Consistent Significance Testing for Nonparametric Regression

Jeff Racine

Department of Economics, BSN 3403
University of South Florida
Tampa, FL, 33620-5500

Abstract. This paper presents a framework for individual and joint tests of significance employing non-parametric estimation procedures. The proposed test is based on nonparametric estimates of partial derivatives, is robust to functional mis-specification for general classes of models, and employs nested pivotal bootstrapping procedures. Two simulations and one application are considered to examine size, power relative to mis-specified parametric models, and to test for the linear unpredictability of exchange rate movements for G7 currencies.

Keywords. Kernel density estimation, inference, pivotal, nested bootstrap.

1. INTRODUCTION

The inability to test hypotheses in a nonparametric framework has remained a source of frustration for many applied researchers and econometricians. The motivation for using nonparametric methods for both estimation and hypothesis testing comes from the fact that employing a mis-specified parametric model for the conditional mean and/or the data generating process will typically result in inconsistent parameter estimates and hypothesis tests that possess asymptotically incorrect size. Nonparametric estimators are consistent under less restrictive assumptions than those required for consistency of parametric estimators. Therefore, in the absence of knowledge regarding the true functional forms of the conditional mean and data generating process, nonparametric methods may be preferred.

The significance test is probably the most frequently used test in applied multivariate regression. In addition, the significance test is often used to confirm or refute economic theories. However, the use of mis-specified parametric models for the purpose of significance testing will typically yield tests which have incorrect size and low power. The likelihood of mis-specification in a parametric framework is high given the fact that most applied researchers choose parametric models on the basis of parsimony and tractability. Significance testing in a nonparametric framework would therefore have obvious appeal given that nonparametric techniques are consistent under much less restrictive assumptions than those required for a parametric approach.

This paper demonstrates how the nonparametric estimation of partial derivatives of an unknown conditional mean can be used for the purpose of significance testing. Of course, such estimates are often of direct interest in themselves. The proposed approach is based on the application of pivotal bootstrapping methods and resolves some important outstanding practical issues regarding hypothesis testing in a nonparametric framework.

In this paper it is assumed that an unknown conditional mean and associated partial derivatives are estimated using the nonparametric kernel estimation technique based on the approaches of Nadaraya (1965) and Watson (1964), henceforth known as the Nadaraya-Watson kernel approach. The proposed approach can be applied with little or no modification to many other nonparametric and semiparametric approaches such as orthogonal series estimators, feedforward neural networks, spline smoothers and so on.

Given the distribution-free nature of the estimation technique, interest lies in conducting significance tests which themselves do not rely on distributional assumptions. Methods for hypothesis testing in a nonparametric kernel framework based on asymptotic results have been recently proposed in Robinson (1991), Robinson (1994), Lavergne and Vuong (1992), and Rilstone (1991). In each of these proposed approaches asymptotic theory is employed to derive the limiting distributions of a test statistic which is based on a nonparametric kernel estimate. Rilstone (1991), for example, uses nonparametric kernel estimators of average derivative functionals for hypothesis testing, derives the asymptotic distribution of the proposed test statistic, and demonstrates that the test has a $\sqrt{n}$ rate of convergence, the same rate as that obtained for a semiparametric model. Robinson (1994) builds on such results for semiparametric averaged derivatives and demonstrates that the rate of convergence of the finite-sample distribution to the normal limit distribution can equal that of standard parametric estimates. For related work in this area the reader is referred to Stoker (1989), Härdle and Stoker (1989), and Powell, Stock and Stoker (1989). Unfortunately, there are a number of drawbacks with such asymptotic-based approaches which arise in practice.

The most troubling aspect of such asymptotic-based testing procedures is that the null distribution of such test statistics does not depend on the bandwidth, while the value of the test statistic depends directly on the bandwidth. This is due in part to the fact that the bandwidth is a quantity which vanishes asymptotically. This is a serious drawback in practice, since the outcome of such asymptotic-based tests tends to be quite sensitive to the choice of bandwidth. This has been noted by a number of authors including Robinson (1991) and Rilstone (1991). Robinson (1991) noted that "substantial variability in the [test statistic] across bandwidths was recorded", which would be most troubling in applied situations due, in part, to numerous competing approaches for data-driven bandwidth choice. For a current overview of data-based bandwidth selection procedures in the context of kernel density estimation see Jones, Marron and Sheather (1992).

This paper resolves issues surrounding the asymptotic-based approaches to hypothesis testing in nonparametric settings by applying pivotal bootstrap resampling. Resampling techniques are employed to obtain the null distribution of a test statistic which is based on nonparametric estimates of partial derivatives. Related work on bootstrapping nonparametric point estimates includes that of Härdle and Marron (1991) who propose the application of a 'wild bootstrap' to obtain error bars for kernel estimators of a conditional mean. For theoretical work on the bootstrap see Bickel and Freedman (1981) who derive the asymptotic validity of the bootstrap for a number of situations including pivotal quantities, empirical and quantile processes, and U-statistics and other von Mises functionals. For a recent survey of resampling methods see Jeong and Maddala (1993).

There are two important and quite distinct reasons why a resampling approach might be preferable to an asymptotic one for the purposes of hypothesis testing in a nonparametric framework. First, given the relatively slow rates of convergence of kernel estimators, the use of resampling techniques may be preferred for small to moderate sample sizes because resampling techniques might be expected to perform better in finite sample situations than an asymptotic-based counterpart (Efron 1983). This rate-of-convergence related problem has been discussed by many authors, for instance Mammen (1992) (pp. 4-5) who emphasizes the poorness of the asymptotic approximations in this context for moderate sample sizes. To quote, "Often the asymptotic distributions of these functionals cannot be calculated explicitly or explicit approximations are so poor that, typically, they cannot be used in practice for moderate sample sizes."

Another reason for using resampling techniques is that there exist cases in which a resampling approach works under weaker conditions than those necessary for asymptotic approximations to hold. Bickel and Freedman (1983) demonstrate this result for linear models to which the bootstrap has been applied. Such results justify the use of bootstrapping through appeal to robustness arguments in the sense of being robust to departures from assumptions underlying the modeling procedure.

The remainder of this paper proceeds as follows. The test statistic and algorithms to obtain its null distribution are presented in Section 2. Simulation results and applications are considered in Section 3, while Section 4 summarizes and concludes.

2. A NONPARAMETRIC SIGNIFICANCE TEST

The most commonly used regression-based hypothesis test is the test of significance. Rejecting or failing to reject the null is often used as evidence confirming or refuting a theory, and can have important practical and theoretical implications. Applied researchers typically choose parametric models on the basis of tractability and ease of interpretation, not on the basis of any prior knowledge regarding the unknown DGP. Such models are typically mis-specified to some degree. It is well known that hypothesis tests based on functionally mis-specified parametric regression models will have asymptotically incorrect size. That is, the probability of a Type I error will not equal the assumed nominal value regardless of the sample size. In addition, hypothesis tests based on mis-specified models will suffer from low power and thereby fail to detect departures from the null.

In the absence of knowledge regarding the true functional form for the conditional mean nonparametric Nadaraya-Watson Kernel methods will be used which are robust to functional mis-specification among the class of twice-continuously differentiable functions. The important features of the proposed approach are that the null distribution of the nonparametric-based test will have correct size and the test will have power in the direction of the class of twice-continuously differentiable alternatives.

2.1. **The Test Statistic.** Let $f(Y,X)$ denote the joint density of a set of random variables of interest (Y,X) where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p, p \in \mathbb{N}$ and let (y_i, x_i) be a realization of (Y,X). The density of Y conditional on X will be denoted $g(Y|X) = f(Y,X)/f_1(X)$ where $f_1(X)$ denotes the marginal density of X. The conditional mean of Y with respect to X is defined as $E(Y|X) = \int y\, g(y|X)\, dy$. The gradient of the conditional mean with respect to the conditioning variables is defined as $\nabla E(Y|X) = \partial E(Y|X)/\partial X$ where $\nabla E(Y|X) \in \mathbb{R}^p$. It is assumed that the reader is interested in a model of the form $M(x_i) = E(Y|x_i)$ and in partial derivatives of the form $\beta(x_i) = \nabla E(Y|x_i)$ where $|_{x_i}$ is taken to mean conditional on the random vector X assuming the realization $x_i \in \mathbb{R}^p$.

For notational simplicity, partition the vector X into two parts, the variables whose significance is to be tested $X_{(j)}$ and all other conditioning variables $X_{(-j)}$ excluding $X_{(j)}$. The partitioned matrix of conditioning variables is written as $X = (X_{(-j)}, X_{(j)})$ where $X_{(-j)} \in \mathbb{R}^{p-j}$ and $X_{(j)} \in \mathbb{R}^j$. If the conditional mean $E(Y|X)$ is independent of a variable or group of variables in question, $X_{(j)}$, then the true but unknown vector of partial derivatives of the conditional mean of the dependent variable with respect to these variables is zero. This condition for independence of $E(Y|X)$ and $X_{(j)}$ is stated as

$$E(Y|X) \perp X_{(j)} \Leftrightarrow \frac{\partial E(Y|X)}{\partial X_{(j)}} = 0 \quad a.s. \tag{1}$$

where $\partial E(Y|X)/\partial X_{(j)} \in \mathbb{R}^j$ and where $\perp$ denotes orthogonality or independence.

Nonparametric estimation techniques yield partial derivatives which are permitted to vary over their domain. Contrast this with parametric multivariate linear regression techniques in which the partial derivative is typically assumed to be constant over its domain. This has implications for the type of test statistic used in a nonparametric context. In particular, tests must be formulated to detect whether a partial derivative equals zero over the entire domain of each variable in question.

Noting explicitly that the partial derivatives vary over their domain, the null hypothesis can be stated in terms of the vector of partial derivatives of the conditional mean as

$$\begin{aligned} H_0 :&\quad \frac{\partial E(Y|X)}{\partial X_{(j)}} = 0 \text{ for all } x \in X \\ H_A :&\quad \frac{\partial E(Y|X)}{\partial X_{(j)}} \neq 0 \text{ for some } x \in X \end{aligned} \tag{2}$$

Since a test statistic in this context must necessarily involve some aggregate measure of the derivative over its domain, an aggregate L_2 norm measure will be used. This norm is adopted based on power considerations.

Using this L_2 aggregate based on the unknown derivatives, the null and alternative hypotheses can be stated as

$$
\begin{aligned}
H_0: \quad \lambda &= E\iota' \left[\frac{\partial E(Y|X)}{\partial X_{(j)}}^2 \right] = 0 \\
H_A: \quad \lambda &= E\iota' \left[\frac{\partial E(Y|X)}{\partial X_{(j)}}^2 \right] > 0
\end{aligned} \tag{3}
$$

where ι denotes a unit vector of length j, and $\partial E(Y|X)/\partial X_{(j)}{}^2$ is intended to mean that this is a vector of squared derivatives. If the null hypothesis is true then λ will be identically equal to zero. Otherwise, λ will exceed zero.

The proposed test statistic is obtained by constructing the sample analogue of Equation (3) in which the unknown derivatives are replaced with nonparametric estimates, $\hat{\beta}(x_i) \in \mathbb{R}^j$, $i = 1, \ldots, n$. The resulting test statistic will be denoted by $\hat{\lambda}$ and is written as

$$
\hat{\lambda} = n^{-1} \sum_{i=1}^{n} \sum_{h=j}^{p} \left[\hat{\beta}_h(x_i) \right]^2 . \tag{4}
$$

The finite sample properties of this test statistic are not known at this time, however it is fully expected that this will yield a consistent test under standard regularity conditions.

2.2. **Obtaining The Null Distribution of $\hat{\lambda}$.** To conduct tests based on the proposed statistic a sampling distribution under the null must be obtained. One option at this point would be to work on obtaining an asymptotic approximation to this distribution. Robinson (1991) employed asymptotic approximations for his analytically simpler semiparametric average derivative and noted that "substantial variability in the [test statistic] across bandwidths was recorded", which is troubling for reasons outlined in Section 1. Against this backdrop a resampling approach is pursued. As will be seen, the resulting test will have correct size and will be extremely insensitive to very large deviations from the optimal bandwidth.

The sampling distribution of $\hat{\lambda}$ under the null will be estimated using Efron's bootstrap (Efron 1983). Percentiles of the test statistic under the null can then be obtained from this estimated distribution, and one-sided tests can then be performed by comparison of the test statistic with the appropriate percentile obtained from the estimated distribution.

2.3. **Pivotal Resampling.** Recent modifications of the bootstrap are known to give more reliable percentiles than the standard bootstrap. The best approaches (as argued by Beran (1988) and Hall (1986)) are known as *pivotal* methods (also known as *percentile-t* methods). A statistic is (asymptotically) pivotal if its limiting distribution does not depend on unknown quantities (Hall 1992, p. 83). The general idea is that instead of bootstrapping a raw statistic $\hat{\theta}$, a studentized statistic $(\hat{\theta} \perp \theta)/s(\hat{\theta})$ is bootstrapped where $s(\hat{\theta})$ is a consistent estimate of the standard error of $(\hat{\theta} \perp \theta)$. For most applications, $s(\hat{\theta})$ is a $\sqrt{n}$ consistent estimator. The bootstrap does a better job of estimating the distribution of a pivotal statistic than it does a non-pivotal one, and this pivotal bootstrap approach been shown to be asymptotically superior to non-pivotal bootstrap approaches (Hall 1988). In particular it has been shown (Beran 1988) that the pivotal bootstrap distribution coincides through order $\sqrt{n}$ with the Edgeworth expansion of the exact finite sample distribution. In addition, Beran (1988) has demonstrated that critical values obtained from a pivotal approach will result in tests with finite sample sizes closer to the nominal size than tests based on asymptotic critical values.

There are two potential applications of pivoting for the test statistic at hand. First, our statistic is an average of (squared) pointwise derivative estimates $\hat{\beta}_h(x_i)$. These estimates can be pivoted by pointwise dividing by their standard errors based on asymptotic approximations, $SE(\hat{\beta}_h(x_i))$. The test statistic based on the pivoted derivative estimates would therefore be given by

$$\hat{\lambda} = n^{-1} \sum_{i=1}^{n} \sum_{h=j}^{p} \left[\frac{\hat{\beta}_h(x_i)}{SE(\hat{\beta}_h(x_i))} \right]^2 . \tag{5}$$

Secondly, the statistic $\hat{\lambda}$ can be pivoted by dividing by its standard error. The asymptotic distribution of the proposed statistic $\hat{\lambda}$ is not known, hence an estimator of the statistic's variance based on asymptotic results cannot be used. An estimate of the statistic's variance can, though, be computed via resampling (I am most grateful to an anonymous referee for noting this point). Efron and Tibshirani Efron and Tibshirani (1993, p. 162) note that "standard error formulae exist for very few statistics, and thus... [for] more complicated statistic[s]... we would need to compute a bootstrap estimate of standard error for each bootstrap sample [which] implies two nested levels of bootstrap sampling." Following this approach a nested pivotal bootstrap procedure is applied to estimate the null distribution of

$$\hat{t} = \frac{\hat{\lambda}}{SE(\hat{\lambda})} \tag{6}$$

as opposed to $\hat{\lambda}$, where $SE(\hat{\lambda})$ is the estimated standard error of $\hat{\lambda}$ which is itself obtained via nested resampling. It will be seen that pivoting the derivative estimates and the test-statistic itself will yield a test procedure which is remarkably insensitive to bandwidth choice. Section 3.1 and Appendix A present simulation results which demonstrate the improvements in empirical size from pivoting both with and without the pivoting of the derivative estimates.

There is an issue of the bias of the nonparametric estimator $\hat{\beta}_h(x_i)$ which could be raised since, though $\hat{\beta}_h(x_i)$ is consistent, it is biased in small samples. There are three common approaches to addressing this problem. The first involves explicit bias correction, the second involves the use of higher order kernels, while the third involves undersmoothing of the estimate. It will be seen that the third approach will have obvious advantages in this context. This is due to the fact that the proposed test is insensitive to the choice of bandwidth, therefore if one is concerned with the adverse effects of bias one can simply choose to undersmooth the estimated derivatives and this will not adversely affect the size of the test. This being said, either of the three approaches mentioned above may be utilized if one is concerned with the adverse consequences of small sample bias.

Finally, there is the question of how valid the pivotal bootstrap procedure is in this context. The above modifications of the standard bootstrap utilize improvements known to be the best currently available. Horowitz (1991) bootstrapped a smoothed maximum score estimator which, like the kernel estimator, is not $\sqrt{n}$ consistent. His Monte Carlo evidence suggests that critical values based on the percentile-t are much more accurate than those obtained from first-order asymptotic theory. The validity of the pivotal bootstrap in this context can be checked via Monte Carlo results, and this issue is addressed in Section 3.1.

2.4. **The Resampling Algorithm.** The algorithm presented here is for the case of *iid* random variables. Extension to the case of general stationary observations would follow by replacing the bootstrap below with the more sophisticated resampling procedure found in Künsch (1989), while the proposed test procedure would remain unchanged.

The bootstrapping algorithm for obtaining the null distribution of the test statistic $\hat{t}$ proceeds as follows:

1. Estimate the 'restricted' conditional mean $E(Y|x_{(-j)i}, \bar{x}_{(j)i})$. The resulting fitted conditional mean is denoted $\hat{M}(x_{(-j)i}, \bar{x}_{(j)i})$, $i = 1, \ldots, n$. Note that since the null is $E(Y|X) \perp X_{(j)}$, this

restricted conditional mean does not vary with the variables whose significance are to be tested ($X_{(j)}$) since they are held constant at their means for all $i = 1, \ldots, n$.

2. Generate residuals $\hat{\epsilon}_i = y_i \perp \hat{M}(x_{(-j)i}, \bar{x}_{(j)i})$, $i = 1, \ldots, n$, then re-center them around the value zero (Freedman 1981). Note that these residuals are constructed under the null.
3. Generate the empirical distribution $\hat{F}$ which has probability mass $1/n$ at $\hat{\epsilon}_i$. That is,

$$\hat{F}: \text{ mass } \frac{1}{n} \text{ at } \hat{\epsilon}_i, \quad i = 1, \ldots, n.$$

4. Draw a 'bootstrap residual sample' from $\hat{F}$ by sampling *with replacement* from $\hat{F}$ and call this bootstrap sample $\{\epsilon_i^*\}_{i=1}^n$.
5. Generate a 'null bootstrap data sample' with dependent variable generated from

$$y_i^* = \hat{M}(x_{(-j)i}, \bar{x}_{(j)i}) + \epsilon_i^*, \quad i = 1, \ldots, n$$

The bootstrap sample will be $\{y_i^*, x_i\}_{i=1}^n$, where the conditioning variables are those from the original sample and therefore contain both $X_{(j)}$ and $X_{(-j)}$.

6. Obtain the bootstrap estimators $\hat{\beta}(x_i)^*$ and $SE(\hat{\beta}(x_i)^*)$ using $\{y_i^*, x_i\}_{i=1}^n$. Using these bootstrap estimators, calculate $\hat{t}^* = \hat{\lambda}^*/SE(\hat{\lambda}^*)$ where $SE(\hat{\lambda}^*)$ is obtained according to Section 2.5 below. This will yield one bootstrap replication 'null' value of the test statistic $\hat{t}$ under H_0.
7. Independently repeat Step 6 a large number of times obtaining bootstrap replications $\hat{t}_1^*$, $\hat{t}_2^*$, ..., $\hat{t}_{B_1}^*$ where B_1 is the number of bootstrap replications.

2.5. **Pivoting the Test Statistic via Resampling.** The standard error of the test statistic denoted by $\sigma_{\hat{\lambda}}$ is required, an estimate of which can be obtained via nested resampling. Resampling proceeds from the $\{Y, X\}$ pairs used to compute a given value of the statistic, and proceed as follows:

1. For a sample $\{Y, X\}$ used to compute the test statistic $\hat{\lambda}$, draw a resample maintaining the (Y_i, X_i') pairs. That is, resample $Z_i' = (Y_i, X_i')$. Call a given resample $\{Y^*, X^{*'}\}$.
2. Given the resample $\{Y^*, X^{*'}\}$, compute $\hat{\lambda}^*$, the test statistic based on this resample.
3. Repeat steps 1 and 2 B_2 times, and call the resampled test statistics $\hat{\lambda}_1^*$, $\hat{\lambda}_2^*$, ..., $\hat{\lambda}_{B_2}^*$. Note that since the variance is being estimated rather than tail percentiles, a fairly small number of resamples will be required.
4. Given the B_2 resampled values of a given value of the test statistic, compute their standard deviation, and call this $SE(\hat{\lambda})$.
5. The pivotal value of a given value of the test statistic will therefore be given by $\hat{t} = (\hat{\lambda} \perp \lambda_0)/SE(\hat{\lambda}) = \hat{\lambda}/SE(\hat{\lambda})$.

This approach can be applied for both the test statistic and those values of the test statistic computed under the null. The pivoted value of the test statistic is denoted by $\hat{t} = \hat{\lambda}/SE(\hat{\lambda})$, and the pivoted values of the test statistic under the null by $\hat{t}_i^* = \hat{\lambda}_i^*/SE(\hat{\lambda}_i^*), i = 1, 2, \ldots, B_1$.

2.6. **Decision Rules for the Proposed Test.** Having computed and pivoted the test statistic $\hat{\lambda}$ and having obtained the empirical sampling distribution of this test statistic under the null, the $(1 \perp \alpha)$ percentile $t_{1-\alpha}^*$ can be obtained where $t_{1-\alpha}^*$ is that value of t such that

$$Pr[t > t_{1-\alpha}^*] = \alpha \tag{7}$$

A test of size α can therefore be conducted by obtaining the null distribution for $\hat{t}$ as outlined above and then determining whether $\hat{t} > t_{1-\alpha}^*$. If so H_0 is rejected, otherwise we fail to reject H_0.

2.7. **Appropriate Number of Bootstrap Resamples and Bootstrap Pivot Resamples.** The number of bootstrap replications is always context dependent. If one uses the percentile method for obtaining confidence intervals or empirical p-values, a large number of replications might be necessary to get reasonable accuracy in the tails of the distribution. However, if one is not obtaining tail percentiles but is simply estimating low-order moments, then a small number of replications can suffice. For recent work on the appropriate number of bootstrap replications, see Hall and Titterington (1989).

The proposed approach adopts the percentile method and, in addition, requires the estimation of a variance for pivoting. Therefore, in this context 1,000 replications are recommended for obtaining the tail percentiles, while 100 replications are recommended for estimation of the variance. Clearly, the higher the number of replications the better, but these suggested values appear sufficient to get extremely good accuracy of empirical size across a wide range of settings and sample sizes.

Although intuitively one would expect the bootstrap to provide consistent estimates of the distributions of the test statistics considered in the paper, this has yet to be rigorously proven. The results of Hall and others would indicate that this is the case.

3. APPLICATIONS

3.1. **Empirical Size.** As noted in Section 1, the nonparametric asymptotic-based testing procedures suffer from the fact that the outcome of the tests are sensitive to the choice of bandwidths since the null distributions do not depend on the bandwidth. The proposed test should not suffer from this drawback since the null distribution depends explicitly on the bandwidth. In this section the empirical size of the proposed test is examined when the bandwidths deviate significantly from their optimal values. Given the inherent sampling variability of data-driven bandwidth selection procedures this is perhaps the single most important practical issue to be addressed and it can only be examined via simulation.

All computations which follow were performed on a 90mhz Pentium. Source code was written in ANSI C, and was compiled using gcc 2.6.3. The multivariate Gaussian kernel was used throughout. Tests were conducted with nominal sizes of $\alpha = 0.01, 0.05, 0.10$. The sample size was set at $n = 50$. There were $1,000$ bootstrap replications ($B_1 = 1,000$), 100 pivot ($B_2 = 100$) replications, and $1,000$ Monte Carlo replications. The following DGP was simulated for the experiment:

$$y_i = \sin(2\pi x_{1i}) + \epsilon_i \tag{8}$$

A variable X_2 was generated which was unrelated to $E(Y|X)$. The data for X_1 and X_2 were distributed $U[0,1]$ and the disturbance term was distributed $N(0, 0.25)$.

The estimated model was of the form

$$y_i = E[Y|x_{1i}, x_{2i}] + \epsilon_i \tag{9}$$

We wish to test whether the variable X_2 is significant or not, and the null is $H_0 : E(Y|X) \perp X_2$.

For what follows, the scaling factor for the bandwidth for variable j refers to the constant c_j in the formula for the optimal bandwidth for the kernel employed, $c_j \sigma_j n^{-1/(4+p)}$, where p is the number of conditioning variables (in this case $p = 2$) and where σ_j denotes the standard deviation of X_j. The unknown constant c_j depends on the joint distribution of X and on the kernel function. However, this constant can be obtained by data-driven methods such as leave-one-out cross-validation (CV). For an overview of CV see Stone (1974).

For this simulated DGP, bandwidth choice via CV was investigated to determine the likely range of values for c_1 and c_2 which would be encountered in a practical setting. For the 1,000 simulated data sets for which CV was applied the mean of c_1 was 0.24, and the mean of c_2 was 5×10^5. These mean values will be referred to as the 'optimal values' for the following simulations.

Given the need to use data-driven methods for bandwidth selection for almost all practical settings, and given the inherent sampling variability in bandwidths obtained by such methods, it is highly desirable that the outcome of any proposed test not depend on bandwidth choice. Therefore, the empirical size of the proposed test was calculated for $0.18 \le c_1 \le 0.30$, and for $0.5 \le c_2 \le 10.0$. Note that for $c_2 > 10.0$ the results do not differ quantitatively from those for $c_2 = 10.0$ and hence are not reported. These ranges for the bandwidths include the likely range of values which would be chosen by CV for this DGP.

Table 1 below considers the effects of deviations of the bandwidths from their optimal values on empirical size of the proposed test. Scaling is linear in its effect, therefore, going from $c_2 = 1.0$ to $c_2 = 5.0$ represents a 500% increase in the bandwidth. The upper left values in Table 1 denote empirical size when the conditional mean is dramatically undersmoothed, while those in the lower right corner correspond to oversmoothing. Boldface entries denote empirical size for the optimal bandwidth, while values marked with an asterisk differ significantly from nominal size at the 1% level. Appendix B presents some of the estimated conditional means for the range of bandwidths found in the table below in order to convey the effect of the range of bandwidths considered.

Nominal Size: 0.01

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.01	0.01	0.02	0.01
0.24	0.01	0.01	0.02	**0.01**
0.30	0.01	0.01	0.02	0.01

Nominal Size: 0.05

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.05	0.07*	0.06	0.05
0.24	0.06	0.06	0.06	**0.05**
0.30	0.06	0.05	0.06	0.05

Nominal Size: 0.10

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.11	0.14*	0.11	0.10
0.24	0.10	0.12	0.11	**0.10**
0.30	0.12	0.10	0.10	0.10

Table 1: Empirical sizes of the proposed test, $\hat{\alpha}$

For this example, CV yields a range of bandwidths for which nominal size does not differ significantly from the actual size. These values are found in the rightmost column of Table 1. Large deviations of the bandwidths from their optimal values leave the test's size virtually unaffected, highlighting the practical appeal of the proposed test given the need to use data-driven bandwidth selection techniques in practice. These results demonstrate that the proposed test is remarkably *insensitive* to the choice of the bandwidth, unlike the asymptotic tests.

3.2. **Mis-Specification and Power.** Suppose you are presented with a sample of data $\{y_i, x_{1i}, x_{2i}\}$ of size $n = 100$. In the absence of prior knowledge about the true DGP the following linear regression model is estimated using the method of least squares.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i \tag{10}$$

Coefficient	Estimate	Standard Error	t-statistic
$\hat{\beta}_0$	-0.015677	0.41287	-0.037971
$\hat{\beta}_1$	-0.33105	0.69603	-0.47562
$\hat{\beta}_2$	-0.013421	0.69603	-0.019282
$\hat{\beta}_3$	0.61423	1.1734	0.52346
R^2	-0.0217439		
$F_{3,96}$	0.297721		
$\log L$	-157.679		

Table 2: Summary of Parametric Test Results

The parameters β_1, β_2, and β_3 are not significant based on this model, either individually or jointly. Based on this model it is concluded that $E[Y|X] \perp X_1, X_2$.

Now the above hypotheses is tested using the proposed test. The following table summarizes the results of three tests - one for $E[Y|X] \perp X_1$, one for $E[Y|X] \perp X_2$, and one joint test for $E[Y|X] \perp X_1, X_2$. Bandwidths were chosen via leave-one-out CV.

Variable	$\hat{t}$	$\hat{t}^c_{0.95}$	$\hat{p}$
X_1	6.57	2.81	$p < 0.001$
X_2	3.45	2.26	$p < 0.001$
X_1, X_2	10.1	4.00	$p < 0.001$

Table 3: Summary of Proposed Test Results

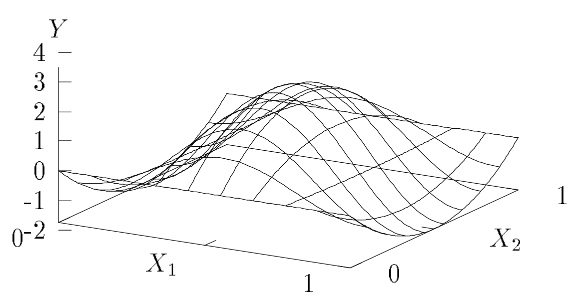
The results are clear rejection that X_1, X_2, and (X_1, X_2) jointly are independent of the conditional mean of Y.

Now suppose that the true state of nature is revealed. The true DGP for the data used above is given by

$$\begin{aligned} y_i &= E(Y|x_{1i}, x_{2i}) + \epsilon_i \\ &= 8.0 \times \cos(2.0\pi x_{1i}) \times (x_{2i}^2 \perp x_{2i}) + (x_{2i} \perp x_{2i}^2) + \epsilon_i \end{aligned} \tag{11}$$

where $X_1 \sim U[0,1]$, $X_2 \sim U[0,1]$ and $\epsilon_i \sim N(0, 0.5)$. Note that this is a highly nonlinear function and is twice continuously differentiable. This DGP is graphed below.

Figure 1: Actual DGP



Having had the true state of nature revealed to us, it is seen that the linear parametric model is clearly mis-specified. In fact, the estimated linear model is a horizontal plane through the data, hence the estimated parameters/derivatives will be close to zero. In this case the mis-specification of the functional form of the conditional mean $E[Y|x_1, x_2]$ has led to tests which are not asymptotically valid and which possess incorrect size and low power regardless of the sample size. In this example, the parametric model would lead one to believe that the conditioning variables X_1 and X_2 do not help explain movements in the dependent variable Y. In fact, quite the opposite is true. Clearly the parametric model above would fail the most simple specification test, however, it is very unlikely that the true model would be found in practice. Hence tests based on any other parametric model other than Equation (11) would not be, strictly speaking, valid.

The point to be made is that parametric models will always be mis-specified to some degree, and a practitioner might likely encounter situations such as that above in which inference based on the parametric model is incorrect and hence misleading.

3.3. **Testing for the Unpredictability of Exchange Rates.** The market efficiency hypothesis applied to foreign exchange rates is typically interpreted to mean that there is no information contained in past percentage changes in exchange rates which can be used to predict future percentage changes in exchange rates. This hypothesis is referred to as 'unpredictability of exchange rates'.

The linear unpredictability of exchange rates has a long history going back to early work on efficient markets such as that by Fama (1965) and Cootner (1964). In addition, conditional heteroskedasticity in exchange rates has been repeatedly documented (Diebold 1988). The typical parametric characterization of exchange rate dynamics has been that of linear conditional means with nonlinearities working through the conditional variance in the form of autoregressive conditional heteroskedasticity (ARCH) and related effects.

Recent work by Diebold and Nason (1990) has questioned two related aspects of this parametric approach. First, there is the question of whether the conditional mean is truly linear. Second, there is the question of whether the ARCH effects may be an artifact of neglected nonlinearities in the conditional mean. Since both of these questions are concerned with potential mis-specification of the conditional mean process, this would appear to be a good application for the proposed test. Diebold and Nason (1990) do not attempt any direct testing and they simply compare out-of-sample predictions of locally weighted regression (LWR) (Cleveland, Devlin and Grosse 1988) versus a parametric random walk specification. The approach presented in this paper goes beyond the work of Diebold and Nason (1990) and allows us to actually test the market efficiency hypothesis without assuming the functional form for the conditional mean.

Following the methodology of Diebold and Nason (1990), data for nominal weekly dollar spot rates for the G7, Friday average, (S_t) were collected from Citibase. All data are measured in cents per unit of foreign currency. Each series contains 636 observations, and each series begins 1/4/1980. Following Diebold and Nason (1990), interest focuses on percentage exchange rate changes $\Delta \log S_t$, thereby avoiding potential problems associated with estimation of nonstationary regression functions and highly collinear conditioning variables. It is worth noting that this transformed exchange rate series $\Delta \log S_t$ may not in fact be an *iid* series, however, for this application I shall proceed under the assumption that it is. Diebold and Nason (1990) consider lag structures of one, three, and five lags. Their findings were unaffected by using different lag structures, and they conclude that "Our findings bode poorly for recent conjectures that exchange rates contain nonlinearities exploitable for enhanced point prediction".

Results of the proposed test of the hypothesis $H_0 : E[\Delta \log S_t|\Delta \log S_{t-1}] \perp \Delta \log S_{t-1}$ corresponding to the one lag structure of Diebold and Nason (1990) are given in the following table. Bandwidths were chosen via leave-one-out CV. Estimated critical values for $\alpha = 0.05$ are given along with the value of the test statistic and the empirical p-value. Graphs of the estimated conditional means are found in Appendix C.

Country	$\hat{t}$	$\hat{t}^c_{0.95}$	$\hat{p}$
Canada	2.92	1.52	$p < 0.01$
France	3.54	1.41	$p < 0.01$
Germany	3.69	1.46	$p < 0.01$
Italy	3.48	1.52	$p < 0.01$
Japan	4.53	1.38	$p < 0.01$
UK	4.16	1.38	$p < 0.01$

Table 4: Predictability Test for G7 Exchange Rates

For each series for the period considered, the hypothesis $E[\Delta \log S_t | \Delta \log S_{t-1}] \perp \Delta \log S_{t-1}$ was rejected at all conventional levels. The estimated derivatives average from 0.2 to 0.3, and there is strong evidence of a significant and positive relationship between $E[\Delta \log S_t | \Delta \log S_{t-1}]$ and $\Delta \log S_{t-1}$ given the outcome of the proposed test for this data.

Given this statistically significant rejection of the null for all six series, the root mean squared prediction error (RMSPE) was then computed for the one-step forecasts based on both the random walk hypothesis (RW) and the nonparametric forecasts (NP). The model was fit on the first $T = 500$ observations, and ex-ante one-period forecasts were computed for the remaining observations in the series which were not included in the estimation sample. That is, given $\Delta \log S_t$, the fitted model was used to forecast $\Delta \log S_{t+1}$, $t = 501, 502, \ldots, 633$. The bandwidth was selected by leave-one-out CV for the observations on which the model was fit, making this a completely ex-ante approach. The following table presents RMSPE for the RW and NP forecasts.

Canada		France		Germany	
RW	NP	RW	NP	RW	NP
0.004471	0.004277	0.013248	0.012848	0.013915	0.013479

Italy		Japan		UK	
RW	NP	RW	NP	RW	NP
0.012802	0.012346	0.011611	0.010959	0.014016	0.013587

Table 5: One-Step Forecast RMSPE for G7 Exchange Rates

The significant nonlinear relationship detected by the proposed test can, in this instance, be exploited for improved one-step-ahead forecasting over that obtained assuming that $\Delta \log S_t$ follows a random walk. These results were robust to a very wide range of fitting/evaluation splits in the series. These results suggest that, for the exchange rate series considered and for the time-period considered, percentage changes in weekly exchange rates possess small but significant nonlinear persistence which can be exploited for one-step-ahead forecasting. These results run counter to the findings of Diebold and Nason (1990). Whether these findings can be exploited accounting for both risk and transactions costs remains an open question.

4. CONCLUSION

The test of significance is probably the most widely used test statistic in the context of multivariate regression. Its importance stems partly from the fact that the significance test is often used to confirm or refute theories and so incorrect size or low power would have important practical and theoretical implications. In this paper the test of significance is considered in the context of nonparametric kernel regression. The approach taken is based on the application of resampling methods and resolves some important outstanding practical issues regarding hypothesis testing in a nonparametric framework.

The motivation for using nonparametric instead of parametric methods for both estimation and hypothesis testing derives from the fact that employing a mis-specified parametric model will typically result in inconsistent parameter estimates and significance tests possessing both asymptotically incorrect size and low power. The utility of nonparametric estimation techniques is due to the fact that they are robust to functional mis-specification for a wide class of data generating processes. Hypothesis tests based on such models do not therefore suffer from the adverse effects of functional mis-specification.

The proposed test statistic is based on nonparametric estimates of derivatives of an unknown conditional mean with respect to the conditioning variables. A resampling technique known as nested pivotal bootstrapping is used to derive the null distribution of the test statistic.

Competing approaches in the context of nonparametric regression have been based on derivations of the asymptotic or limiting distribution of similar test statistics. The application of resampling techniques resolves one extremely troublesome aspect of tests based on limiting distributions, that the test statistics' value depends on a bandwidth while the limiting distribution does not, and hence the outcome of a test based on limiting distributions is highly sensitive to bandwidth choice. The test statistic proposed in this paper and the associated null distribution depends explicitly on the bandwidth, and the proposed test is therefore remarkably insensitive to the choice of bandwidth. Furthermore, the empirical size of the test based on cross-validated bandwidths does not differ significantly from the nominal size for the simulation undertaken and is expected to be the case in general.

The main contributions of the proposed approaches are threefold. First, the proposed significance test has correct size and in addition possesses power in the direction of the class of twice-continuously differentiable alternatives. Second, both the test statistic and its null distributions depend explicitly on the bandwidth, a feature lacking if the null distributions are derived using asymptotic theory. Third, it is believed that the test statistic has the same rate of convergence as those based on parametric models due to the form of averaging employed in the construction of the statistics, though this is beyond the scope of this paper and remains the subject of ongoing research.

This paper represents part of an ongoing project whose goal is that of working towards a unified approach to estimation, inference, and hypothesis testing in a nonparametric context. The test of significance is widely used and can have important practical and theoretical implications, but clearly there is much to be done before a sound, unified, and workable nonparametric framework exists.

5. Acknowledgments

I would like to thank Hal White and Clive Granger for their numerous insightful comments and suggestions, and two anonymous referees whose comments were most helpful. In addition, I would like to thank participants at the Econometric Society meetings, the CESG meetings, and participants at the Econometrics Workshops at the University of California San Diego. All errors remain, of course, my own.

Appendix A. Empirical Sizes of Non-Pivotal Test Statistics

It is clear that the pivotal approach taken in this paper is more computationally intensive than the standard bootstrap. The question arises as to the practical benefit of the application of pivotal bootstrapping for the situation at hand. To answer this question, the empirical size for two test statistics is considered. The first is the unpivoted statistic given in Equation (4), while the second is the pivoted version of Equation (4). Note that the proposed statistic involves first pivoting the derivative estimates and then considering a pivoted version of this pivoted statistic.

Table A therefore gives the empirical size for the statistic for which no pivoting occurs,

$$\hat{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(x_i)_j\right]^2 \tag{12}$$

Table A

Nominal Size: 0.01

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.07*	0.14*	0.10*	0.10*
0.24	0.02	0.07*	0.04*	**0.04***
0.30	0.01	0.05*	0.03*	0.04*

Nominal Size: 0.05

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.22*	0.34*	0.29*	0.28*
0.24	0.12*	0.22*	0.18*	**0.18***
0.30	0.09*	0.17*	0.13*	0.15*

Nominal Size: 0.10

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.36*	0.50*	0.44*	0.43*
0.24	0.23*	0.37*	0.30*	**0.29***
0.30	0.18*	0.29*	0.24*	0.23*

Table A.1: Empirical Size.

Note that the empirical sizes for the raw statistic are so poor as to render the test statistic in this form virtually unuseable.

Table B gives the empirical size for the statistic for which there is no pivoting of the derivative estimates, but the statistic $\hat{\hat{\lambda}}$ given in Equation (12) is pivoted, that is,

$$t = \frac{\hat{\hat{\lambda}}}{SE(\hat{\hat{\lambda}})} \tag{13}$$

Table B

Nominal Size: 0.01

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.01	0.01	0.01	0.01
0.24	0.01	0.01	0.01	**0.01**
0.30	0.01	0.01	0.01	0.01

Nominal Size: 0.05

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.06	0.03*	0.04	0.03*
0.24	0.04	0.04	0.04	**0.04**
0.30	0.05	0.04	0.05	0.05

Nominal Size: 0.10

$c_1 \vert c_2$	0.5	1.0	5.0	10.0
0.18	0.11	0.07*	0.08	0.08
0.24	0.10	0.08	0.09	**0.08**
0.30	0.11	0.09	0.10	0.09

Table A.2: Empirical Size.

Note that pivoting the statistic given in Equation (4) yields a dramatic improvement in empirical size. These results are just slightly worse than that for the proposed statistic.

Appendix B. Degree of Smoothing and Size

One of the main contributions of this paper is the fact that the proposed test is remarkably insensitive to bandwidth choice. The estimates plotted below are those for the center row of Table 1 in Section 3.1. The empirical sizes for tests based on each of these estimated conditional means differ by at most 2%, yet clearly these differ greatly in the amount of smoothing occurring. The estimates range from a severely undersmoothed estimate ($c_1 = 0.24, c_2 = 0.5$) to an appropriately smoothed one ($c_1 = 0.24, c_2 \geq 10$). Note that for all of these cases, the empirical size is very close to the nominal size. Finally, again note that when using CV, the empirical and nominal sizes do not differ significantly, and these values are to be found in the rightmost column of Table 1.

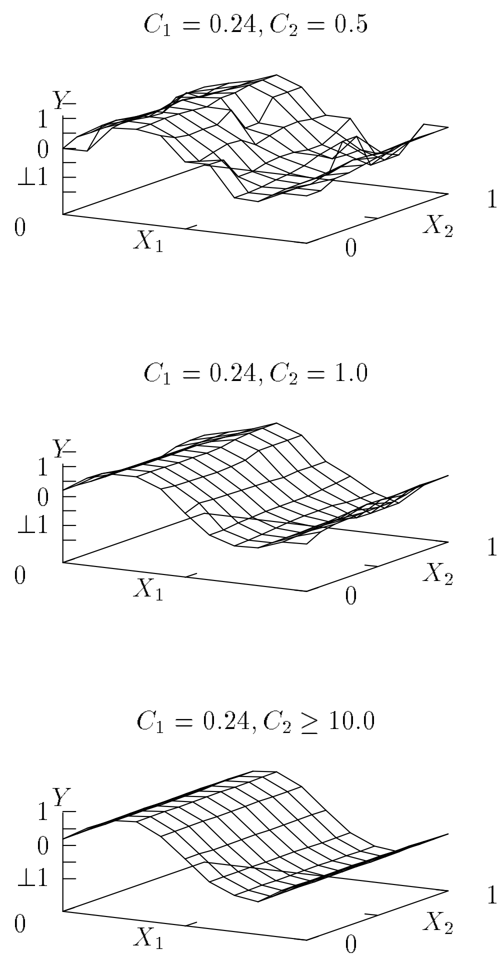


Figure B.1: Range of Bandwidths and Degree of Smoothing.

Appendix C. Degree of Smoothing and Size

The following graphs present the data and kernel estimates of the conditional mean $E[\Delta \log S_t | \Delta \log S_{t-1}]$ for G7 exchange rates for the case of one lag.

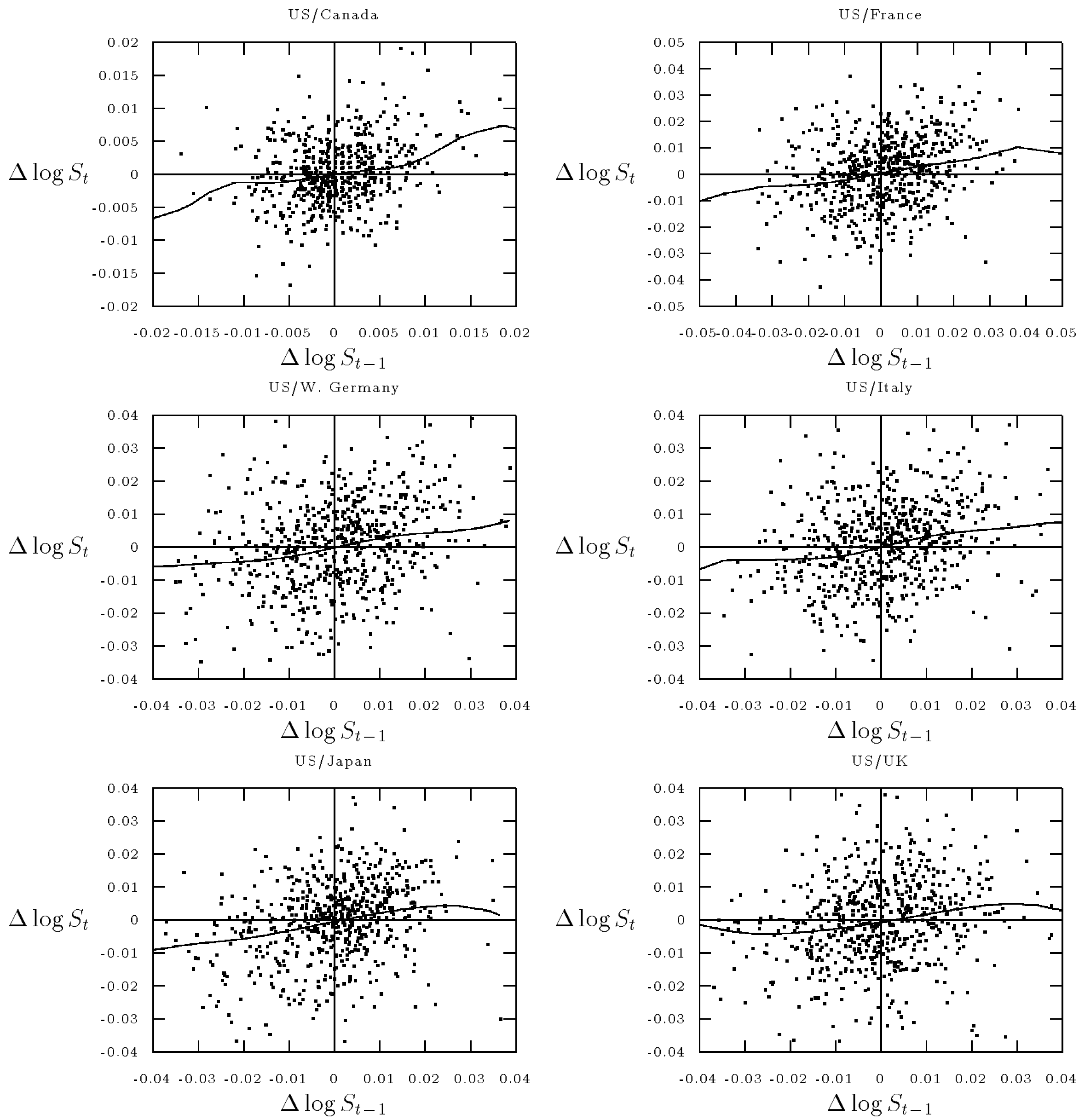


Figure C.1: Kernel Estimates of the Conditional Mean for G7 Exchange Rates.

Note that there appears to be a common small and positive nonlinear relationship present between $\Delta \log S_t$ and $\Delta \log S_{t-1}$ for all series.

References

Beran, R. (1988), 'Prepivoting test statistics: A bootstrap view of asymptotic refinements', *Journal of the American Statistical Association* **83**, 687–697.

Bickel, P. and Freedman, D. (1981), 'Some asymptotic theory for the bootstrap', *Annals of Statistics* **9**(6), 1196–1217.

Bickel, P. and Freedman, D. (1983), Bootstrapping regression models with many parameters, *in* 'A Festschrift for Erich Lehmann', Wadsworth, Belmont, California, pp. 24–48.

Cleveland, W., Devlin, S. and Grosse, E. (1988), 'Regression by local fitting: Methods, properties, and computational algorithms', *Journal of Econometrics* **37**, 87–114.

Cootner, P. (1964), *The Random Character of Stock Market Prices*, MIT Press, Cambridge, MA.

Diebold, F. X. (1988), *Empirical Modeling of Exchange Rate Dynamics*, Springer-Verlag, New York.

Diebold, F. X. and Nason, J. A. (1990), 'Nonparametric exchange rate prediction', *Journal of International Economics* **28**, 315–332.

Efron, B. (1983), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19103.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York, London.

Fama, E. F. (1965), 'The behavior of stock market prices', *Journal of Business* **38**, 34–105.

Freedman, D. A. (1981), 'Bootstrapping regression models', *The Annals of Statistics* **9**, 1218–1228.

Hall, P. (1986), 'On the bootstrap and confidence intervals', *The Annals of Statistics* **14**, 1431–1452.

Hall, P. (1988), 'Theoretical comparison of bootstrap confidence intervals', *The Annals of Statistics* **16**, 927–953.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer Series in Statistics, Springer-Verlag, New York.

Hall, P. and Titterington, D. (1989), 'The effect of simulation order on level accuracy and power of monte carlo tests', *Journal of the Royal Statistical Society* **B 51**, 459–467.

Härdle, W. and Marron, S. (1991), 'Bootstrap simultaneous error bars for nonparametric regression', *Annuls of Statistics* **19**, 778–796.

Härdle, W. and Stoker, T. (1989), 'Investigating smooth multiple regression by the method of average derivatives', *Journal of the American Statistical Association* **84**, 986–995.

Horowitz, J. L. (1991), 'Bootstrap based critical values for the information matrix test', *Working Paper #91-21, University of Iowa* .

Jeong, J. and Maddala, G. S. (1993), 'A perspective on application of bootstrap methods in econometrics', *Handbook of Statistics* **11**, 573–610.

Jones, M., Marron, J. and Sheather, S. (1992), 'Progress in data-based bandwidth selection for kernel density estimation', *Department of Statistics Mimeo Series, Chapel Hill, North Carolina, #2088* .

Künsch, H. R. (1989), 'The jackknife and the bootstrap for general stationary observations', *The Annals of Statistics* **17**(3), 1217–1241.

Lavergne, P. and Vuong, Q. (1992), 'Nonparametric selection of regressors: The nonnested case', *Mimeo, INRA-ESR Toulouse* .

Mammen, E. (1992), *When Does Bootstrap Work? Asymptotic Results and Simulations*, Springer-Verlag, New York.

Nadaraya, E. A. (1965), 'On nonparametric estimates of density functions and regression curves', *Theory of Applied Probability* **10**, 186–190.

Powell, J. L., Stock, J. L. and Stoker, T. M. (1989), 'Semiparametric estimation of index coefficients', *Econometrica* **57**(6), 1403.

Rilstone, P. (1991), 'Nonparametric hypothesis testing with parametric rates of convergence', *International Economic Review* **32**, 209–227.

Robinson, P. M. (1991), 'Consistent nonparametric entropy-based testing', *Review of Economic Studies* **58**, 437–453.

Robinson, P. M. (1994), 'The normal approximation for semiparametric averaged derivatives', *Econometrica* **63**(8), 667–680.

Stoker, T. M. (1989), 'Tests of additive derivative constraints', *Review of Economic Studies* **56**, 535–552.

Stone, C. J. (1974), 'Cross-validatory choice and assessment of statistical predictions (with discussion)', *Journal of the Royal Statistical Society* **36**, 111–147.

Watson, G. S. (1964), 'Smooth regression analysis', *Sanikhya* **26:15**, 175–184.