"مقارنة لمقدر نموذج تجميعي معمّم بطريقة تكرارية و غير تكرارية مع صيغة مقترحة"

د.ظافر حسين رشيد
أستاذ في قسم الإحصاء
جامعة بغداد

د.عمر عبد المحسن علي
مدرس في قسم الإحصاء
جامعة بغداد

الخلاصة

يعدّ النموذج التجميعي المعمّم GAM من الممهدات متعددة المتغيرات حديثة الإستعمال في الأنحدار اللامعلمي تلافياً للوقوع في مشكلة البعدية. ولذا فقد تم تسليط الضوء على طرائق تقدير GAM سواءً بأسلوب تكراري أو غير تكراري، مع تقديم أسلوب بصيغة مقترحة، ومقارنة نتائج الطرائق جميعها بتوظيف المحاكاة. وأعتمد البحث على الشرائح التمهيدية كممهدات خطية، وبإعتبار الأخطاء العشوائية عند الحالة الكاوسية.

Abstract

Generalized Additive Model (GAM) has been considered as a multivariate smoother that appear recently in Nonparametric Regression Analysis to avoid falling in Curse of Dimensionality. We pay attention to the techniques of estimating GAM either by iterative or non-iterative approach in addition to suggested approach. Finally, we compare the three approaches using simulations. Smoothing splines have been taken under consideration as a linear smoother. The research devoted for Gaussian situation of errors.

1.1 المقدمة وهدف البحث

عندما يراد تحليل ظاهرة ما بإستعمال تحليل الإنحدار، مع عدم توفر فكرة عن طبيعة سلوك تلك الظاهرة، أو عدم تحديد توزيعها الإحتمالي، فسيتم ترك البيانات لتفصح عن ذاتها بذاتها لعدم إمكانية تكوين صيغة داليّة مسبقة عن الظاهرة، وستكون الإستعانة بتحليل الإنحدار اللامعلمي. وأبرز أداة لامعلمية حديثة يمكن توظيفها عند حالة أحادي المتغيرات (أي وجود متغير توضيحي واحد) هي الممهد Smoother.

وبشكل أدق، لو كان هناك متغيرين توضيحيين قيد الدراسة ولم يتوفر تصوّر كامل عن كيفية تأثيرهما في المتغير المعتمد، فإن النموذج التجميعي المعمّم Generalized Additive Model: GAM سيكون حلاً معقولاً بإعتباره حالة معممّة للممهدات الأحادية. إذ سيقوم GAM بتجاهل حد التفاعل Interaction بين فضائي المتغيرين وسيركز فقط على التأثيرات الرئيسة Main effects لهما وبشكل تجميعي Additively. وبذلك سيجمع GAM بين قوة التمثيل Representation

للبيانات من جهة، مع تفادي الوقوع في مشكلة البعدية Curse of Dimensionality من جهة أخرى. ويعتبر GAM ممهداً شاملاً Global يجمع في مضمونه ممهدات موضعية Local.
ويهدف البحث الى تسليط الضوء على إمكانية إستعمال تقدير GAM – والذي يعبر عن الظاهرة بشكل تجميعي لمركباتها الرئيسة، وبطرائق تكرارية أوغير تكرارية ومقارنتهما مع صيغة مقترحة.

1.2 الشرائح التمهيدية Smoothing Splines

تبرز الشرائح التمهيدية كمقدر يراد به مطابقة البيانات بشكل جيد من جهة ، ويكون على قدر معين من التمهيد من جهة أخرى. ويتم إيجاد المقدر الأمثل لها عن طريق تصغير معيار المربعات الصغرى الجزائية لكل دالة إنحدار g لامعلمية بحيث أن: $g \in W_2^m[a,b]$ ، إذ أن $W_2^m[a,b]$ يمثل فضاء Sobolev، وكما يأتي:

$$\sum_{i=1}^{n}(Y_i - g(t_i))^2 + \lambda\int_a^b (g^{(m)}(t))^2\,dt \qquad \text{... (1)} \quad \lambda > 0 \quad ;$$

m: تشير الى المشتقة من الدرجة m^{th} للدالة g(t).
λ: تدعى بالمعلمة التمهيدية .
وسيدعى المقدر بالشريحة التمهيدية من الدرجة m^{th}. وعند m=2، سيتم الحصول على الشريحة التمهيدية التكعيبية. وبشكل خاص، عندما تكون a=0 وb=1، سينتج الشريحة التكعيبية الطبيعية Natural Cubic Spline: NCS. وسيكون الشرط الضروري على الدالة g أن تكون قابلة للأشتقاق مرتين وتكون هناك إمكانية إجراء التكامل لمربع المشتقة الثانية لها. ويمكن إستعمال طريقة الجزاء غير الممهد Roughness Penalty لإيجاد الحل الأمثل لمجموع المربعات الجزائية للمقدار الآتي[3]:

$$\sum_{i=1}^{n}\big(Y_i - g(t_i)\big)^2 + \lambda\int_0^1 (g''(t))^2\,dt \qquad \text{... (2)}$$

وبإستعمال صيغة المصفوفات :

$$(Y-g)'(Y-g) + \lambda g'Kg \qquad \text{... (3)}$$

وبهذا يكون مقدر الشريحة التمهيدية، كالآتي :

$$\hat{g} = SY \qquad \text{... (4)}$$

S: مصفوفة تمهيد متماثلة ومن درجة n×n :

$$S = (I + \lambda K)^{-1} \qquad \text{... (5)}$$

K: مصفوفة جزاء تربيعية متماثلة ومن درجة n×n:

$$K = QR^{-1}Q' \quad \ldots (6)$$

إذ أن:

$$\int g''(t)^2 dt = \gamma' Q' g = \gamma' R \gamma$$

بإستعمال الشرط $Q'g=R\gamma$، يتم الحصول على [3],[1]:

$$\int g''(t)^2 dt = g' Q R^{-1} Q' g = g' K g$$

γ: متجه المشتقات الثانية للدوال (g_i)، ومن درجة 1×(n-2).

Q: مصفوفة tridiagonal بدرجة، (n-2)×n. ويتم إحتساب عناصرها كالآتي[3],[1]:

حيث أن:

$$\left.\begin{aligned} q(j-1,j) &= h^{-1}(j-1) \\ q(j,j) &= -h^{-1}(j-1) - h^{-1}(j) \\ q(j+1,j) &= h^{-1}(j) \\ q(i,j) &= 0, \forall |i-j| \geq 2 \end{aligned}\right\}$$

i=1,2, ... ,n ; j=2,3, ... , n-1

h: يمثل متجه المسافات (الفروق) بين أي مشاهدتين متتاليتين، حيث أن:

$h_i=t_i-t_{i-1}$; i=2,3, ... ,n

R: مصفوفة tridiagonal متماثلة وبدرجة (n-2)×(n-2). ويتم إحتساب عناصرها كالآتي[3],[1]:

$$\left.\begin{aligned} r(i,i) &= (1/3)\left[h(i-1)+h(i)\right]; i = 2,3,\ldots,n-1 \\ r(i,i+1) &= r(i+1,i) = (1/6)h(i); i = 2,3,\ldots,n-2 \\ r(i,j) &= 0; \forall |i-j| \geq 2 \end{aligned}\right\}$$

حيث أن: i=2,3, ... ,n-1 ; j=2,3, ... ,n-1

ويلاحظ أن الغرض من الحد الثاني من المعادلة (2) هو الموازنة بين هدفين مختلفين هما: مدى مطابقة المنحنى للنموذج، وتقدير المنحنى بحيث لايكون فيه إنقلاب سريع ومفاجئ. ويلاحظ كذلك أن الحد الثاني والذي يمثل حد الجزاء لايتأثر بإضافة حد ثابت الى الدالة g. ويذكر أن العقد هنا هي البيانات نفسها، فيتم تعيين عقدة عند كل مشاهدة. ولذا قد يتم تعريف الشريحة التمهيدية في بعض الأحيان على أنها من الممهدات كاملة الرتبة Full Rank Smoothers. وتصنف الشرائح التمهيدية كذلك على أنها من الممهدات الخطية المتماثلة Symmetric Linear Smoothers لأن مصفوفة التمهيد S متماثلة، وخطية لأنه يمكن كتابة التقدير كما في المعادلة (4). وتعدّ الشرائح التمهيدية كذلك من عائلة الممهدات التقليصية Shrinking Smoothers، في حين لايكون الممهد Lowess –مثلاً– كذلك[3]. ويذكر بأن الشرائح التمهيدية عموماً هي من نوع ممهدات الأنحدار Regression type smoothers أو ماتدعى Scatterplots Smoothers، وهي بذلك تختلف عن

غيرها من الممهدات، كممهدات السلاسل الزمنية والتي تعتمد على Sequences تمثل المسافات بين البيانات[4].

2.2 النموذج التجميعي المعمّم Generalized Additive Model (GAM)

يعدّ GAM أحد الحلول العملية خاصة عند دمجه بأسلوب Backfitting وذو أثر بالغ في تجاوز مشكلة البعدية وغيرها من المشاكل الأخرى المرتبطة بها بصورة مباشرة أو غير مباشرة. وأهم ميزة للـ GAM هي تحليل دالة متعددة المتغيرات الى تجزئة Decomposition حاصل جمع مركبات أحادية البعد. ولذا يعتبر GAM كحل وسط بين الأنحدار المتعدد وبين مطابقة سطح بعدة أبعاد، وذلك بأستخدام البواقي الجزئية Partial Residuals[1].

كما ويعدّ GAM الحالة اللامعلمية المطورة للـ Generalized Linear Model:GLIM ويتم الحصول عليه بمجرد إستبدال Linear Predictor : $\eta = \sum_{j=1}^{p} X_j \beta_j$ بحد آخر لامعلمي هو Additive Predictor [9],[7],[6] : $\eta = \sum_{j=1}^{p} g_j(t_j)$ ؛ حيث :

$$... (7)\ y_i = \sum_{j=1}^{p} g_j(t_j) + \varepsilon_i$$

ε_i : أخطاء مستقلة عن ال t_{ij}'s ، بحيث : $E(\varepsilon) = 0;\quad E(t\varepsilon) = 0;\quad Var(\varepsilon) = \sigma_\varepsilon^2;$

g_j : متجه دالة مجهولة تعبر عن المتغير التوضيحي t_j. وسيتم إعتبارها شرائح تمهيدية لهذا البحث.

ويقوم GAM في الحقيقة بدمج فكرة التقريب التجميعي Additive Approximation مع فكرة التمهيد. ومن الجدير بالذكر، ان هذا البحث سيقتصر على البيانات عند الحالة الكاوسية فقط.

3.2 طرائق تقدير GAM

هنالك طرائق عديدة للحصول على تقدير للـ GAM، سيتم التطرق لأهمها في أدناه.

1.3.2 أسلوب Backfitting التكراري

ويعدّ [Friedman & Stuetzle , 1981] أول من قدم لهذه الطريقة. إذ تعتمد على الأسلوب التكراري. فلمطابقة النموذج (7)، سيكون التوقع الشرطي لكل k من الt's، كالآتي[3]:

$$E\left[Y - \alpha - \sum_{j=1}^{p} g_j\left(t_j | t_k\right) \right] = g_k(t_k)$$

أما خطوات الطريقة ، فهي[5] :

I. البدء مع قيم أولية :

$$\alpha = average(y_i)$$
$$g_j = g_j^0 \qquad j = 1,2,...,p;$$

II.القيام بالدورة الـ j^{th} : **j=1,2, … ,p, and again 1,2, … ,p, … ;**

$$g_j = S_j\left(Y - \alpha - \sum_{k\neq j}^{p} g_k \middle| t_j\right) \quad \text{... (8)} \qquad \textbf{k,j=1,2,...,p.}$$

$$S_j = (I + \lambda_j K_j)^{-1} \quad \text{... (9)}$$
$$K_j = Q_j R_j^{-1} Q_j'$$

وهي مشابهة لما جاء في المعادلة (6)، ولكن يعاد إيجادها لكل j^{th} من المتغيرات التوضيحية.

III.يتم تكرار الخطوة (II) ، حتى تستقر الدوال المفردة $g_j(t_j)$، بدون تغيير عند الدورة m، وحسب معيار التقارب [3]:

$$\Delta_m\left(g_j^m, g_j^{m-1}\right) = \sum_{j=1}^{p}\left(g_j^m - g_j^{m-1}\right)^2$$

ويراد بهذه الطريقة مطابقة الدوال آنياً، إذ أن عملية التحسين والتعديل لقيم g_j بالخطوات المتتابعة سوف تحذف تأثيرات كل المتغيرات الأخرى من الـY قبل تمهيد هذه الأخيرة – والتي ستدعى بالبواقي الجزئية Partial Residuals – المقابلة للt_j. وتعرف البواقي الجزئية بأنها القيم المطابقة لكل دالة زائداً البواقي الإجمالية من النموذج التجميعي، بشرط أن تكون: $E\{g_j(t_j)\}=0$; j=1,2,....,p.

إذ سيعتمد Y على كل الـt_j's ماعدا واحداً منها t_k :(k≠j) ، فيتم إعتباره ثابتاً بالنسبة للبقية. ويتم إعتبارأن t_k متعامدة مع باقي المتغيرلت التوضيحية t_j's الأخرى.

ويلاحظ كذلك، أن مصفوفة التمهيد S ستعتمد على الـ t's وليس على الـ Y's، بشكل يشابه الحالة المعلمية. بمعنى أن:

$$\hat{\beta} = HY$$
$$H = X(X'X)^{-1}X'$$

ويتم إستبدال مصفوفة التمهيد S_j بمصفوفة تمهيد أخرى، ولتكن S_j^* ، بحيث تقوم بعملية التمهيد أولاً، ثم تقوم بحذف متوسط التمهيد للحصول على صفة الـ Identifiable للتقديرات. وبهذا ستكون[3]:

$$S_j^* = (I - 11'/n)S_j \quad \text{... (10)} \qquad \textbf{;} \quad \textbf{j=1,2,...,p}$$

1:عبارة عن متجه n×1 كل عناصره واحد .

وعندها يدعى هكذا ممهد بـ "الممهد الممركز" Centered Smoother[2],[9]. وبهذا يتم ضمان أنه في كل مرحلة من مراحل عملية التقدير Backfitting التكرارية سيكون للـ $\hat{f}_j$'s متوسط صفري.

ولذا عند تعميم المعادلة(2) سيتم الحصول على تقدير GAM، لتصغير[3]:

$$\sum_{i=1}^{n}\left\{y_i-\alpha-\sum_{j=1}^{p}g_j(t_j)\right\}^2+\sum_{j=1}^{p}\lambda_j\int\left\{g_j''(t_j)\right\}^2dt \qquad \ldots (11)$$

يلاحظ على الدالة(11) بأنها تكون جزائية بثابت منفصل λ_j لكل حد. وهذا الثابت يعود الى تحديد تمهيد الدالة في الحل.

ويمكن إعادة كتابة(11)، وبصيغة المصفوفات كالآتي [3],[6]:

$$(y_i-\alpha-\sum_{j=1}^{p}g_j)'(y_i-\alpha-\sum_{j=1}^{p}g_j)+\sum_{j=1}^{p}\lambda_j g_j'K_j g_j \qquad \ldots (12)$$

K_j: مصفوفات جزاء لكل t_j، كما وردت في المعادلة (6).

فإذا أجري التفاضل على المعادلة (12) بالنسبة للg_k ومساواته بالصفر سنحصل على :

$$\hat{g}_k=S_k(y-\alpha-\sum_{j\neq k}^{p}\hat{g}_j) \qquad \ldots (13)$$

إذ أن:

$$S_k=(I+\lambda_k K_k)^{-1} \quad j,k=1,2,\ldots,p. \quad , \qquad \ldots (14)$$

ويمثل المقدار $(y-\alpha-\sum_{j\neq k}^{p}\hat{g}_j)$: البواقي الجزئية من التمهيد. وعلى هذه الشاكلة سيتم تكرار(إعادة) عملية التمهيد للبواقي الجزئية. وتكون الحاجة الى O(np) من العمليات الحسابية.

2.3.2 **<u>أسلوب حل النظام غير التكراري</u>**

إذ يتبين من تسميتها أنها لاتحتاج الى عملية تكرارية لإيجاد الحل (التقدير) الأمثل. بل يكفي دورة loop واحدة للحل دفعة واحدة. وهي طريقة تعتمد أصلاً على إعتماد وجهة نظر أخرى للمعادلة(13). فلو كانت مجموعة المعادلات التقديرية بالصيغة[3] **الآتية:**

$$\begin{pmatrix} I & S_1 & S_1 & \ldots & S_1 \\ S_2 & I & S_2 & \ldots & S_2 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ S_p & S_p & S_p & \ldots & I \end{pmatrix}\begin{pmatrix} g_1 \\ g_2 \\ . \\ . \\ . \\ g_p \end{pmatrix}=\begin{pmatrix} S_1y \\ S_2y \\ . \\ . \\ . \\ S_py \end{pmatrix} \qquad \ldots (15)$$

إذ ستكون الحاجة هنا الى $O(\{np\}^3)$ من العمليات الحسابية.ويمكن صياغتها كحل لنظام متسّق Consistent **من المعادلات الطبيعية التقديرية بأسلوب Gauss-Seidel وبشكل غير تكراري.**

فعند حالة ثنائي المتغيرات(p=2):

$$\hat{g}_1=(I-S_1S_2)^{-1}S_1(I-S_2)Y \qquad \ldots (16)$$

$$\hat{g}_2=(I-S_1S_2)^{-1}S_2(I-S_1)Y \qquad \ldots (17)$$

ولإثبات الإتجاه الآخر لممهدات غير الشرائح التمهيدية والتي قد تكون مصفوفة التمهيد فيها Singular أو قريبة من حالة الـ Singular، فإن مصغرات معيار المربعات الصغرى الجزائية ستكون:

$$(y-\sum_{j=1}^{p} g_j)'(y-\sum_{j=1}^{p} g_j)+\sum_{j=1}^{p} g_j'(S_j^- - I)g_j$$

لكل $g_j \in R(S_j)$ **، وستكون هي الأخرى حلولاً للمعادلات التقديرية**(15).

3.3.2 مقترح إستعمال MGCV

من المألوف إستعمال معيار GCV الإعتيادي كأسلوب آلي في إختيار المعلمة التمهيدية في حالة أحادي المتغيرات. اما في حالة ثنائي المتغيرات أو متعدد المتغيرات بشكل عام، يمكن إستعمال مايدعى معيارMGCV، لأختيار P من المعلمات التمهيدية دفعة واحدة.

إذ يفترض تحديد قيم المعلمات التمهيدية مسبقاً من قبل الباحث وذلك بالإستناد الى الخبرة أو الى شكل رياضي معين بدون الحاجة الى معايير آلية مثل معيار الGCV وغيره. ولذا قام الباحثان بإستعمال معيارMGCV لإختيار المعلمات التمهيدية من البيانات وبشكل آلي، ومن ثم توظيفها في الشرائح التمهيدية، وأخيراً تقدير GAM. أما صيغة هذا المعيار فهي كما في أدناه:

$$\min MGCV(\lambda_1,\lambda_2,...,\lambda_p)=\frac{\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\alpha-\sum_{j=1}^{p}\hat{g}_{j,\lambda_j}(t_{ij})\right\}^2}{\left(1-\left[1+\sum_{j=1}^{p}\{trS_j(\lambda_j)-1\}\right]/n\right)^2} \qquad \ldots (18)$$

وأن قيم λ_j's المثلى هي التي تحقق ذلك.

4.2 إختيار المعلمة التمهيدية Choosing Smoothing Parameter

من المعروف إنه يمكن إختيار قيمة المعلمة التمهيدية عبر أحد الأسلوبين الآتيين:

I. الأسلوب الشخصي Subjective Method

وفيه يتم تحديد قيمة λ بشكل مسبق ومن خلال إحدى الأساليب التالية مما سيجعل الممهدات من النوع الخطي [8],[3] **:**

أ. رسم البواقي [Residuals Graph] وتحديد أفضل نقطة على الرسم.

ب. إستعمال الصيغة في المعادلة الآتية[4]:

$$\lambda = \frac{p}{1-p} \quad ; \qquad 0 \le p \le 1 \quad \ldots(19)$$

ج. حساب درجات الحرية المكافئة Equivalent Degrees of Freedom: EDF للمعلمات، وتثبيتها، ثم البحث عن قيمة المعلمة التمهيدية المقابلة لدرجة الحرية هذه.

د.يتم تحديدها من خلال الخبرة بما يتلائم والظواهر المدروسة.

II. <u>الأسلوب الموضوعي (الآلي)</u> Objective (Automatic) Method

إذ تترك البيانات لتحدد بذاتها قيمة المعلمة التمهيدية (المثلى) عن طريق تصغير معيار معين يعتمد بدوره على البواقي. والذي سيحصل في حقيقة الأمر، هو إختزال مسألة إختيار النموذج الى مسألة إختيار المعلمة التمهيدية له والأكتفاء بها. ومن أشهر هذه المعايير: , AIC(λ) , GML(λ) , GCV(λ) , $C_p(\lambda)$. ومن المعروف أن قيم المعلمة التمهيدية المستحصلة بهذه الإحصاءات أو المعايير تكون غير وحيدة Not-Unique**. وكمثال على ذلك: المعيار MGCV الوارد في المعادلة (18).**

ومما تجدر الإشارة إليه أن الباحثين غالباً مايستعملون الأسلوب الشخصي (I) في تحديد قيم المعلمات التمهيدية λ_j's عند التحليل اللامعلمي – متعدد المتغيرات –. وقد تمّ قتراح إستعمال أسلوب آخر في هذا البحث هو الأسلوب الآلي (II) لإختيار المعلمات التمهيدية عن طريق المعيار MGCV والمبين في المعادلة (18)، في محاولة لتقليل وقت تنفيذ عمل البرامج على الحاسوب، وهذا مالم تشهده البحوث الخاصة بالتمهيد عموماً، وتحليل GAM بوجه خاص.

3 <u>الجانب التجريبي</u>

تم تنفيذ تجارب المحاكاة بإستخدام برنامج MATLAB 7.5 ، وبتوظيف ثلاثة أحجام للعينات وكما يأتي: (n=50) ،(n=100) و (n=200). وبتكرارات(Replicates=300) لكل تجربة محاكاة. وكالآتي:

I. <u>المتغيرات التوضيحية</u>

وهي الt_j's تتوزع توزيع منتظم قياسي مستقل ، أي: $t_j \sim U(0,1)$, $j=1,2$

II.**<u>الأخطاء العشوائية</u>**

يتم توليد الأخطاء العشوائية بإستعمال طريقة Box-Muller، وبتوزيع طبيعي. أي أن:

$$e_i \sim N(0,\sigma^2) \quad , i=1,2,...,n$$

وتم تناول ثلاث مستويات من التباين على أساس Signal to Noise Ratio: S.N.R. [12]:

تباين عالي High Noise : $\sigma=(1/2)*$Function Range

تباين متوسط Medium Noise : $\sigma=(1/4)*$Function Range

تباين واطيء Low Noise : $\sigma=(1/8)*$Function Range

إذ أن σ يمثل الإنحراف المعياري للخطأ e.

III. **<u>المتغير المعتمد</u>**

يتم توليد المتغير Y_i مباشرة من خلال النماذج المستخدمة في تجارب المحاكاة، وذلك بإستخدام دوال الإنحدار بدلالة المتغيرين التوضيحيين اللذان تم توليدهما في الفقرة (I) أعلاه، مضافاً اليهما الأخطاء العشوائية والتي تم توليدها في الفقرة (II)، ولكل نموذج من النماذج قيد البحث.

1.3 **<u>النماذج المستعملة في المحاكاة</u>**

تم إستقاء بعض النماذج من بحوث منشورة فعلاً. وتم إستعمال GAM لصياغة مركبات كل من النماذج اللامعلمية بحيث تكون:

$$Y_i = \eta_i(t_{1i}, t_{2i}) + e_i \quad \ldots(20)$$

$$\eta_i(t_{1i}, t_{2i}) = g_1(t_{1i}) + g_2(t_{2i})$$

وبذا كانت مركبات كل نموذج كما يأتي:

1.**<u>النموذج الأول</u>** [10]**:**

$$g_1(t_{1i}) = Sin(4\pi t_{1i})$$

$$g_2(t_{2i}) = t_{2i}^3$$

إذ أن هاتين الدالتين من النوع Spatially homogeneous.

2.<u>النموذج الثاني</u> [10] **:**

$$g_1(t_{1i}) = t_{1i}^3$$

$$g_2(t_{2i}) = \exp\{-400(t_{2i}-0.6)^2\} + \frac{5}{3}\exp\{-500(t_{2i}-0.75)^2\} + 2\exp\{-500(t_{2i}-0.9)^2\}$$

إذ يتم تكوين هذا النموذج من جمع دالتين حيزيتين:إحداهما متجانسةSpatially homogeneous وهي الدالة الأولى، مع دالة غير متجانسة Spatially heterogeneous وهي الدالة الثانية.

3. النموذج الثالث[6] :

$$g_1(t_{1i}) = \frac{Cos(2.5\pi t_{1i})}{1+3t_{1i}^2}$$

$$g_2(t_{2i}) = \begin{cases} -2t_{2i} & t_{2i} < 0.6 \quad if \\ -1.2 & t_{2i} \geq 0.6 \quad if \end{cases}$$

وهو يجمع صيغة لاخطية مثلثية Cosine للدالة الأولى، مع صيغة خطية تغير سلوكها الى صيغة ثابتة عندما ($t_{2i} \geq 0.6$) للدالة الثانية.

2.3 معايير مقارنة النماذج

هناك عدد من المعايير لقياس مقدار الجودة في تقدير دالة الإنحدار اللامعلمية $\hat{g}(t)$، والتي تم تناولها آنفاً، مع الأخذ بنظر الإعتبار تنوع النماذج(الظواهر)التي يراد توظيف الشرائح التمهيدية لتمثيلها. ومن هذه المعايير (والتي يتم إختيار الأصغر منها عندما يراد قياس الأفضلية):

1.**معيار AMSE** [11]

ويمثل معدل متوسط مربعات الخطأ [Average of Mean Squared Error] ، وبالصيغة:

$$AMSE = n^{-1}\sum_{i=1}^{n} E\{\eta_i - \hat{\eta}_i\}^2 \quad \ldots (21)$$

2. **معيار GCV** [4]

من المعروف عن هذا المعيار إنه غالباً مايستعمل في تحديد قيمة (أو قيم) المعلمة (أو المعلمات) التمهيدية، ضمن مايعرف بالأساليب الآلية [21] **المعدة لهذا الغرض، والمبينة في المعادلة** (18)**. وهو يعتمد بصورة أساسية على درجات الحرية المكافئة EDF .**

فللمهد j^{th} **فأن درجة حريته تكون:** $df_j = trace(S_j) - 1 \quad ; \quad j = 1,2,\ldots,p$

وبالتالي يحقق نتائجاً أقل تحيزاً من MSE [3]**. وستستعمل الصيغة:**

$$MGCV = \frac{\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \alpha - \sum_{j=1}^{p}\hat{g}_j(t_{ij})\right\}^2}{\left(1 - \left[1 + \sum_{j=1}^{p}\{trS_j - 1\}\right]\Big/ n\right)^2} \quad \ldots (22)$$

3. **معيار AMDSE** [10]

ويمثل معدل وسيط مربعات الخطأ [Average of Median Squared Error]، وبالصيغة:

$$AMDSE = E\left(Median\{\eta_i - \hat{\eta}_i\}^2\right) \quad \ldots (23)$$

4. **معيار Concurvity** [8]

وهو معيار يبين فيهامعرفة قوة الـ Concurvity، وبالصيغة:

$$Concurvity = E(Corr.(\hat{Y}, \hat{Y}_{new})) \quad \ldots (24)$$

$$\hat{Y}_{new} = \hat{Y} + N(0, \sigma^2)$$

3.3 تنفيذ تجارب المحاكاة

لكل نموذج من النماذج الواردة في الفقرة أعلاه، تم القيام بما يأتي :

1. توليد المتغيرات: تم توليد المتغيرين t_1 و t_2 التوضيحية ليتوزع كل منهما توزيعاً منتظماً قياسياً U(0,1)، مع توليد أخطاء عشوائية تتبع توزيع Gaussian.

2. طريقة Backfitting: تم إختيار المعلمات التمهيدية وفق الأسلوب الشخصي المشار إليه في المعادلة (19)، ومن ثم إيجاد تقدير GAM بالصيغة (9).

3. الطريقة المباشرة: تم إختيار المعلمات التمهيدية وفق الأسلوب الشخصي المشار إليه في المعادلة (19)، ومن ثم إيجاد تقدير GAM كما في المعادلتين (16) و(17) معاً كحل للمعادلات الطبيعية.

4. إقتراح: تم إقتراح إستعمال أسلوب آلي لإختيار المعلمات التمهيدية لم تشهده البحوث المتعلقة بالـ GAM من قبل، كمحاولة لإختزال الزمن اللازم لتفيذ تجارب المحاكاة مما يؤدي الى فائدة برمجية، وكما في الصيغة (18) متمثلةً بالمعيار MGCV المتعدد، وتطبيق المعادلة (19) لإيجاد مقدّر GAM وبأسلوب تكراري مشابه للـ Backfitting.

4. النتائج

النموذج الأول من نتائج الجدول (1)، يلاحظ أن أسلوب backfitting أفرز نتائج أفضل عند أحجام العينات (n=50,100) إستناداً الى أغلب المعايير، عدا المعيار AMDSE عند حجم العينة (n=50) بالمقارنة مع الطريقة المباشرة. في حين سجل أسلوب MGCV المقترح تقدماً ملحوظاً على الطريقتين الأخريين فيما عداه، وعند حجم عينة (n=200). وبقيت نتائج الطريقة المباشرة Direct متأخرة قليلاً عن الأسلوبين الآخرين لاسيما عند (n=200). ويلاحظ ان المعيار Concurvity يزداد مع نقصان σ.

النموذج الثاني يتضح من الجدول (2) لنتائج التحليل من أن طريقة backfitting أظهرت أفضلية في نتائج المعايير AMSE وGCV وAMDSE عند أحجام العينات (n=50,100). أما المعيار Concurvity فقد تفوقّت backfitting فيه بصورة مطلقة ولجميع أحجام العينات، على ما سواها من الطريقتين الأخريين، إذ يتناسب عكسياً مع حجم العينة ولجميع الطرائق، ولكنه يزداد مع نقصان σ. في حين أظهر أسلوب MGCV المقترح أفضلية لكل المعايير ماعدا Concurvity وعند حجم العينة (n=200). وجاءت في المرتبة الأخيرة الطريقة المباشرة Direct مقارنةً بالأسلوبين الآخرين. ويلاحظ في هذا النموذج أن معيار الـ concurvity يزداد مع نقصان σ أيضاً.

النموذج الثالث فمن نتائج الجدول (3)، يلاحظ أن أسلوب backfitting أفرز نتائج أفضل عند أحجام العينات (n=50,100) إستناداً الى جميع المعايير. في حين سجل أسلوب MGCV المقترح تقدماً ملحوظاً على الطريقتين الأخريين ما عدا معيار Concurvity ومعيار AMSE عند (σ=1/2)، وعند حجم عينة (n=200). ويلاحظ كذلك أن معيارConcurvity يزداد مع نقصان σ.

5. الإستنتاجات

بناءً على ماجاء في الجانب التجريبي، توصل الباحثان الى مجموعة من الإستنتاجات التي من الممكن إدراجهاعلى ضوء نتائج التحليل للتجارب المقامة وبالإستناد الى النماذج قيد البحث، وكما يأتي:

1. ظهرت طريقة MGCV المقترحة في التحليل متفوقةً على ماسواها عند حجم (n=200) عند مقارنة المعايير الثلاثة AMSE و GCV و AMDSE.

2. أظهر المعيار concurvity طريقة backfitting في التحليل متقدمة على ماسواها، وذلك لأسباب تعود لتكنيك الطريقة في إيجاد فضاءات متعامدة لتلافي مشكلة تعدد المنحني.

3. يزداد معيار Concurvity مع نقصان مستويات σ ولجميع النماذج المدروسة.

4. أظهر معيار AMDSE أقل حساسية للشواذ من المعايير الأخرى.

5. تبدأ قيم المعايير – ماعدا Concurvity – بالتقارب عند مستويات التباين المتوسطة والواطئة مع تثبيت حجم العينة.

المصادر العربية:

1. علي، عمر عبد المحسن؛ (2007) ؛ "مقارنة مقدرات النماذج التجميعية المعممّة بإستخدام الشرائح التمهيدية عند تحليل الأنحدار اللامعلمي وشبه المعلمي"؛ أطروحة دكتوراه في الأحصاء، كلية الأدارة والأقتصاد، جامعة بغداد.

المصادر الاجنبية:

2. **Aerts, M.; Claeskens,G.; and Wand, M.P.; (2002); "Some theory for Penalized Spline Generalized additive models"; J.Statist.Plann.Inference, Vol. 103, pp. 455-470.**
3. **Buja, A.; Hastie, T.J. and Tibshirani, R.J ; (1989); " Linear smoothing and additive models" (with discussion); Annals of Statistics; Vol.17; pp. 453-555.**
4. **Eubank, R.L.; (1988); "Spline smoothing and nonparametric regression"; Marcel Dekker, New-York.**
5. **Hastie, T.J. and Tibshirani, R.J;(1987); "generalized additive models: some applications"; JASA; Vol.82, No.398, pp. 371-386.**
6. **Hastie, T.J and Tibshirani, R.J.; (1990); "Generalized additive models"; Chapman and Hall, London .**
7. **Hastie, T.J. and Tibshirani, R.J.; (1996); "Generalized additive models"; Encyclopedia of statistical sciences :Chapter "GAM"; Elsevier .**
8. **He, Shui; (2004); "Generalized additive models for data with concurvity: statistical issues and a novel model fitting approach" ; Ph.D. Dissertation, School of public health, University of Pittsburgh.**
9. **Lin, X. and Zhang, D.;(1999); "Infrence in generalized additive mixed models by using smoothing splines"; J.R.Statist.Soc., B., Vol.61, Part 2, pp. 381-400.**
10. **Ruppert, D.; and Carroll, R.J.; (2000); "Spatially-Adaptive Penalties for Spline Fitting "; Australian & New-Zealand J.Statist., Vol.42, No.2, pp. 205-223.**
11. **Wahba, G; (1990); "Spline models for observational Data"; SIAM: Society for Industrial and Applied Mathematics Philadelphia, PA.**
12. **Wood, S.; Kohn, R.; Shively, T.; Jiang, W.; (2002); "Model Selection in Spline Nonparameteric Regression"; J.R.Statist.Soc.B., Vol. 64, Part1, pp. 119-139.**

الملحق

جدول (1)

معايير تقدير GAM بالأسلوب التكراري وغير التكراري والمقترح، ولمختلف احجام العينات للنموذج الأول

		AMSE			GCV			AMDSE			Concurvity		
Method	n \ σ	1/2	1/4	1/8	1/2	1/4	1/8	1/2	1/4	1/8	1/2	1/4	1/8
GAM Backfitting	50	0.548170	0.536052	0.533049	0.475217	0.464741	0.462144	0.368265	0.372320	0.373821	0.299255	0.479824	0.692371
	100	0.564693	0.557057	0.555670	0.528187	0.520579	0.519287	0.401190	0.401346	0.401811	0.223248	0.358964	0.557423
	200	0.575700	0.573208	0.572590	0.565325	0.562903	0.562304	0.429654	0.431804	0.431558	0.123371	0.203830	0.338259
GAM Direct	50	0.548379	0.536082	0.533028	0.475398	0.464767	0.462127	0.367926	0.371704	0.373595	0.303184	0.485320	0.699081
	100	0.565336	0.559487	0.558023	0.528829	0.523379	0.522014	0.401253	0.402895	0.402788	0.227068	0.362551	0.558214
	200	0.590709	0.588120	0.587450	0.579848	0.577345	0.576697	0.434359	0.435169	0.435117	0.126385	0.207460	0.344402
Prop. GAM with MGCV	50	0.573810	0.547589	0.540719	0.505943	0.482848	0.476788	0.370865	0.363176	0.362324	0.442761	0.623053	0.802633
	100	0.581699	0.568155	0.564909	0.550584	0.537792	0.534730	0.390931	0.392748	0.393614	0.358835	0.545451	0.751490
	200	0.574871	0.569752	0.567723	0.563406	0.561167	0.559984	0.417794	0.418739	0.418316	0.126235	0.203452	0.330099

جدول (2)

معايير تقدير GAM بالأسلوب التكراري وغير التكراري والمقترح، ولمختلف احجام العينات للنموذج الثاني

		AMSE			GCV			AMDSE			Concurvity		
Method	n \ σ	1/2	1/4	1/8	1/2	1/4	1/8	1/2	1/4	1/8	1/2	1/4	1/8
GAM Backfitting	50	0.376166	0.364268	0.361355	0.326044	0.315770	0.313256	0.196323	0.194271	0.194367	0.251501	0.388663	0.590128
	100	0.377250	0.371754	0.370472	0.352942	0.347822	0.346626	0.205775	0.205071	0.204622	0.176486	0.280373	0.453794
	200	0.374407	0.371798	0.371141	0.367610	0.365076	0.364438	0.210092	0.208748	0.208386	0.098400	0.153860	0.265852
GAM Direct	50	0.376704	0.364257	0.361128	0.326511	0.315760	0.313060	0.194976	0.192952	0.193425	0.263907	0.408526	0.618170
	100	0.380264	0.374478	0.373046	0.355997	0.350606	0.349272	0.205683	0.204094	0.203857	0.189242	0.298506	0.483064
	200	0.477161	0.474564	0.473890	0.470461	0.467950	0.467298	0.212221	0.211329	0.211467	0.101249	0.164117	0.282760
Prop. GAM with MGCV	50	0.395710	0.369996	0.363663	0.348928	0.326244	0.320659	0.198415	0.190969	0.187582	0.502841	0.603699	0.761275
	100	0.388641	0.375321	0.372188	0.367816	0.355231	0.352270	0.200891	0.196408	0.194402	0.399009	0.505572	0.688285
	200	0.373831	0.371588	0.371045	0.367096	0.364918	0.364391	0.207152	0.206264	0.206704	0.169385	0.199375	0.291966

جدول (3)

معايير تقدير GAM بالأسلوب التكراري وغير التكراري والمقترح، ولمختلف احجام العينات للنموذج الثالث

		AMSE			GCV			AMDSE			Concurvity		
Method	n \ σ	1/2	1/4	1/8	1/2	1/4	1/8	1/2	1/4	1/8	1/2	1/4	1/8
GAM Backfitting	50	0.359577	0.347301	0.344238	0.311616	0.301001	0.298351	0.178571	0.172771	0.171526	0.275227	0.418288	0.616693
	100	0.363571	0.357772	0.356331	0.340059	0.334653	0.333310	0.175661	0.174110	0.173564	0.198885	0.314402	0.498724
	200	0.369017	0.366470	0.365845	0.361822	0.359349	0.358741	0.178820	0.177693	0.177614	0.099867	0.154193	0.259775
GAM Direct	50	0.359673	0.347324	0.344244	0.311700	0.301021	0.298357	0.178664	0.172672	0.171467	0.276516	0.420425	0.619000
	100	0.367842	0.362012	0.360558	0.344410	0.338977	0.337621	0.179590	0.178048	0.177495	0.199917	0.316370	0.501073
	200	0.529736	0.527398	0.526853	0.523742	0.521480	0.520952	0.319585	0.318770	0.318776	0.100342	0.154986	0.261515
Prop. GAM with MGCV	50	0.380271	0.354244	0.347755	0.335301	0.312353	0.306637	0.190088	0.178651	0.175507	0.391869	0.548048	0.739625
	100	0.373511	0.359992	0.352549	0.353562	0.340794	0.333459	0.181818	0.175805	0.171279	0.324806	0.478639	0.687608
	200	0.365333	0.363219	0.362666	0.358620	0.356565	0.356028	0.175349	0.174143	0.170328	0.166725	0.179096	0.259850