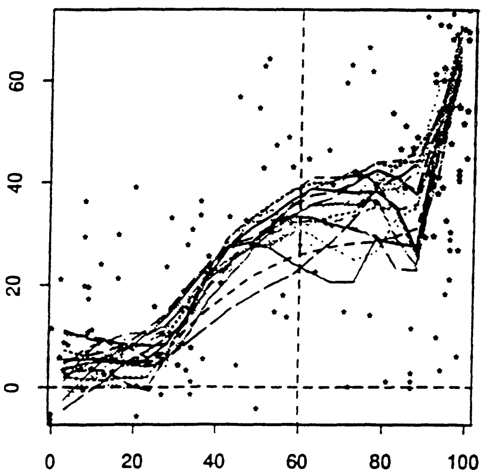

FIG. 2. *The first* 20 *bootstrap lowess curves*; *the sharp break at* 0.85 *seen in the original lowess curve is validated by the bootstrap replications.*

to the readers. Lowess is probably better for this situation. Figure 2 shows the first 20 of the 2,000 bootstrap lowess curves. The sharp break in the response function at $x = 0.85$ is a dependable feature of the replications. It is easy to quantify "dependable" with a bootstrap confidence interval for, say, $\gamma_{\text{break}} = \log((\theta_{100} - \theta_{85})/(\theta_{85} - \theta_{50}))$.

Without making too much of this small example, it does illustrate some encouraging trends in modern data analysis: more flexible fitting techniques than ordinary least squares polynomial regression; better confidence intervals than $\widehat{\theta} \pm 1.645\widehat{\sigma}$; and attention to "difficult" but interesting parameters like γ_{break}.

Theoreticians naturally focus on pathologies, which test a theory to its limits. Real applications tend less to be pathological than clumsy, awkward and difficult, as illustrated by the cholesterol example. In other words, they do not easily fit the simple mathematical models of classical statistical analysis. Computer-intensive methods like the bootstrap greatly extend the range of classical methods, and this is the way I believe that they will most dramatically affect 21st century statistics. Young's knowledgeable delineation of the limits of current bootstrap theory should not obscure an important fact: that these limits are already wide enough to permit a much more flexible approach to statistical practice.

Comment

Patricia M. Grambsch, Mary Kathryn Cowles and Thomas A. Louis

Young's review provides an informative history of the development of the bootstrap and discusses recent developments. We let others comment on technical issues, and briefly discuss Young's warnings related to the bootstrap. His principal worry is that the bootstrap invites mispractice by many users in that it has the reputation of an all-purpose procedure that will provide at least approximately valid inferences. Developers and generators of the bootstrap literature understand that, as with all statistical procedures, the bootstrap performs extremely well in many contexts (basically those where large-sample Gaussian asymptotics hold), but can fall on its face in nonregular contexts. Embellishments such as bias-correction and the nested bootstrap have improved small and moderate sample performance, but bring with them additional complications and decisions. Also, they strongly refute Efron's original claim that the bootstrap is "A statistical procedure devoid of intellectual content"!

Although Young's concerns are valid, are they any more compelling for the bootstrap than for other procedures such as the t-test, multiple regression or the Cox model? Our answer is both yes and no. Any statistical procedure frequently used will be frequently abused. Availability in a user-friendly computing package facilitates use and abuse. So, Young's criticisms unfairly single out the bootstrap. On the other hand, especially in its nonparametric, vanilla form, the bootstrap is relatively easy to apply to a limitless class of problems. All one has to do is decide on the sampling unit (or not decide and just get on with it), put the relevant data on actual or symbolic tokens and let the Monte Carlo run. Unlike the t-test, regression or Cox model, there are no explicit or implicit limits to the models or methods that comprise the "black box" around which one bootstraps. Sometimes the bootstrap will provide valid assessments of

Patricia M. Grambsch and Thomas A. Louis are faculty in the Division of Biostatistics, School of Public Health, University of Minnesota, Box 197 Mayo, Minneapolis, Minnesota 55455. Mary Kathryn Cowles is a faculty member in Preventive and Societal Medicine, University of Nebraska Medical Center, 600 South 42nd Street, Omaha, Nebraska 68198.

the properties of a procedure (the black box), but it will not rescue an inappropriate or suboptimal one. Articles on the bootstrap have appeared in *Science* and *The New York Times*. Because of its apparent general applicability, ease of use and reputation, the bootstrap has caught on like wildfire in fields ranging from genetics to geology. So, Young's special concern regarding the bootstrap is quite valid.

TABLE 1
Bias

	Span = 0.40		Span = 0.75	
X	Loess algorithm	Bootstrap estimate	Loess algorithm	Bootstrap estimate
0	0.28	0.11	0.65	−0.10
1/8	−0.35	−0.19	−0.67	0.01
1/4	0	0	0.06	0.10
3/8	0.35	0.23	0.91	0.12
1/2	0	0	0	0

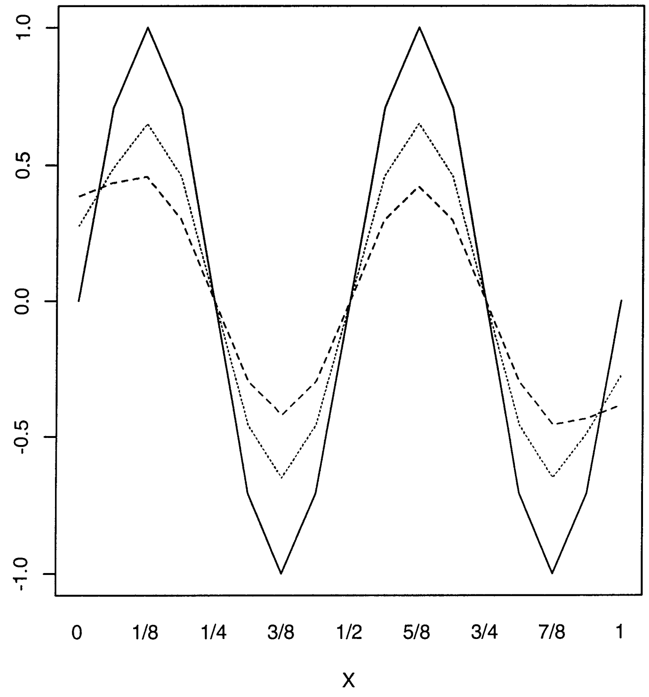

FIG. 1. *The solid line shows the sine curve at the* 17 *equally spaced points on* [0, 1]. *The dotted curve shows the result of linear least squares loess (span* = 0.40) *applied to those points on the sine curve. The dashed curve shows the result of applying loess to the dotted curve. Thus, the difference between the solid and dotted curves gives the bias for the loess algorithm. The difference between the dotted and dashed curves gives the mean bootstrap bias.*

AN EXAMPLE OF MISPRACTICE

As an example of how easy it is to misuse the bootstrap, consider the problem of producing a confidence interval or band for a *loess* curve (Chambers and Hastie, 1992; Cleveland and Devlin, 1988). We simulated data from the sine-wave model considered by Härdle and Bowman (1988):

$$Y_i = \sin(4\pi X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where the ε_i's are iid $N(0, \sigma^2)$.

We used 17 points, with the X_i's evenly spaced on [0, 1], and fit a linear loess smoother to each data set using two bandwidths: span = 0.75 (near the S-plus algorithm's default value) and span = 0.40 (better to adapt to curvature). Then, we computed point estimates and 90% confidence intervals at five X-values: $0, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}$, using four standard bootstrap techniques [see Efron and Tibshirani (1993) for definitions]:

1. the type I bootstrap percentile interval, based on resampled residuals from the loess fit that are rescaled to account for degrees of freedom lost in fitting;
2. residuals as in technique 1 with the bias-corrected percentile intervals;
3. a type II bootstrap percentile method, based on resampled (X, Y) pairs;
4. the bias-corrected confidence interval for these pairs.

Each bootstrap used 1,000 samples, and simulation estimates are based on 1,000 data sets.

The loess algorithm produces a biased estimate at many X-values, and the bias varies considerably. Due to the large bias in the loess estimate, the bootstrap estimate of bias is very poor (see Table 1 and Figure 1). Note especially that, in the absence of measurement error, there is a considerable discrepancy between the true bias (the difference between the dotted curve and the solid curve) and the bootstrap estimate of the bias (the difference between the dashed curve and the dotted curve). Since the bias is associated with X-values, the residuals are not even approximately exchangeable, and the type I bootstrap does not properly account for variability. The type II bootstrap cannot rescue the situation. All coverage probabilities differ substantially from the nominal values. Table 2 shows results for span = 0.4; results for span = 0.75 showed even greater departure from the nominal probabilities.

As pointed out the nonparametric smoothing literature (see Härdle and Bowman, 1988), for nonlinear and oscillating curves, the span or bandwidth must be adapted to local curvature. Our example suggests that the curve that generates the bootstrap samples must be fit with a substantially smaller span or bandwidth than is desirable for the curve estimate being evaluated. This approach reduces the bias in the curve generating the bootstrap samples and reveals the bias and nonexchangeability of residuals associated with the curve estimate being evaluated.

TABLE 2
Percent coverage for nominal 90% *confidence interval**

	Bootstrap method							
	$\sigma = 0.1$				$\sigma = 0.5$			
X	T1	T1 BC	T2	T2 BC	T1	T1 BC	T2	T2 BC
0	74	94	50	50	83	88	65	63
1/8	0	7	81	50	39	76	70	67
1/4	100	100	100	94	96	89	90	90
3/8	0	3	74	48	33	76	67	66
1/2	100	100	100	95	97	90	91	90

*T1 means type I percentile bootstrap; T1 BC is type I percentile bias-corrected; T2 is type II bootstrap; T2 BC is type II bias-corrected as described in text.

To confuse matters further, in using the bootstrap to pick a bandwidth for a kernel density estimate, the model generating bootstrap data must be an oversmoother. Failure to recognize these subtleties will result in very poor inferences.

Applying the smoother and then the bootstrap is a breeze (our simulations caused the breeze to blow 1,000 times), and we were able to commit mispractice with practically no effort. There are many other examples where hidden problems with the bootstrap will occur unless one is especially knowledgeable and careful.

Our response to Young's paper and to our example is a call to action. The statistical profession needs to communicate the good news, the bad news and the "no news yet." The bootstrap will succeed for a broad class of models and data structures. It will fail in others; sometimes it can be rescued by modifications that attend to the structure of the problem. We need to communicate what we know about the procedure's strengths and weaknesses and to identify situations where we do not yet know the answers. This communication must reach current and potential users and thus must appear in a broad array of journals and other information sources. As we learn more, information needs to be updated. Of course, the same recommendations hold for all statistical procedures, but the attraction of the bootstrap makes the need most acute.

ACKNOWLEDGMENTS

Partial support was provided by Grant RO1 CA51932 from the National Cancer Institute.

Comment

David Hinkley

INTRODUCTION

This is a timely article. It is likely to appear in print about the same time as first reviews of the excellent introductory book by Efron and Tibshirani (Efron and Tibshirani, 1993), a book which should allay some of the impatience and scepticism that I sense in the sophisticated user community about the bootstrap as a practical tool. We are also beginning to see the first wave of software products which claim to do bootstrap analysis: some of these are embarrassingly naive. Let us hope for more good applications-oriented books and better software products.

I think that Alastair Young has done an excellent job of highlighting the key theoretical developments and has suggested some sensible steps for further research. Much of what I have to say will complement his assessment and will focus on a few practical points.

WHEN DOES BOOTSTRAP WORK?

This question comes up twice in the paper, in the context of nonparametric bootstrapping of a point estimator. The first time we are given a succinct mathematical characterization which is clearly useless to even the best applied statistician. The second time

David Hinkley is Professor of Statistical Science and Head of the Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England.