# A permutation test for randomness with power against smooth variation

Fernando Tusell[1]

[1] Departamento de Estadística y Econometría. Facultad de CC.EE. y Empresariales, Avda. del Lehendakari Aguirre, 83, 48015 BILBAO (Spain). E-mail: `etptupaf@bs.ehu.es`.

## Abstract

A permutation test for the white noise hypothesis is described, offering power against a general class of smooth alternatives. Simulation results show that it performs well, as compared with similar tests available in the literature, in terms of power. An example demonstrates its use in a particular problem in which a test for randomness was sought without any specific alternative.

**Keywords:** randomness; serial independence; permutation tests; computer intensive tests; smoothing; complexity.

# 1 Introduction

With widely available cheap computing power and the parallel improvement in algorithms, approaches that would have been infeasible a few years back are now within the reach of the desktop personal computer. This paper presents one example. We propose to test the null hypothesis $Y = m + \epsilon$ where $m$ is a constant and $\epsilon$ random noise versus $Y = f(x) + \epsilon$, where $f(.)$ is a function required only to be "smooth." The test statistic can be regarded as a penalized goodness of fit criterion. If the fit is significantly better than could be expected for random noise, the null is rejected. Significance is assessed by means of a permutation test.

The structure of the paper is as follows: Section 2 introduces notation, standard methods and results in smoothing and describes the test. Section 3 reviews some related tests, which then are compared to ours in a small simulation in Section 4. More extensive results are available from the author: we have tried to summarize the results in a few typical situations. Section 5 ends with some remarks about the proposed test.

# 2 The test

Let $(y_i, x_i), i = 1, \ldots, n$ be a sample from a bivariate variable, and assume that $Y = f(x) + \epsilon$, where $\epsilon \overset{\text{i.i.d.}}{\sim} (0, \sigma^2)$. Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ The design points $\mathbf{x}^T = (x_1, \ldots, x_n)$ are assumed fixed. Usually they will be equispaced points, but this is not a requirement. If $f(x)$ were a linear function of $x$ we would have the standard linear model with one regressor; but a much more flexible class of functions can be postulated, requiring only regularity conditions such as continuity of $f(.)$ and perhaps of some of its derivatives.

Several methods have been put forward for the estimation of $f(.)$. Kernel estimators approximate $f(x)$ by a weighted average of the values $y_i$ for $x_i$ "close" to $x$. A bandwidth parameter $h$ controls the weights —see for example Hart (1997) or Härdle (1990) among many comprehensive references.

Splines provide a different (although not unrelated: see Silverman (1984)) smoothing method. A cubic spline with nodes at $\psi_1, \ldots, \psi_k$ is a piecewise polynomial with first two continuous derivatives and a third derivative which is constant within each interval $(\psi_i, \psi_{i+1})$. It can be shown that the minimizer of

$$L(\mathbf{y}, \mathbf{x}, \lambda; g) = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g''(x)]^2 \, dx \qquad (1)$$

over functions $g(x)$ with two continuous derivatives, is a cubic spline with nodes at $(x_1, \ldots, x_n)$. The first term in (1) requires $g(x)$ to pass close to the points $(x_i, y_i)$, while the second penalizes departures from linearity, thus forcing $g(x)$ to be smooth. The parameter $\lambda$ takes up the role of $h$ in kernel estimators, and specifies the desired trade-off between goodness of fit and smoothness. We can choose $\lambda$ from the data, a popular method being generalized cross-validation (GCV) —see Craven and Wahba (1979). We denote $\lambda_{\text{GCV}}$ the smoothing parameter chosen by GCV, and

$$L_{\text{GCV}} = \min_{g} L(\mathbf{y}, \mathbf{x}, \lambda_{\text{GCV}}; g) \qquad (2)$$

the corresponding minimum value of (1).

Both kernels and the minimizer of (1) give smoothed estimates of the form $\hat{\mathbf{y}} = S_\lambda \mathbf{y}$, where $S_\lambda$ is a matrix dependent on the smoothing parameter ($\lambda$ or $h$) and the design points, but not on $\mathbf{y}$.

We are interested in testing:

$$\begin{array}{rll} & H_0: & Y = m + \epsilon \\ \text{versus} & H_a: & Y = f(x) + \epsilon, \end{array}$$

where $\epsilon$ is white noise, $m$ is a constant, and $f(x)$ is only required to be "smooth." Our alternative then includes trends and fluctuations with changing amplitude and phase, among others.

In the absence of any pattern, the sequence $\{y_i\}$ should fluctuate randomly about a fixed level. If we fit a cross-validated spline to $(y_i, x_i)$, $i = 1, \dots, n$, the fitted curve should be nearly a straight horizontal line: in a sense, a curve of lowest possible complexity, because there is no structure to adapt to. On the other hand, under the alternative, the spline might pick up the smooth function $f(x)$ and hence display greater complexity.

A natural idea then is to measure how much the complexity of the fitted spline deviates from what would be expected when fitting white noise. If the deviation is sufficiently large, we would reject the hypothesis of white noise.

The smoothing parameter chosen by generalized cross-validation, $\lambda_{\text{GCV}}$, can be regarded as a proxy for the complexity of the fitted spline, and used as a test statistic: the larger $\lambda_{\text{GCV}}$, the smoother ("less complex") the curve is. We could also use $\operatorname{tr}(S_\lambda)$), the number of "equivalent parameters", as a test statistic: see for example Hastie and Tibshirani (1991) for the rationale of equating $\operatorname{tr}(S_\lambda)$ to the number of equivalent parameters.

As an alternative, and more in keeping with the ideas in Rissanen (1989) on minimum description length modelling (MDL), we can use as a test statistic $L_{\text{GCV}}$, defined in (2). The first term measures the deviation of the data from the fits and the second the complexity of the "model" —here a non-parametric one. Besides being easier to rationalize as a penalized goodness of fit criterion or analog to a MDL criterion, $L_{\text{GCV}}$ has also been found preferable to $\lambda_{\text{GCV}}$ in terms of performance: $\lambda_{\text{GCV}}$ is very unstable. In the following we refer only to $L_{\text{GCV}}$.

Other criteria for choosing $\lambda$ can be used, besides GCV. Hurvich et al. (1998) give a detailed comparison of several such criteria. Using their notation, all of them can be written in the form

$$\log(\hat{\sigma}^2) + \phi(H) \tag{3}$$

where $\hat{\sigma}^2$ is the mean square residual, $H$ is the smoothing matrix (our $S_\lambda$), and $\phi(H)$ the penalty term. GCV minimizes (3) with $\phi(H) = -2\log(1 - \operatorname{tr}(H)/n)$, while a nonparametric version of AIC (assuming normality) uses $\phi(H) = 2\operatorname{tr}(H)/n$ and Rice's $T$ criterion uses $\phi(H) = -\log(1 - 2\operatorname{tr}(H)/n)$ (see Rice (1984)). Criterion AIC$_C$, a modified version of AIC, uses $\phi(H) = 1 + 2(\operatorname{tr}(H) + 1)(n - \operatorname{tr}(H) - 2)^{-1}$ (see Hurvich et al. (1998)). Fitting a spline that minimizes (3) for different choices of $\phi(H)$ above gives different "optimal" smoothings and corresponding minimum values $L_{\text{AIC}}$, $L_{\text{T}}$, etc., that can be used in place of $L_{\text{GCV}}$ in the test. All four criteria give very similar results in the simulation reported below.

In order to obtain a yardstick against which to measure the value of the test statistic, we resort to resampling after permutation. Since for realistic sizes of $n$ complete enumeration of all the $n!$ permutations is out of the question, we simulate a few hundreds or thousands. The proposed test is therefore described as follows:

**Permutation test.**

1. For the given sample $(y_i, x_i)$, $i = 1, \ldots, n$ compute $L_{\text{GCV}}$.
2. For $k = 1, \ldots, K$ do the following:
   (a) Shuffle the $y_i$ to generate $(y_{k(i)}, x_i)$, where $k(i)$, $i = 1, \ldots, n$, is a random permutation of the first $n$ integers.
   (b) Fit a cubic spline to the $(y_{k(i)}, x_i)$ and compute $L_{\text{GCV}}^{(k)}$
3. Reject the hypothesis of randomness at the approximate $100\alpha\%$ level of significance if $L_{\text{GCV}}$ is among the $\lfloor \alpha K \rfloor$ smallest $L_{\text{GCV}}^{(k)}$.

## 3 Related tests

There is an extensive literature on nonparametric smoothing and lack of fit tests: see for instance Hart (1997). Among the earlier proposals is von Neumann (1941). The von Neumann ratio

$$\sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 / s^2, \tag{4}$$

where $s^2$ is the sample variance of the $y_i' s$, can be used as a test for white noise with good power against smooth alternatives.

Raz (1990) proposes a test for no-effect in a nonparametric regression setting somewhat different from ours. He considers a situation in which both $m$ and $\sigma^2$ are known without error when the null hypothesis of no effect is true. Defining

$$\begin{aligned} Q_1 &= \sum_{i=1}^{n} \left( \hat{f}(x_i) - \overline{Y} \right)^2 & (5) \\ Q_2 &= \sum_{i=1}^{n} \left( Y_i - \hat{f}(x_i) \right)^2 & (6) \\ Q_3 &= \sum_{i=1}^{n} \left( Y_i - \overline{Y} \right)^2, & (7) \end{aligned}$$

he goes on to define by analogy with the usual linear regression statistics the generalized statistics

$$\begin{aligned} F_1 &= (n - g_1) Q_1 / \left[ (g_1 - 1)(Q_3 - Q_1) \right] & (8) \\ F_2 &= g_1^* (Q_3 - Q_2) / \left[ (n - g_1^* - 1) Q_2 \right], & (9) \end{aligned}$$

where $g_1 = \text{trace}(S_\lambda {S_\lambda}^T)$ and $g_1^* = \text{trace}\left[ (I - S_\lambda)(I - S_\lambda)^T \right]$. In linear regression, with the smoothing matrix $S_\lambda$ replaced by the familiar “hat” matrix $H = X(X^T X)^{-1} X^T$, $g_1 = p$, the number of parameters, $g_1^* = n - p$ and both $F_1$ and $F_2$ reduce to the usual

F test statistic. Raz (1990) uses as a test statistic $R = Q_1/\sigma^2$, a monotonic function of $F_1$ whose distribution is easier to approximate. In the linear regression case, under the null hypothesis of no effect, $R$ would follow a chi square distribution. In the non-parametric case, its distribution is approximated by a scaled chi square by matching the first two moments to those of $R$ (which, in the situation studied by Raz (1990), can be computed exactly).

One important departure of our test from Raz's, besides the use of a different test statistic, is that Raz (1990) considers a fixed amount of smoothing, while we optimize the smoothing. This would affect the distribution of Raz's $R$ statistic, much in the same way that choosing the best model in linear regression renders inadequate the use of the $F$ test for testing the no-effect hypothesis.

Buckley (1991) builds on work by Cox et al. (1988) to show that *cusum* type tests are locally most powerful for a wide class of alternatives. Let $E_p(x)$ be the space generated by polynomials $\sum_{k=0}^{p-1} \alpha_k x^k$ evaluated at the design points $x_i$, $i = 1, \ldots, n$. To test the adequacy of a polynomial regression model we could consider alternatives given by

$$Y_i = \sum_{k=0}^{p-1} \alpha_k x_i^k + bf(x_i) + \epsilon_i, \tag{10}$$

for $i = 1, \ldots, n$ and some $f(.) \notin E_p(x)$. In matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + b\mathbf{f} + \boldsymbol{\epsilon}, \tag{11}$$

where $\mathbf{X}$ is a $n \times p$ matrix whose $i, j$ term is $x_i^{j-1}$. Under the assumption of normality the usual F test for $H_0 : b = 0$ is most powerful against the alternative (10) especified by $\mathbf{f}$. If we do not have a specific alternative, no uniformly most powerful test exists. However, considering $\mathbf{f} \sim N(0, V)$, a locally most powerful test exists (see Cox et al. (1988)). Choosing $V$ conveniently, we can impose smoothness on $\mathbf{f}$. Buckley (1991) proposes to use $V = R^-$, where $R = D^T D$ and $D$ is the matrix that takes $p$-order differences, i.e.,

$$D\mathbf{f} = \nabla^p \mathbf{f}. \tag{12}$$

(We assume equally spaced points; otherwise, divided differences must be used, which adds some complexity but no essential changes.) Then,

$$\mathbf{f}^T R \mathbf{f} = \mathbf{f}^T D^T D \mathbf{f} = \sum_{i=1}^{n-p} \left(\nabla^p f(x_i)\right)^2. \tag{13}$$

Expressions such as (13) have been much used as penalty terms in splines, and are a reasonable way to penalize departure from a $p$-order polynomial (for which (13) is identically zero).

Let $B = I_n - X(X^T X)^{-1} X^T$ and $D$, $S$ be respectively $p$-fold difference and cusum matrices, defined by

$$D = \tilde{D}_{n+p-1} \cdots \tilde{D}_n \tag{14}$$

$$S = \tilde{S}_{n+p-1} \cdots \tilde{S}_n, \tag{15}$$

with $\tilde{D}_m$, $\tilde{S}_m$ defined (for $m \geq 2$) as the $(m-1) \times m$ matrices

$$\tilde{D}_m = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \qquad \tilde{S}_m = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 0 \end{bmatrix}.$$

Buckley (1991) then shows that the locally most powerful test uses the statistic $\hat{\sigma}^2_S/\hat{\sigma}^2_D$, where

$$\begin{aligned} \hat{\sigma}^2_S &= \mathbf{y}^T BS^T SB\mathbf{y} / \operatorname{tr}(BS^T SB) & (16) \\ \hat{\sigma}^2_D &= \mathbf{y}^T D^T D\mathbf{y} / \operatorname{tr}(D^T D) & (17) \end{aligned}$$

Moreover, when $p = 1$ he shows the relationship between his proposed most powerful *cusum* test and the von Neumann test, which can be expressed as a similar ratio less sensitive to departures from the null hypothesis.

Aerts et al. (1998) propose a test for goodness of fit to a parametric function, $\gamma(x_i; \theta_1, \dots, \theta_p)$. Basically, they consider a family of functions such as

$$\gamma(x_i; \theta_1, \dots, \theta_{p+r}) \quad = \quad \gamma(x_i; \theta_1, \dots, \theta_p) + \sum_{j=1}^{r} \theta_{p+j} u_j(x_i), \tag{18}$$

where the $u_j(x)$ functions span a suitably large space of functions (they could be Hermite polynomials or trigonometric functions, for instance). Their setup is slightly more complex in that it includes a link function which is inessential for our purposes. To assess the fit of their baseline model $\gamma(x_i; \theta_1, \dots, \theta_p)$, they propose to fit models with an increasing number of parameters starting at $p$, and use a modified AIC criterion (MAIC) to choose among them. If the criterion chooses a model with $q > p$ parameters, that is taken as evidence of a substantial departure from the baseline model.

Clearly, taking $p = 1$ and $\gamma(x_i; \theta_1) = \mu$ the proposal in Aerts et al. (1998) specializes to a method to test for random fluctuation about the mean against a smooth alternative. The nature of the functions $u_j(x)$, $j = 1, \dots, r$, specify the nature of alternatives considered, which can be quite general. Their method tests whether the inclusion of additional parameters is justified. In their case, those parameters appear explicitly multiplying the functions $u_j(x)$, while in our test statistic the complexity of the (nonparametric) "model" is taken up by the second term of (1).

The advantage for the method in Aerts et al. (1998) is that there is distributional theory (at least asymptotic) to aid in the realization of the test, while we have to resort to permutation. The disadvantages we see (compared to our method) are: i) The inclusion of a term such as $\theta_{p+\ell} u_\ell(x)$ has a global effect across all values of $X$ —the test appears less able to pick *local* departures from the baseline model than ours—; and ii) The order of inclusion of terms in the right hand side of (18) is undefined. It may happen that term $\theta_{p+\ell} u_\ell(x)$ points to a highly relevant departure from the baseline model, and yet, if $\ell$ is large, the improvement in the fit is not enough to warrant inclusion of all terms $\theta_{p+1} u_1(x), \dots, \theta_{p+\ell} u_\ell(x)$.

We cannot attempt a complete survey of all related tests in the literature. A good source of pointers is Hart (1997). We also would like to mention Ramil and González-Manteiga (1998) and González-Manteiga and Cao (1993), close in spirit to Raz (1990). Chapter 6 in Bowman and Azzalini (1997) contains also a discussion on tests for the "no effect" hypothesis against a smooth alternative. Their test statistic is similar to

Raz's. Rather than using a fixed smoothing, they propose to plot the observed $p$-value against the smoothing parameter $h$ for different values of $h$. Thus they obtain a "significant trace", a useful diagnostic of the sensitivity of the $p$ value to the amount of smoothing used.

# 4 Simulation results and an example

## 4.1 Power simulation

We have performed a simulation study to gain some insight into the performance of the test and obtain at least a crude approximation of its power. In the same simulation runs we computed various other test statistics described in Section 3, for comparison. The artificial samples $\{y_i\}, i = 1, \dots, n$ used were generated as follows:

$$y_i \quad = \quad g_i + \epsilon_i, \tag{19}$$

with $g_i$ having one of the following alternative definitions (constant $\ell$ is defined below).

**Sine curve:**

$$g_i \quad = \quad \ell \sin\left(\frac{3\pi(i-1)}{4n}\right). \tag{20}$$

**Temporary shift signal:**

$$g_i \quad = \quad \ell\delta(i/n), \tag{21}$$

where

$$\delta(x) = \begin{cases} 1 & \text{if } x \in [0.10, 0.20] \\ 0 & \text{otherwise.} \end{cases}$$

**Quadratic trend:**

$$g_i \quad = \ell \times \left(\tfrac{i-1}{n}\right)^2. \tag{22}$$

The constant $\ell$ was chosen in each case so as to achieve the desired signal-to-noise ratio, defined as:

$$\text{SNR} \quad = \quad 100 \times \frac{\sum_{i=1}^n (g_i - \overline{g})^2}{\sigma_\epsilon^2}.$$

Function (20) is meant to exemplify smooth oscillating variation. Function (21) is taken as an example of a sudden temporary shift, while function (22) is an example of a smooth trend. In (19), $\epsilon_i$ is Gaussian noise with zero mean and unit variance and $i = 1, \dots, n$.

The test is performed as described in the Section 2. Sample sizes $n = 128, 256$ and 512 were used, typical of moderate to fairly large data sets. The SNR takes values from 0.0 to 0.30 in steps of 0.05, then up to 0.50 in steps of 0.10. We investigated the power of the test at the conventional significance level $\alpha = 0.05$. Artificial samples were generated, and the test statistics computed. Then each sample was shuffled 100 times and the test statistics recomputed each time, to obtain the approximate critical value for $L_{\mathrm{GCV}}$ (and $L_{\mathrm{AIC}}$, etc.) under the null of no signal. The procedure was repeated 1000 times for each type of signal, SNR value and sample size. More extensive results than reported here are available from the author.

For the method proposed by Aerts et al. (1998), we fit trigonometric polynomials of increasing frequency, starting with the lowest frequency terms. Since they form an orthogonal basis, all coefficients $\theta_{p+j}$ in (18) can be computed at once. This can be done particularly fast using the FFT algorithm for $n$ equal to the product of small primes (hence the choice in our simulations of $n = 2^d$).

The first four columns in Table 1 to 3 show the empirical power obtained in one thousand replications using the proposed permutation test with four different smoothings (which respectively minimize $L_{\mathrm{GCV}}$, $L_{\mathrm{AIC}}$, $L_{\mathrm{T}}$ and $L_{\mathrm{AICC}}$). The following four columns show the corresponding empirical power for the other tests mentioned in Section 3. For all test statistics, the power has been estimated comparing the test statistic with the values obtained in 100 random permutations of the sample, except for the von Neumann test, for which the normal asymptotic distribution has been used. For Raz's test, the smoothing necessary has been fixed arbitrarily at what we felt a reasonable value. For sample lengths of $n = 128$, 256 and 512, respectively $k = 17$, 33 and 65 contiguous ordinates have been averaged with a rectangular kernel.

In Table 1 we see that the test in Aerts et al. (1998) gives slightly better power figures than our's and Raz's. Indeed, the sine wave is the signal that the test in Aerts et al. (1998) is particularly likely to deal well with. The other two tests give somewhat inferior power. We expected the test in Buckley (1991) to be a very strong contender, at least for small SNR, due to its locally most powerful property. However, it does not appear to be better than the others for the small SNR's investigated, and has less power for large SNR.

It is apparent that the proposed method gives very similar results with any of the four smoothing selectors used. That was to be expected: as Hurvich et al. (1998) points out, for small $\mathrm{tr}(H)/n$,

$$2\,\mathrm{tr}(H)/n \approx -2\log(1 - \mathrm{tr}(H)/n) \approx -\log(1 - 2\,\mathrm{tr}(H)/n), \tag{23}$$

so all three criteria impose penalties which are quite similar whenever $\mathrm{tr}(H)/n$ is small. This will be typically the case.

The patterns in Table 1 mostly recur in Table 2 and Table 3, except for the fact that the relative performance of the Aerts et al. (1998) method seems now weaker. This is understandable, since the alternatives in Tables 2 and 3 are harder to approximate with few parameters in (18) than is the alternative in Table 1.

Our proposed test does work quite well overall, and so does Raz's. To the credit of the last, it is much cheaper in the sense that it uses fixed smoothing. Our proposed test optimizes the smoothing by generalized cross validation (or any other similar criterion), making it much more expensive to compute.

Since the smoothing in Raz's test was fixed arbitrarily and knowing the underlying signal, it is of interest to consider how much better or worse the results can be if we apply a substantially different smoothing. Table 4 shows the empirical power obtained with four widely different smoothings and $n = 256$. Although there are differences in power, Raz's test seems quite resistant to changes in the smoothing.

The simulation was carried on a Digital Alpha machine running OSF1 version 4.0, and was coded in FORTRAN. Standard (64 bit) double precision was used, rather than the extended (128 bit) double precision the compiler defaults to. We used the routines

Table 1: Sine curve signal. Empirical power for the $\alpha = 0.05$ tests in one thousand experiments, shuffling each sample 100 times to obtain approximate critical values, for different signal-to-noise ratios. $SNR = 0$ corresponds to the null.

| **SNR** | **Proposed** | | | | **Buckley** | **Raz** | **Aerts et al.** | **von Neumann** |
|---|---|---|---|---|---|---|---|---|
| **%** | **GCV** | **AIC** | **T** | **AICC** | | | | |
| | Sample size $n = 128$ | | | | | | | |
| 0 | 0.046 | 0.061 | 0.046 | 0.047 | 0.046 | 0.048 | 0.040 | 0.060 |
| 5 | 0.070 | 0.074 | 0.077 | 0.081 | 0.057 | 0.085 | 0.093 | 0.066 |
| 10 | 0.224 | 0.220 | 0.222 | 0.221 | 0.176 | 0.218 | 0.276 | 0.108 |
| 15 | 0.446 | 0.428 | 0.442 | 0.437 | 0.304 | 0.450 | 0.514 | 0.183 |
| 20 | 0.694 | 0.680 | 0.678 | 0.674 | 0.477 | 0.697 | 0.781 | 0.297 |
| 25 | 0.913 | 0.890 | 0.875 | 0.855 | 0.672 | 0.919 | 0.945 | 0.445 |
| 30 | 0.982 | 0.974 | 0.937 | 0.930 | 0.808 | 0.985 | 0.993 | 0.638 |
| 40 | 0.999 | 0.996 | 0.986 | 0.979 | 0.958 | 1.000 | 1.000 | 0.925 |
| 50 | 1.000 | 0.998 | 0.999 | 1.000 | 0.995 | 1.000 | 1.000 | 0.989 |
| | Sample size $n = 256$ | | | | | | | |
| 0 | 0.050 | 0.049 | 0.050 | 0.050 | 0.049 | 0.046 | 0.040 | 0.044 |
| 5 | 0.141 | 0.139 | 0.142 | 0.141 | 0.100 | 0.123 | 0.151 | 0.065 |
| 10 | 0.435 | 0.415 | 0.440 | 0.444 | 0.275 | 0.406 | 0.490 | 0.123 |
| 15 | 0.814 | 0.782 | 0.823 | 0.827 | 0.545 | 0.797 | 0.859 | 0.214 |
| 20 | 0.968 | 0.963 | 0.959 | 0.960 | 0.775 | 0.968 | 0.985 | 0.406 |
| 25 | 0.995 | 0.994 | 0.988 | 0.989 | 0.905 | 0.997 | 0.997 | 0.649 |
| 30 | 1.000 | 0.999 | 0.990 | 0.989 | 0.972 | 1.000 | 1.000 | 0.829 |
| 40 | 1.000 | 1.000 | 0.999 | 0.999 | 1.000 | 1.000 | 1.000 | 0.993 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | Sample size $n = 512$ | | | | | | | |
| 0 | 0.049 | 0.049 | 0.053 | 0.053 | 0.049 | 0.043 | 0.049 | 0.058 |
| 5 | 0.232 | 0.231 | 0.233 | 0.233 | 0.167 | 0.208 | 0.262 | 0.074 |
| 10 | 0.768 | 0.753 | 0.765 | 0.765 | 0.485 | 0.728 | 0.796 | 0.145 |
| 15 | 0.984 | 0.974 | 0.984 | 0.986 | 0.822 | 0.980 | 0.989 | 0.299 |
| 20 | 1.000 | 0.998 | 1.000 | 1.000 | 0.970 | 1.000 | 1.000 | 0.600 |
| 25 | 1.000 | 0.999 | 0.999 | 0.999 | 0.996 | 1.000 | 1.000 | 0.868 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.980 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 2: Temporary shift signal. Empirical power for the $\alpha = 0.05$ tests in one thousand experiments, shuffling each sample 100 times to obtain approximate critical values, for different signal-to-noise ratios. $SNR = 0$ corresponds to the null.

| **SNR** | **Proposed** | | | | **Buckley** | **Raz** | **Aerts et al.** | **von Neumann** |
|---|---|---|---|---|---|---|---|---|
| **%** | **GCV** | **AIC** | **T** | **AICC** | | | | |
| | Sample size $n = 128$ | | | | | | | |
| 0 | 0.046 | 0.061 | 0.046 | 0.047 | 0.046 | 0.048 | 0.040 | 0.060 |
| 5 | 0.288 | 0.274 | 0.304 | 0.307 | 0.069 | 0.259 | 0.238 | 0.103 |
| 10 | 0.873 | 0.847 | 0.840 | 0.821 | 0.147 | 0.820 | 0.743 | 0.344 |
| 15 | 0.992 | 0.987 | 0.956 | 0.939 | 0.221 | 0.996 | 0.990 | 0.787 |
| 20 | 0.999 | 0.999 | 0.990 | 0.989 | 0.308 | 1.000 | 1.000 | 0.972 |
| 25 | 1.000 | 1.000 | 1.000 | 1.000 | 0.391 | 1.000 | 1.000 | 0.998 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 0.489 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 0.688 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 0.784 | 1.000 | 1.000 | 1.000 |
| | Sample size $n = 256$ | | | | | | | |
| 0 | 0.050 | 0.049 | 0.050 | 0.050 | 0.049 | 0.046 | 0.040 | 0.044 |
| 5 | 0.604 | 0.569 | 0.615 | 0.617 | 0.096 | 0.511 | 0.475 | 0.121 |
| 10 | 0.993 | 0.988 | 0.974 | 0.976 | 0.205 | 0.986 | 0.971 | 0.515 |
| 15 | 1.000 | 0.999 | 0.996 | 0.996 | 0.397 | 1.000 | 1.000 | 0.939 |
| 20 | 1.000 | 1.000 | 0.999 | 0.999 | 0.609 | 1.000 | 1.000 | 1.000 |
| 25 | 1.000 | 1.000 | 1.000 | 1.000 | 0.781 | 1.000 | 1.000 | 1.000 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 0.888 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 0.975 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 0.995 | 1.000 | 1.000 | 1.000 |
| | Sample size $n = 512$ | | | | | | | |
| 0 | 0.049 | 0.049 | 0.053 | 0.053 | 0.049 | 0.043 | 0.049 | 0.058 |
| 5 | 0.872 | 0.866 | 0.874 | 0.880 | 0.126 | 0.837 | 0.765 | 0.186 |
| 10 | 1.000 | 1.000 | 0.999 | 0.999 | 0.406 | 1.000 | 1.000 | 0.755 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 0.708 | 1.000 | 1.000 | 0.997 |
| 20 | 1.000 | 1.000 | 1.000 | 1.000 | 0.886 | 1.000 | 1.000 | 1.000 |
| 25 | 1.000 | 1.000 | 1.000 | 1.000 | 0.981 | 1.000 | 1.000 | 1.000 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 0.998 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

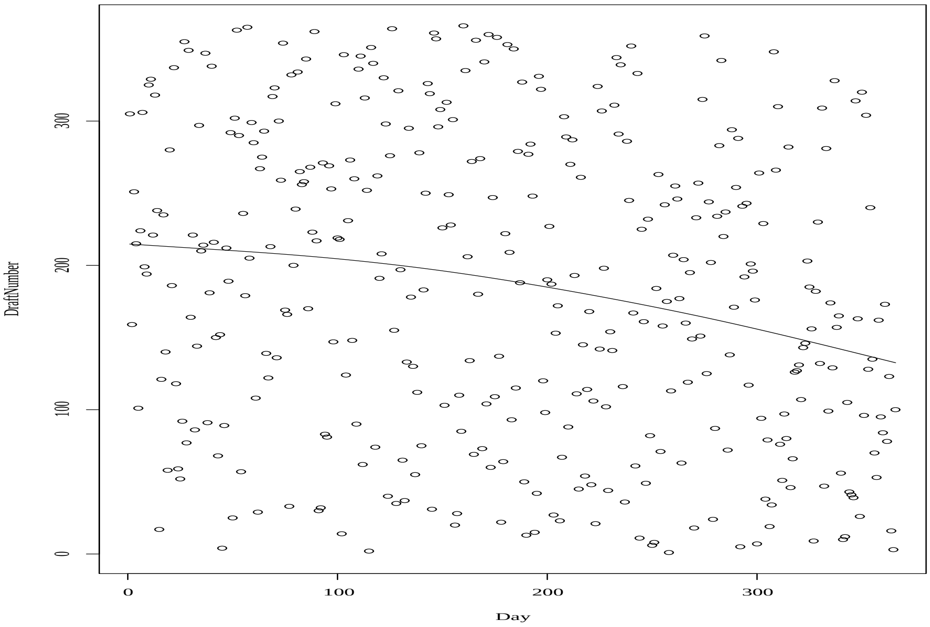
Table 3: Quadratic trend signal. Empirical power for the $\alpha = 0.05$ tests in one thousand experiments, shuffling each sample 100 times to obtain approximate critical values, for different signal-to-noise ratios. $SNR = 0$ corresponds to the null.

| **SNR** | **Proposed** | | | | **Buckley** | **Raz** | **Aerts et al.** | **von Neumann** |
|---|---|---|---|---|---|---|---|---|
| **%** | **GCV** | **AIC** | **T** | **AICC** | | | | |
| | Sample size $n = 128$ | | | | | | | |
| 0 | 0.046 | 0.061 | 0.046 | 0.047 | 0.046 | 0.048 | 0.040 | 0.060 |
| 5 | 0.121 | 0.127 | 0.116 | 0.116 | 0.066 | 0.129 | 0.117 | 0.095 |
| 10 | 0.458 | 0.460 | 0.430 | 0.409 | 0.140 | 0.503 | 0.382 | 0.285 |
| 15 | 0.850 | 0.853 | 0.800 | 0.786 | 0.208 | 0.874 | 0.751 | 0.667 |
| 20 | 0.996 | 0.992 | 0.976 | 0.970 | 0.296 | 0.997 | 0.954 | 0.934 |
| 25 | 0.999 | 0.999 | 0.998 | 0.998 | 0.415 | 1.000 | 0.995 | 0.995 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 0.487 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 0.676 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 0.816 | 1.000 | 1.000 | 1.000 |
| | Sample size $n = 256$ | | | | | | | |
| 0 | 0.050 | 0.049 | 0.050 | 0.050 | 0.049 | 0.046 | 0.040 | 0.044 |
| 5 | 0.247 | 0.243 | 0.243 | 0.240 | 0.085 | 0.243 | 0.208 | 0.109 |
| 10 | 0.848 | 0.847 | 0.828 | 0.826 | 0.177 | 0.872 | 0.733 | 0.478 |
| 15 | 0.998 | 0.996 | 0.988 | 0.987 | 0.334 | 1.000 | 0.988 | 0.911 |
| 20 | 1.000 | 1.000 | 0.999 | 0.999 | 0.536 | 1.000 | 1.000 | 0.997 |
| 25 | 1.000 | 1.000 | 1.000 | 1.000 | 0.715 | 1.000 | 1.000 | 1.000 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 0.837 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 0.965 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 0.992 | 1.000 | 1.000 | 1.000 |
| | Sample size $n = 512$ | | | | | | | |
| 0 | 0.049 | 0.049 | 0.053 | 0.053 | 0.049 | 0.043 | 0.049 | 0.058 |
| 5 | 0.465 | 0.465 | 0.458 | 0.456 | 0.109 | 0.516 | 0.393 | 0.168 |
| 10 | 0.995 | 0.995 | 0.994 | 0.994 | 0.352 | 0.995 | 0.986 | 0.691 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 0.636 | 1.000 | 1.000 | 0.996 |
| 20 | 1.000 | 1.000 | 1.000 | 1.000 | 0.846 | 1.000 | 1.000 | 1.000 |
| 25 | 1.000 | 1.000 | 1.000 | 1.000 | 0.961 | 1.000 | 1.000 | 1.000 |
| 30 | 1.000 | 1.000 | 1.000 | 1.000 | 0.994 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 50 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 4: Empirical power using Raz's test and four different smoothings. A uniform kernel has been used and $k = 17, 33, 65$ and $129$ contiguous values have been averaged, except at the ends. Rows for large SNR which made the empirical power 1 for all $k$ have been omited.

| **SNR** | **Number of contiguous ordinates averaged** | | | |
|---|---|---|---|---|
| **%** | $\mathbf{k = 17}$ | $\mathbf{k = 33}$ | $\mathbf{k = 65}$ | $\mathbf{k = 129}$ |
| | Sine signal with $n = 256$ | | | |
| 0 | 0.049 | 0.055 | 0.062 | 0.061 |
| 5 | 0.106 | 0.123 | 0.143 | 0.147 |
| 10 | 0.297 | 0.421 | 0.515 | 0.556 |
| 15 | 0.685 | 0.810 | 0.896 | 0.905 |
| 20 | 0.908 | 0.959 | 0.979 | 0.978 |
| 25 | 0.986 | 0.999 | 1.000 | 1.000 |
| | Time limited shift signal with $n = 256$ | | | |
| 0 | 0.049 | 0.055 | 0.062 | 0.061 |
| 5 | 0.403 | 0.525 | 0.645 | 0.694 |
| 10 | 0.969 | 0.990 | 0.994 | 0.998 |
| | Quadratic signal with $n = 256$ | | | |
| 0 | 0.049 | 0.055 | 0.062 | 0.061 |
| 5 | 0.266 | 0.240 | 0.243 | 0.195 |
| 10 | 0.904 | 0.877 | 0.787 | 0.659 |
| 15 | 0.999 | 0.998 | 0.988 | 0.942 |
| 20 | 1.000 | 1.000 | 0.999 | 0.997 |

Figure 1: Draft selection number by day of year in the 1970 U.S. Lottery Draft. Superimposed: a cubic spline fit with smoothing chosen by GCV.



described in Hutchinson (1986) for the computation of cross validated splines. For other general purpose routines (like FFT and random number generation) we turned to the public domain part of the PORT3 library (©Lucent Technologies, Inc., available from Netlib). Code for the different tests to compare was also written in FORTRAN. Summarization of the results and the only graph in the paper were done in S-PLUS (©MathSoft, Inc.).

## 4.2 An illustration

Figure 1 shows the results of the 1970 U.S. Lottery Draft, with a spline superimposed (the smoothing was chosen using GCV). A lottery was held to determine the order in which men in military age would be called into active service. An urn was set with one ball for each day of the year, then all balls were removed and the order of extraction (the "draft number") recorded. The fairness of the procedure was later challenged on the grounds that the allocated draft numbers did not appear to be sufficiently "random"; men born in early months appeared to have been assigned significantly larger draft numbers than those born later in the year. Subsequent inquiry suggested that balls entered the urn in day order, and the urn was insufficiently shaken.

To test for randomness against a smooth alternative, we can use the test described in Section 2. The value obtained for $L_{\text{GCV}}$ is 10683.36. It is the second smallest out of five hundred replications, permuting each time the $y_i$'s. Therefore, the null hypothesis of no trend can be quite confidently rejected, even at the 0.01 level of significance.

## 5 Some remarks

The test described in Section 2 can be used directly as such (see for instance Gutiérrez and Tusell (1997)) or as a test for a flat spectrum, by fitting a cross-validated spline to a periodogram. Since autocorrelation in residuals translates to non-flat spectra, this would yield a test against (fairly general) autocorrelation.

Finally, we would like to point out that the proposed test nicely adapts to situations in which there are missing and unequally spaced data, so far as the exchangeability assumption implied by a permutation test is met.

**Acknowledgments.** Jeffrey Hart offered helpful hints and a draft of the paper Aerts et al. (1998). Eva Ferreira and the participants at the EC2 - Simulation Methods in Econometrics (Firenze, 1996) made some comments. One referee and one associate editor were quite helpful with their criticism. To all of them my thanks. Any errors and obscurities are my own. This work was partly supported by the Spanish MEC (grants PB95-0346 and PB98-0149).

## References

Aerts, M., Claeskens, G., and Hart, J. (1998). Testing the fit of a parametric function. Unpublished manuscript.

Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford Univ. Press.

Buckley, M. (1991). Detecting a smooth signal: Optimality of cusum based procedures. *Biometrika*, 78, 253–262.

Cox, D., Koh, E., Wahba, G., and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Annals of Statistics*, 16, 113–119.

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, 31, 377–403.

González-Manteiga, W. and Cao, R. (1993). Testing the Hypothesis of a General Linear Model using Nonparametric Regression Estimation. *Test*, 2, 161–188.

Gutiérrez, J. and Tusell, F. (1997). Suicides and the lunar cycle. *Psychological Reports*, 80, 243–250.

Hart, J. (1997). *Nonparametric Smoothing and Lack of Fit Tests*. New York: Springer Verlag.

Hastie, T. and Tibshirani, R. (1991). *Generalized Additive Models*. London: Chapman & Hall, second edition.

Härdle, W. (1990). *Smoothing Techniques with implementations in S*. New York: Springer Verlag.

Hurvich, C. F., Simonoff, J. S., and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Ser. B*, 60, 271–293.

Hutchinson, M. F. (1986). Cubic Spline Data Smoother. *ACM Transactions of Mathematical Software*, 12, 150–153, software available at `http://www.acm.org/calgo/contents/`.

Ramil, L. and González-Manteiga, W. (1998). Chi square goodness-of-fit tests for polynomial regression. *Commun. Statistics - Simulation*, 27, 229–258.

Raz, J. (1990). Testing for No Effect When Estimating a Smooth Function by Nonparametric Regression: A Randomization Approach. *Journal of the American Statistical Association*, 85, 132–138.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12, 1215–1230.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Silverman, B. W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *Annals of Statistics*, 12, 898–916.

von Neumann, J. (1941). Distribution of the ratio of the mean squared successive difference to the variance. *Annals of Mathematical Statistics*, 12, 367–395.