Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 307 a 319

The Size Problem of Bootstrap Tests when the Null is Non- or Semiparametric

El problema del tamaño de los contrastes bootstrap cuando la hipótesis nula es No- o semiparamétrica

Jorge Barrientos-Marín[1,a], Stefan Sperlich[2,b]

[1]Departamento de Economía, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia

[2]Institut für Statistik und Ökonometrie, Georg August Universität Göttingen, Göttingen, Germany

Abstract

In non- and semiparametric testing, the wild bootstrap is a standard method for determining the critical values of tests. If the null hypothesis is also semi- or nonparametric, then we know that at least asymptotically oversmoothing is necessary in the pre-estimation of the null model for generating the bootstrap samples. See Härdle & Marron (1990, 1991). However, in practice this knowledge is of little help. In this note we highlight that this bandwidth choice problem can become quite serious. As an alternative, we briegly discuss the possibility of subsampling.[1]

***Key words*:** Bandwidth choice, Bootstrap tests, Nonparametric specification tests.

Resumen

En contrastes no- y semiparamétricos el *wild-bootstrap* es un método estándar para la determinación de los valores críticos de los estadísticos de contrastes. Si la hipótesis nula es no o semiparamétrica, sabemos que al menos asintóticamente es necesaria una sobre-suavización en la pre-estimación del modelo bajo la nula para generar las muestras bootstrap, ver por ejemplo Härdle & Marron (1990, 1991).

No obstante, en la práctica este conocimiento es de poca o ninguna ayuda. En este artículo, ponemos de manifiesto que el problema de la selección de la banda de suavidad para procedimientos de contraste puede ser muy serio. Como alternativa, discutimos brevemente la posibilidad de usar submuestras.

***Palabras clave*:** ancho de banda, contrastes de especificación no-paramétricos, contrastes bootstrap.

[a]Profesor. E-mail: jbarr@economicas.udea.edu.co

[b]Professor. E-mail: stefan.sperlich@wiwi.uni-goettingen.de

[1]The authors gratefully acknowledge very helpful comments of two anonymous referees as well as financial support from the Spanish MTM2008-03010 and the Deutsche Forschungsgemeinschaft FOR916.

1. Introduction

In both applied and mathematical statistics, non- and semiparametric specification testing is still quite a popular research field. Unfortunately, only a few papers address the problem of choosing an appropriate smoothing parameter. This a problem is fundamental for the reasonable use of these methods. There has been a growing amount of literature on adaptive testing where the adaptiveness refers to the smoothness of the alternative and deals with the smoothing of the test or the alternative.

However, these papers typically concentrate on testing problems where the null hypothesis is fully parametric. Here we are interested in testing qualitative restrictions, i.e. where the null hypothesis is semi- or nonparametric; think e.g. of additivity tests. When bootstrap is used to determine the critical value, these tests entail at least one more parameter choice problem: pre-estimating the model under the null hypothesis to later generate the bootstrap samples. This is necessary as in most cases the bandwidths for the estimation and the bootstrap should have different rates. See Härdle & Marron (1990, 1991). As in practical applications this problem has hardly been addressed, in most published procedures for testing or constructing confidence bands with a semi- or nonparametric null hypothesis, there is no guarantee that the bands meet the nominal coverage probability. This has been confirmed in the work of Dette, von Lieres, Wilkau & Sperlich (2005). In the latter paper, the problem is avoided by using subsampling.

To study the problem outlined in more detail, we concentrate on the problem of testing additivity. We limit ourselves to two test statistics proposed in Dette et al. (2005) and Roca & Sperlich (2007) but we extended this to different modifications including subsampling. The aim is not to find the most efficient additivity test or to propose new ones. Our focus is only directed at highlighting the size problem when the null hypothesis and the resampling method are non- or semiparametric. After a review of the additivity tests considered here, we study some of the typically proposed procedures for bandwidth choice. Unfortunately, we have not found a generally valid method. Our conclusion is that further research is necessary to find a proper bootstrap bandwidth.

2. Estimators and test statistics for additive models

Assume we face (not necessarily) independent and identically distributed (i.i.d.) data $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$, where

$$Y_i = m(X_i) + u_i \qquad i = 1, 2, \ldots, n, \tag{1}$$

with $m : \mathbb{R}^d \to \mathbb{R}$ an unknown function of interest, $m(x) = E(Y \mid X = x)$, and u_i i.i.d. random errors with $E[u_i] = 0$ and finite variance $\sigma^2(x_i)$. The internalized

Nadaraya-Watson estimator is defined as

$$\widehat{m}_k(x) = \sum_{i=1}^{n} v_k(x, X_i) Y_i, \text{ with } v_k(x, X_i) = \left(\widehat{f}_k(X_i)\right)^{-1} \mathbf{K}_k(x - X_i) \qquad (2)$$

where $\widehat{f}_k(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_k(X_j - X_i)$ is a kernel density estimator with a multiplicative kernel, i.e. for $w = (w_1, \ldots, w_d) \in I\!R^d$ we think of $\mathbf{K}_k(w) = \prod_{\alpha=1}^{d} K_k(w_\alpha)$, $K_k(w_\alpha) = k^{-1} K\left(w_\alpha k^{-1}\right)$. Commonly, the kernel is assumed to be Lipschitz continuous with compact support and $\int |K(x)|\, dx < \infty, \int K(x)dx = 1$. Furthermore, k is the bandwidth, assumed to go to zero for sample size n going to infinity, but nk_n^d going to infinity. Let V_k be the $n \times n$ matrix whose (j, i) element is $v_k(X_j, X_i)$, then $\widehat{m}(x) = V_k(x)\, Y$.

We are interested in the additive model, which we write in terms of

$$E(Y \mid X = x) = m_S(x) = \psi + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \qquad (3)$$

where we set $E_{X_\alpha}\{m_\alpha(X_\alpha)\} = \int m_\alpha(x) f_\alpha(x) dx = 0\ \forall \alpha$ for identification. Here, m_α, $\alpha = 1, \ldots, d$ are the marginal impact functions for each regressor. Therefore, ψ is a constant equal to the unconditional expectation of Y. Writing $m(X) = m_\alpha(X_\alpha) + m_{-\alpha}(X_{-\alpha})$ where $X_{-\alpha}$ is the vector X of all explanatory variables without X_α, i.e. $X_{-\alpha} = \left(X_{i1}, \ldots, X_{i(\alpha-1)}, X_{i(\alpha+1)}, \ldots, X_{id}\right)$, we can use the identification condition directly to estimate m_α. The so called marginal integration idea is based on that for x_α fix we have

$$E_{X_{-\alpha}}[m(x_\alpha, X_{-\alpha})] = \int m(x_\alpha, x_{-\alpha}) f_{-\alpha}(x_{-\alpha}) \prod_{\beta \neq \alpha} dx_\beta = \psi + m_\alpha(x_\alpha)$$

Substituting for $m(\cdot)$ a nonparametric pre-estimator such as the one given in (2), a sample average for the expectation, and for ψ simply $\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} y_i$ gives

$$\widehat{m}_\alpha(x_\alpha) = \sum_{i=1}^{n} w_{\alpha h}(x_\alpha, X_{i\alpha})\, Y_i$$

where for a bandwidth h (the one fixing the smoothness of our H_0 model)

$$w_h(x_\alpha, X_{\alpha i}) = K_h(x_\alpha - X_{i\alpha}) \frac{\widehat{f}_{-\alpha}(X_{i,-\alpha})}{\widehat{f}(X_{i\alpha}, X_{i,-\alpha})} \qquad (4)$$

Finally, we set $\widehat{m}_S(X_j) = \widehat{\psi} + \sum_{\alpha=1}^{d} \widehat{m}_\alpha(X_{j\alpha})$ for each $j = 1, 2, \ldots, n$. Note that defining $W_h = \sum_{\alpha=1}^{d} W_{\alpha h}(x_\alpha)$ with $W_{\alpha h}(x_\alpha)$ being the $n \times n$ matrices with $w_{\alpha h}(X_j, X_i)$ as elements, one has $\widehat{m}_S(x) = \psi + W_h(x)\, Y$.

As mentioned before, we do not introduce new testing procedures but rather study two modified statistics which have already been studied in the above mentioned papers, and which performed excellently in the study by Roca & Sperlich

(2007) though in a different context. The null hypothesis of interest is $H_0 : m(\cdot) = m_S(\cdot)$ versus $H_1 : m(\cdot) \neq m_S(\cdot)$. We consider the following two test statistics:

$$\tau_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}\left(X_i\right) - \widehat{m}_S\left(X_i\right)\right)^2 w(X_i)$$

$$\tau_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}\mathbf{K}_k\left(X_i - X_j\right)\left(Y_j - \widehat{m}_S\left(X_j\right)\right)\right]^2 w(X_i)$$

where $\widehat{e}_i = Y_i - \widehat{m}_S\left(X_i\right)$, i.e. the residuals under the null hypothesis, and $\widehat{u}_i = Y_i - \widehat{m}\left(X_i\right)$, the residuals without restrictions. We included also a weight function $w(\cdot)$ which typically is just used for trimming at the boundaries or regions where data are sparse. Note that in our simulation study we will make use of the trimming at the boundaries. Obviously, τ_1 calculates directly the integrated squared difference between the null and alternative models. Alternatively, τ_2 seeks to mitigate the bias problem inherited from the estimate $\widehat{m}$, which suffers from the "curse of dimensionality". In Dette et al. (2005) it is proved that, for both tests τ_j, the $nk^{\frac{d}{2}}\left(\tau_j - \mu_j\right)$ converge under the null to a normal variable with mean zero and variances v_j^2 for $j = 1, 2$ with

$$\mu_1 = E_{H_0}\left\{\tau_1\right\} = \frac{1}{nk^d}\int \sigma^2(x)w(x)dx \int \mathbf{K}^2(x)dx + o\left(\frac{1}{nk^d}\right)$$

$$\mu_2 = E_{H_0}\left\{\tau_2\right\} = \int \left(\mathbf{K} * \mathbf{K}\right)^2(x)\,dx \int \sigma^2(x)f^2(x)w(x)\,dx$$

and

$$v_1^2 = Var_{H_0}\left\{\tau_1\right\} = 2\int \sigma^4(x)w^2(x)dx \int \left(\mathbf{K} * \mathbf{K}\right)^2(x)\,dx$$

$$v_2^2 = Var_{H_0}\left\{\tau_2\right\} = \int \sigma^4(x)f^4(x)w^2(x)\,dx$$

All tests have been proven to be consistent in the sense that under the alternative they converge with n to infinity. Let us also mention that we have studied many more test statistics, e.g. those given in Dette et al. (2005) or Roca & Sperlich (2007) but not presented here. These, however, showed even less satisfactory performance, so we have skipped them in our presentation.

3. The resampling

Asymptotic expressions are of little help in practice for several reasons: Bias and variance contain unknown expressions which have to be estimated nonparametrically, and the convergence rate is quite slow for large d. For this reason, it is common to use resampling-mostly bootstrap-methods to approximate the critical value for the particular sample statistic. These can be bootstrap methods or subsampling procedures. Unfortunately, for the bootstrap it is not known how to

choose the smoothing parameter in practice for the pre-estimation of the model that is used to generate the bootstrap samples. From theory, it is known that one should somewhat oversmooth.

We give the general bootstrap procedure first and then discuss the details:

1. With bandwidth h, calculate the estimate $\widehat{m}_S$ under the null hypothesis of additivity and its resulting residuals $\widehat{e}_i$, $i = 1, \ldots, n$.

2. With bandwidth k, calculate the estimator $\widehat{m}$ for the conditional expectation without the additivity restriction, and the corresponding residuals $\widehat{u}_i$, $i = 1, \ldots, n$.

3. With the results from step 1 and 2, we can calculate our test statistics τ_1 and τ_2.

4. Repeat step 1 with a bandwidth h_b. We call the outcome $\widehat{m}_S^b$, respectively $\epsilon_i = Y_i - \widehat{m}_S^b(X_i)$, $i = 1, \ldots, n$.

5. Draw random variables e_i^* with $E[(e_i^*)^j] = u_i^j$ (respectively $\widehat{e}_i^j$ or ϵ_i^j, see discussion below) for $j = 1, 2, 3$ (respectively $j = 1, 2$, see below again). Set $Y_i^* = \widehat{m}_S^b(X_i) + e_i^*$, $i = 1, \ldots, n$, i.e. generate wild bootstrap samples. Repeat this B times. This defines B different bootstrap samples $\{(X_i, Y_i^{*,b})\}_{i=1}^n$, $b = 1, \ldots, B$.

6. For each bootstrap sample from steps 4 and 5, calculate the test statistics $\tau_j^{*,b}$, $j = 1, 2$, $b = 1, \ldots, B$. Then, for each test statistic τ_j, $j = 1, 2$, the critical value is approximated by the corresponding quantiles of the distribution of the B bootstrap analogues: $F^*(u) = \frac{1}{B} \sum_{b=1}^{B} I\{\tau_j^{*,b} \leq u\}$. Recall that they are generated under the null hypothesis.

In step 2, the bandwidth k has simply to obey the different assumptions required for each specific test. It can be chosen in such a way that it maximizes the power of the test for a given size. Therefore, different from Dette et al. (2005) we apply the adaptive testing approach introduced in Spokoiny (1998, 1996). He considers simultaneously a family of tests $\{\tau^k, k \in \mathfrak{K}\}$, where $\mathfrak{K} = \{k_1, k_2, \ldots, k_P\}$ is a finite set of reasonable bandwidths. The theoretical maximal number P depends on n, but is of no practical relevance. For details, see Horowitz & Spokoiny (2001). They define

$$\tau^{\max} = \max_{k \in \mathfrak{K}} \frac{\tau^k - E_0[\tau^k]}{Var^{1/2}[\tau^k]}$$

where $E_0[\cdot]$ indicates the expectation under H_0. A particularity of the resampling analogues of $\tau^{\max}$ is that one first needs to calculate the resampling statistics $(\tau^k)^{*,b}$ for all $k \in \mathfrak{K}$ to afterwards get $(\tau^{\max})^{*,b}$. Note that for each k, the empirical moments of the resampling statistics $(\tau^k)^{*,b}$ can be used as a substitute for $E_0[\tau^k]$, respectively $Var^{1/2}[\tau^k]$, in practice.

In step 5, the wild bootstrap (see Härdle & Mammen 1993) it is let open which residuals should be taken $\widehat{u}_i$, $\widehat{e}_i$ or ϵ_i. While theory says clearly that the best

power can be reached when taking the residuals of the alternative, i.e. $\widehat{u}_i$, our simulations (not shown) confirm the findings of Dette et al. (2005) that in practice ϵ_i should be taken. Next, it is often sufficient if we allow for heteroscedasticity of an unknown form using $e_i^* = \varepsilon_i \epsilon_i$, where the ε_i are i.i.d., drawn either from the golden-cut distribution, i.e.

$$\varepsilon_i = \begin{cases} -(\sqrt{5}+1)/2 & \text{with probability } \ p = (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } \ 1-p \end{cases}$$

or from the Gaussian normal $N(0,1)$. This answers the question up to order the moment of the bootstrap errors have to coincide with the residual moments. In the simulation section, we will compare golden-cut with Gaussian bootstrap.

In step 4, bandwidth h_b has to be chosen along the arguments of Härdle & Marron (1990, 1991): For the mean of $\widehat{m}_h(x) - m(x)$ under the conditional distribution of $Y_1, \ldots, Y_n \mid X_1, \ldots, X_n$, respectively of $\widehat{m}_h^*(x) - \widehat{m}_{h_b}(x)$ under the conditional distribution of $Y_1^*, \ldots, Y_n^* \mid X_1, \ldots, X_n$, it is well known that

$$E^{Y|X}(\widehat{m}_h(x) - m(x)) \approx h^2 \frac{\mu(K)}{2} m''(x) \tag{5}$$

$$E^{*}(\widehat{m}_h^*(x) - \widehat{m}_h(x)) \approx h^2 \frac{\mu(K)}{2} \widehat{m}''_{h_b}(x) \tag{6}$$

where $\mu(K) = \int u^2 K(u) du$. Obviously, we need that $\widehat{m}''_{h_b}(x) - m''(x) \to 0$. The optimal bandwidth h_b for estimating the second derivative must to be larger (in rates) than bandwidth h for estimating the function itself. We can even give the optimal rate. For example, the optimal rate to estimate m''_S is of the order $n^{-1/9}$ (instead of $n^{-1/5}$), an observation we make use of in our simulation studies. There it will be seen that the typical comment *h_b has to be oversmoothing*, is unhelpful in practice. Intuitively, one may think that a proper choice for h_b depends strongly on h. This might be true numerically, looking at equations (5) and (6) in the asymptotics the "h-effect" seems to cancel out as long as h/h_b goes to zero (a necessary condition for the consistency of bootstrap inference here) for n going to infinity. As one wants check whether the best possible additive model is an adequate fit, one therefore can concentrate on those bandwidth selectors for h which aim to optimize $\widehat{m}_S$ like cross validation or some plug-in methods do.

After all, it might be interesting to also have a look at subsampling as an alternative to bootstrapping (see Politis, Romano & Wolf 1999). Neumeyer & Sperlich (2006) introduce subsampling in a slightly context, other than that we discuss here, because there the bootstrap failed. There exists an automatic choice of the adequate subsample size m. As we remodeled this method to serve as a procedure for finding h_b, we introduce subsampling and the automatic choice of the subsample size m in more detail:

Let $\mathcal{Y} = \{(X_i, Y_i) \mid i = 1, \ldots, n\}$ be the original sample, and denoted by $\tau(\mathcal{Y})$ the original statistic calculated from this sample, leaving aside index $j = 1, 2, 3$ for a moment. To determine the critical values we need to approximate

$$Q(z) = P\left(n\sqrt{k^d}\tau(\mathcal{Y}) \leq z\right) \tag{7}$$

Recall that under H_0 this distribution converges to an $N(\mu_j, v_j^2)$, for μ_j and v_j, $j = 1, 2$, see above. For finite sample size n, drawing B subsamples $\mathcal{Y}_b$ -each of size m- we can approximate Q under H_0 by

$$\widehat{Q}(z) = \frac{1}{B}\sum_{b=1}^{B} I\left(m\sqrt{k_m^d}\tau^{k_m}(\mathcal{Y}_m) \leq z\right) \tag{8}$$

Note that the awkward notation comes from we have to adjust all bandwidths for the new sample size m. For example, imagine $k = k_0 \cdot n^{-\delta}$ for k_0 being constant. Then, τ^{k_m} is calculated like τ but with bandwidth $k_m = k_0 n^{\delta} m^{-\delta}$.

Certainly, under the alternative H_1, both $n\sqrt{k^d}\tau(\mathcal{Y})$ and $m\sqrt{k_m^d}\tau^{k_m}(\mathcal{Y}_m)$ converge to infinity. When demanding $m/n \to 0$ guarantees that $n\sqrt{k^d}\tau(\mathcal{Y})$ converges (much) faster to infinity than the subsample analogues. Then, $\widehat{Q}$ underestimates the quantiles of Q, which yields the rejection of H_0.

The optimal m is actually a function of the level α. Again, one applies resampling methods: Draw some pseudo sequences $\mathcal{Y}^{*,l}$, $l = 1, \ldots, L$ of $\mathcal{Y}$ of size n with the same distribution as $\mathcal{Y}$. For the desired level α, test $H_0^* : m(x) - m_S(x) = \widehat{m}(x) - \widehat{m}_S(x)$ the same way as you want to test $H_0 : m(x) = m_S(x)$, i.e. applying your particular test statistic to H_0^* and using subsampling. From the L repetitions you can determine the empirical rejection level (estimated size) for your given α. Now, find an m such that this empirical rejection level is $\approx \alpha$. In practice, you choose from a grid of possible m the one whose estimated rejection level for H_0^* is closest to α from below. Note that H_0^* is always true up to an estimation error that should be almost the same as in your original test. The only drawback of this procedure is the enormous computational effort. For further details and examples, see Politis et al. (1999) or Delgado, Rodríguez & Wolf (2001).

4. Simulation results

We give here only a summary of our large simulation study. The model considered is as follows: As in Dette et al. (2005), we draw $n = 100$ i.i.d. $X \in \mathbb{R}^3$ with

$$X_i \sim N(0, \Sigma_X) \quad \text{with} \quad \Sigma_X = \begin{pmatrix} 1 & 0.2 & 0.4 \\ 0.2 & 1 & 0.6 \\ 0.4 & 0.6 & 1 \end{pmatrix}$$

to generate

$$Y_i = X_{1,i} + X_{2,i}^2 + 2\,\sin(\pi X_{3,i}) + vX_{2,i}X_{3,i} + e_i, \quad i = 1, \ldots, n$$

with i.i.d. standard normal errors e_i, $v = 0$ being an additive separable model, or $v = 2$ for an alternative.

In both test, statistics we use the weighting function $w(\cdot)$ for a possible trimming: We cut the outer 5% or nothing (0%) of the sample, where "outer" refers to the tails of the explanatory variables. This is done to get rid of the boundary

effects in the statistics. To speed up our simulation studies, the presented results are calculated from 250 replications using only 200 bootstrap samples (or subsamples respectively). We used the multiplicative quartic kernel throughout but note that we know from our simulations in Dette et al. (2005) as well as from three years simulation experiences for the studies in Barrientos (2007), that the results change hardly for larger bootstrap samples.

We first looked for an average cross validation bandwidth h, which turned out to be $h_{opt} = 0.78$ for the direction of interest, and $6h_{opt}$ for the nuisance directions, cf. Dette et al. (2005). This was done not only for computational reasons but also because otherwise the size of the tests would also depend on the randomness induced by the estimation of h. For the k-adaptive test procedure, k ran over an equispaced grid of 10 bandwidths from $k_{\min} = 0.1 \cdot range(X_1)$ to $k_{\max} = range(X_1)$.

We will study now the results for several choices of h_b with different bootstrap generating methods, i.e. golden-cut vs. Gaussian bootstrap errors. To have h_b as a function of h, to take also into account $h/h_b \to 0$, and validate the rate $n^{-1/9}$ (motivated above) we set $h_b = hn^{1/5-1/\kappa}$ and try different $\kappa \leq 9$.

Table 1 shows the results for the k-adaptive bootstrap tests. We compare the size and power for different h_b, golden-cut vs Gaussian bootstrap, trimming boundary effects vs no trimming, and finally also a bit τ_1 vs τ_2 (though the latter is not the aim of this paper).

First, the results basically show that the size problem is not solved simply by different smoothing in the pre-estimation. Oversmoothing, in contrast to the theoretical findings, seems to go in the wrong direction, at least for τ_1. In particular, the hope that the ideas of Härdle & Marron (1990, 1991) (see equations (5) and (6)) might give us a hint or even provide a rule of thumb for the choice of h_b is not confirmed here.

Second, following to some extent the findings of (Delgado et al. 2001), we find a clear improvement for the Gaussian compared to the golden-cut bootstrap. Actually, when using the golden-cut method, then τ_1 does not hold the size for several h_b (κ respectively). Even worse, it rejects more often under H_0 than it does under H_1. This phenomenon is not observed for the simpler Gaussian wild bootstrap.

Third, boundary effects seem not to be the reason of our size and power problems. Surely, we get different numerical results for different weighting (i.e. trimming) functions, but cutting at the boundaries does not substantially change our general findings.

Finally, it is obvious that τ_2 outperforms τ_1 throughout. When recalling the motivation of the construction of τ_2, cf. Neumeyer & Sperlich (2006) and Roca & Sperlich (2007), it is obvious that the size problem comes from the bias rather from the variance. Or, in other words, bootstrap can capture pretty well the variance of a statistic but not its bias. There are two possible reasons for the surprising fact that τ_1 sometimes rejects more under H_0 than under H_1. First, while it is clear that the bias distorts the rejection level, it is not clear in what direction; moreover, the distortion effect certainly changes with the true underlying data generation

TABLE 1: Rejection levels of the two k-adaptive test statistics with and without trimming. Critical values are determined with golden-cut respectively Gaussian wild bootstrap, using $h_b = hn^{1/5-1/\kappa}$ for the pre-estimation.

			Golden Cut				Gaussian Residuals			
			H_0 $(v=0)$		H_1 $(v=2)$		H_0 $(v=0)$		H_1 $(v=2)$	
Trim	α%	κ	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
0%	5	4	.000	.024	.016	.364	.000	.024	.004	.440
		5	.012	.020	.016	.332	.008	.024	.020	.380
		6	.056	.020	.028	.344	.056	.020	.100	.376
		7	.136	.024	.044	.332	.124	.028	.168	.368
		8	.196	.016	.072	.360	.172	.020	.244	.384
		9	.244	.016	.088	.388	.216	.016	.320	.396
	10	4	.000	.068	.064	.572	.012	.088	.068	.672
		5	.024	.052	.068	.464	.024	.060	.076	.508
		6	.100	.048	.084	.440	.032	.036	.076	.492
		7	.188	.040	.092	.452	.036	.036	.096	.464
		8	.252	.040	.104	.468	.056	.036	.108	.488
		9	.308	.040	.124	.476	.068	.036	.132	.508
5%	5	4	.004	.024	.060	.352	.004	.020	.008	.420
		5	.024	.020	.048	.324	.028	.020	.040	.348
		6	.112	.016	.068	.336	.096	.020	.144	.360
		7	.180	.016	.100	.316	.164	.020	.236	.348
		8	.276	.012	.132	.348	.216	.016	.328	.360
		9	.360	.012	.152	.372	.292	.016	.436	.388
	10	4	.016	.072	.108	.568	.064	.088	.120	.664
		5	.036	.048	.100	.460	.052	.052	.108	.500
		6	.164	.044	.116	.428	.068	.036	.104	.460
		7	.256	.036	.144	.448	.092	.036	.144	.460
		8	.356	.036	.184	.460	.120	.032	.176	.484
		9	.432	.036	.228	.476	.128	.040	.224	.504

process. Second, making the tests $k-$adaptive entails a normalization by the estimated variance. In the unfortunate situation where the variance estimation is getting larger, the power of the test decreases. Both effects together lead here to the counter-intuitive performance of τ_1.

In the last section we introduced subsampling as an alternative resampling method to bootstrap. Therefore, we also provide a simulation study where the critical values are approximated by subsampling, trying several subsample sizes m. Recall that the different subsample sizes have a similar effect here like it has the choice of h_b for bootstrap tests. The results are given in Table 2 for k-adaptive tests. For τ_1 we see here basically the same bad behavior we observed when using golden-cut bootstrap to determine the critical values. In contrast, τ_2 seems to

TABLE 2: Rejection levels of the two k-adaptive test statistics with and without trimming. Critical values are determined with subsampling, using subsamples of sizes m.

			H_0 $(v=0)$		H_1 $(v=2)$	
Trim	α%	m	τ_1	τ_2	τ_1	τ_2
0%	5	80	.000	.000	.000	.000
		70	.000	.000	.000	.000
		60	.056	.000	.020	.036
		50	.276	.020	.076	.292
		40	.516	.212	.168	.732
	10	80	.000	.000	.000	.000
		70	.020	.000	.016	.000
		60	.272	.008	.072	.144
		50	.584	.104	.256	.644
		40	.816	.476	.480	.912
5%	5	80	.000	.000	.000	.000
		70	.000	.000	.000	.000
		60	.016	.000	.000	.032
		50	.060	.020	.012	.276
		40	.152	.216	.024	.712
	10	80	.000	.000	.000	.000
		70	.004	.000	.000	.000
		60	.060	.008	.016	.164
		50	.200	.092	.024	.636
		40	.380	.460	.120	.908

work-though with less power than we observed when using Gaussian bootstrap, cf. Table 1.

Recall that our main focus is the size distortion of resampling tests. Therefore our last two studies are about the automatic choice of m in subsampling and h_b in (Gaussian) bootstrap, respectively.

A quite time consuming simulation study evaluating the automatic choice of m indicates that this procedure does unfortunately not work at all. Nevertheless, our last study is to apply this idea for getting an automatic choice of h_b. In order to do so, we first have to adjust the procedure for an automatic choice of the subsample size m to now find an adequate bootstrap bandwidth h_b.

This can be done as follows, described here in detail for τ_2. To make notation and calculation easier, we consider the non-k-adaptive version but fix $k = range(X_1)/2$. Let now $\{Y_i^*, x_i^*\}_{i=1}^n := \mathcal{Y}^*$ be a member of the pseudo sequence introduced above. Then, for testing $H_0^* : m(x) - m_S(x) = \widehat{m}(x) - \widehat{m}_S(x)$ with

sample $\mathcal{Y}^*$, an analogue to τ_2 would be

$$\tau_2^\sharp = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}\mathbf{K}_h(X_i^* - X_j^*)\left\{Y_j^* - \widehat{m}_S(X_j^*)\right\}\right.$$
$$\left. - \mathbf{K}_h(X_i^* - X_j)\left\{Y_j - \widehat{m}_S(X_j)\right\}\right]^2 w(X_i^*) \quad (9)$$

Other statistics are thinkable certainly, e.g.

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}\mathbf{K}_h(X_i - X_j^*)\{Y_j^* - \widehat{m}_S(X_j^*)\}\right.$$
$$\left. - \mathbf{K}_h(X_i - X_j)\{Y_j - \widehat{m}_S(X_j)\}\right]^2 w(X_i)$$

but they should all be asymptotically equivalent to (9). The procedure was performed with only $L = 100$ pseudo samples $\mathcal{Y}^*$. As the results varied widely we were forced either to enlarge L considerably or to reduce σ_e considerably. For computational reasons we decided on the second option and repeated the study with $\sigma_e = 0.1$.

Some results are summarized in Table 3. As et can be seen, this time we emphasize the possibility of undersmoothing much more. You first have to look at $\tau_2^\sharp$ to find the κ giving the rejection level closest to $\alpha = 5\%$ from below. Here, this is always $\kappa = 3$. Note that this might also change depending on the trimming, α, sample size, etc. It is important to understand that the lines of τ_2^* can always be calculated, i.e. without knowing the true data generating process. Therefore we call this method fully automatic. Now look at the lines for τ_2, the test of interest. Obviously, $\kappa = 3$ is indeed the best possible choice; it has the strongest power among all κ respecting the nominal level. This could be taken as indicating that our suggestion for selecting h_b works. Unfortunately, this method does not work that well for all possible α; specifically, it becomes quite incorrect for $\alpha \geq 10\%$. Even worse, it did not work for τ_1 (not shown).

5. Conclusions

Our main focus is the bootstrap and its size distortion in practice when the sample size is small or moderate. These points are illustrated along the popular problem of additivity testing. Naturally, one looks for an optimal trade-off between controlling for size under the null hypothesis H_0 and maximizing power. Even though these problems have already been discussed and studied in theory, as yet, it is unclear how to set the smoothing parameter for the bootstrap prior estimates in practice. We show that theory is not just unhelpful here; at present, a reasonable application of bootstrap tests of these kinds is questionable.

TABLE 3: Rejection levels of τ_2 and $\tau_2^\sharp$ for $\alpha = 5\%$, with and without trimming, using Gaussian bootstrap with $h_b = hn^{1/5-1/\kappa}$ for the pre-estimation, and $k = range(X_1)/2$.

			κ						
Trim			1	2	3	4	5	6	7
H_0	0%	$\tau_2^\sharp$	.012	.063	.028	.030	.032	.031	.029
$(v = 0)$		τ_2	.680	.392	.032	.012	.012	.012	.016
	5%	$\tau_2^\sharp$	.012	.062	.028	.030	.032	.031	.029
		τ_2	.676	.380	.024	.012	.012	.012	.020
H_1	0%	$\tau_2^\sharp$	.001	.019	.042	.022	.015	.011	.009
$(v = 2)$		τ_2	.972	.932	.632	.380	.272	.260	.264
	5%	$\tau_2^\sharp$	.001	.019	.042	.023	.015	.011	.010
		τ_2	.968	.936	.620	.368	.260	.252	.264

Further, we have shown that subsampling is an interesting alternative to bootstrap which in addition provides a procedure for the analogue problem of subsample size choices.

Finally we introduced the idea of extending the procedure of subsample size selection to smoothing parameter (h_b) selection in bootstrap testing problems. However, further research is necessary to provide reliable procedures for the nonparametric testing problems considered here.

[Recibido: marzo de 2010 — Aceptado: octubre de 2010]

References

Barrientos, J. (2007), Some Practical Problems of Recent Nonparametric Procedures: Testing, Estimation and Application, Tesis doctoral, Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, España.

Delgado, M. A., Rodríguez, J. M. & Wolf, M. (2001), 'Subsampling Cube Root Asymptotics with an Application to Manski's MSE', *Economics Letters* **73**, 241–250.

Dette, H., von Lieres, C., Wilkau, C. & Sperlich, S. (2005), 'A Comparison of Different Nonparametric Method for Inference on Additive Models', *Journal of Nonparametric Statistics* **17**, 57–81.

Härdle, W. & Mammen, E. (1993), 'Comparing Nonparametric Versus Parametric Regression Fits', *Annals of Statistics* **21**(1926-1947).

Härdle, W. & Marron, J. S. (1990), 'Semiparametric Comparison of Regression Curves', *Annals of Statistics* **18**, 63–89.

Härdle, W. & Marron, J. S. (1991), 'Bootstrap Simultaneous Bars For Nonparametric Regression', *Annals of Statistics* **19**, 778–796.

Horowitz, J. L. & Spokoiny, V. (2001), 'An adaptive, rate-optimal test of parametric mean-regression model against a nonparametric alternative', *Econometrica* **69**, 599–631.

Neumeyer, N. & Sperlich, S. (2006), 'Comparison of separable components in different samples', *Scandinavian Journal of Statistics* **33**, 477–501.

Politis, D. N., Romano, J. P. & Wolf, M. (1999), *Subsampling*, Springer Series in Statistics, Springer-verlag, New York.

Roca, J. & Sperlich, S. (2007), 'Testing the Link when the Index is Semiparamtric - A Comparison Study', *Computational Statistics and Data Analysis* **12**, 6565–6581.

Spokoiny, V. (1996), 'Adaptive Hypothesis Testing using Wavelets', *Annals of Statistics* **24**, 2477–2498.

Spokoiny, V. (1998), 'Adaptive and spatially adaptive testing of a nonparametric hypothesis', *Mathematical Methods of Statistics* **7**(245-273).