Developed the Work in the Gamma test with Application

Manal .S.Hamde

Department of Operations Research, College of Computers Science and Mathematics , University of Mosul, Iraq.

(**Received:** 3 / 10 / 2010 ---- **Accepted:** 11 / 5 / 2011)

Abstract:

In this paper, developed the work in the gamma test has been done with application of the smooth functions. Where a developed work by added the terms A_k , B_k ,C_k ,,D_k and E_k to function the gamma test , when the gamma test is used for non-linear of time series of analysis and can be expressed relation ship between input /output data set of the from (x(i) ,y(i)) ($1 \le i \le M$).And after we application the gamma test in the smooth function $f(x) = \sin(4\pi x) + \cos(2\pi x)$ which indicates that the development of work are better.

1. Introduction-:

This work has been developed to test the Gamma[2] and the introduction of the limits fixed by the terms of our work in the theory chapter to clarify the method and derived based on some of the definitions on the subject and write an algorithm and the work of a flowchart with steps easily applied either in the chapter were taken as examples of some functions of the application of the test

2. The Gamma test

For any ordered set of points T ={$x_1,...,x_M$}, the unique kth nearest neighbour of any point $x_i \in T$ is defined inductively as follows. The first nearest neighbour of x_i is that point $x_{j1} \in T \setminus \{x_i\}$ having minimal distance from x_i and minimal index j_1.The second nearest neighbour of x_i is that point $x_{j2} \in T \setminus \{x_i; x_{j1}\}$ having minimal distance from x_i and minimal index j_2. In general, the kth nearest neighbour of x_i is that point $x_{jk} \in T \setminus \{x_i ,x_{j1} ,....,x_{jk-1}\}$ having minimal distance from x_i and minimal index j_k. Ordering equidistant points by their indices avoids some awkward consequences of other possible definitions. We remark that nearest-neighbour lists for the first k nearest neighbours can be found in O(M log M) time using techniques developed in[3]. Let $x_{N[i,k]}$ denote the kth [2] nearest neighbour of x_i in the set {$x_1,...,x_M$}. The Gamma test is based on the statistics

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} \left| x_{N[i,k]} - x_i \right|^2 \quad ..(1)$$

and

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} \left| y_{N[i,k]} - y_i \right|^2 \quad ....(2)$$

where $y_{N[i,k]}$ is the output value associated with $x_{N[i,k]}$ (note that $y_{N[i,k]}$ is not necessarily the kth nearest neighbour of y_i in output space) [6]. Using

$$y = f(x) + r \quad ..(3)$$

and the continuity of the unknown function f , it can be shown that

$$\gamma_M(k) \to \text{var(r) in probability as } \delta_M(k) \to 0 \quad ..(4)$$

However, this limit cannot be computed directly, since $\delta_M(k)$ cannot be made arbitrarily small for any finite dataset { (x_i; y_i); $1 \le i \le M$}.The Gamma test employs a linear-regression technique to estimate this limit as $\delta_M(k) \to 0$. This is based on the assertion that the relationship between the points ($\delta_M(k)$, $\gamma_M(k)$) is approximately linear in probability for M sufficiently large. For some `constant' A(M, k), defined in (12) and that is independent of the particular sample {$x_1,...,x_M$}, we show that

$$\gamma_M(k) = \text{var(r)} + A(M,k)\,\delta_M(k) + O(\delta_M(k)) + e_M \text{ as } M \to \infty \quad .(5)$$

where the convergence is in probability and e_M represents the sampling error (this is approximately of order $O(1/\sqrt{M})$ as $M \to \infty$). For some fixed integer p, we compute the pairs ($\delta_M(k)$, $\gamma_M(k)$ for $1 \le k \le p$ and perform linear regression on these points. The intercept gamma test (Γ) of the regression line [4].

3. Assumptions

We select to state our assumptions as clearly as possible. The class of models f The domain of possible models f : $C \to R$ is restricted to the class of continuous functions having bounded first partial derivatives over the input space $C \subset R^m$ [5]. In particular, we remark that the Gamma test is not directly applicable to problems involving categorical data Let $\nabla f(x)$ denote the first partial derivatives of f at the point x, defined by

$$\nabla f(x) = \left(\frac{\partial f}{\partial x(i)}\right)_{i=1}^{m} \qquad (6)$$

where x (i) are the ith components of x, respectively. Let H (C) denote the convex hull of C. Clearly, f can be extended to a continuous function having bounded first partial derivatives over H (C), and we require that there exist constants $b_1 > 0$ such that,

$$\text{for all } H(C) \in x \quad |\nabla f(x)| < b_1 \quad .... \quad (7)$$

4. Theorem.

Let f : $R^m \to R$ be continuous and have bounded first and second partial derivatives on a closed bounded set $C \subset R^m$ that contains no isolated points. Let r be a real random variable with mean zero and having its first four moments bounded. Let the points $x_1,.....,x_M$ in $C \subset R^m$ and $r_1,.....,r_M$ in R be independently selected according to the probability density functions ϕ and Ψ , respectively, and write

$$y_i = f(x_i) + r_i \quad ... (8)$$

Denote the kth nearest neighbour of x_i by $x_{N[i,k]}$ and define

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |x_{N[i,k]} - x|^2 \quad ..(9)$$

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} |y_{N[i,k]} - y_i|^2 \quad ..(10)$$

Then, for every k > 0, we have [1]

$$\gamma_M(k) = \text{var}(r) + A(M,k)\delta_{M(k)} + o(\delta_M(k)) + O(1/M^{1/2-k}) \quad (11)$$

with probability greater than 1- $O(1/M^{2k})$ as $M \to \infty$, where A(M, k) is defined by

$$A(M,k) = \frac{\varepsilon_\phi(((x_{N[i,k]} - x_i).\nabla f(x_i))^2)}{2\varepsilon_\phi(|x_{N[i,k]} - x_i|^2} \quad .(12)$$

and satisfies the condition

$$0 \le A(M,k) \le \frac{1}{2} b_1^2 < \infty \quad .. (13)$$

where b_1 is the upper bound on the gradient of f over C as defined in(4) . The Gamma-test algorithm is shown in algorithm (6) [4].

5. Decomposition of the problem

Let x_i' denote the kth nearest neighbour of x_i and consider

$$\frac{1}{2}(y_i' - y_i)^2 = \frac{1}{2}((r_i' - r_i) + (f(x_i') - f(x_i)))^2 \quad \ldots(14)$$

Where $y_i' = f(x_i') + r_i'$ is the output value associated with x_i' . By (7), f has bounded first partial derivatives at each point in H(C), so we can apply the second mean-value theorem to f. Thinking of x_i as a column vector, this implies that there exists some point $\varepsilon_i \in H(C)$ on the line segment joining x_i' and x_i such that

$$f(x_i') - f(x_i) = (x_i' - x_i).\nabla f(x_i) + T_f(x_i, x_i') \quad (15)$$

Where

$$T_f(x_i, x_i') = \frac{1}{2}(x_i' - x_i)^T Hf(\varepsilon_i)(x_i' - x_i) \quad .(16)$$

Substituting (15) back into (14), we get

$$\frac{1}{2}(y_i' - y_i)^2 = \frac{1}{2}((r_i' - r_i) + (x_i' - x_i).\nabla f(x_i) + T_f(x_i, x_i'))^2 \quad ..(17)$$

Note that since C is of finite diameter, by (7) we have

$$\left.\begin{array}{l}(x_i' - x_i).\nabla f(x_i) = O(|x_i' - x_i|) \\ T_f(x_i, x_i') = O(|x_i' - x_i|^2)\end{array}\right\} \quad as |x_i' - x_i| \to 0 \quad ..(18)$$

With this in mind, summing both sides of (17) over 1 ≤ i ≤M and dividing by M, we arrive at an expression for $\gamma_M(k)$ given by

$$\gamma_M(k) = \text{var}(r) + A_k + B_K + C_k + D_k + E_k \quad ..(19)$$

$$A_k = \frac{1}{2M}\sum_{i=1}^{M}(r_{N[i,k]} - r_i)^2 - \text{var}(r) \quad \ldots(20)$$

$$B_K = \frac{1}{M}\sum_{i=1}^{M}(r_{N[i,k]} - r_i)(x_{N[i,k]} - x_i).\nabla f(x_i) \quad \ldots(21)$$

$$\tilde{C}_K = \frac{1}{2M}\sum_{i=1}^{M}((x_{N[i,k]} - x_i).\nabla f(x_i))^2 \quad \ldots(22)$$

$$D_K = \frac{1}{M}\sum_{i=1}^{M}(r_{N[i,k]} - r_i)T_f(x_i, x_{N[i,k]}) \quad ..(23)$$

$$E_K = \frac{1}{M}\sum_{i=1}^{M} O(|x_{N[i,k]} - x_i|^3) \quad .(24)$$

By definition of $\delta_M(k)$, it follows that $E_k = o(\delta_M(k))\ as\ M \to \infty$ and hence

$$\gamma_M(k) = \text{var}(r) + A_k + B_K + \tilde{C}_k + D_k + o(\delta_M(k))\ \ as\ \ M \to \infty \quad (25)$$

The proof of theorem in section (4)is based on the idea that the sample means defined in (20)-(23) behave very much as if they were sample means of independent identically distributed variables in that they converge (in probability) to their expected values $as\ \ M \to \infty$ We shall show in lemma (1) in section (6) that the expected value of each of the terms A_k, B_k and D_k is zero. Moreover, by definition of A(M, k) in (12), the expected value of $C_k - A(M,k)\delta_M(k)$ is also zero.

Thus we define

$$C_k = \tilde{C}_k - A(M,k)\delta_M(k) \quad \text{, i.e.}$$

$$C_K = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{2}((x_{N[i,k]} - x_i).\nabla f(x_i))^2 - A(M,k)|x_{N[i,k]} - x_i|^2) \quad (26)$$

so that

$$\gamma_M(k) = \text{var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + A_k + B_K + C_k + D_k \quad (27)$$

To establish theorem in section (4), we need to show that each of the terms A_k, B_k, C_k and D_k converge (in probability) to their expected value (zero) $as\ M \to \infty$, and establish the appropriate rates of convergence.

6. **Lemma(1)** :- $E(A_k) = E(B_k) = E(C_k) = E(D_k) = 0$ [4].

7. Algorithm. The Gamma test.

Procedure Gamma Test (data), {data is an array of points (x_i , y_i) (1 ≤ i ≤ M) where x Є R^m and y Є R }

for i = 1 to M do {compute nearest-neighbour list for each data point x_i -this can be done in O(M log M) time } for k = 1 to p do{ typically p ≈ 10 }

compute N[i, k], where $x_{N[i,k]}$ is the kth nearest neighbour of x_i

end

end

{ if multiple outputs, do the remainder for each output} [1]

for k = 1 to p do

compute $\delta_M(k)$ as in (9)

compute $\gamma_M(k)$ as in (10)

end

for k = 1 to p do

compute N[i, k], where $\Gamma_{N[i,k]}$ is the kth nearest neighbour of Γ_i

end

for k = 1 to p do

compute A_k as in(20)

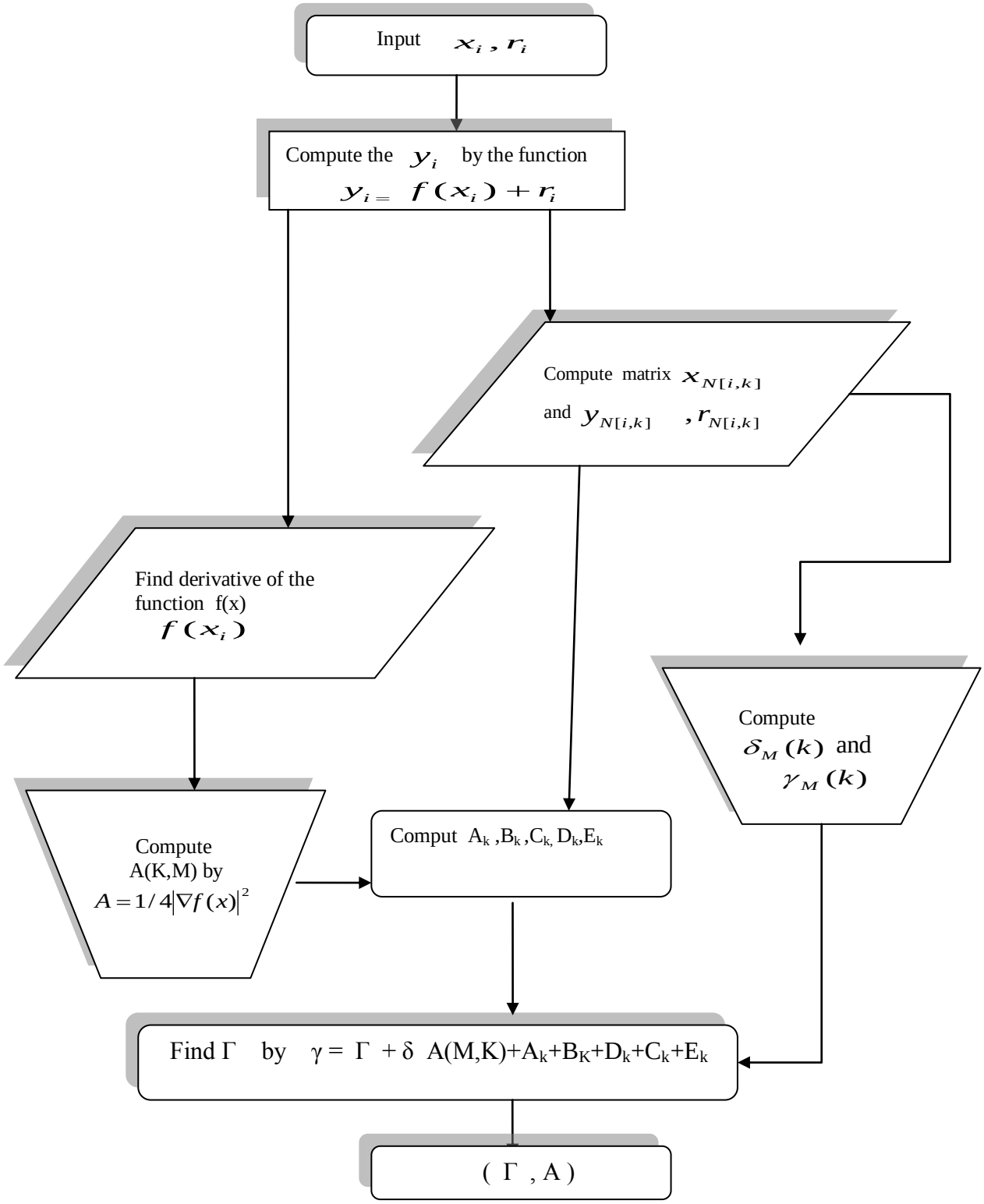
compute B_k as in(21)

compute $\hat{C}_k$ as in(22)

compute C_k as in(26)
compute D_k as in(23)
compute E_k as in(24)
end

perform least-squares fit on coordinates {
($\delta_M(k)$, $\gamma_M(k)$), A_k, B_k, C_k , D_k , E_k
$1 \le k \le p$ } obtaining
(say) $\gamma = \Gamma + A\delta$ +A+B+C+D+E
return (Γ , A)

flow chart



Applied side :-

Example:- consider the following
f(x)=sin(4Πx) +cos(2Πx)

Here x is randomly generated numbers from a uniform distribution with noise in the range [0,1] and [-1,1],of length M=1000 for the figures(1), figures (2) .

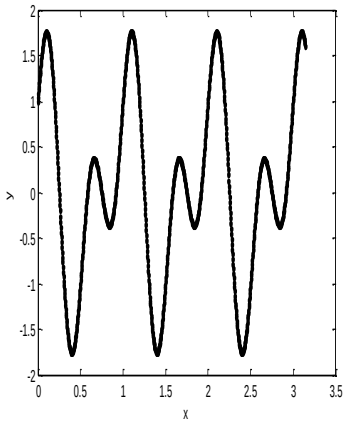


Figure1:- the smooth function with sampled "noisy" points in the [0,1]

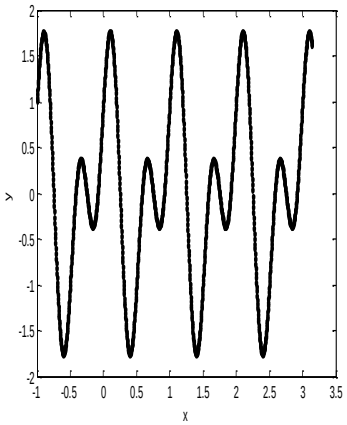


Figure2:- the smooth function with sampled "noisy" points in the [-1,1]

The first case:- Using the following equations (20) ,(21),(22),(23),(24) through the design of programs to Matlab we find A_k ,B_k, $\tilde{C}_k$,D_k ,E_k, and Using the equations(19) ,through the program Excel for the function during the periods [0,1], [-1,1] we get the following table (1) and table(2):-

Table(1) : Best results form a complete feature in the [0,1] by using the equations(19) .

m	Γ	$\gamma_M(k)$	E_k	D_k	$\tilde{C}_k$	B_k	A_k
1	0.0554	0	0	0	0	0	-0.0554
2	0.0529	0.0004	-0.0002	-0.0001	0.0018	0.001	-0.055
3	0.0471	0.0015	-0.0015	-0.0006	0.0067	0.0037	-0.0539
4	0.04	0.0034	-0.0044	-0.002	0.0139	0.0079	-0.052
5	0.0328	0.0059	-0.0091	-0.0042	0.0225	0.0134	-0.0495
6	0.0261	0.009	-0.0152	-0.0073	0.0319	0.0199	-0.0464
7	0.0205	0.0127	-0.0224	-0.0112	0.0413	0.0272	-0.0427
8	0.0154	0.0169	-0.0297	-0.0156	0.05	0.0352	-0.0384
9	0.0108	0.0217	-0.0366	-0.0202	0.0577	0.0437	-0.0337
10	0.0059	0.0269	-0.0422	-0.0246	0.0637	0.0526	-0.0285

Table(2) : Best results form a complete feature in the [-1,1], by using the equations(19).

m	Γ	$\gamma_M(k)$	E_k	D_k	$\tilde{C}_k$	B_k	A_k
1	0.0554	0	0	0	0	0	-0.0554
2	0.0529	0.0004	-0.0002	-0.0001	0.0018	0.001	-0.055
3	0.0471	0.0015	-0.0015	-0.0006	0.0067	0.0037	-0.0539
4	0.04	0.0034	-0.0044	-0.002	0.0139	0.0079	-0.052
5	0.0328	0.0059	-0.0091	-0.0042	0.0225	0.0134	-0.0495
6	0.0261	0.009	-0.0152	-0.0073	0.0319	0.0199	-0.0464
7	0.0205	0.0127	-0.0224	-0.0112	0.0413	0.0272	-0.0427
8	0.0154	0.0169	-0.0297	-0.0156	0.05	0.0352	-0.0384
9	0.0108	0.0217	-0.0366	-0.0202	0.0577	0.0437	-0.0337
10	0.0059	0.0269	-0.0422	-0.0246	0.0637	0.0526	-0.0285

for a more complex time series, once the embedding dimension has been fixed the Gamma test can be used to select an appropriate irregular embedding using a full embedding search- in much the same way that selected features in the figures(4),figures(5).

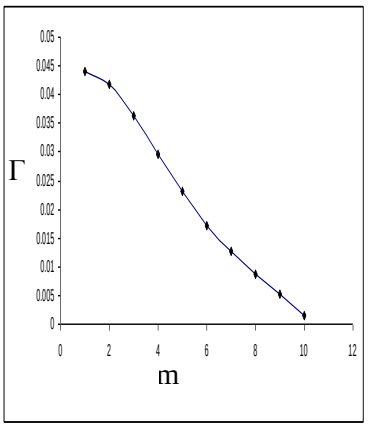


Figure 3: False nearest neigh bour embedding for function with no added noise . the graph shows that an embedding dimension m=4,Γ=0.0296 in[0,1]

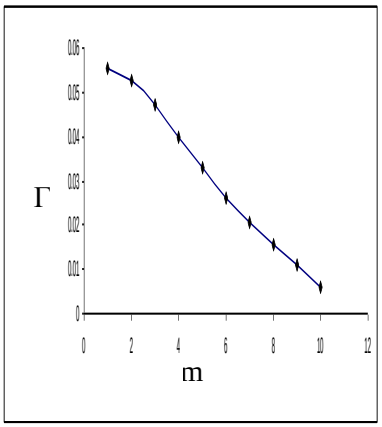


Figure 4: False nearest neigh bour embedding for function with no added noise . the graph shows that an embedding dimension m=4,Γ=0.04 in [-1,1]. [-1,1].

The second case:- Using the following equations (20) ,(21),(22),(23),(24) ,(26) through the design of programs to Matlab we find A_k ,B_k, C_k,D_k ,E_k , and Using the equations(27) ,through the program Excel for the function during the periods [0,1], [-1,1] we get the following table (3) and table(4):-

Table(3) : Best results form a complete feature in the [0,1] by using the equations(27).

m	Γ	$\gamma_M(k)$	E_k	$\delta_M(k)$ * A(M, k)	$\delta_M(k)$	A(M , k)	D_k	C_k	B_k	A_k
1	-19.6208556	0	0	0.0196556	0.001	19.6556	0	0	0	-0.0554
2	-19.18323543	0.0004	-0.0002	0.0194255	0.0010099	19.2351	-0.0001	-0.0176	0.001	-0.055
3	-18.69869905	0.0015	-0.0015	0.0194609	0.0010382	18.7448	-0.0006	-0.0128	0.0037	-0.0539
4	-18.17902452	0.0034	-0.0044	0.0197409	0.0010836	18.2179	-0.002	-0.0058	0.0079	-0.052
5	-17.59182205	0.0059	-0.0091	0.0201771	0.0011449	17.6235	-0.0042	0.0023	0.0134	-0.0495
6	-16.97707144	0.009	-0.0152	0.0207509	0.0012205	17.002	-0.0073	0.0111	0.0199	-0.0464
7	-16.31980195	0.0127	-0.0224	0.0213927	0.0013093	16.339	-0.0112	0.0199	0.0272	-0.0427
8	-15.60102881	0.0169	-0.0297	0.0220187	0.0014101	15.615	-0.0156	0.028	0.0352	-0.0384
9	-14.87035605	0.0217	-0.0366	0.0226348	0.0015212	14.8796	-0.0202	0.0351	0.0437	-0.0337
10	-14.08557443	0.0269	-0.0422	0.0231326	0.0016418	14.0898	-0.0246	0.0406	0.0526	-0.0285

Table(4) : Best results form a complete feature in the [-1,1] by using the equations(27).

m	Γ	$\gamma_M(k)$	E_k	$\delta_M(k)$ * A(M k)	$\delta_M(k)$	A(M , k)	D_k	C_k	B_k	A_k
1	-19.6959192	0	0	0.01972	0.001	19.7192	0	0	0	-0.044
2	-19.26690865	0.0004	-0.0002	0.0195	0.0010099	19.3076	-0.0001	-0.0177	0.0009	-0.0437
3	-18.79108356	0.0015	-0.0015	0.01955	0.0010382	18.8262	-0.0006	-0.0127	0.0034	-0.0428
4	-18.27902161	0.0034	-0.0045	0.01984	0.0010836	18.3075	-0.0017	-0.0058	0.0073	-0.0413
5	-17.69903367	0.0059	-0.0092	0.02029	0.0011449	17.721	-0.0037	0.0024	0.0123	-0.0393
6	-17.09059911	0.0091	-0.0155	0.02088	0.0012205	17.1066	-0.0065	0.0114	0.0183	-0.0367
7	-16.43874742	0.0128	-0.0229	0.02154	0.0013093	16.4501	-0.0101	0.0203	0.025	-0.0337
8	-15.72499422	0.0171	-0.0304	0.02218	0.0014101	15.7323	-0.014	0.0285	0.0324	-0.0303
9	-14.99844224	0.0219	-0.0375	0.02282	0.0015212	15.002	-0.0182	0.0359	0.0402	-0.0264
10	-14.21556265	0.0272	-0.0433	0.02295	0.00161418	14.2168	-0.0222	0.0407	0.0484	-0.0222

for a more complex time series, once the embedding dimension has been fixed the Gamma test can be used to select an appropriate irregular embedding using a full embedding search- in much the same way that selected features in the figures(5),figures(6).

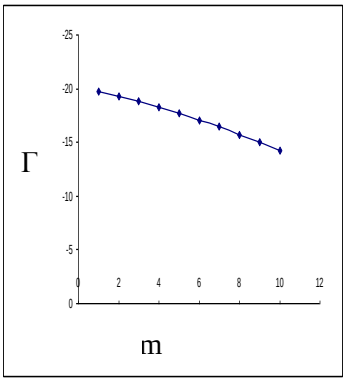


Figure 5: False nearest neigh bour embedding for function with no added noise . the graph shows that an embedding dimension m=4, Γ=-18.27902161 in [0,1].

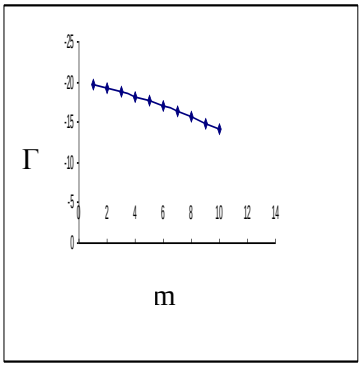


Figure 6: False nearest neigh bour embedding for function with no added noise . the graph shows that an embedding dimension m=4, Γ= -18.17902452 in [-1,1].

Conclusions and Recommendations

1. We find that the value of the Gamma test not significantly affected by changing times through tabulator (1), (2), (3), (4).
8. We find that the value of the test has been affected by many Gamma presence and amount of A(M,k) through the tabulator (1), (2), (3), (4)
3. Through graphics (3), (4), (5), (6) is to identify embedding dimension m= 4.
4. We recommend that the same method is used for a function that contains the noise.
5. We recommend that the same method is used to Related Mnglleren or more.

References

1. Antonia J. Jones, Steve Margetts, Peter Durrant &Alban P.M. Tsui.," Non-linear modeling and chaotic neural networks", Department of Computer Science, Cardiff University. Cardiff CF24 3XF, U.K.2002.
2. Aziz. M. Maysoon, Asmaa Abd.A. & Manal S.Hamada ,"Using the gamma test for nonlinear of time series analysis with application", Department of Mathematic, college of computers science and Mathematics, University of Mosul, Iraq .The second conference on Mathematical sciences 22-23,2008 Jordan.ISSN:1818-8532.
3. Bentley, J.L, "Multidimensional binary search trees used for associative search" ,Commun ACM 18,309-517.1975.
4. Dafydd E.& Antonia J. Jones," A proof of the Gamma test ", Department of Computer Science, Cardiff University ,Po Box 916, Cardiff CF24 3XF,UK,Published online 13 september2002 ,THE POYAL SOIETY
5. Kathleen T.Alligood,Tim D.Sauer, James A. Yorke,"Chaos an Introduction to dynamical systems (Text books in mathematical sciences)",New York .1996.
6. Jones, A. J., Evans, D., Margetts, S. & Durrant, P. J.," The Gamma test. In Heuristic and optimization for knowledge discovery", (ed. R. Sarker, H. Abbass & C. Newton)",Hershey, PA: Idea Group Publishing. 2001.

تطوير اختبار كاما مع التطبيق

منال سالم حمدي

قسم بحوث العمليات ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

(تاريخ الاستلام: 3 / 10 / 2010 ---- **تاريخ القبول:** 11 / 5 / 2011)

الملخص

في بحثنا هذا تم تطورنا اختبار كاما مع اخذ دالة ملساء وتطبيق هذا الاختبار عليها بفترتين مختلفتين وبإضافة المقادير A_k , B_k ,C_k ,D_k , E_kعلى الاختبار كاما ، حيث ان اختبار كاما هو تحليل للسلاسل الزمنية غير الخطية وكذلك توضيح العلاقة بين مجموعات العناصر الداخلة والخارجة بالشكل (x(i) ,y(i)) (1≤i≤M) وبعد تطبيق الاختبار على الدالة التالية :-

$$f(x) = \sin(4\pi x) + \cos(2\pi x)$$

تبين إن تطوير هذا الاختبار كاما أدى إلى حصولنا على نتائج أفضل .