample mean of exponential data. Other details are the same as for Figure 10. *Comment*: all under-cover badly at the upper end for small n — the bootstrap-t does the best and percentile the worst, improving much more slowly for larger sample sizes than do the other methods.

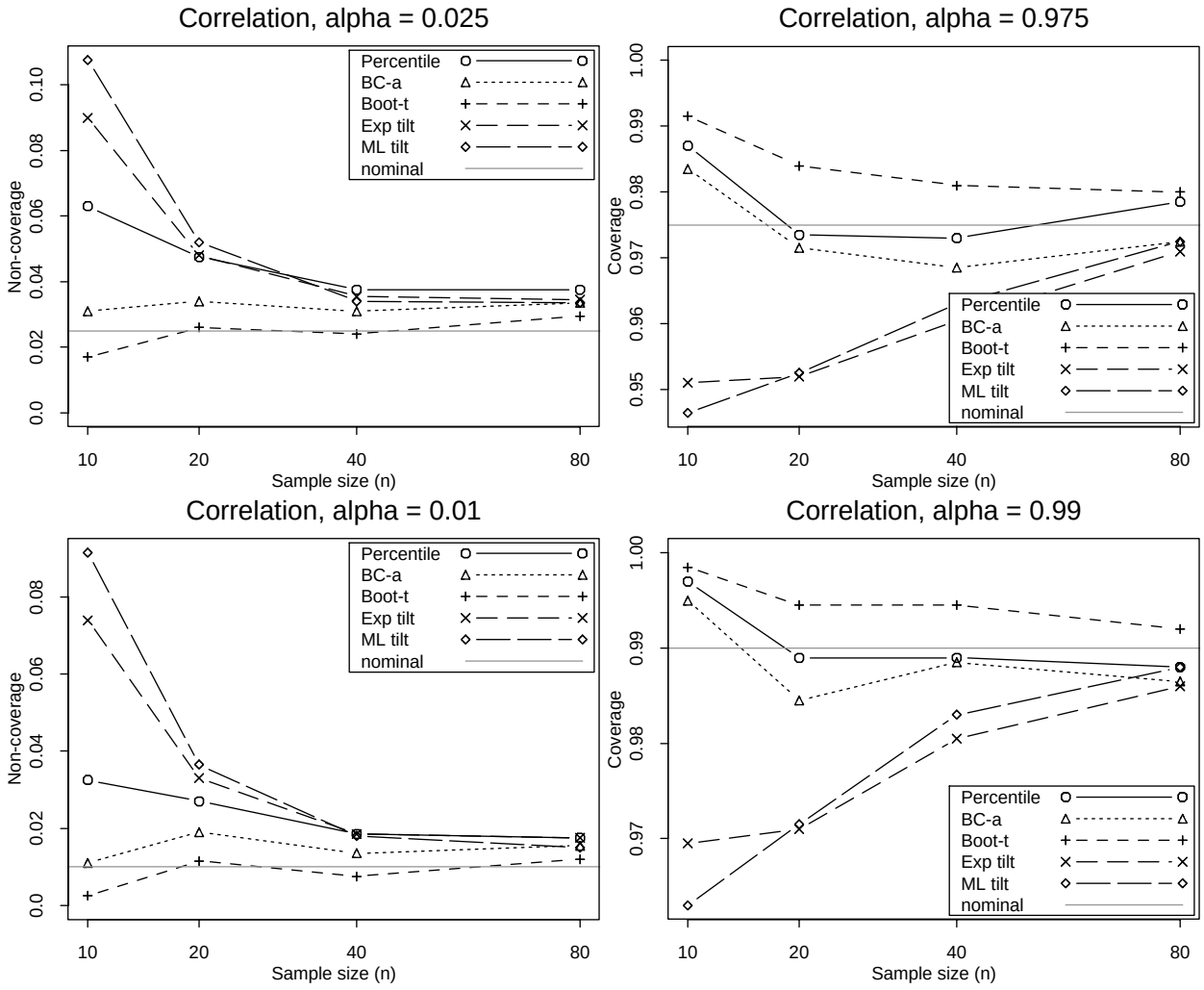

Figure 12: Coverage Accuracy. One-sided confidence intervals for the correlation for bivariate normal data with correlation $(1/2)^{1/2}$. Other details are the same as for Figure 10. *Comment*: both tilting methods undercover badly for small n.

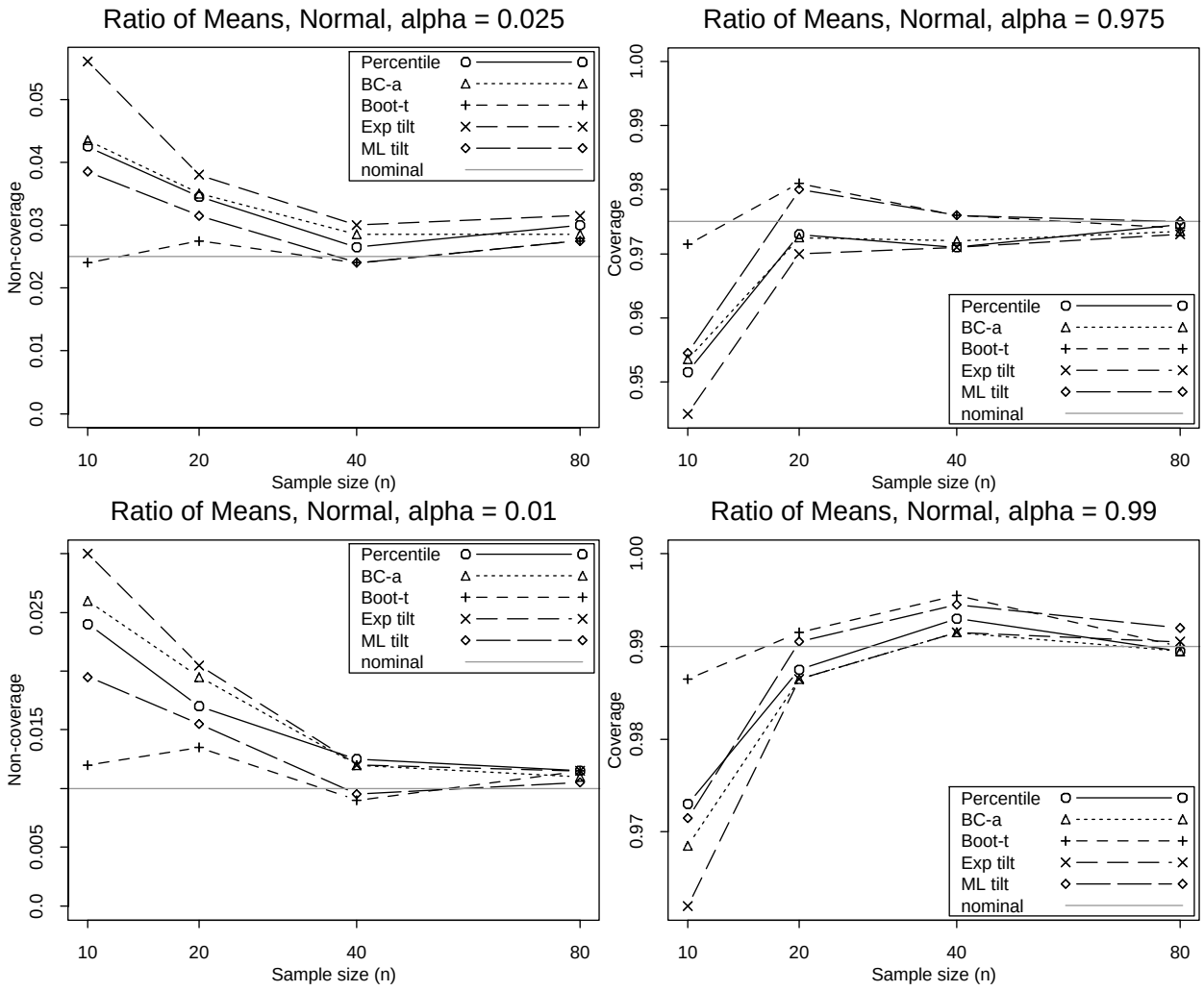

Figure 13: Coverage Accuracy. One-sided confidence intervals for the ratio of means for bivariate normal data (uncorrelated, bivariate mean $(3, 9)$, variance 1) Other details are the same as for Figure 10. *Comment*: all undercover for small n; the bootstrap-t does the best and exponential tilting the worst.

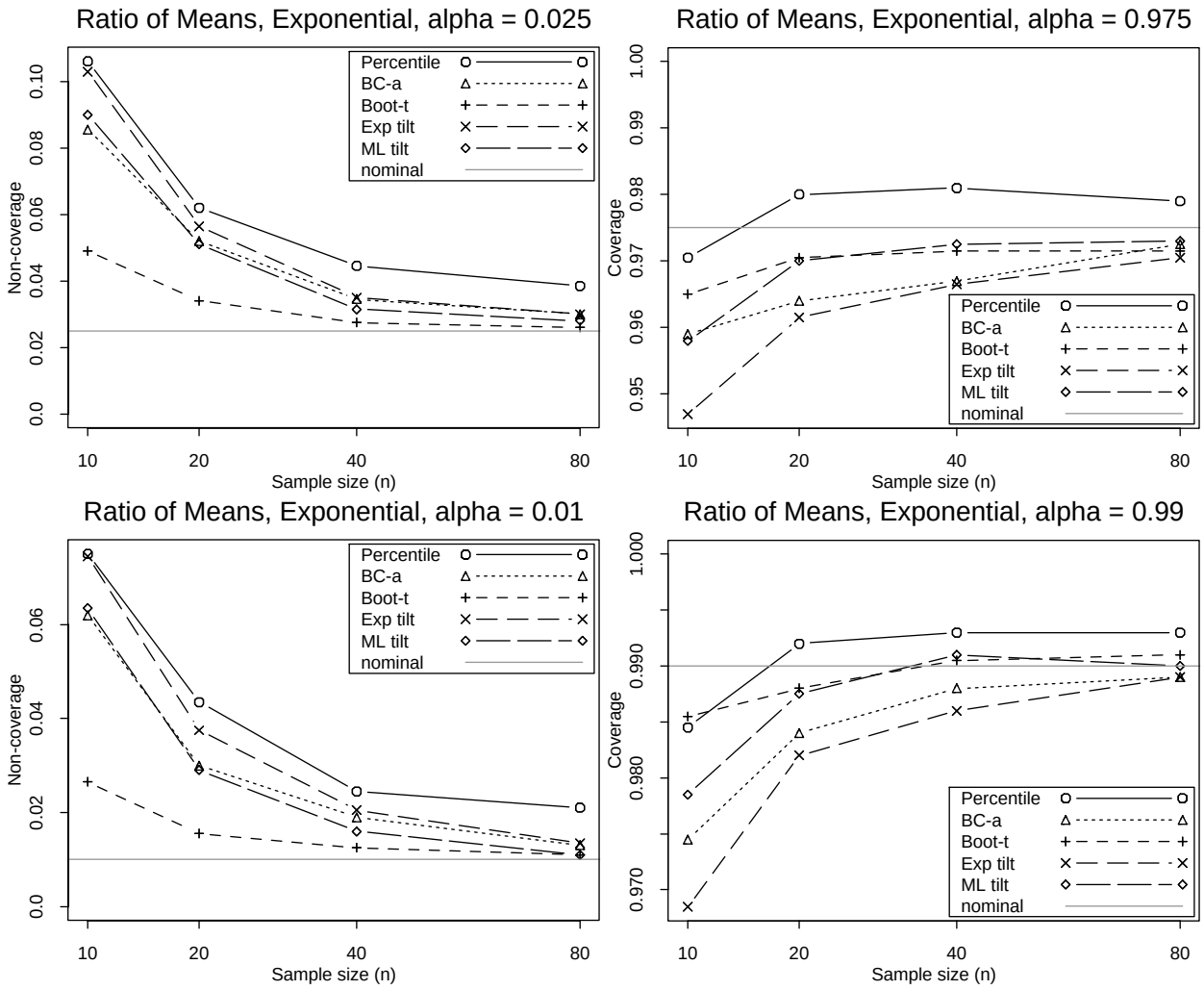

Figure 14: Coverage Accuracy. One-sided confidence intervals for the ratio of means for exponential data (independent, minimum values for x and y are 0 and 2, respectively, standard scale). Other details are the same as for Figure 10. *Comment*: all undercover badly for small n at the lower end; the bootstrap-t does the best and percentile method the worst, improving more slowly than other methods as n increases.

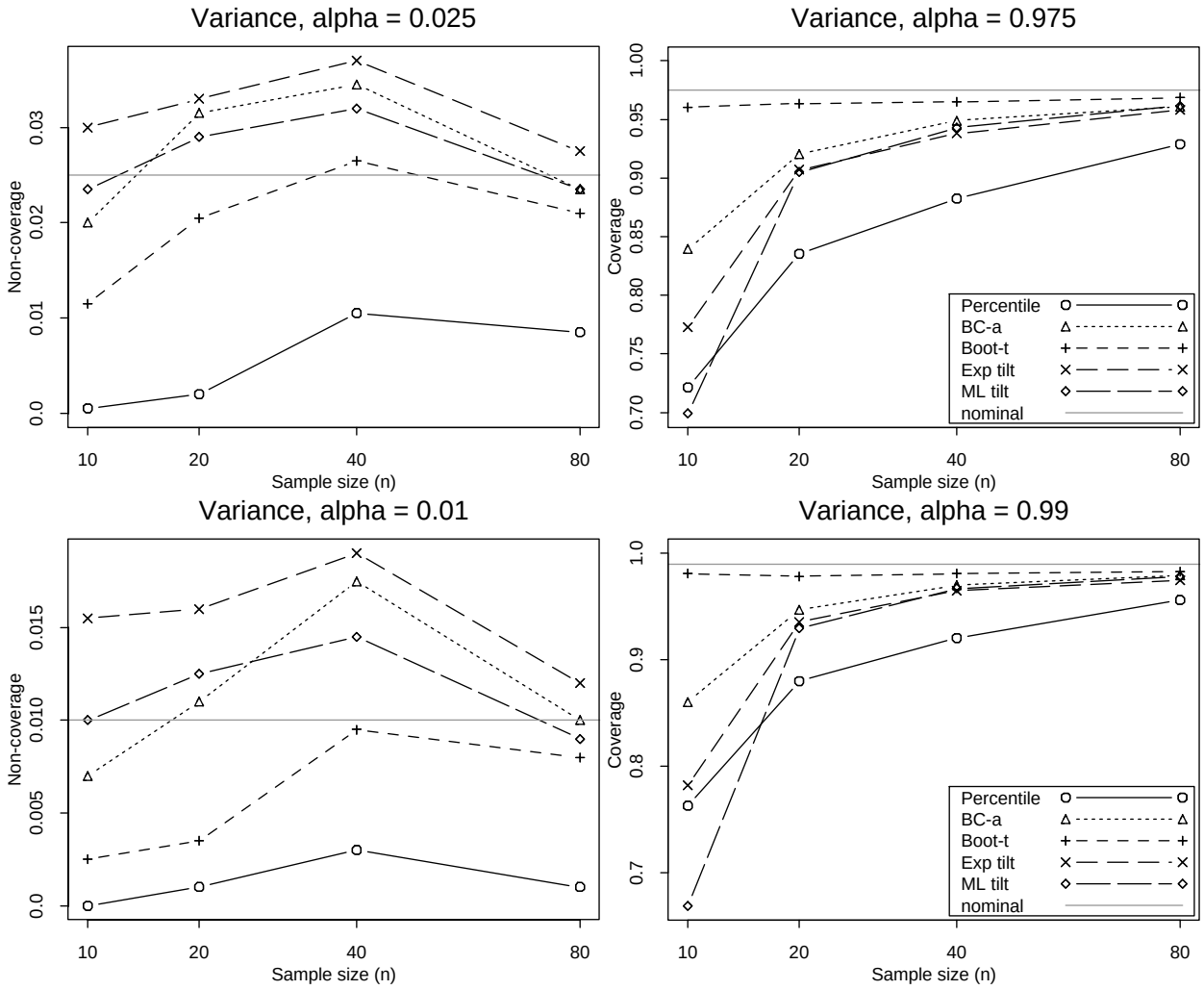

Figure 15: Coverage Accuracy. One-sided confidence intervals for the variance of normal data. Other details are the same as for Figure 10. *Comment*: all undercover very badly for small n at the upper end; the bootstrap-t does the best. The tilting intervals break down for $n = 10$. The percentile method improves much more slowly than other methods as n increases.

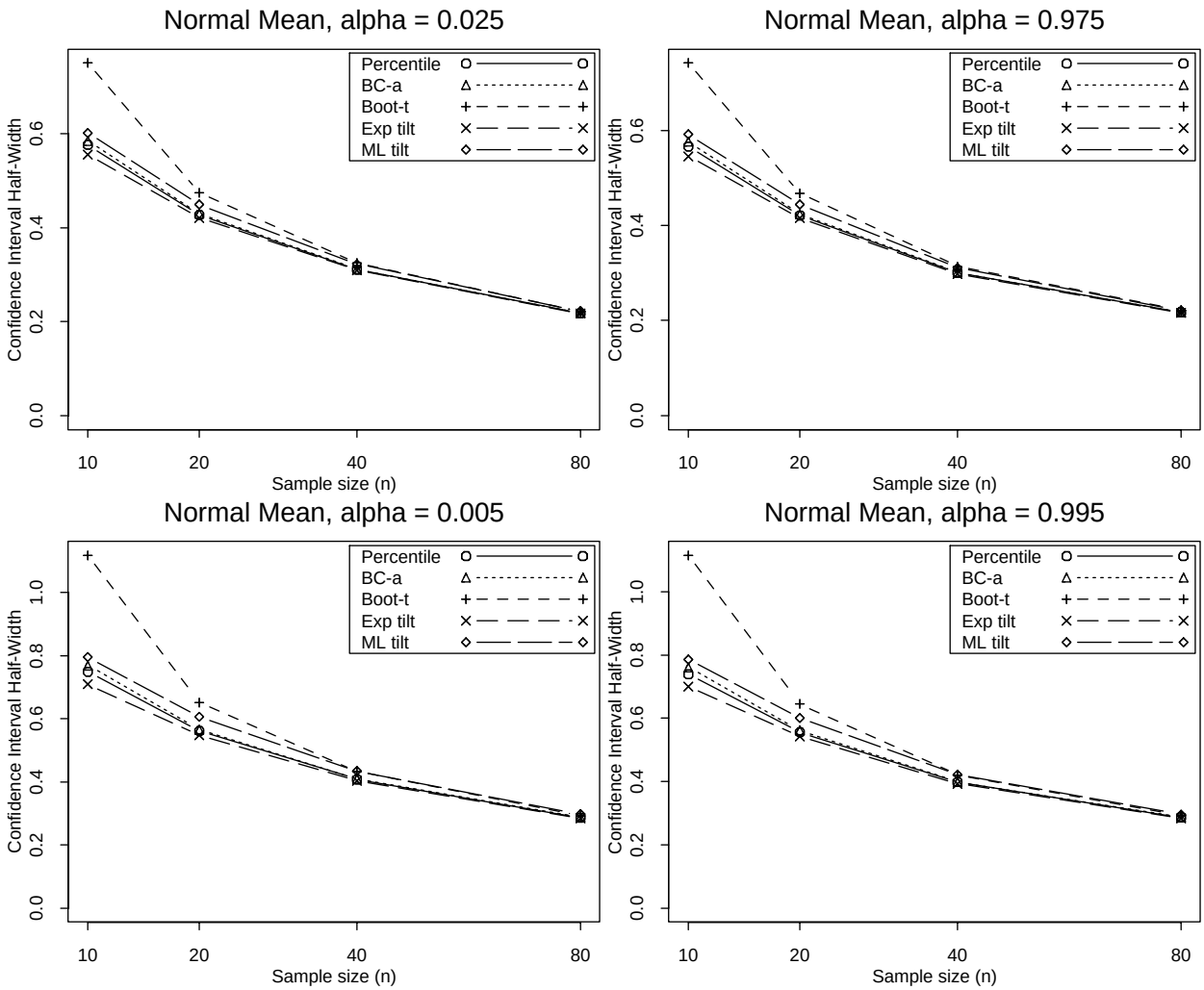

Figure 16: Confidence Interval Length. Average length of one-sided confidence intervals (distance from the estimate to the endpoint of the interval) for the sample mean of normal data. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated and one of each kind of bootstrap confidence interval was generated, using $B = 200$ bootstrap samples for the tilting intervals and $B = 1999$ bootstrap samples for the bootstrap percentile, bootstrap-t, and BC-a intervals. *Comment*: bootstrap-t intervals are the longest.

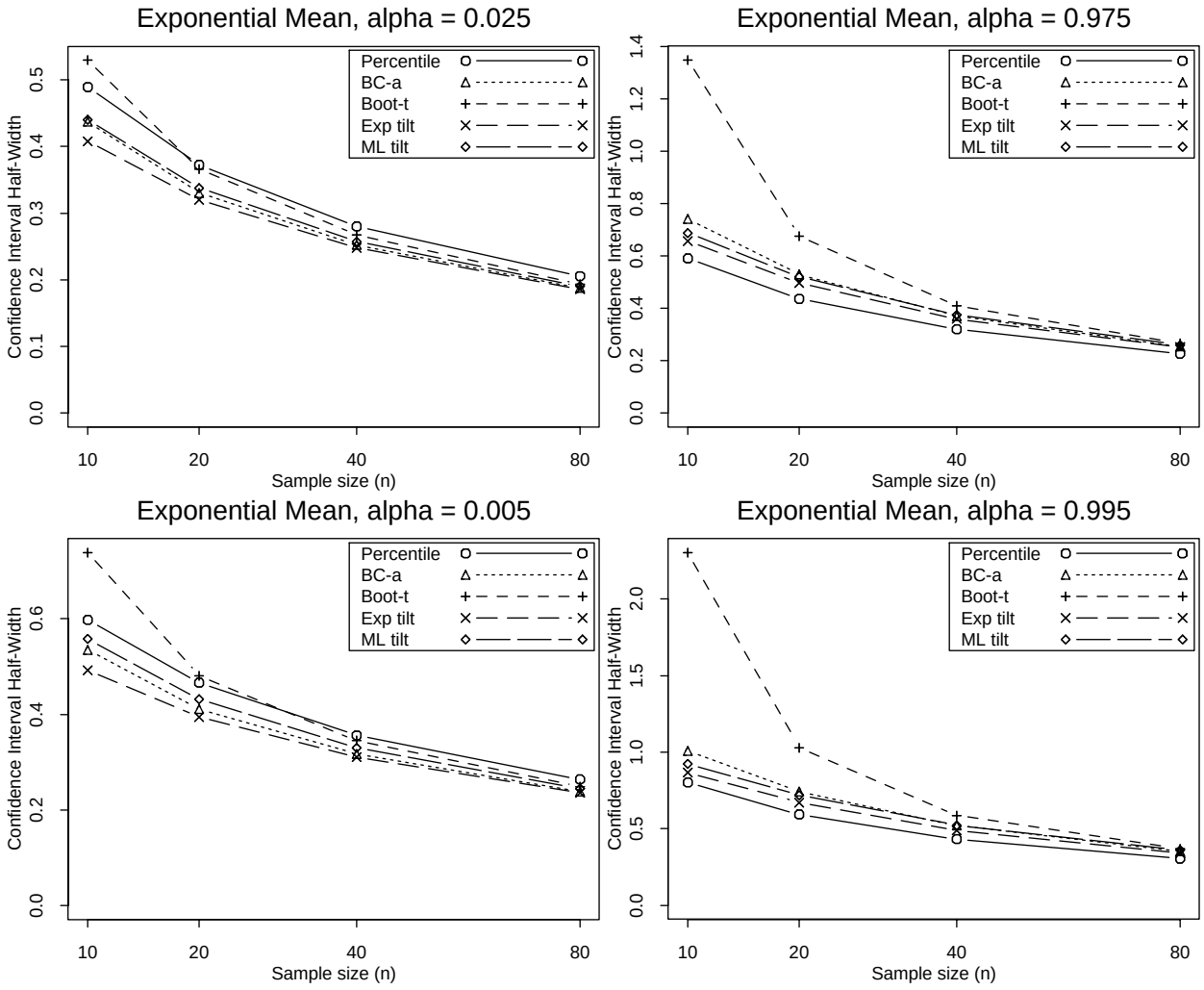

Figure 17: Confidence Interval Length. One-sided confidence intervals for the sample mean of exponential data. Other details are the same as for Figure 16. *Comment*: bootstrap-t intervals are the longest.

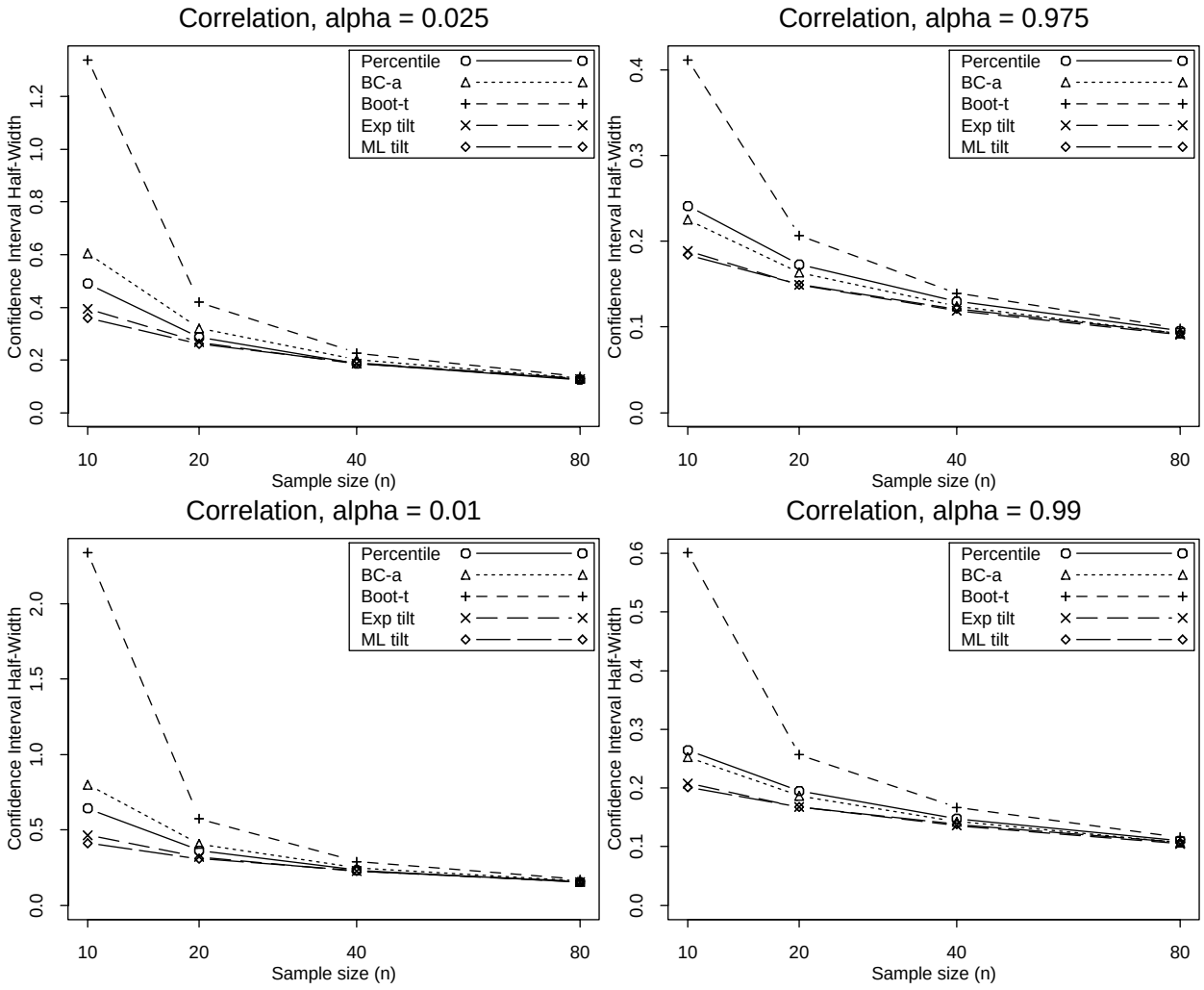

Figure 18: Confidence Interval Length. One-sided confidence intervals for the correlation for bivariate normal data with correlation $(1/2)^{1/2}$. Other details are the same as for Figure 16. *Comment*: bootstrap-t intervals are the longest.

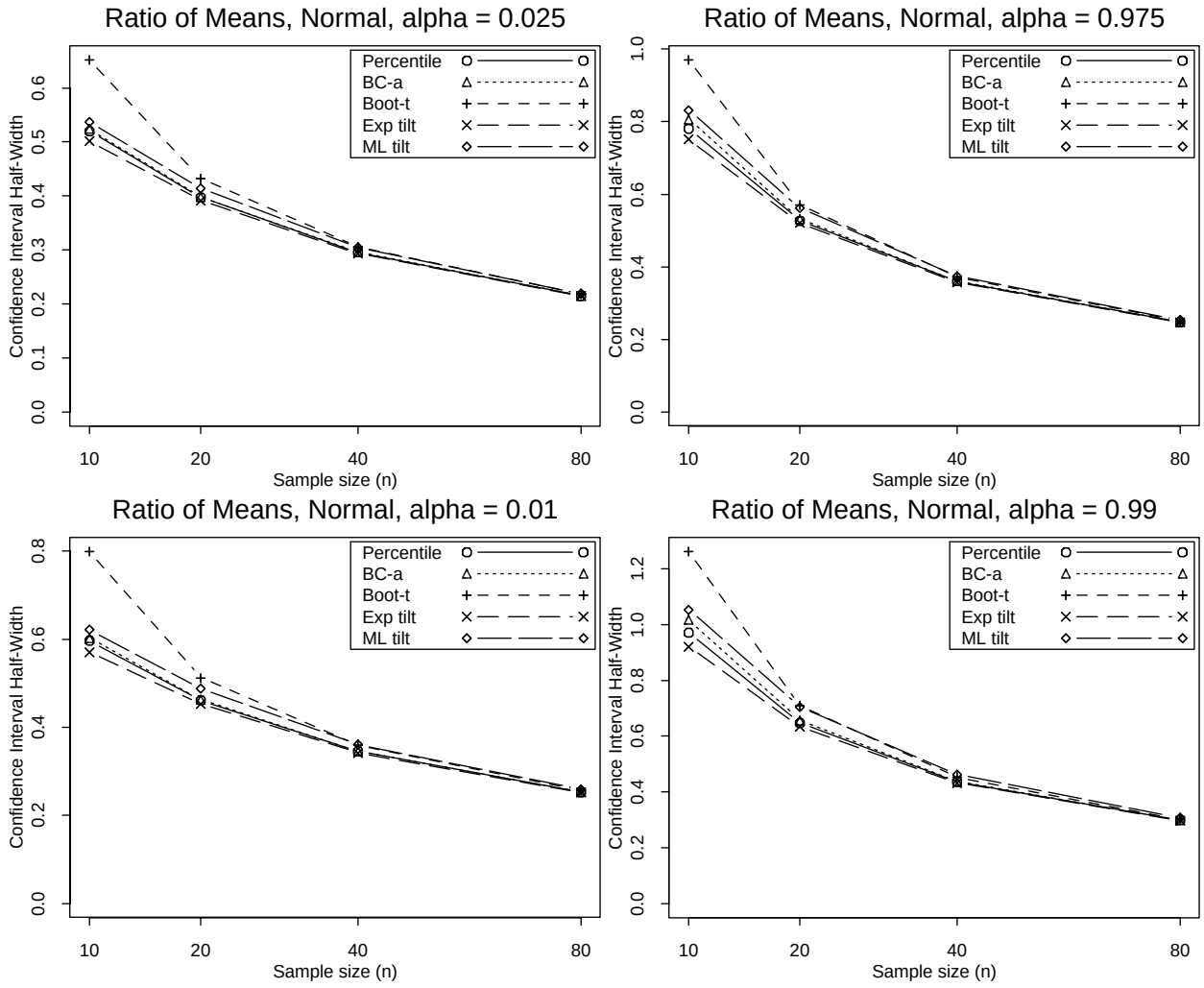

Figure 19: Confidence Interval Length. One-sided confidence intervals for the ratio of means for bivariate normal data (uncorrelated, bivariate mean $(3, 9)$, variance 1) Other details are the same as for Figure 16. *Comment*: bootstrap-t intervals are the longest.

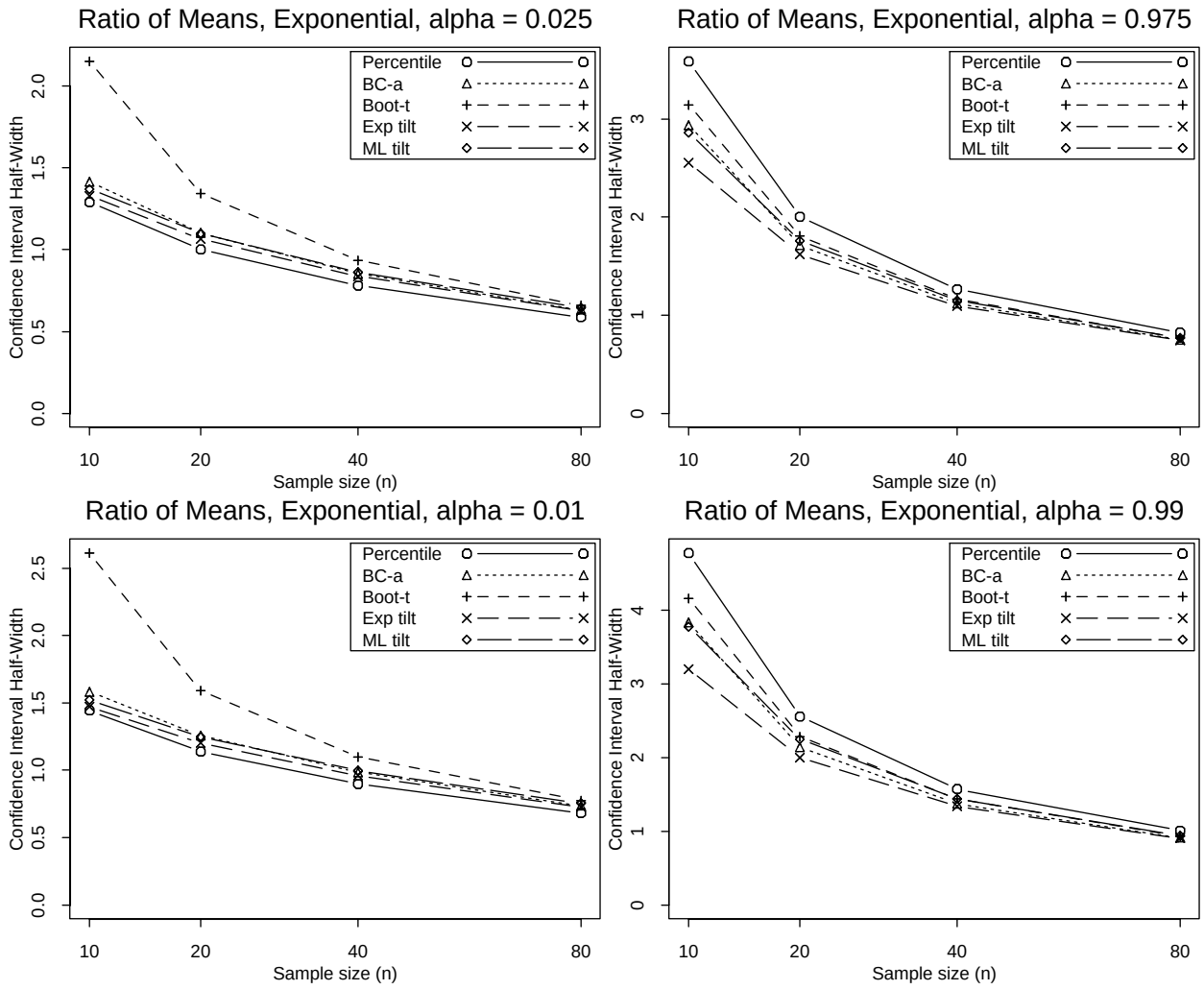

Figure 20: Confidence Interval Length. One-sided confidence intervals for the ratio of means for exponential data (independent, minimum values for x and y are 0 and 2, respectively, standard scale). Other details are the same as for Figure 16. *Comment*: bootstrap-t intervals are the longest overall, though percentile intervals are longer at the upper end.

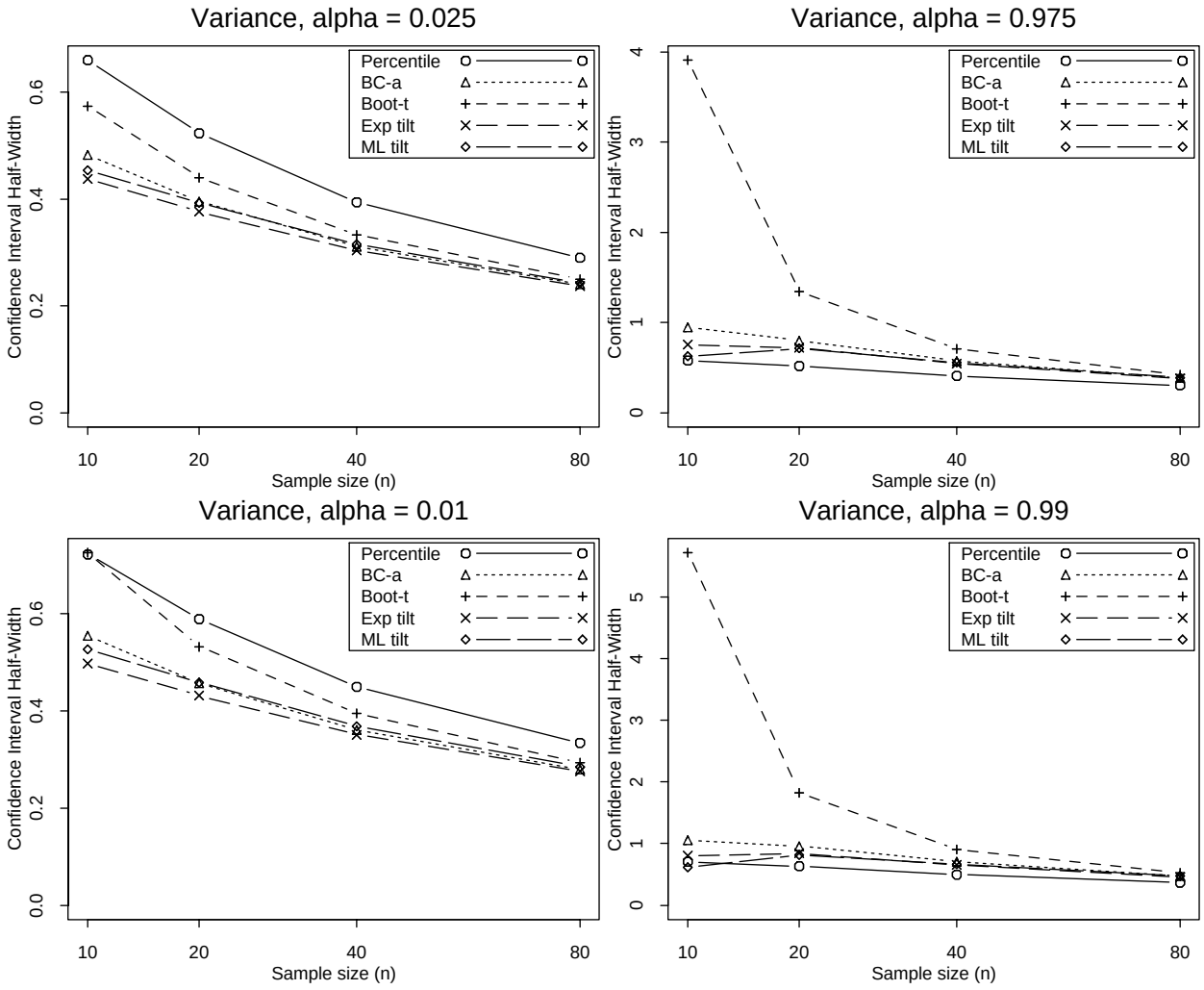

Figure 21: Confidence Interval Length. One-sided confidence intervals for the variance of normal data. Other details are the same as for Figure 16. *Comment*: bootstrap-t intervals are the longest, to an extraordinary degree.

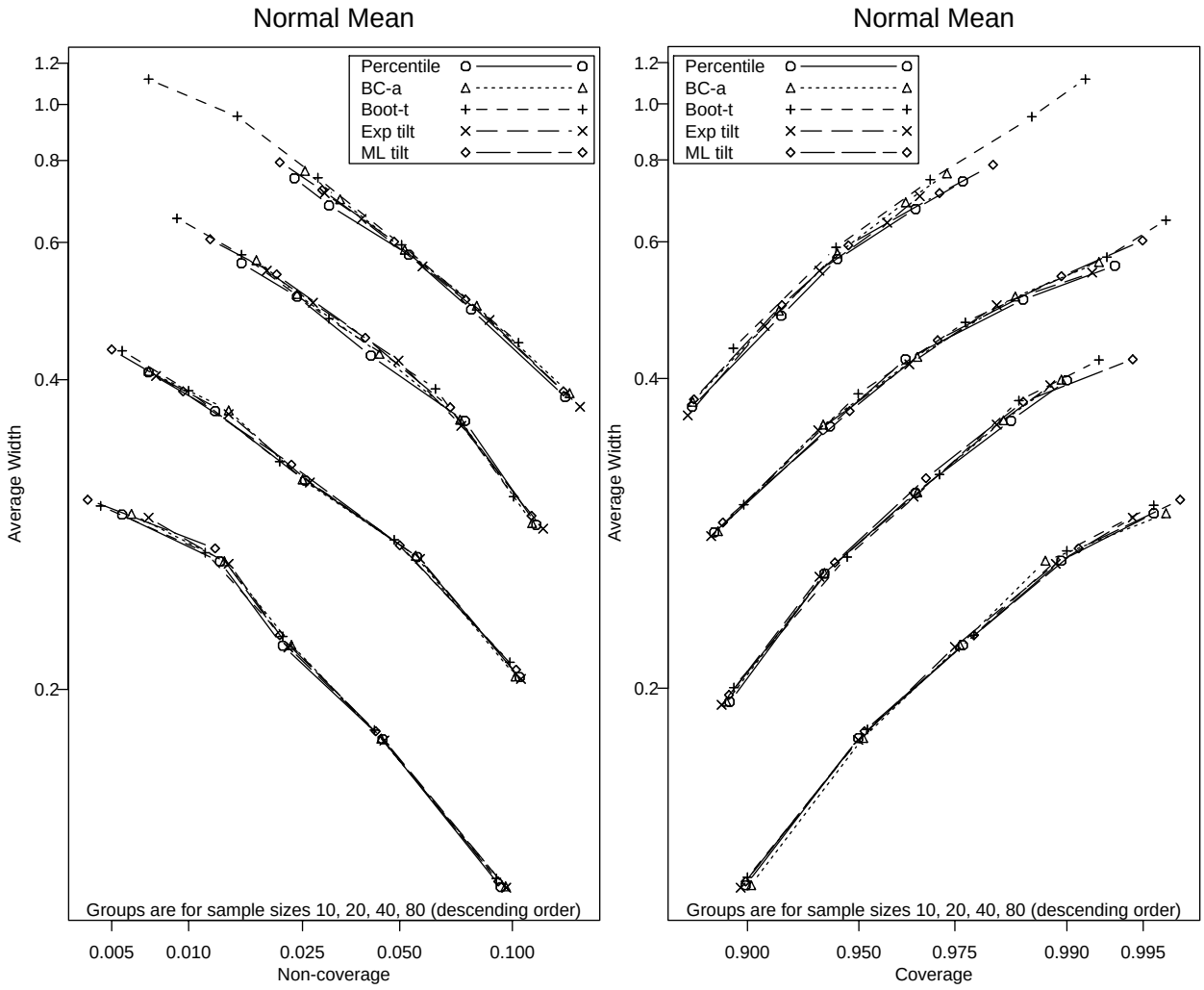

Figure 22: Length/Coverage relationship. Average length of one-sided confidence intervals (distance from the estimate to the endpoint of the interval) vs. Coverage probability for the sample mean of normal data. Tick marks on the x-axis are at the nominal coverage values. Ideally each point would be above the corresponding tick mark. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated and one of each kind of bootstrap confidence interval was generated, using $B = 200$ bootstrap samples for the tilting intervals and $B = 1999$ bootstrap samples for the bootstrap percentile, bootstrap-t, and BC-a intervals.

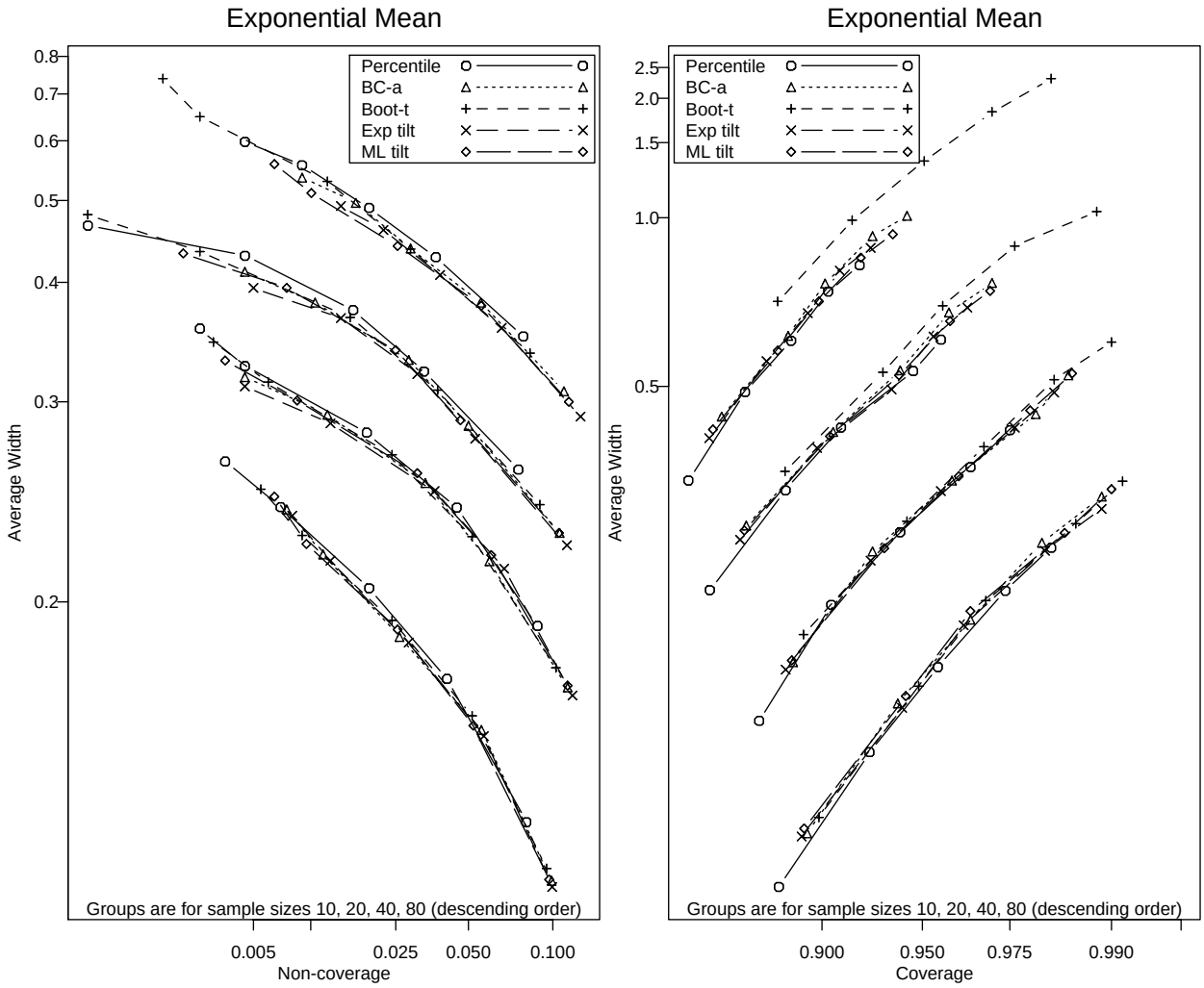

Figure 23: Length/Coverage relationship. One-sided confidence intervals for the sample mean of exponential data. Other details are the same as for Figure 22. *Comment*: bootstrap-t intervals are longer for equivalent coverage.

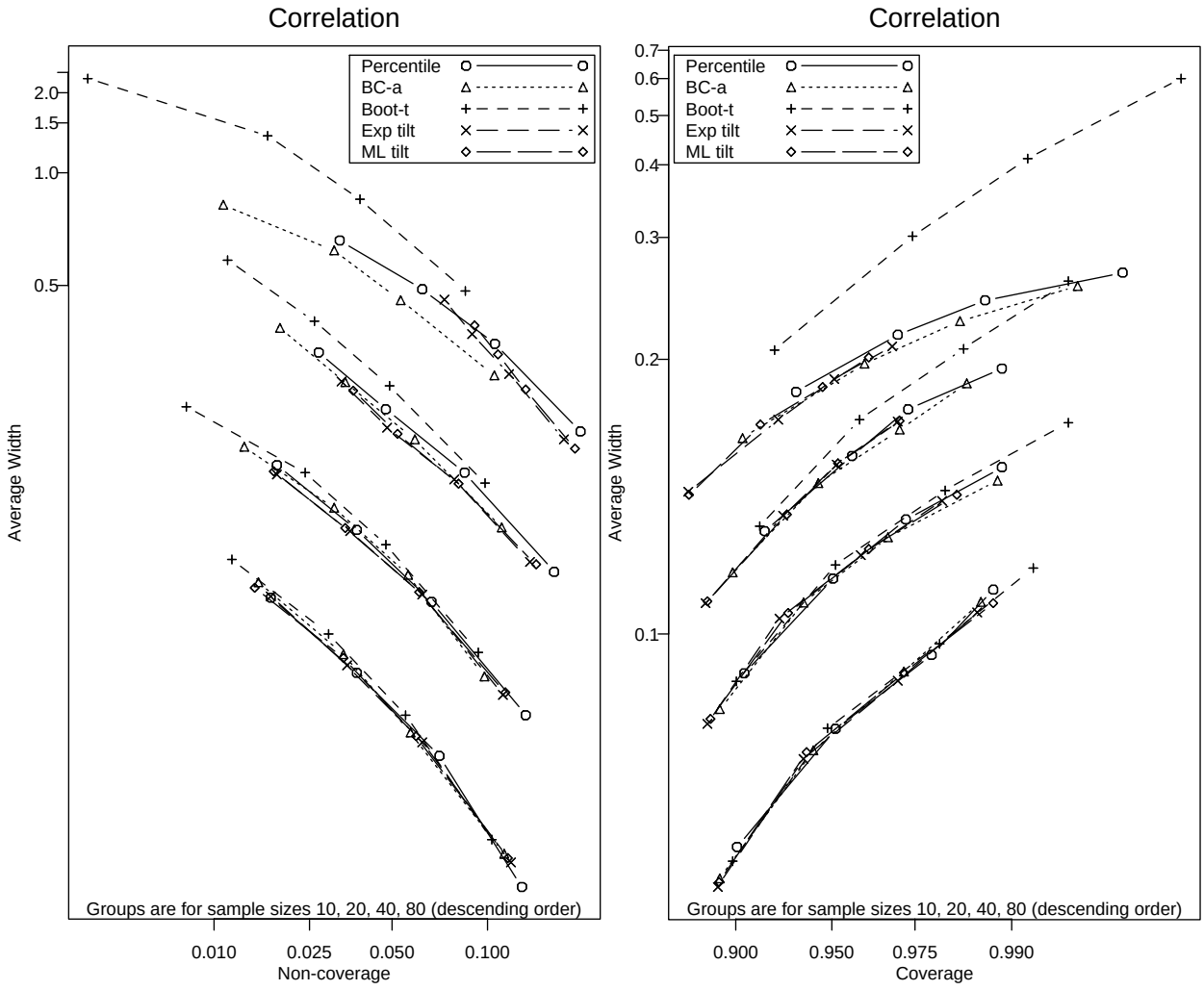

Figure 24: Length/Coverage relationship. One-sided confidence intervals for the correlation for bivariate normal data with correlation $(1/2)^{1/2}$. Other details are the same as for Figure 22. *Comment*: bootstrap-t intervals are substantially longer for equivalent coverage.

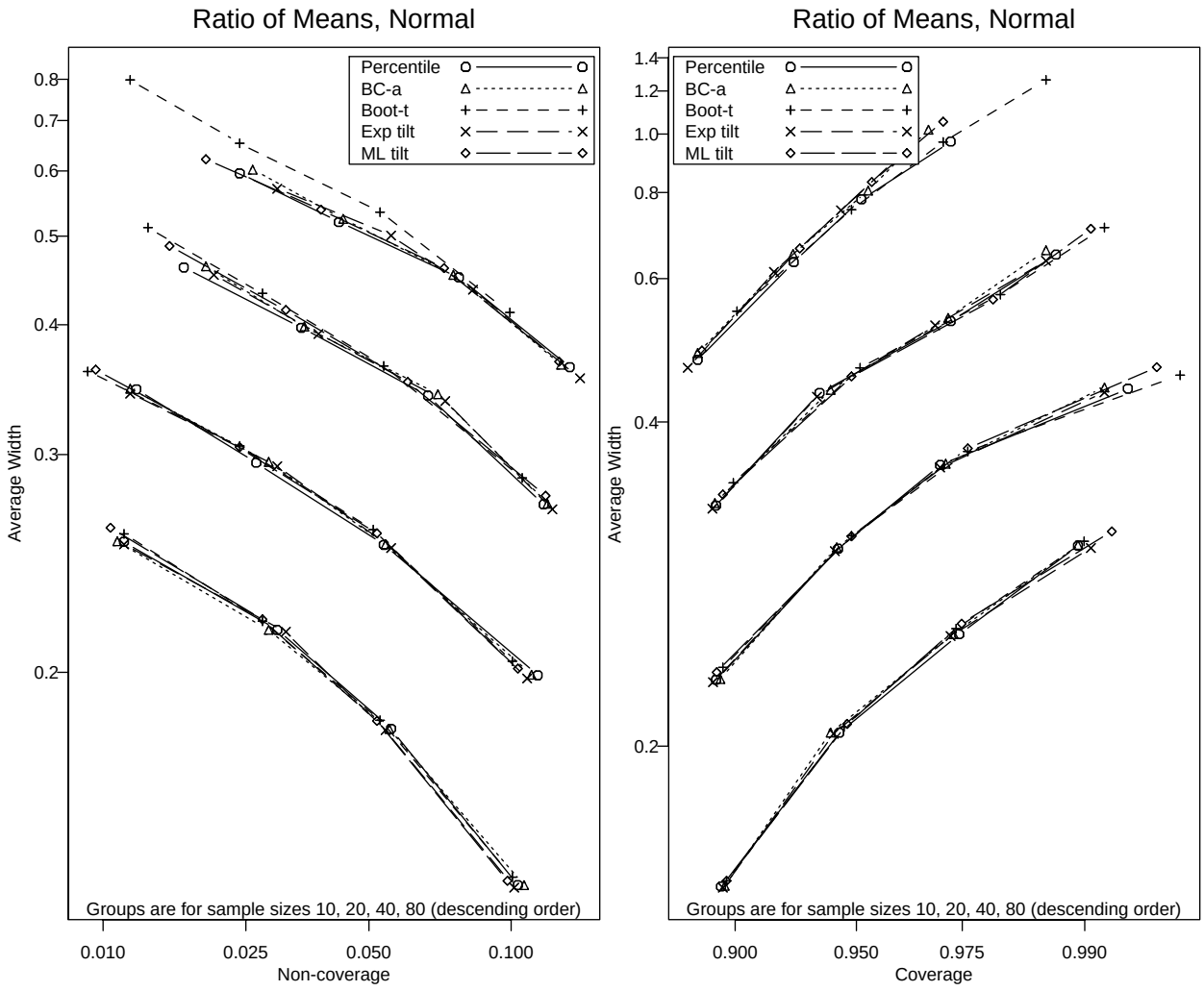

Figure 25: Length/Coverage relationship. One-sided confidence intervals for the ratio of means for bivariate normal data (uncorrelated, bivariate mean $(3, 9)$, variance 1) Other details are the same as for Figure 22. *Comment*: bootstrap-t intervals are longer for equivalent coverage, for small samples at the lower end.

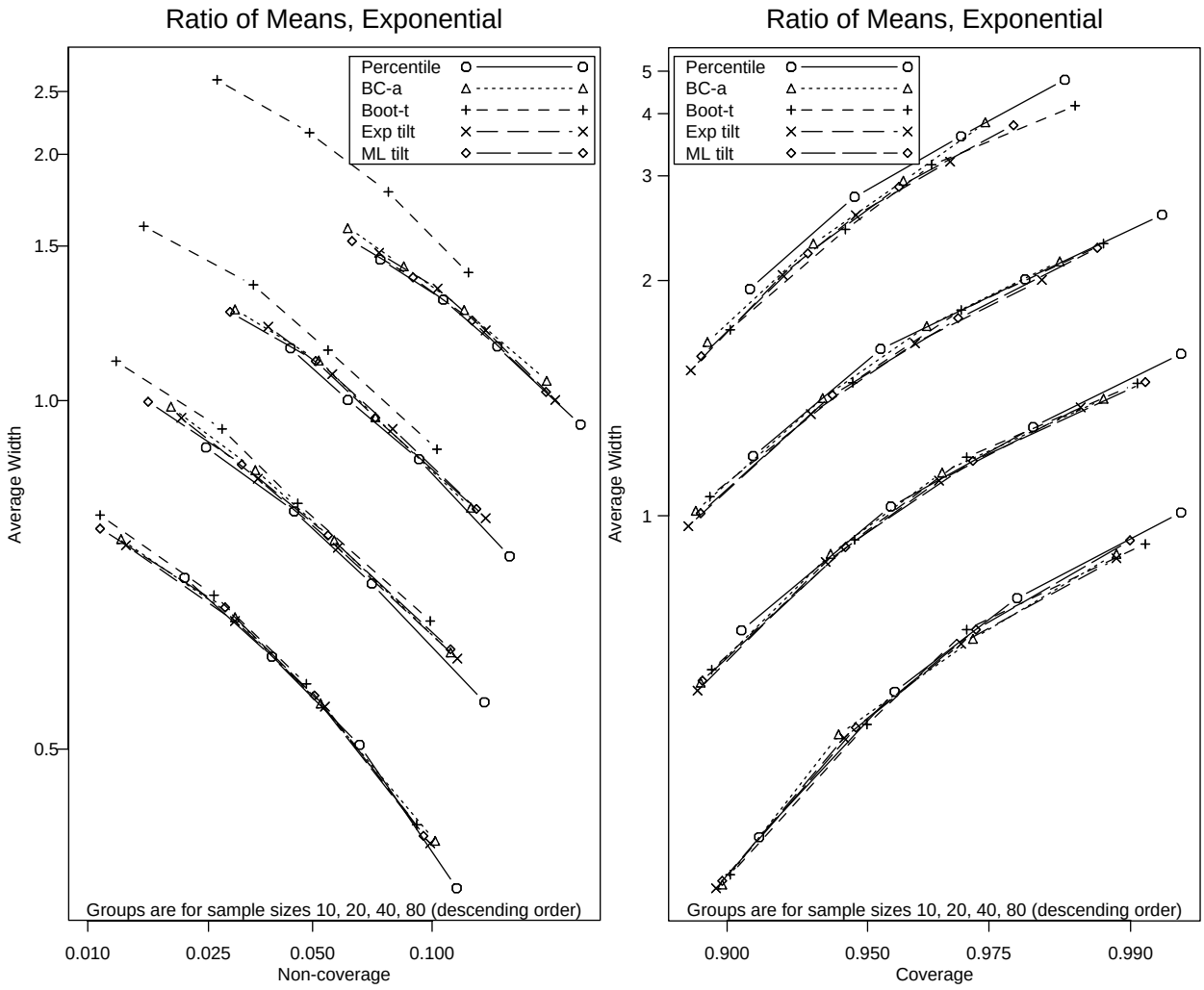

Figure 26: Length/Coverage relationship. One-sided confidence intervals for the ratio of means for exponential data (independent, minimum values for x and y are 0 and 2, respectively, standard scale). Other details are the same as for Figure 22. *Comment*: bootstrap-t intervals are longer for equivalent coverage, at the lower end; percentile intervals are longer at the upper end.

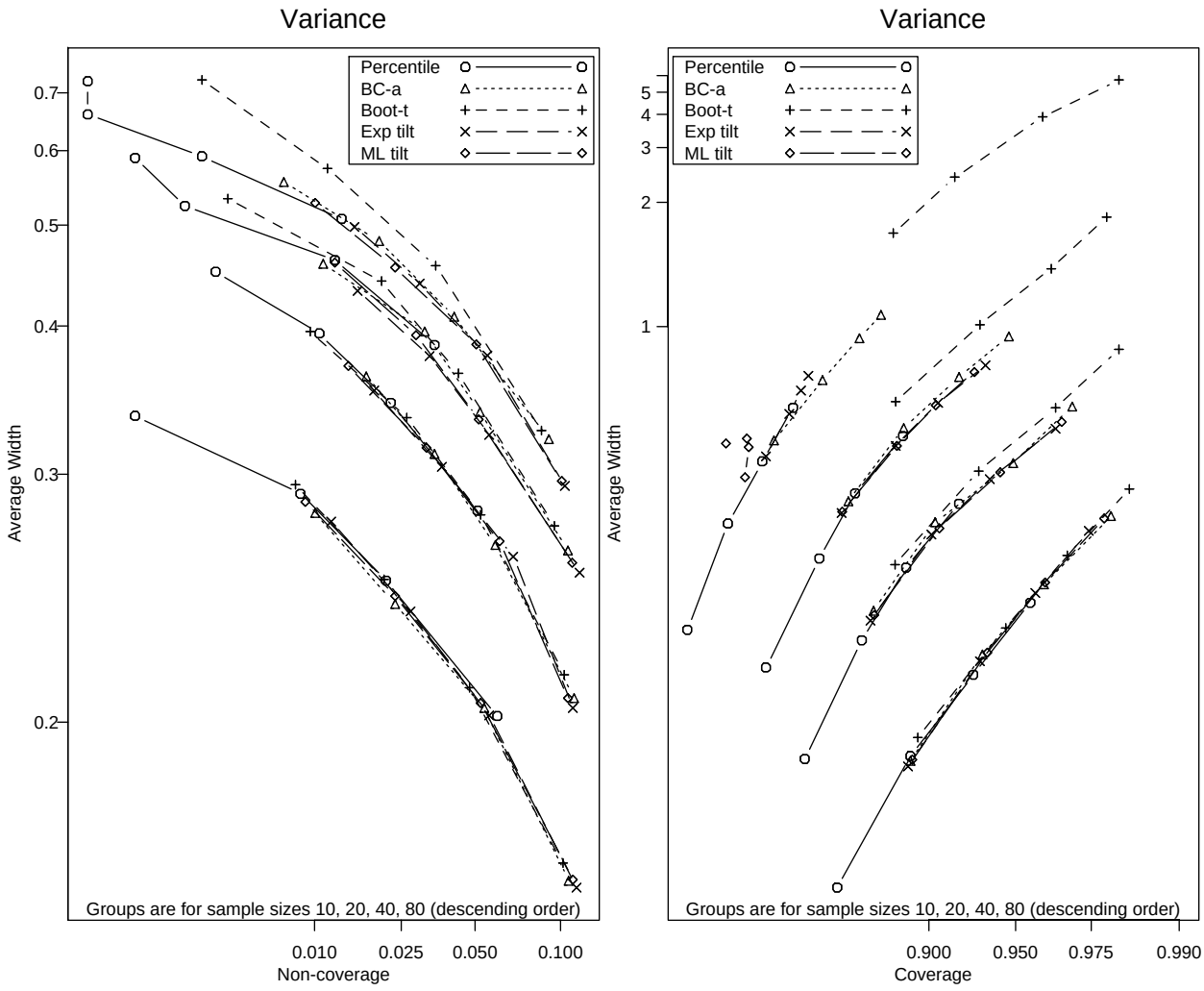

Figure 27: Length/Coverage relationship. One-sided confidence intervals for the variance of normal data. Other details are the same as for Figure 22. *Comment*: bootstrap-t intervals are longer for equivalent coverage, but have much better coverage at the upper end. ML tilting breaks down, failing to increase actual coverage as nominal coverage increases.

5.1 Computational Issues

Exponential tilting offers some computational advantages over ML tilting — the domain for τ is bounded (which simplifies numerical root-finding procedures), and the exponential form makes computing products of weights easier. We have found approximations which are useful as initial values when numerically solving for τ.

Initial values for ML tilting are more difficult — the most effective initial values we have found involve first solving (8) using exponential tilting, then obtaining the starting value for ML tilting by a transformation of the τ from exponential tilting.

While exponential tilting is faster, for reasons outlined in this section and because of lower Monte Carlo variability (see Figure 1), we still recommend ML tilting because of its better coverage accuracy.

6 Updating Derivatives

The second major difference among tilting families (5) is whether a single set of derivatives is used, or whether derivatives are updated as τ changes. In general, using derivatives (6) 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 by (Shao and Tu 1995). these more conservative inferences are usually more accurate.

However, using derivatives that implicitly depend on τ 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 τ.

One compromise involves 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 (Hesterberg 1995a), and in empirical likelihood by (Wood et al. 1996).

Figure 28 shows an example where this works well. The coverage probabilities obtained using one-step updated derivatives are slightly higher, and the confidence intervals are not longer on average.

Figure 29 shows an example where one-step updated derivatives perform very poorly. The right panel here is for the same problem as the right panel of Figure 27, where we see that ML tilting seems to break down when $n = 10$. One-step updating fails to remedy this — in fact it does much worse! However, a fairly simple optimization procedure does work fairly well. Let $\mathbf{U}_0$ and $\mathbf{U}_1$ be the derivatives obtained without updating and with one-step updating. We use

$$\mathbf{U}_\lambda = \lambda \mathbf{U}(\mathbf{p}_0) + (1-\lambda)\mathbf{U}(\mathbf{p}_1). \tag{17}$$

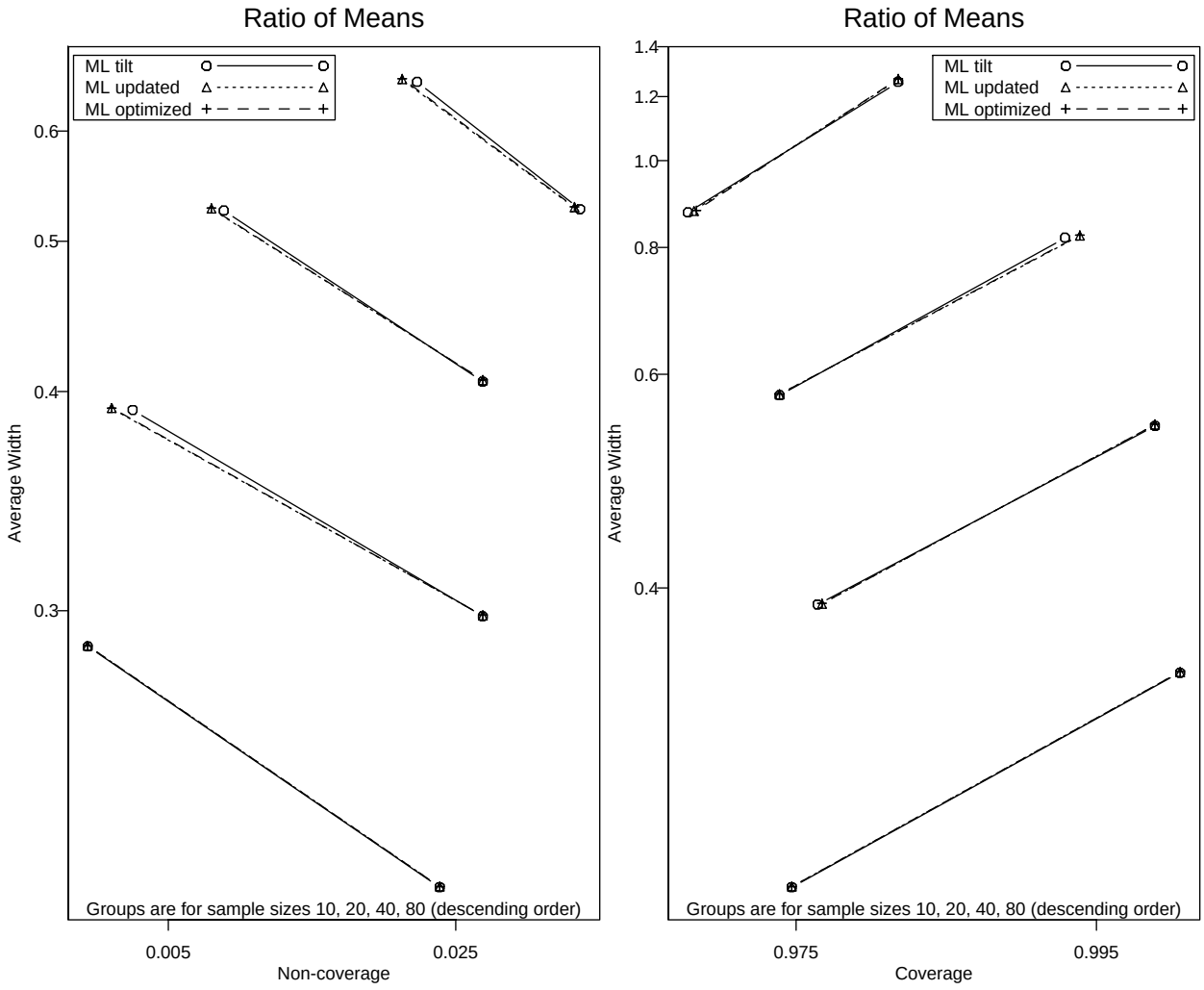

Figure 28: Length/Coverage relationship. Average length of one-sided confidence intervals (distance from the estimate to the endpoint of the interval) vs. Coverage probability for the the ratio of means of bivariate normal data with mean (3,9). Results are for ML tilting, without updating, with one-step updating, and with a 1-dimensional optimization procedure. Tick marks on the x-axis are at the nominal coverage values. Ideally each point would be above the corresponding tick mark. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated and one of each kind of bootstrap confidence interval was generated, using $B = 200$ bootstrap samples for the tilting intervals and $B = 1999$ bootstrap samples for the bootstrap percentile, bootstrap-t, and BC-a intervals.

We choose a number of values of λ between 0 and 1, calculate the ML tilting interval using each corresponding $\mathbf{U}_\lambda$, then use the widest interval. This is a quick-and-dirty method of maximizing the coverage of the interval with respect to λ.

One-step updating also performs badly when the statistic is a bivariate correlation. Additional updating, without optimization, fails for both correlation and the sample variance; in fact the derivatives diverge, varying more wildly with successive iterations. Figure 30 shows an example dataset where divergence occurs for the bivariate correlation. Consider are tilting to obtain a lower limit, with $\tau < 0$. In the left panel, observation 1 has the most negative value of $U_i(\mathbf{p}_0)$; adding weight to this point *initially* decreases the weighted correlation faster than adding weight to any other point. The values of U_i are shown by arrows, with arrows pointing downward indicating negative values of U. Using those values of U_i as derivatives for exponential tilting for a one-sided 99% confidence interval gives the probabilities $\mathbf{p}_\tau$ which are shown as the areas of circles. Note that observation 1 receives the largest weight. Updating by computing the gradient from this set of weights yields the derivatives shown by the right set of arrows; now observation 5 has the most negative derivative; adding more weight to that point causes the fastest decrease in correlation *from this weighted distribution*. However, using that set of derivatives for tilting from scratch causes too much weight to be placed on observation 5 and too little on observation 1. Updating additional times would cause the gradients to oscillate with increasing magnitudes, with the largest negative derivative alternating between observations 1 and 5.

Various solutions to this oscillatory behavior are possible. Our preferred solution is the the optimization procedure (17), because it is relatively cheap — the optimization requires evaluating the statistic θ a number of times, but does not require additional gradients (beyond the two sets already computed).

Other solutions to this oscillatory behavior are possible. Solutions adapted from procedures for numerically solving differential equations (e.g. Runge-Kutta) would be more in spirit with the continuous updating implied by (5), but would require additional gradients, which is expensive if the gradients must be obtained numerically. Cheapest of all would be to let λ depend on the angle between the two sets of derivatives — when the angle is large we would choose λ to be large. But the angle alone does not contain enough information about the geometry of the problem to determine the optimal λ; the geometry varies from problems like the ratio of means, where no dampening is ideal, to problems like correlation and the upper tail for sample variance, where dampening is needed.

Updating is only necessary with small samples and nonlinear statistics. The degree of nonlinearity can be diagnosed using the correlation of $\hat{\theta}_b^*$ and the linear approximation

$$L_b^* = n^{-1} \sum M_i^* U_i. \tag{18}$$

In general the residuals obtained from regressing $\hat{\theta}_b^*$ against L^* are of order $O(1/n)$ (and the ratio of residual to total variance is $O(1/\sqrt{n})$, so that in large samples, where gradients are more expensive to compute, the problem is effectively more linear (Hesterberg 1995a).

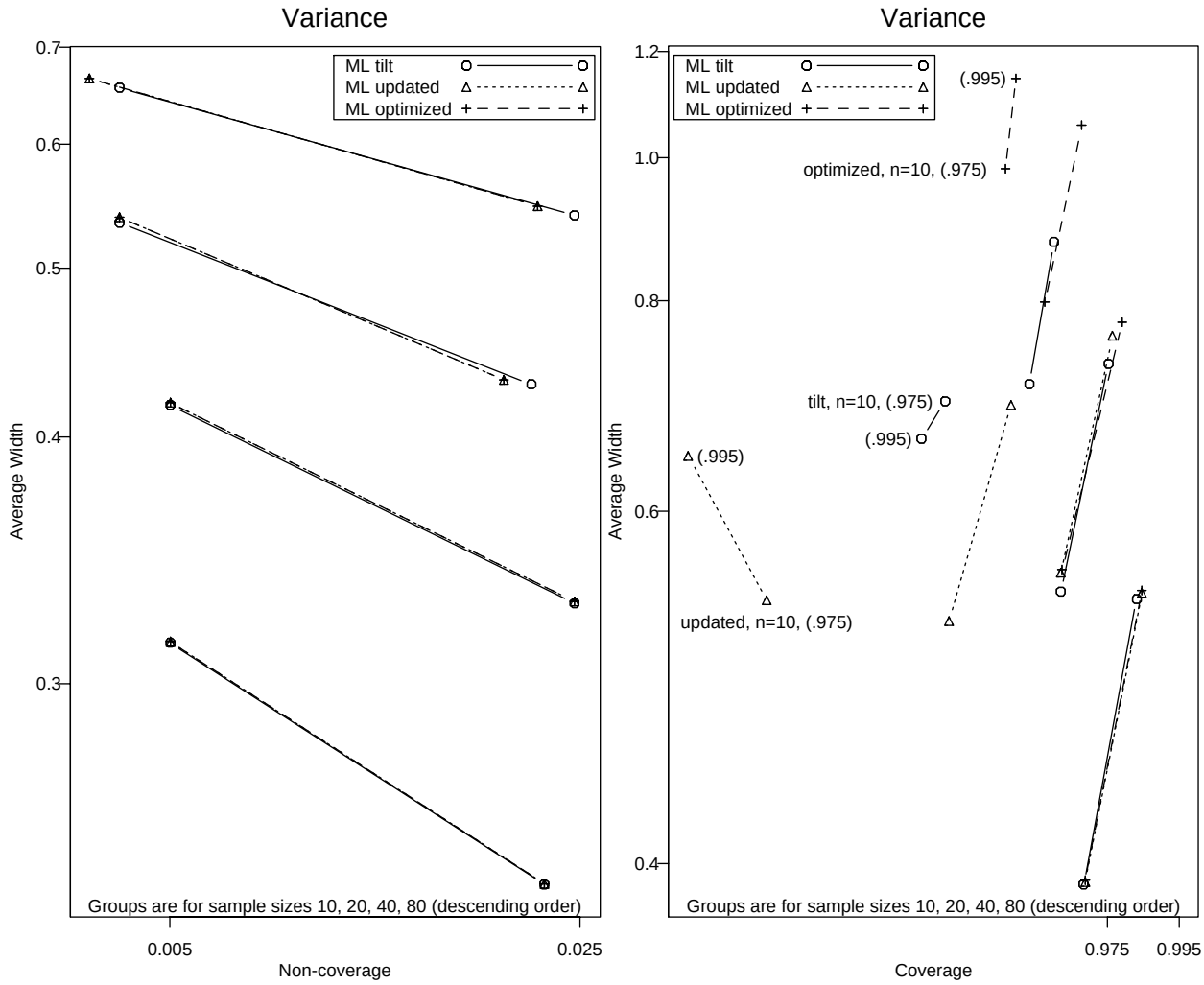

Figure 29: Length/Coverage relationship. One-sided confidence intervals for the variance of normal data. Other details are the same as for Figure 28.

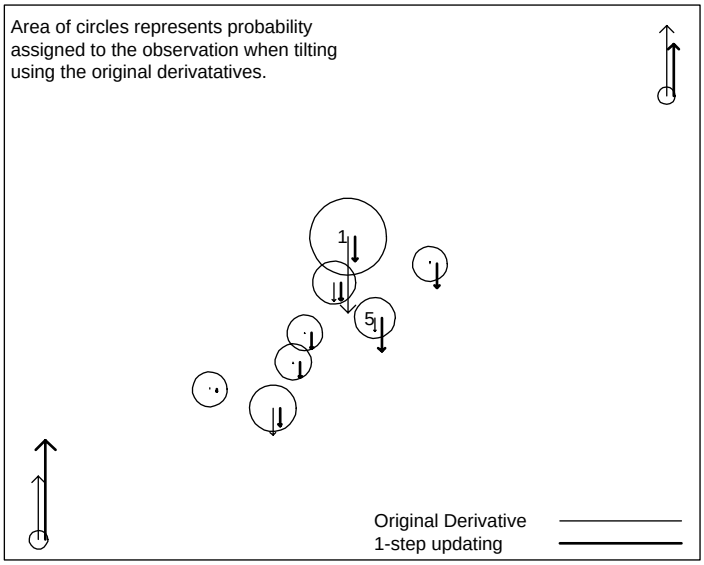

Figure 30: Bivariate dataset. The center of each circle is one observation. The left arrows represent the gradient of the correlation computed at the unweighted empirical distribution. The areas of the circles show the probability assigned to each point after tilting for a lower 99% interval. The right arrows show the gradient computed at the weighted distribution. Observations 1 and 5 are labeled; these have the the most negative derivative in the original and updated gradients, respectively. Divergent oscillation occurs if the process of tilting and computing new gradients is repeated, with observations 1 and 5 alternating in receiving the largest weight.

We propose to use the correlation to diagnose whether updating is necessary. We have not performed simulations to test this.

Linear approximations and Updating Tilting requires the gradient $\mathbf{U}(\mathbf{p}_0)$, and additional gradients if updating methods are used. In some cases analytical derivatives are available. Otherwise gradients may be evaluated by the jackknife and other finite-difference methods, which require evaluating the function an additional n times, or regression methods, which are accurate if B is substantially larger than n (Efron 1990; Hesterberg 1995a). Hesterberg and Ellis (1999) describe a regression procedure which does not require that B be larger than n.

Tilting and Empirical Likelihood $\mathcal{F}_4$ corresponds to maximum likelihood estimation subject to a null hypothesis, and is the family used in empirical likelihood (EL) inference (Owen 1988; Owen 1990; Hall and La Scala 1990); both limit support to the observed values and find the restricted maximum likelihood vector of probabilities. But where EL inference is based on an asymptotic approximation, that the log-likelihood-ratio statistic $-2\log((1/n)^n/\sum_{i=1}^n p_i)$ is approximately distributed as a χ^2_1 random variable in bootstrap tilting the probability (8) is estimated by sampling. Davison et al. (1992) study bootstrap likelihood and EL, and discuss relative advantages of EL and bootstrap methods, and Hesterberg (1997) discusses connections between the bootstrap and EL.

7 Coverage-level Adjustments

We have observed that tilting and other bootstrap procedures tend to under-cover. In this section we discuss two ways to adjust the intervals to obtain better coverage probabilities. We demonstrate using tilting intervals, though the same methods can be used with other bootstrap intervals.

Both adjustments are motivated by the usual Student's-t interval for a univariate mean, $\bar{x} \pm t_{n-1,\alpha} s/\sqrt{n}$. We note two features of this interval:

- The standard error is based on $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$ rather than the functional version of the variance $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$, and
- the use of t quantiles rather than normal quantiles.

Both features produce intervals with higher coverage probability. We discuss these in turn. Both cause changes in coverage level of $O(1/n)$, so do not affect the second-order accuracy of bootstrap tilting, BC-a, or bootstrap-t intervals.

The usual sample variance s^2 is unbiased for the population variance (under the usual i.i.d. assumptions). In contrast, bootstrap sampling corresponds to using the slightly smaller $\hat{\sigma}^2$ (Efron and Tibshirani 1993). In (Hesterberg 1999) we discuss two methods of bootstrap

sampling that provide unbiased variances in this setting, and generally provide wider, less-biased bootstrap distributions. One of these, bootknife sampling, is useful for bootstrap tilting. Each bootstrap sample is created by sampling n observations with replacement from a jackknife sample created by omitting one observation from the original data. Across B bootstrap samples, each original observation should be omitted an equal number of times. Simulation results in (Hesterberg 1999) demonstrate that this does provide conference intervals with better coverage, though still short of the nominal coverage.

We find, curiously, that bootknife sampling appears to have relatively little effect on the coverage of tilting intervals. Figures 31 and 32 show that for the bootstrap percentile and BC-a intervals, using bootknife sampling gives higher coverage probabilities, but any effect for tilting is more subtle. This may have to do with the importance sampling implementation, in one of two ways — tilting results may depend more on the center of the design distribution than on its spread, or bootknife sampling in the design distribution may not matter because the target distribution F_τ does not involve bootknife sampling. Further investigation could determine the effect of bootknife sampling in the target distribution as well (this could also be implemented using importance sampling reweighting).

The second feature of Student's t intervals is the use of t quantiles rather than normal quantiles. The usual explanation for this is that using t quantiles gives exact confidence intervals if the underlying distribution is normal. That is somewhat specious, as in practice no distribution is exactly normal. However, the t quantiles do provide a real practical benefit, of providing wider intervals, which tend to be more accurate in a wide variety of applications. We can achieve equivalent results using normal rather than t quantiles, but with adjusted nominal coverage levels $\alpha^\dagger$ such that

$$t_{n-1,\alpha} = z_{\alpha^\dagger}. \tag{19}$$

Using these same adjusted values $\alpha^\dagger$ in place of the nominal α is effective when computing bootstrap intervals. Figures 33–37 illustrate the consistent and dramatic improvement this "$z : t$" adjustment brings about, especially for small sample sizes.

The success of this method suggests that a similar adjustment could be used to capture the first feature of Student's t intervals, the use of s rather than $\hat{\sigma}$. The corresponding adjusted α is found by solving for $\alpha^\ddagger$ in

$$st_{n-1,\alpha} = \hat{\sigma} z_{\alpha^\ddagger} \tag{20}$$

(note that $(n-1)s^2 = n\hat{\sigma}^2$). We have performed no simulations using this adjustment.

8 Summary

In summary,

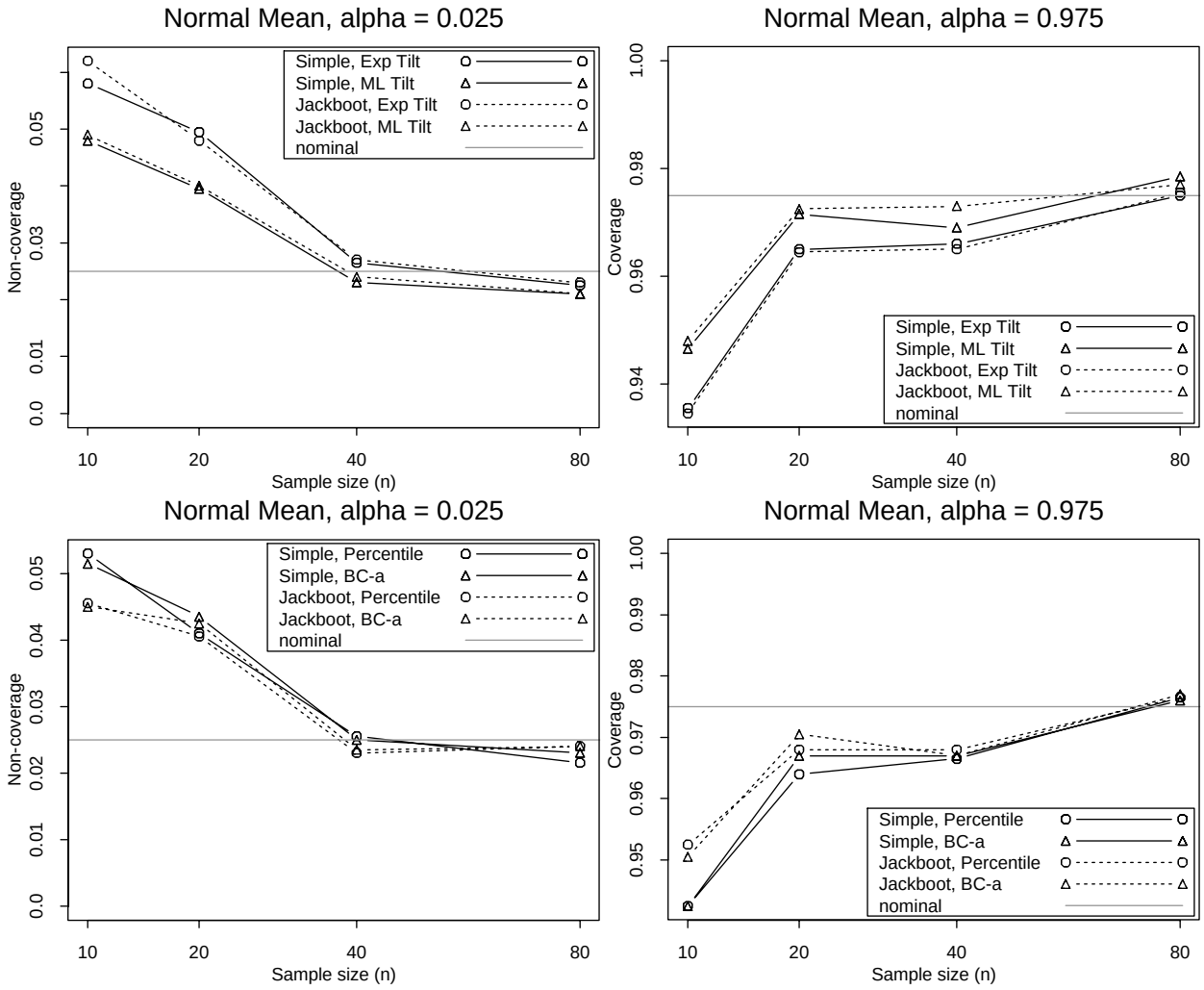

Figure 31: Coverage Accuracy – bootknife sampling. One-sided confidence intervals for the sample mean of normal data. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated and one of each kind of bootstrap confidence interval was generated, using $B = 200$ bootstrap samples for the tilting intervals and $B = 1999$ bootstrap samples for the bootstrap percentile, and BC-a intervals. The standard errors are approximately $(.025 \times .975/2000) = .0035$ for intervals with nominal coverage of 0.025 or 0.975. *Comment*: bootknife sampling improves coverage for percentile and BC-a intervals but shows little effect on tilting intervals.

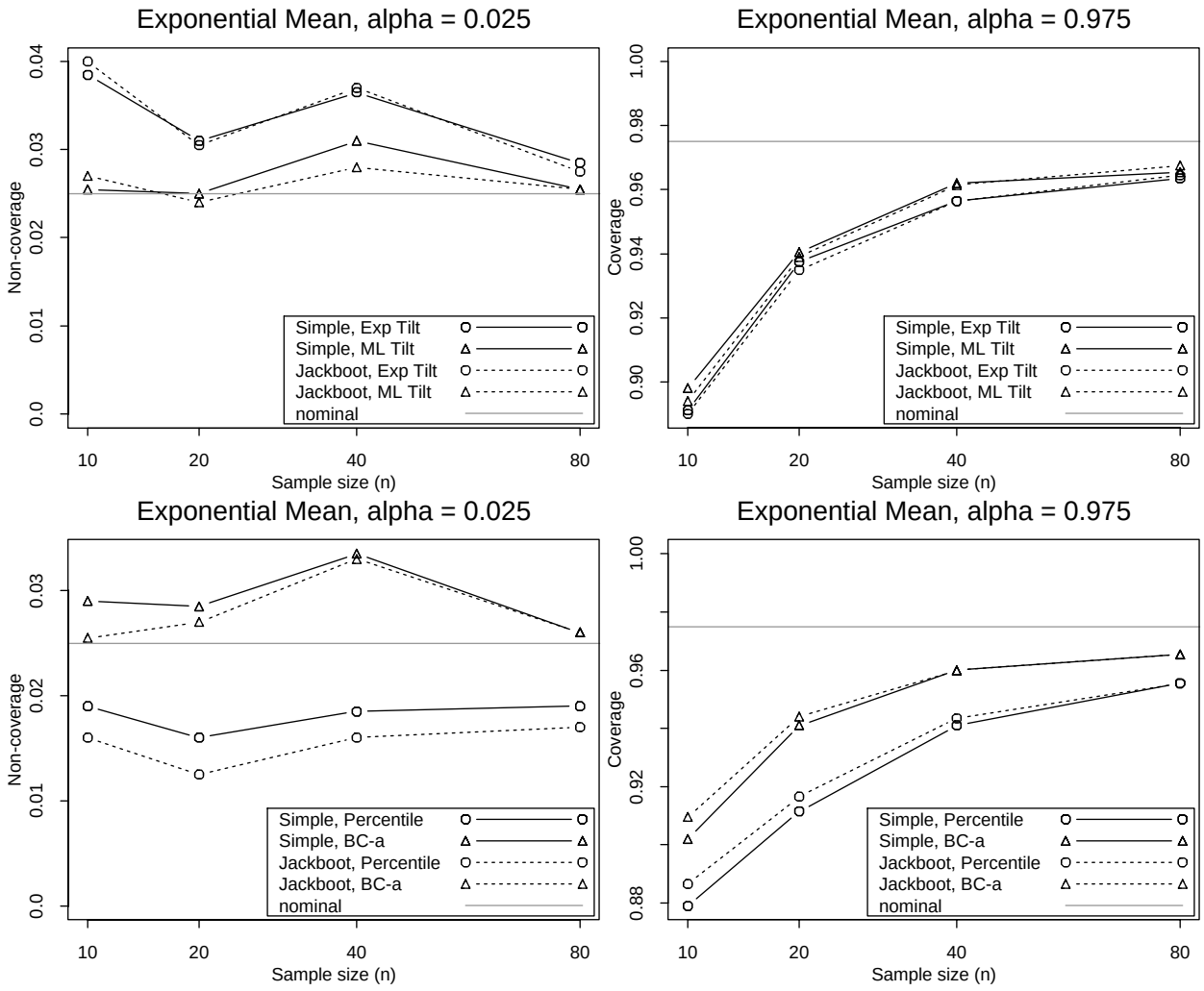

Figure 32: Coverage accuracy – bootknife sampling. One-sided confidence intervals for the sample mean of exponential data. Other details are the same as for Figure 31. *Comment*: bootknife sampling improves coverage for percentile and BC-a intervals but shows little effect on tilting intervals.

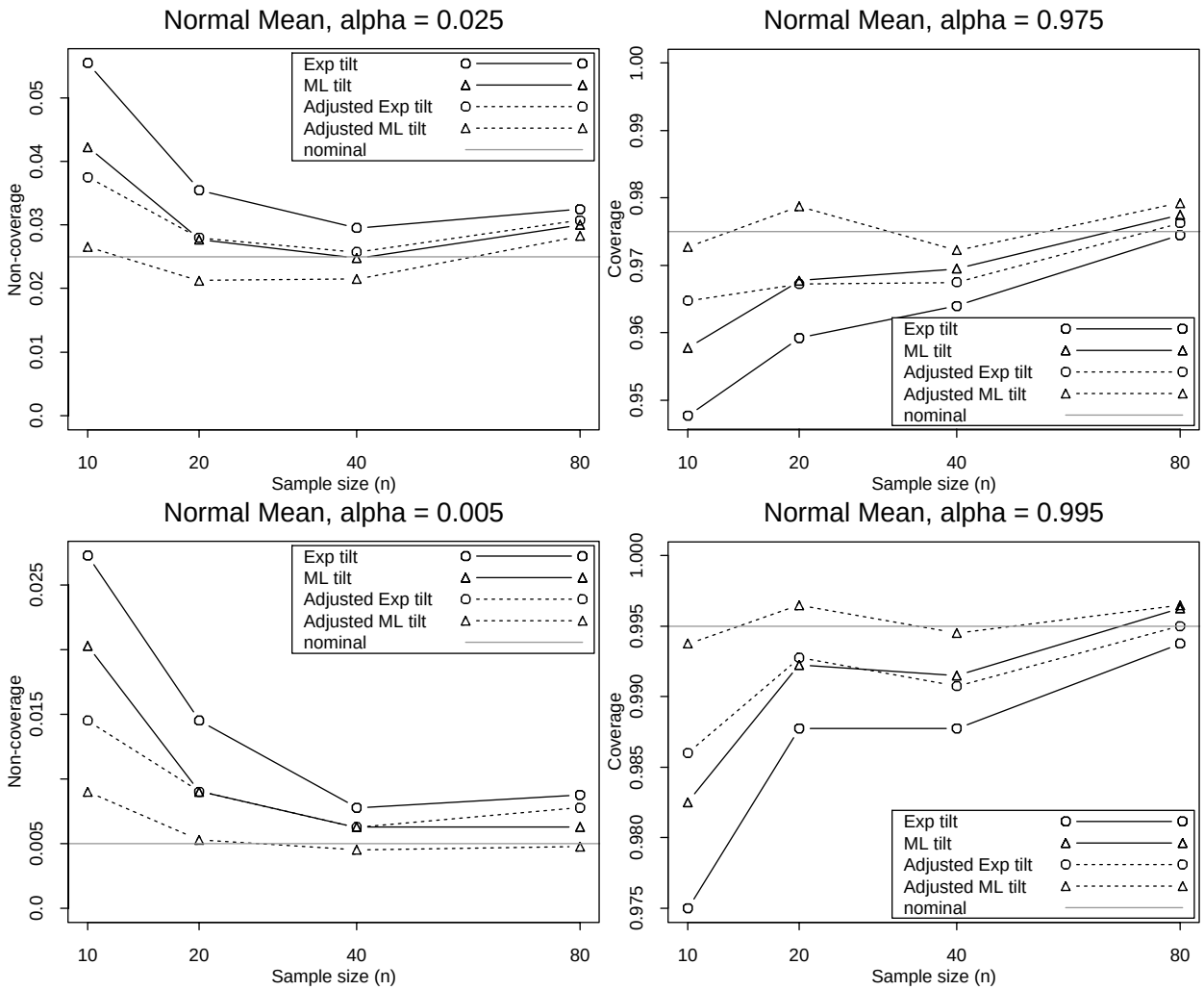

Figure 33: Coverage Accuracy – effect of z : t adjustment of nominal levels. One-sided confidence intervals for the sample mean of normal data. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated and one of each kind of bootstrap confidence interval was generated, using $B = 200$ bootstrap samples. The standard errors are approximately $(.025 \times .975/2000) = .0035$ for intervals with nominal coverage of 0.025 or 0.975. *Comment*: the adjustment improves coverage probabilities, particularly for small samples.

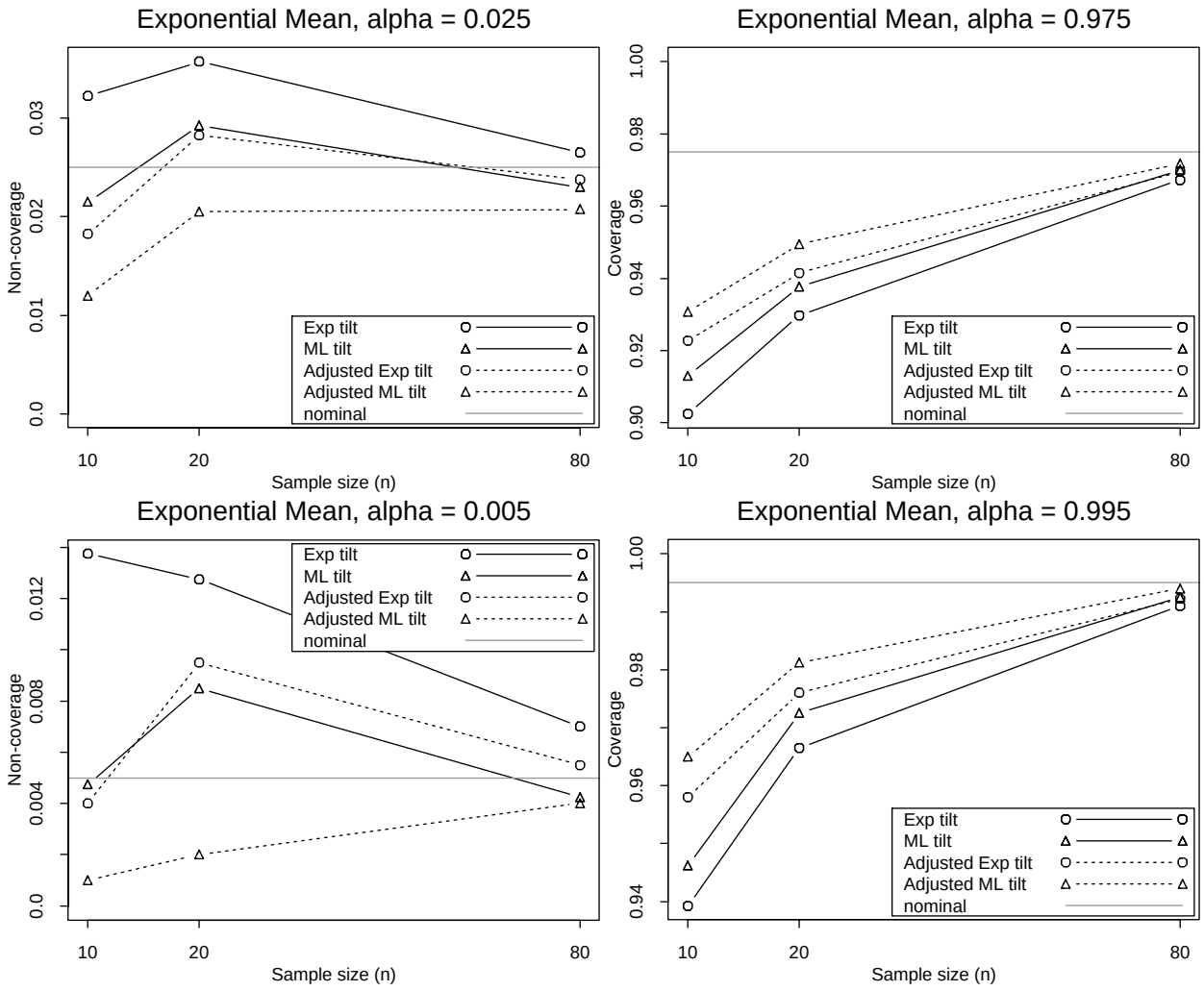

Figure 34: Coverage accuracy – effect of z : t adjustment of nominal levels. One-sided confidence intervals for the sample mean of exponential data. Other details are the same as for Figure 33. *Comment*: the adjustment improves coverage probabilities, particularly for small samples.

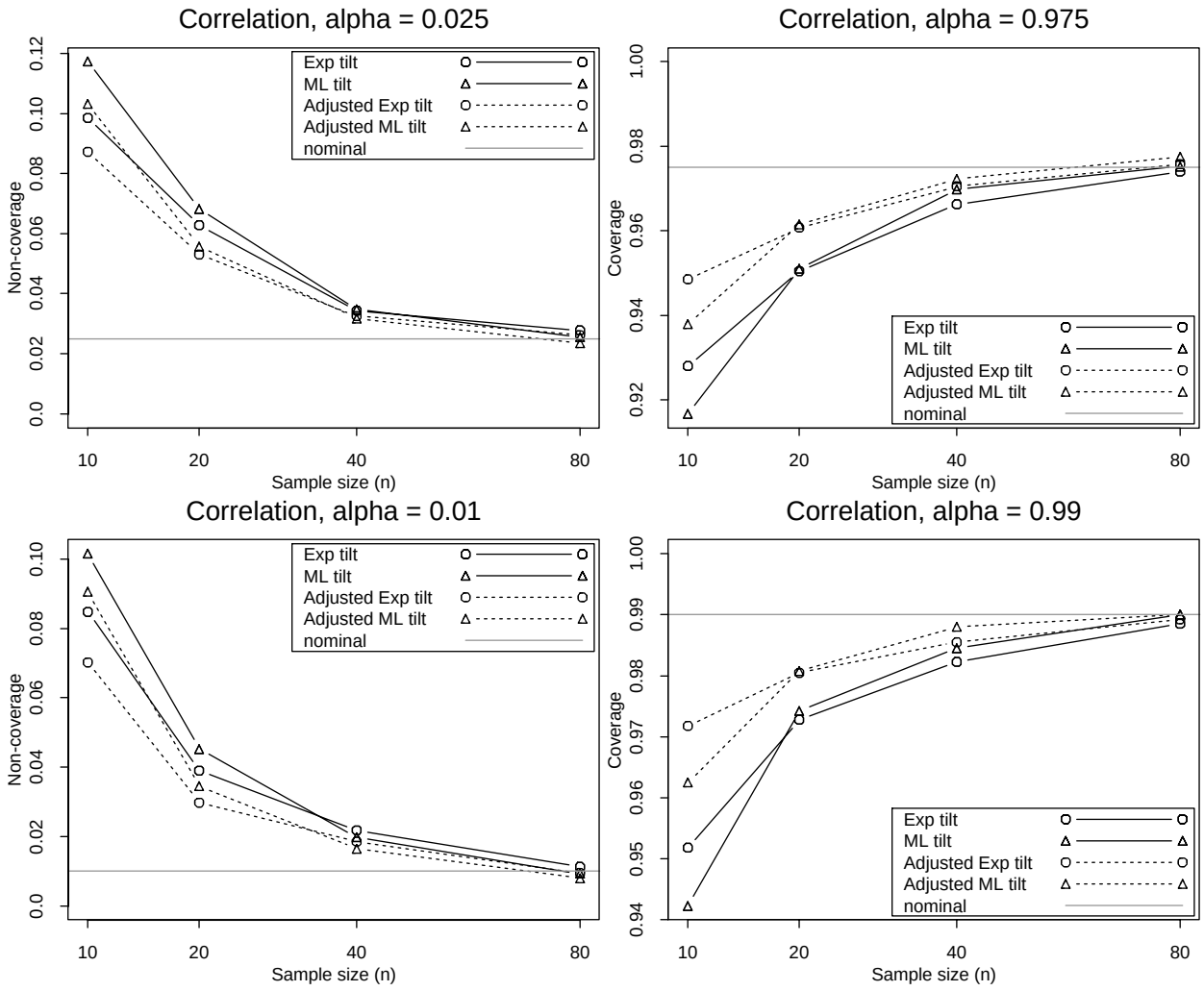

Figure 35: Coverage accuracy – effect of z : t adjustment of nominal levels. One-sided confidence intervals for the correlation for bivariate normal data with correlation $(1/2)^{1/2}$. Other details are the same as for Figure 33. *Comment*: the adjustment improves coverage probabilities, particularly for small samples.

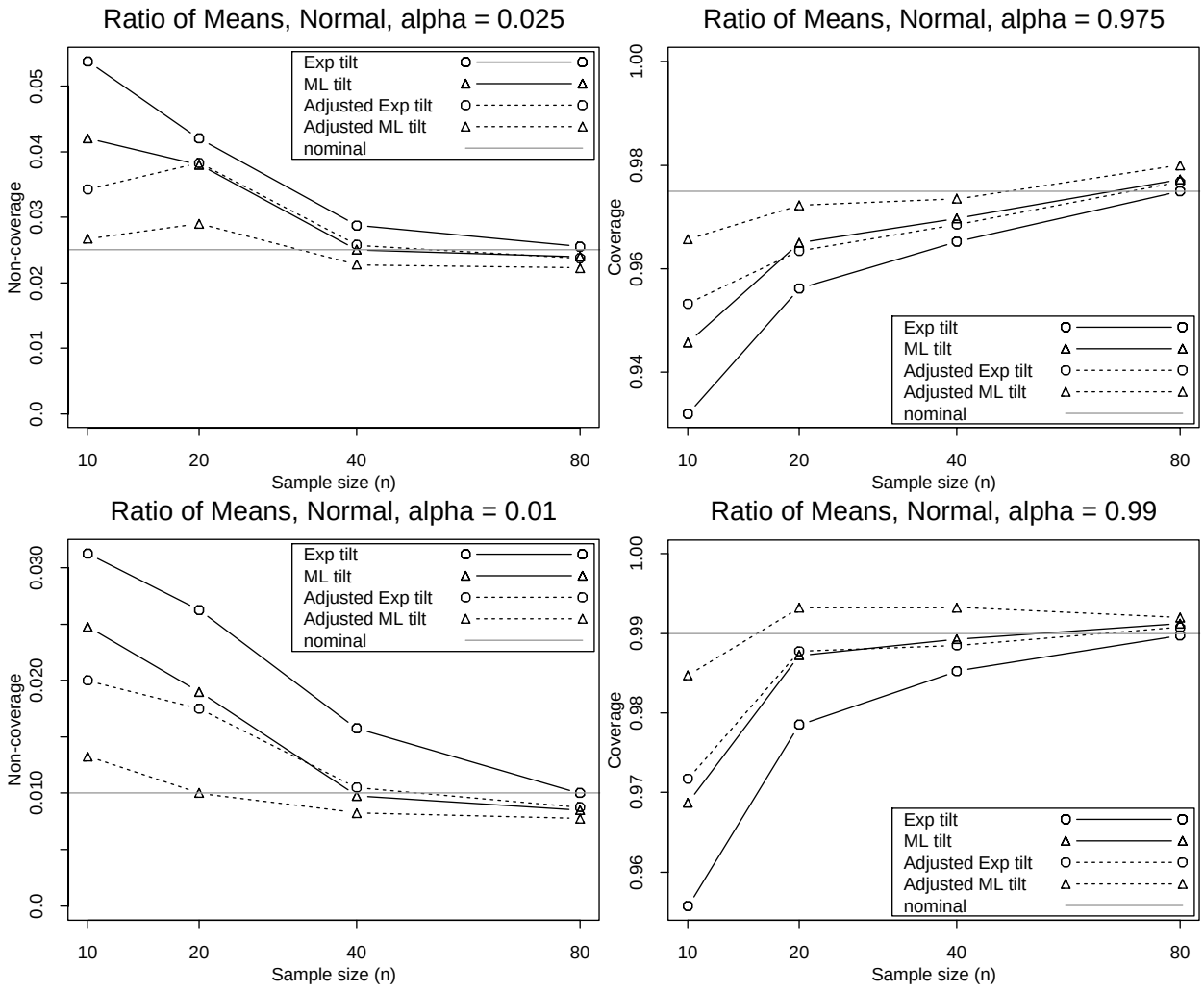

Figure 36: Coverage accuracy – effect of $z : t$ adjustment of nominal levels. One-sided confidence intervals for the ratio of means for bivariate normal data (uncorrelated, bivariate mean $(3, 9)$, variance 1) Other details are the same as for Figure 33. *Comment*: the adjustment improves coverage probabilities, particularly for small samples.

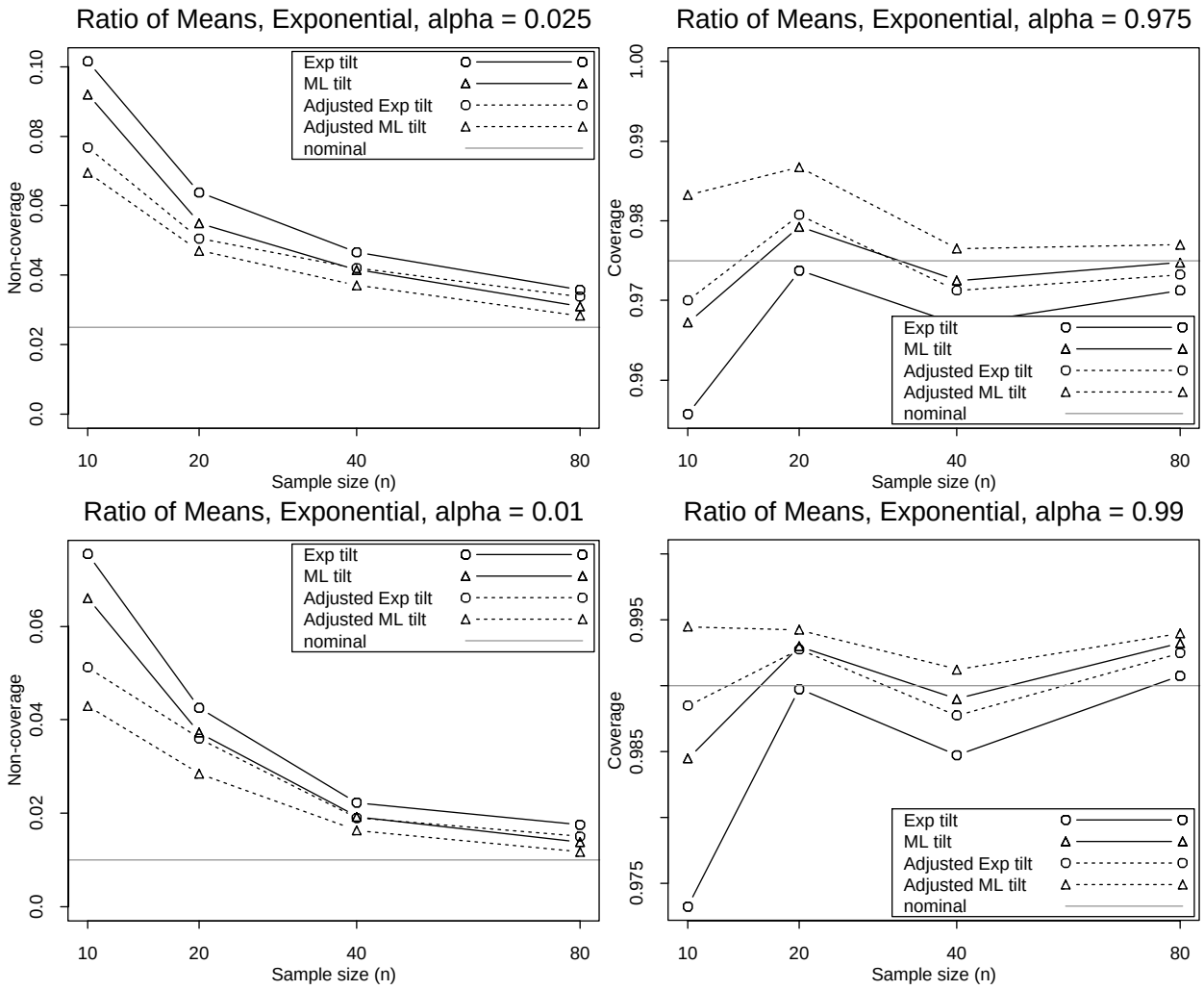

Figure 37: Coverage accuracy – effect of z : t adjustment of nominal levels. One-sided confidence intervals for the ratio of means for exponential data (independent, minimum values for x and y are 0 and 2, respectively, standard scale). Other details are the same as for Figure 33. *Comment*: the adjustment improves coverage probabilities, particularly for small samples.

- Bootstrap tilting confidence intervals and hypothesis are very computationally efficient, 17 or more times more efficient than other bootstrap methods, using an importance sampling reweighting implementation.
- A mixture design distribution makes the Monte Carlo sampling for tilting more robust against nonlinearity — this is helpful for small sample sizes with highly-nonlinear statistics.
- Tilting inferences are second order accurate, an order of magnitude (power of $\sqrt{n}$) more accurate than ordinary Student's-t or bootstrap percentiles (except in special cases like for symmetric distributions).
- In practice the coverage accuracy of bootstrap tilting inferences (using 200 bootstrap samples) is comparable to the best competing methods (using 2000 bootstrap samples), except that in some applications the bootstrap-t interval is more accurate.
- The length of tilting intervals is comparable to other bootstrap intervals, and sometimes substantially shorter than for bootstrap-t intervals, after adjusting for coverage probabilities.
- ML tilting offers more accurate coverage probabilities in general than exponential tilting. The latter offers some implementation benefits, and can be used to obtain starting values for the numerical solutions required for ML tilting.
- Updating derivatives when tilting offers better coverage probabilities. However, simple one-step updating sometimes fails spectacularly. A quick optimization procedure, using linear combinations of two gradients, works much better.
- Adjusting the nominal coverage probabilities based on the difference between normal and t quantiles provides better coverage probability; this is applicable to any bootstrap interval (except the bootstrap-t).
- Bootknife sampling appears to have little effect on bootstrap tilting intervals. Further work is needed to investigate why this is. An alternative would be to adjust nominal coverage probabilities, based on the difference between s and $\hat{\sigma}$ as an estimate of standard error.

Acknowledgments:

This work was supported by NSF SBIR Award No. DMI-9861360. I wish to thank Steve Ellis for comments that improved this report, and Chris Fraley and Shan Jin for assistance in creating the software used here.

References

Beckman, R. J. and McKay, M. D. (1987). Monte Carlo Estimation Under Different Distributions Using the Same Simulation. *Technometrics*, 29:153–160.

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

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Applications*. Cambridge University Press.

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

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

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.

Efron, B. (1981). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, 9:139 – 172.

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.

Garthwaite, P. H. and Buckland, S. T. (1992). Generating Monte Carlo Confidence Intervals by the Robbins-Monro Process. *Applied Statistics*, 41(1):159–171.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall, P. and La Scala, B. (1990). Methodology and Algorithms of Empirical Likelihood. *International Statistical Review*, 58(2):109–127.

Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, 61(1):143–158.

Hammersley, J. M. and Hanscomb, D. C. (1964). *Monte Carlo Methods*. Methuen, London.

Hesterberg, T. C. (1988). *Advances in Importance Sampling*. PhD thesis, Statistics Department, Stanford University.

Hesterberg, T. C. (1995a). Tail-Specific Linear Approximations for Efficient Bootstrap Simulations. *Journal of Computational and Graphical Statistics*, 4(2):113–133.

Hesterberg, T. C. (1995b). Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics*, 37(2):185–194.

Hesterberg, T. C. (1996). Estimates and Confidence Intervals for Importance Sampling Sensitivity Analysis. *Mathematical and Computer Modeling*, 23(8/9):79–86.

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. (1999). Smoothed bootstrap and jackboot sampling. Research Department 87, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109.

Hesterberg, T. C. and Ellis, S. J. (1999). Linear Approximations for Functional Statistics in Large-Sample Applications. Research Department 86, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109.

Hinkley, D. V. (1989). Bootstrap Significance Tests. *Bulletin of the International Statistical Institute*, pages 65–74.

Newton, M. A. and Geyer, C. J. (1994). Bootstrap Recycling: A Monte Carlo Alternative to the Nested Bootstrap. *Journal of the American Statistical Association*, 89(427):905–912.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.

Owen, A. (1990). Empirical likelihood confidence regions. *Annals of Statistics*, 18:90–120.

Reiman, M. I. and Weiss, A. (1986). Sensitivity Analysis Via Likelihood Ratios. In *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289.

Romano, J. P. (1988). A Bootstrap Revival of Some Nonparametric Distance Tests. *Journal of the American Statistical Association*, 83(403):698–708.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, 17:141–159.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

Tukey, J. W. (1987). Configural Polysampling. *SIAM REVIEW*, 29:1–20.

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

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