تقدير معاملات نماذج انحدار شرائح الجزاء باستعمال طريقة (SOP)

م.م.سهاد احمد عبد الله/جامعة بغداد/كلية الإدارة والاقتصاد/ suhad.a@coadec.uobaghdad.edu.iq

P: ISSN : 1813-6729
E : ISSN : 2707-1359
https://doi.org/10.31272/jae.i141.1011

تأريخ أستلام البحث : 2023/4/12 مقبول للنشر بتأريخ :2023/4/5

المستخلص

تلعب تقنيات تحليل الانحدار اللامعلمي دوراً مركزياً في التحليل الإحصائي، حيث تعتبر طريقة انحدار شرائح الجزاء واحدة من أكثر الطرائق المستعملة حالياً لتمهيد البيانات، اذ يمكن من خلالها تقدير الدوال مباشرة من البيانات الصاخبة(التي تحتوي على أخطاء) أو الملوثة (*noisy data*) بدلاً من الاعتماد على نماذج معلمية محددة، وتعتمد طريقة التقدير المستعملة لملائمه نموذج انحدار شرائح الجزاء في الغالب على طرائق المربعات الصغرى (OLS)، والتي من المعروف أنها حساسة للمشاهدات غير النمطية (المتطرفة)، في هذا البحث سيتم تقدير نماذج انحدار شرائح الجزاء (P-spline) المضافة المعممة باستعمال طريقة فصل المصفوفات الدقيقة المتداخلة (SOP) المقترحة من قبل الباحث (*Rodríguez*)، واخرون في عام 2015، والتي تأخذ المشاهدات المتطرفة في الاعتبار، حيث يعتمد التقدير على التكافؤ بين (P-spline) والنماذج المختلطة الخطية، ويتم تقدير معلمات التباين ومعلمات التمهيد بناءً على طريقة الإمكان الاعظم المقيد (REML). ومن اهم الاستنتاجات التي تم التوصل اليها عدم الحاجة الى استعمال طرائق التحسين العددي، كما يمكن دمج طريقة (SOP) بسهولة في تقدير النماذج المختلطة المضافة المعممة (GAMM) مع مجموعات التأثيرات العشوائية المستقلة، فضلاً عن سرعة تطبيق طريقة (SOP) في تنفيذ العمليات الحسابية.

الكلمات المفتاحية: شرائح الاساس، انحدار شرائح الجزاء، طريقة فصل المصفوفات المتداخلة (SOP).

مجلة الادارة والاقتصاد
مجلد 48 العدد 141 / كانون الاول / 2023
الصفحات : 159 - 167

1- المقدمـــــــــة *Introductionction*

أصبحت عملية التمهيد باستعمال دوال الجزاء الطريقة الأكثر شيوعاً لإجراء الانحدار اللامعلمي بسبب سهولة التركيب والمرونة في اختيار العقد وتمهيد المعلمة. مع ذلك ، تعتمد هذه المنهجية على خطوة أساسية وهي اختيار معلمة التمهيد (*smoothing parameter*)، التي تتحكم في المقايضة بين دقة البيانات والتسوية (أو التمهيد)، ويعود استعمال تركيبة من شرائح الانحدار والتي تحتوي على عقد أقل بكثير من حجم العينة، واستعمال شرائح الجزاء إلى الباحث (*O'Sullivan*) الذي استعمل في عام (1986) طريقة الشريحة الاساس المكعبة (*cubic B-spline*) للتقدير في المشاكل المعكوسة. **[10] [8]**

ومع تطور الجوانب النظرية لانحدار شريحة الجزاء، تم ربط الخصائص النظرية لانحدار شرائح الجزاء بخصائص انحدار الشرائح (بدون جزاء) وخطوط التمهيد للنماذج الخطية المعممة. ان طريقة التقدير المستخدمة لملائمة نماذج انحدار شرائح الجزاء، تقلل من مجموع مربعات البواقي الخاضعة لقيود على قاعدة معاملات انحدار الشريحة، حيث نلاحظ أن هذا النهج قد يكون شديد الحساسية لوجود نسبة صغيرة من المشاهدات غير النمطية، وتتمثل إحدى طرائق الحصول على ملائمة أكثر مقاومة لتأثير المشاهدات غير النمطية في البيانات باستبدال مربعات البواقي بدالة خسارة متزايدة ببطء ، كما هو الحال بالنسبة لمقدرات الانحدار-M. **[9] [6]**

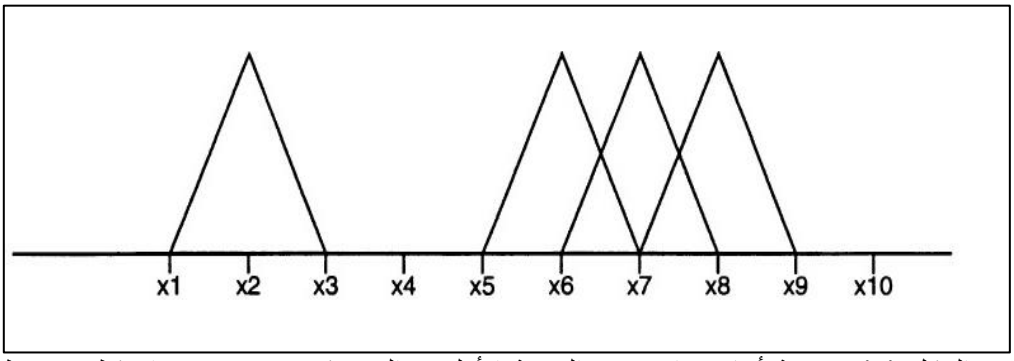
وهناك طريقتان رئيسيتان لتمهيد اختيار المعلمات، الطريقة الاولى تعتمد على تحسين بعض المعايير مثل معايير معلومات اكياكي (*Akaike information criteria*) والذي يشار اليه اختصارا بالرمز (AIC) أو معيار العبور الشرعي المعمم (*generalized cross validation*) والذي يشار اليه اختصارا بالرمز (GCV)، والطريقة الثانية يتم فيها التعامل مع الدالة الممهدة كدالة عشوائية، ومعلمات التمهيد المقدرة بالإمكان الاعظم (ML)، أو الامكان الاعظم للاحتمال المقيّد (REML). وعندما يشتمل النموذج على العديد من الدوال الممهدة، يزداد العبء الحسابي بسرعة مع عدد معلمات التمهيد التي سيتم اختيارها، ويمكن أن يصبح إجراء التصغير غير مستقر. لذلك، تم تطوير العديد من الخوارزميات لتحقيق الاستقرار العددي وتقليل الوقت الحسابي، حيث ان معظم هذه الخوارزميات تقع في إطار عمل معيار العبور الشرعي العام (GCV)، ويستند بعضها إلى مصفوفة العوامل (*matrix factorizations*)، أو تستعمل طريقة نيوتن (*Newton method*) بدلاً من طريقة المربعات الصغرى الموزونة التكرارية (*iterative re-weighted least squares*). **[14] [12]**

في هذا البحث سنتطرق إلى شرح خوارزمية جديدة لتقدير معاملات التمهيد لنموذج الانحدار الخطي العام متعدد الابعاد مع شرائح الجزاء متباينة الخواص (*p-spline*)، على أساس صيغة النموذج المختلط، والتي من خلالها يتم إيجاد تقديرات لكل مكون من مكونات التباين والتي يتم تضمينها بعد ذلك في عملية التكرار (*iterative procedure*) وبالتالي، فإن الخوارزمية ستكون سهلة التنفيذ في الممارسة العملية، كما يمكن استعمال بعض خصائص المقادير الجذرية المشتقة لزيادة سرعة العمليات الحسابية، كما سيتم إجراء مقارنات مع الطرائق البديلة على أساس معيار العبور الشرعي العام (GCV) ووقت الحوسبة.

2- شرائح الأساس *B-splines*

تعتبر تقنية شرائح الاساس (*B-Splines*) مهمة للنمذجة اللامعلمية، لكن اختيار العدد الأمثل ومواضع العقدة يعد امراً معقداً، اذ يمكن استعمال عقد متساوية البعد، لكن عددها الصغير والمنفصل يسمح فقط بالتحكم المحدود في التمهيد والملائمة، تتكون الشريحة الاساس (*B-splines*) من قطع او اجزاء متعددة الحدود، متصلة مع بعضها بعقد داخلية، وكما موضح بالشكل رقم (1) الاتي: **[7] [15]**

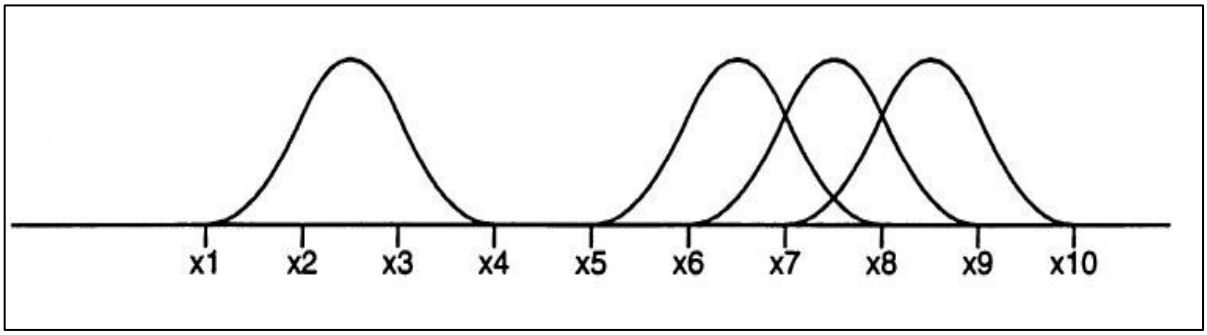
الشكل (1)، يوضح شريحة الاساس من الدرجة الأولى



ويوضح الشكل (1) شريحة أساس واحدة من الدرجة الأولى، والتي تتكون من جزئين او قطعتين خطيتين، الجزء الاول من x1 إلى x2، والأخر من x2 إلى x3. اذ ان العقد هي x1 وx2 وx3. على يسار x1 وإلى

يمين x3، ويكون المقطع B هو صفر. كما نلاحظ في الجزء الأيمن من الشكل (1) يتم عرض ثلاثة شرائح اساس (*B-splines*) أخرى من الدرجة الاولى، كل واحدة مبنية على ثلاث عقد، كما يمكننا بناء مجموعة كبيرة من الشرائح الاساس، من خلال تقديم المزيد من العقد. ويوضح الشكل (2) التالي الشريحة الاساس من الدرجة الثانية، والتي تتكون من ثلاث اجزاء او قطع تربيعية، مرتبطة بعقدتين، حيث تتطابق إحداثيات الاجزاء متعددة الحدود عند نقاط الربط، وتكون المشتقات الأولى متساوية، وتعتمد الشريحة الاساس على أربع عقد متجاورة، x1, ... ,x4. كما نلاحظ في الجزء الأيمن من الشكل (2)، يتم عرض ثلاثة شرائح أساس (*B-splines*) أخرى من الدرجة الثانية. **[7] [5]**

الشكل (2)، يوضح شريحة الأساس من الدرجة الثانية



من الشكلين (1) و (2) نلاحظ أن شرائح الأساس تتداخل مع بعضها البعض، اذ تتداخل الشرائح الاساس من الدرجة الأولى مع نقطتين جارتين، والشرائح الاساس من الدرجة الثانية مع أربعة نقاط متجاورة. بطبيعة الحال، فإن الشرائح الموجودة في أقصى اليسار وأقصى اليمين لها تداخل أقل، عند النقطة x، تكون شريحتان من الدرجة الأولى (او ثلاث من الدرجة الثانية) غير صفرية، ويمكن التعبير عن صيغة مشتقات الشرائح الاساس (*B-splines*) كما يلي:

$$\boldsymbol{l}\sum_{\boldsymbol{j}}\boldsymbol{a_j}\ \acute{\boldsymbol{B}}_{\boldsymbol{j}}\ (\boldsymbol{x};\boldsymbol{p}-\boldsymbol{1}) = \sum_{\boldsymbol{j}}\boldsymbol{a_j}\ \boldsymbol{B_j}\ (\boldsymbol{x};\boldsymbol{p}-\boldsymbol{1}) - \sum_{\boldsymbol{j}}\boldsymbol{a_{j+1}}\ \boldsymbol{B_{j+1}}\ (\boldsymbol{x};\boldsymbol{p}-\boldsymbol{1})$$

$$= -\sum_{\boldsymbol{j}}\Delta\boldsymbol{a_{j+1}}\ \boldsymbol{B_j}\ (\boldsymbol{x};\boldsymbol{p}-\boldsymbol{1}) \qquad .....(\boldsymbol{1})$$

حيث يمثل الرمز $\boldsymbol{l}$ المسافة بين العقد، وان $\Delta\boldsymbol{a_j} = \boldsymbol{a_j} - \boldsymbol{a_{j-1}}$، ويمكن التعبير عن المشتقة الثانية بالصيغة التالية:

$$\boldsymbol{l^2}\sum_{\boldsymbol{j}}\boldsymbol{a_j}\ \acute{\acute{\boldsymbol{B}}}_{\boldsymbol{j}}\ (\boldsymbol{x};\boldsymbol{p}) = \sum_{\boldsymbol{j}}\Delta^{\boldsymbol{2}}\boldsymbol{a_j}\ \boldsymbol{B_j}\ (\boldsymbol{x};\boldsymbol{p}-\boldsymbol{2}) \qquad ..... (\boldsymbol{2})$$

حيث ان $\Delta^{\boldsymbol{2}}\boldsymbol{a_j} = \Delta\Delta\boldsymbol{a_j} = \boldsymbol{a_j} - \boldsymbol{2a_{j-1}} + \boldsymbol{a_{j-2}}$. **[3]**

3- شرائح الجزاء *penalize splines*

تحتوي شرائح الجزاء (*P- Splines*) على عدد من الخصائص المفيدة والمبينة في القسم التالي، هذه الخصائص موروثة جزئياً من شرائح الاساس (*B-splines*)، اذ نلاحظ أن شرائح الجزاء (*P-splines*) لا تُظهر أي تأثيرات حدودية، كما تفعل العديد من أنواع دوال التمهيد اللبية (*kernel smoothing*)، وهذا يعني ان شكل انتشار منحنى الملائمة أو دالة الكثافة الاحتمالية للبيانات، يكون مصحوباً عموماً بالانحناء نحو الصفر. ويمكن لشرائح الجزاء (*P- Splines*) ملائمة بيانات متعددات الحدود، ولنفرض ان (x_i, y_i) معطى، فاذا كانت (y_i) متعددة الحدود في (x) من الدرجة k، فان شرائح الاساس (*B-splines*) من الدرجة (k) او اعلى سوف تتلاءم تماماً مع البيانات.

وينطبق الشيء نفسه على شرائح الجزاء (*P-splines*)، إذا كانت رتبة الجزاء من الدرجة k+1 أو أعلى، مهما كانت قيمة معلمة الضبط (λ)، اذ يمكن أن تحافظ شرائح الجزاء (*P-splines*) على العزوم (*moments*) من البيانات. وبالنسبة للنموذج الخطي الذي يحتوي على شرائح الجزاء من الدرجة (k+1) ودوال جزاء من الرتبة (k+1) أو أعلى، فإن: **[4] [2]**

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{x^k}\ \boldsymbol{y_i} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{x^k}\ \boldsymbol{\hat{y}_i} \qquad .....(\boldsymbol{3})$$

لكل قيم ($\boldsymbol{\lambda}$)، حيث ان $\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{j-1}^{n} \boldsymbol{b}_{ij}\ \widehat{\boldsymbol{a}}_{\boldsymbol{j}}$، هي القيم الملائمة (*fitted values*)، اما بالنسبة لنموذج الانحدار الخطي العام (GLM)، فإن:

$$\sum_{i=1}^{m} \boldsymbol{x}^{\boldsymbol{k}}\, \boldsymbol{y}_{\boldsymbol{i}} = \sum_{i=1}^{m} \boldsymbol{x}^{\boldsymbol{k}}\, \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} \qquad \ldots\ldots (4)$$

هذه الخاصية مفيدة بشكل خاص في سياق تمهيد دوال الكثافة الاحتمالية (pdf)، حيث ستكون مقدرات دالة الكثافة الاحتمالية (المتوسط والتباين) مساويين لمتوسط وتباين البيانات لأي مقدار من التسوية. وهذه الميزة بالمقارنة مع تمهيد الدوال اللبية تضخم التباين بشكل متزايد وتعطي تمهيد أقوى. إن حدود شرائح الجزاء (*P-splines*) التي تتلاءم مع البيانات مع تمهيد قوي هي متعدد الحدود (*polynomial*)، اما بالنسبة للقيم الكبيرة من معلمة الجزاء ($\boldsymbol{\lambda}$)عند جزاء من الرتبة **k**،فأن سلسلة الملائمة ستقترب من متعدد الحدود من الدرجة(**k-1**)، إذا كانت درجة الشرائح الاساس (*B-splines*) تساوي أو أعلى من (k). بالتالي فأن مقدر انحدار شرائح الجزاء ($\hat{\beta}$) يمكن التعبير عنه بالصيغة التالية:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{argmin} \sum_{i=1}^{m} (\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{X}(\boldsymbol{x}_{\boldsymbol{i}})'\, \boldsymbol{\beta})^{\,2} + \boldsymbol{\lambda} \sum_{i=1}^{m} \boldsymbol{\beta}_{\boldsymbol{p+j}}^{2} \qquad \ldots\ldots (5)$$

لكل معلمة جزاء $\boldsymbol{\lambda} > \mathbf{0}$، ولنرمز الى مصفوفة تصميم الشريحة (*spline design matrix*) بالرمز ***F*** حيث ان $\boldsymbol{F} = \{\boldsymbol{X}(\boldsymbol{x}_1)', \ldots, \boldsymbol{X}(\boldsymbol{x}_n)'\}'$، والى متجه الإجابات $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$، والى مصفوفة معاملات الشرائح بالرمز **D** حيث ان $\boldsymbol{D} = \boldsymbol{diag}(\mathbf{0}_{p+1}, \mathbf{1}_k)$ والتي تشير الى جزاء معاملات الشرائح فقط، فان ($\widehat{\boldsymbol{\beta}}$) يمكن التعبير عنه بالصيغة التالية:

$$\widehat{\boldsymbol{\beta}} = (\acute{\boldsymbol{F}}\boldsymbol{F} + \boldsymbol{\lambda}\, \boldsymbol{D})^{-1}\, \acute{\boldsymbol{F}}\boldsymbol{Y} \qquad \ldots\ldots (6) \qquad [1]\ [11]$$

4- طريقة فصل المصفوفات الدقيقة المتداخلة Separation of overlapping precision matrices

مع ظهور بنية هيكلية معقدة أكثر لمصفوفة التغاير المشترك (*variance–covariance matrix*) للنماذج الخطية المختلطة المعممة (*GLMM*)، كان ذلك الدافع وراء دراسة واستعمال طرائق انحدار شرائح الجزاء (*penalized spline regression*)، على الرغم من ذلك، فإن طريقة فصل المصفوفات الدقيقة المتداخلة (*separation of overlapping precision*) والتي يشار اليها اختصاراً بالرمز (SOP)، لا تقتصر على هذا المجال فقط، بل يمكن اعتبارها طريقة تقدير عامة للنماذج الخطية المختلطة المعممة (GLMM) مع مصفوفة دقيقة (*precision matrix*) لهيكل معين، إذ يمكن استعمال طريقة (SOP) كلما كانت هناك جزاءات متعددة تعمل على نفس المعاملات، حيث تعتبر متباينة الخواص (*P-splines*) هو مثال على الجزاءات المتداخلة. [11]

لنفرض لدينا نموذج الانحدار العام التالي:

$$\boldsymbol{g}(\boldsymbol{\mu}_{\boldsymbol{i}}) = \boldsymbol{f}(\boldsymbol{x}_{\boldsymbol{i}}), \qquad \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots\ldots, \boldsymbol{n} \qquad \ldots\ldots (7)$$

حيث ان $\boldsymbol{\mu}_{\boldsymbol{i}} = \boldsymbol{E}(\boldsymbol{y}_{\boldsymbol{i}})$، وان $\boldsymbol{g}(.)$ هي دالة الربط (*link function*)، وان $\boldsymbol{f}(.)$ هي دالة التمهيد المجهولة، ولنفرض ان $\boldsymbol{var}(\boldsymbol{y}_{\boldsymbol{i}}) = \boldsymbol{\phi v}(\boldsymbol{\mu}_{\boldsymbol{i}})$، يتم تقريب الدالة المجهولة $\boldsymbol{f}(\boldsymbol{x}_{\boldsymbol{i}})$ من خلال تركيبة خطية من دوال الأساس (*B-splines*)، أي ان:

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{j=1}^{d} \boldsymbol{\alpha}_{\boldsymbol{j}}\, \boldsymbol{\beta}_{\boldsymbol{j}}(\boldsymbol{x}) \qquad \ldots\ldots (8)$$

بالتالي يمكن الحصول على نموذج شرائح الجزاء باستعمال المصفوفات بالصيغة التالية:

$$\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{B\theta} \qquad \ldots\ldots (9)$$

حيث ان $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_n)^{\boldsymbol{t}}$، وان $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n)^{\boldsymbol{t}}$، وان **B** هي مصفوفة انحدار الشرائح الأساس (*B-splines*) ذات الابعاد (n*d)، أي ان: $\boldsymbol{b}_{\boldsymbol{ij}} = \boldsymbol{B}_{\boldsymbol{j}}(\boldsymbol{x}_{\boldsymbol{i}})$ هي قيمة الشريحة الأساس من الدرجة $\boldsymbol{j}^{\boldsymbol{th}}$ عند $\boldsymbol{x}_{\boldsymbol{i}}$، ويتم تحقيق التسوية او التمهيد من خلال فرض جزاءات على معاملات الانحدار $\boldsymbol{\theta}$ كما في الصيغة التالية:

$$\boldsymbol{\lambda} \sum_{k=q+1}^{d} (\Delta^{q}\boldsymbol{\theta}_{\boldsymbol{k}})^2 = \boldsymbol{\lambda}\boldsymbol{\theta}^{\boldsymbol{t}}\, \boldsymbol{D}_{\boldsymbol{q}}^{\boldsymbol{t}}\, \boldsymbol{D}_{\boldsymbol{q}}\, \boldsymbol{\theta} \qquad \ldots\ldots (10)$$

حيث تمثل $\boldsymbol{\lambda}$ معلمة الضبط، وتمثل $\Delta^{\boldsymbol{q}}$ الاختلافات من الرتبة $\boldsymbol{q}$ على المعاملات المتجاورة، أي ان:

$$\Delta\boldsymbol{\theta}_{\boldsymbol{k}} = \boldsymbol{\theta}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\boldsymbol{k-1}}, \qquad \Delta^2\boldsymbol{\theta}_{\boldsymbol{k}} = \Delta(\Delta\boldsymbol{\theta}_{\boldsymbol{k}}) = \boldsymbol{\theta}_{\boldsymbol{k}} - \mathbf{2}\boldsymbol{\theta}_{\boldsymbol{k-1}} + \boldsymbol{\theta}_{\boldsymbol{k-2}}$$

وتمثل $\boldsymbol{D_q}$ مصفوفة العرض للاختلافات Δ^q، ونلاحظ من المعادلة (10) ان معلمة الضبط ($\boldsymbol{\lambda}$) تنطبق على جميع فروق المعاملات، بغض النظر عن موقعها، وبالتالي فأن النموذج يفترض أن نفس القدر من التمهيد مطلوب عبر المدى الكامل للمتغير المشترك. **[14] [7]**

تعمل شرائح الجزاء التكيفية على تقليل حدة هذا الافتراض من خلال فكرة بسيطة، لاستبدال معلمة التمهيد العامة بمعلمات تمهيد تتنوع مكانياً وفقاً لقيمة المتغير المشترك، يمكن تحقيق ذلك من خلال تحديد معلمة تمهيد مختلفة لكل اختلاف في المعامل، اي ان:

$$\sum_{\boldsymbol{k=q+1}}^{\boldsymbol{d}} \boldsymbol{\lambda_{k-q}}\,(\Delta^{\boldsymbol{q}}\boldsymbol{\theta_k})^2 = \boldsymbol{\theta}^{\boldsymbol{t}}\,\boldsymbol{D_q^t}\,\boldsymbol{diag}(\boldsymbol{\lambda})\,\boldsymbol{D_q}\,\boldsymbol{\theta} \qquad \ldots\ldots (\mathbf{11})$$

حيث أن $\boldsymbol{\lambda} = (\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \ldots, \boldsymbol{\lambda_{d-q}})^{\boldsymbol{t}}$. أن هذا الأسلوب يعني ضمنياً ان العديد من معلمات التمهيد تعامل مثل اختلافات المعامل **(أي ان، $\boldsymbol{d-q}$)**، مما قد يؤدي إلى عمليات حسابية غير مستقرة وغير ممهدة، وبالنظر إلى طبيعة الموقع (*local*) والرتبة (*ordered*) لاختلافات المعامل، فانه يمكننا نمذجة معلمات التمهيد $\boldsymbol{\lambda_k}$ كدالة تمهيد للموقع $\boldsymbol{k}$ واستعمال شرائح الأساس (*B-splines*) لهذا الغرض (اذ لا يُفترض هنا أي جزاء)، حيث ان:

$$\boldsymbol{\lambda} = \boldsymbol{\Upsilon}\,\boldsymbol{\Omega} \qquad \ldots\ldots (\mathbf{12})$$

حيث ان $\boldsymbol{\Upsilon}$ مصفوفة انحدار شرائح الأساس (*B-splines*) ذات الابعاد p * (d-q) مع p<(d-q)، وان $\boldsymbol{\Omega} = (\boldsymbol{\Omega_1}, \boldsymbol{\Omega_2}, \ldots, \boldsymbol{\Omega_p})^{\boldsymbol{t}}$، تمثل متجه المعلمات الممهدة، بالتالي يمكن إثبات أن شرائح الجزاء التكيفية المبينة في المعادلة (11) هي:

$$\boldsymbol{\theta}^{\boldsymbol{t}}\left(\sum_{\boldsymbol{l=1}}^{\boldsymbol{p}} \boldsymbol{\omega_l}\,\boldsymbol{D_q^t}\,\boldsymbol{diag}(\boldsymbol{\Upsilon_l})\,\boldsymbol{D_q}\right)\boldsymbol{\theta} \qquad \ldots\ldots (\mathbf{13})$$

حيث تشير $\boldsymbol{\omega_l}$ الى العمود *l* في المصفوفة $\boldsymbol{\Upsilon}$، كما نلاحظ أنه في ظل دوال الجزاء التكيفية، يتم جزاء جميع المعاملات من خلال معلمات تمهيد متعددة، أي أن هناك جزاءات متداخلة. [13]

ويمكن إجراء تقدير لنموذج شرائح الجزاء المبين في المعادلة (*9*) الخاضع للجزاءات التكيفية المحددة في المعادلة (*13*) بناءً على العلاقة بين شرائح الجزاء والنماذج المختلطة. اذ يؤدي هذا الاسلوب إلى ظهور مصفوفات دقة قطرية (*diagonal precision matrices*). وباستعمال تحويل (*Eilers*)، يمكن التعبير عن النموذج (*9*)بالصيغة التالية: **[6] [1]**

$$\boldsymbol{g(\mu)} = \boldsymbol{B\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} \qquad \ldots\ldots (\mathbf{14})$$

حيث ان $\boldsymbol{X} = [\mathbf{1_n}\,|\boldsymbol{x}|\ldots|\boldsymbol{x}^{(q-1)}|]$ ، وان $\boldsymbol{Z} = \boldsymbol{B}\,\boldsymbol{D_q^t}\,(\boldsymbol{D_q}\,\boldsymbol{D_q^t})^{-1}$، وتصبح مصفوفة الدقة (الجزائية) لمتجه المعاملات العشوائية (الجزائية) $\boldsymbol{\alpha}$ بالصيغة التالية:

$$\boldsymbol{G^{-1}} = \frac{\mathbf{1}}{\boldsymbol{\phi}}\boldsymbol{T^t}\,\boldsymbol{P_{Ad}}\,\boldsymbol{T} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{p}} \boldsymbol{\sigma_l^{-2}}\,\boldsymbol{diag}(\boldsymbol{\Upsilon_l}) = \sum_{\boldsymbol{l=1}}^{\boldsymbol{p}} \boldsymbol{\sigma_l^{-2}}\,\widetilde{\mathbf{A}}_{\boldsymbol{l}} \qquad \ldots\ldots (\mathbf{15})$$

حيث ان $T = D_q^t\,(\boldsymbol{D_q}\,\boldsymbol{D_q^t})^{-1}$ ، و $\widetilde{\mathbf{A}}_{\boldsymbol{l}} = \boldsymbol{diag}(\boldsymbol{\Upsilon_l})$ ، و $\boldsymbol{\sigma_l^2} = \boldsymbol{\phi}/\boldsymbol{\omega_l}$ وان $\boldsymbol{P_{Ad}} = \boldsymbol{\lambda D_q^t\,D_q}$، لذلك فان مصفوفة الدقة (*precision matrix*) تكون تركيبة خطية من معلمات الدقة $\boldsymbol{(\sigma_l^{-2})}$، وبالتالي يمكن استعمال طريقة (SOP). [11] [4]

5- التمهيد الأمثل (التسوية المثلى) *Optimal Smoothing*

الآن بعد أن أصبح بالإمكان التأثير بسهولة على تمهيد منحنى الملائمة (*fitted curve*) مع معلمة الضبط ($\boldsymbol{\lambda}$)، نحتاج إلى طريقة مناسبة لاختيار القيمة المثلى لمعلمة الضبط ($\boldsymbol{\lambda}$)، سيتم استعمال معيار معلومات كياكي (AIC)، لتصحيح الاحتمالية اللوغاريتمية (*log-likelihood*) للنموذج الملائم (*fitted model*) لعدد المعلمات الفعالة، وبدلاً من الاحتمالية اللوغاريتمية، سيكون من الاسهل استعمال الانحراف (*deviance*)، بالتالي يمكن تعريف (AIC) بالصيغة التالية:

$$\boldsymbol{AIC(\lambda)} = \boldsymbol{dev(y;\,\alpha,\lambda)} + \mathbf{2} * \mathbf{dim}(\boldsymbol{\alpha},\boldsymbol{\lambda}) \qquad \ldots\ldots (\mathbf{16})$$

حيث يمثل المصطلح $\mathbf{dim}(\boldsymbol{\alpha},\boldsymbol{\lambda})$ البعد الفعال للمنحنى الملائم (أي أثر للمصفوفة الأكثر تمهيد)، ويمثل المصطلح $\boldsymbol{dev(y;\,\alpha,\lambda)}$ الانحراف (*deviance*). وفي بعض النماذج الخطية العامة (GLM) يكون مقياس البيانات معرف، كما هو الحال بالنسبة للبيانات التي تتبع توزيع بواسون، والبيانات التي تتبع توزيع ذي الحدين، اذ يمكن حساب الانحراف مباشرةً. اما بالنسبة للبيانات الخطية، فانه يلزم تقدير التباين، وتتمثل إحدى

الطرائق في أخذ تباين القيم المتبقية من $(\hat{y}_i)$ التي يتم حسابها عندما تكون $(\lambda = 0)$ والتباين $(\hat{\sigma}_0^2)$، ويكون معيار معلومات كياكي بالصيغة التالية:

$$AIC = \sum_{i=1}^{m} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_0^2} + 2\,tr(H) - 2m\,ln\,\hat{\sigma}_0 - m\,ln\,2\pi \qquad \ldots\ldots (17)$$

يعد اختيار التباين هذا اعتباطياً إلى حد ما، لأنه يعتمد على عدد العقد (*knots*) فقط. بدلاً من ذلك يمكن الاستناد بصورة عامة الى معيار التحقق من الصحة او معيار العبور الشرعي (*cross-validation*)، والذي يشار اليه بالرمز (CV) بالصيغة التالية: **[13] [6]**

$$CV(\lambda) = \sum_{i=1}^{m} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2 \qquad \ldots\ldots (18)$$

حيث تمثل h_{ii} العناصر القطرية للمصفوفة H ذات الابعاد (*m*m*)، ويمكن التعبير عن معيار التحقق من الصحة العام او معيار العبور الشرعي العام (*general cross-validation*)، والذي يشار اليه اختصاراً بالرمز (GCV) بالصيغة التالية:

$$GCV(\lambda) = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{(m - \sum_{i=1}^{m} h_{ii})^2} \qquad \ldots\ldots (19)$$

بشكل عام الفرق بين كلا المقدارين صغير، وان أفضل λ هي القيمة التي تقلل من مقدار معيار العبور الشرعي (CV) أو معيار العبور الشرعي العام (GCV)، ويعتبر تباين القيم المتبقية (تباين البواقي) عند معلمة الضبط (λ) المثلى الاختيار الأفضل والمناسب لاستعماله في تقدير التباين (σ_0^2) **لحساب** $\mathbf{AIC}(\lambda)$، ومن الناحية العملية يفضل التعامل مع الاصدارات المعدلة من معيار العبور الشرعي $\mathbf{CV}(\lambda)$ ومعيار العبور الشرعي العام $\mathbf{GCV}(\lambda)$، مع قيم يمكن تفسيرها على أنها تقديرات للانحراف المعياري للعبور الشرعي، إذ أن:

$$\overline{CV(\lambda)} = \sqrt{CV(\lambda)/m} \quad , \quad \overline{GCV(\lambda)} = \sqrt{m\,GCV(\lambda)} \qquad \ldots\ldots (20)$$

وتعتبر تقديرات الانحراف المعياري للعبور الشرعي مفيداً للمقارنة بين عدة نماذج ملائمة ولأعداد مختلفة من العقد ورتب الجزاء، في حين أن معلمة الضبط (λ) يمكن أن تختلف عبر مدى كبير من القيم وليس لها جاذبية بديهية واضحة. **[12] [6]**

6- وصف تجربة المحاكاة وتحليل النتائج *Simulation Study & Analysis results*

تعرف المحاكاة بانها عملية تمثيل او تقليد للواقع الحقيقي باستعمال نماذج معينة، وكثيرا ما نجد في الواقع الحقيقي ان هناك عمليات تكون معقدة الفهم والتحليل لذلك فمن الافضل ان نوصف هذه العمليات بصورة مشابهة للصور الحقيقية بنماذج معينة. ان اول مراحل استعمال اسلوب المحاكاة هو توليد المتغيرات العشوائية، اذ تم توليد البيانات وفقاً للصيغة التالية:

$$y_i = \sin(4/x_i) + 1.5 + \varepsilon_i\,, \qquad i = 1,\ldots,n \qquad \ldots\ldots (21)$$

حيث ان $x_i \sim U[0,1]$ وان $\varepsilon_i \sim N[0, 0.04]$، وتم تنفيذ تجربة المحاكاة من خلال توظيف دالة دوبلر (*doppler function*)، اذ تم توليد عينة بحجم (n=500)، وتكرار التجربة مائة مره **(m=100)** للحصول على درجة عالية من التجانس، أيضا تم توليد المتغير معتمد **(y)**، من خلال النماذج المستعملة في تجربة المحاكاة وذلك باستعمال دالة دوبلر (*doppler function*) بدلالة المتغيرات التوضيحية التي تم توليدها مضافاً اليها الأخطاء العشوائية، وفي جميع الحالات، تم استعمال (200) شريحة أساس مكعبة (*cubic B-splines*) لتمثيل الدالة الممهدة، بالاشتراك مع الفروق من الدرجة الثانية، وبالنسبة لطريقة شرائح الجزاء التكيفية (*adaptive p-splines*)، تم استعمال (20) مقطعاً من شريحة الاساس مكعبة متباعدة بشكل متساوٍ لتسوية معلمات التمهيد (انظر المعادلة (12))، حيث يتم اختيار هذه القيم لتوفير مرونة كافية للنموذج، وكانت النتائج موضحة بالجدول (1) الاتي:

الجدول رقم (1) نتائج تطبيق انحدار شرائح الجزاء باستعمال طريقة (SOP).

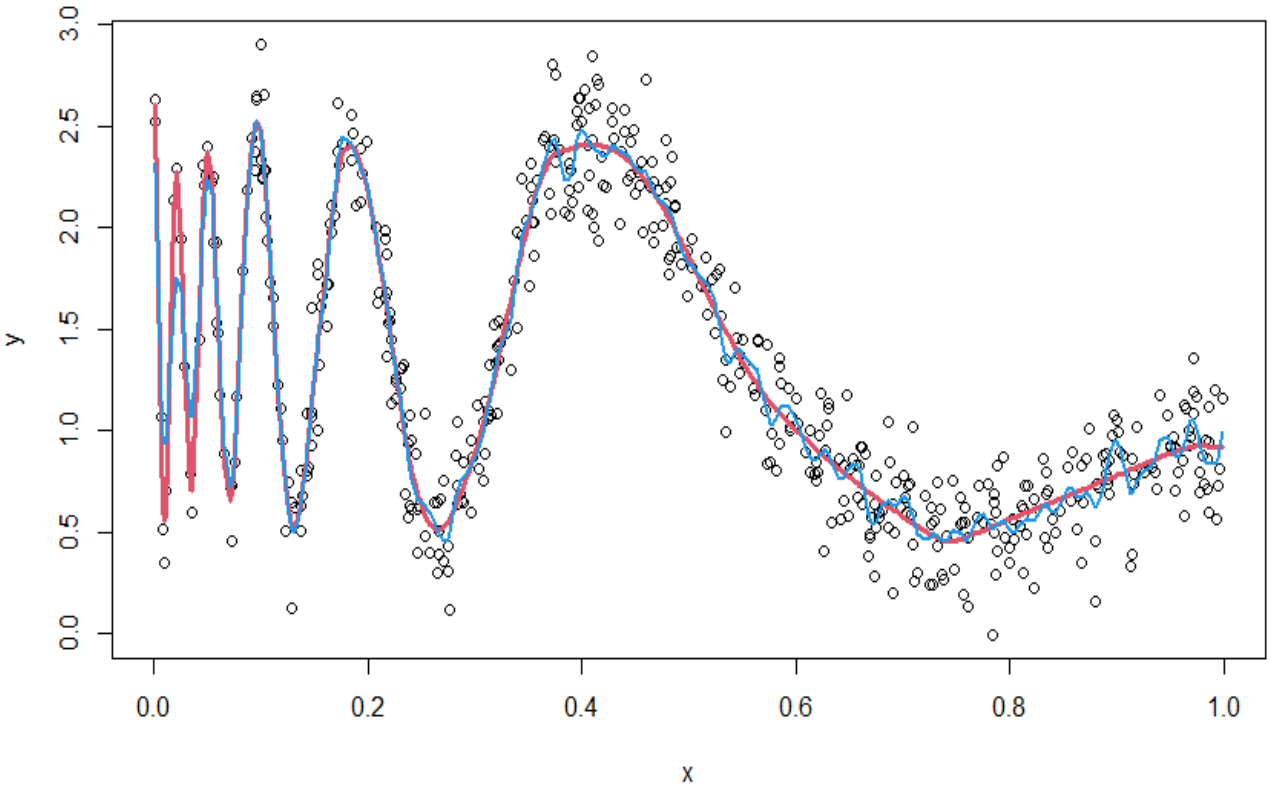
	α	β	R^2	Dev.exp.	AIC	GCV
with adaptive smoothing	**1.268**	**5.1013**	**0.91**	**92.7%**	**250.485**	**18.179**
Without adaptive smoothing	**1.269**	**4.9944**	**0.861**	**91.1%**	**250.485**	**17.232**

نلاحظ من الجدول (1) أن قيمة ثابت الانحدار الناتجة من تطبيق انحدار شرائح الجزاء باستعمال طريقة فصل المكونات الدقيقة (SOP) مع تمهيد تكيفي قد بلغت (α=1.268)، وبلغت قيمة معلمة الميل $(\beta = 5.1013)$،

أما قيمة معامل التحديد فقد بلغت (R=0.91)، بما يدل على وجود تحيز في تقدير المعالم، وان النموذج يوضح ما نسبته (92.7%) من الانحرافات وهي الأعلى إذ ما تمت مقارنتها بطريقة فصل المكونات الدقيقة (SOP) بدون تمهيد تكيفي، وبلغت قيمة معيار كياكي (AIC=250.485)، وبلغت قيمة معيار العبور الشرعي العام (GCV=18.179). كما نلاحظ أن قيمة ثابت الانحدار الناتجة من تطبيق انحدار شرائح الجزاء باستعمال طريقة فصل المكونات الدقيقة (SOP) بدون تمهيد تكيفي قد بلغت (α=1.269)، وبلغت قيمة معلمة الميل ($\beta = 4.9944$)، أما قيمة معامل التحديد فقد بلغت (R=0.861)، وان النموذج يوضح ما نسبته (91.1%) من الانحرافات، وبلغت قيمة معيار كياكي (AIC=250.485)، وبلغت قيمة معيار العبور الشرعي العام (GCV=17.232) بما يدل على أنها الطريقة الأفضل لتقدير المعالم.

ويوضح الشكل (3) التالي المنحنيات المقدرة بناءً على طريقة فصل المكونات الدقيقة (SOP) بدون جزاءات تكيفية والتي تم تمييزها باللون الازرق، والمنحنيات المقدرة بناءً على طريقة فصل المكونات الدقيقة (SOP) باستعمال الجزاءات التكيفية والتي تم تمييزها باللون الاحمر، اذ نلاحظ انه يمكن تحديد (5) دورات لدالة دوبلر باستعمال الجزاءات التكيفية ، من ناحية أخرى فإن النهج غير التكيفي قادراً على تحديد (3) دورات فقط، كما يقدم تقديرات متذبذبة للغاية، خاصة على الجانب الأيمن من المجال المتغير.

الشكل (3)، يوضح المنحنيات المقدرة باستعمال طريقة فصل المكونات الدقيقة (SOP)



7- الاستنتاجات Conclusions

- تم تحليل البيانات الملوثة (*noisy data*) التي تم توليدها من خلال اسلوب المحاكاة، باستعمال طريقة تقدير حديثة نسبياً وهي طريقة فصل المصفوفات الدقيقة المتداخلة (SOP)، وعلى عكس طرائق التقدير الاخرى، فإن طريقة (SOP) تكون مناسبة للنماذج الخطية المختلطة المعممة، حيث تكون مصفوفة الدقة خطية في معلمات الدقة، وشائعة عند إعادة صياغة النماذج الجزاء الممهدة كنماذج مختلطة خطية (معممة).
- توفر طريقة (SOP) عند التقارب، تقدير لدرجات الحرية الجزئية الفعالة المرتبطة بكل مكون ممهد / عشوائي في النموذج، وهذا مهم بشكل خاص عند التعامل مع نماذج التمهيد، حيث تعطي درجات الحرية الفعالة لكل مصطلح ممهد صورة مهمة حول تعقيد الدالة الملائمة (*fitted function*).
- نستنتج من خلال تنفيذ تجربة المحاكاة أن طريقة (SOP) تضمن تقديرات غير سلبية لمعلمات التباين (*variance parameters*) ومناقشة ظروف التجربة التي بموجبها تكون هذه التقديرات إيجابية تماماً، كما أظهرت طريقة (SOP) تفوقاً من حيث أوقات الحوسبة باستخدام لغة البرمجة الإحصائية (R) مقارنةً بطريقة (*B-splines*).

المصادر

[1] Currie, I. D., & Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. Statistical Modelling, 2(4), 333-349.
[2] de Boor, C. (1977). Package for calculating with B-splines. SIAM J. Numer. Anal. 14 441{472}.
[3] de Boor, C. (1978). A Practical Guide to Splines. Springer, Berlin.
[4] Djeundje, V.A., Currie, I.D.: Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. Electron. J. Stat. 4, 1202–1224 (2010).
[5] Durban, M., Harezlak, J., Wand, M.P., Carroll, R.J.: Simple fitting of subject-specific curves for longitudinal data. Stat. Med. 24(8), 1153–1167 (2005).
[6] Eilers, P.H.C.: Discussion of Verbyla et al. J. R. Stat. Soc. Ser. C (Appl. Stat.) 48, 300–311 (1999).
[7] Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. Statistical science, 11(2), 89-121.
[8] Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Stats.*, 35:73{101}.
[9] Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2007). Some asymptotic results on generalized penalized spline smoothing. Research Report KBI-0733, ORSTAT, K.U. Leuven.
[10] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Stat. Science*, 1:505{527}. With discussion.
[11] Rodríguez-Álvarez, M. X., Lee, D. J., Kneib, T., Durbán, M., & Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. Statistics and Computing, 25(5), 941-957.
[12] Wahba, G. (1990). Spline models for observational data. Society for industrial and applied mathematics.
[13] Wood, S.N.: Fast stable direct fitting and smoothness selection for generalized additive models. Biometrics **62**, 1025–1036 (2008).
[14] Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B **73**, 3–36 (2011).
[15] Wood, S.N., Scheipl, F., Faraway, J.J.: Straightforward intermediate rank tensor product smoothing in mixed models. Stat. Compute. **23**, 341–360 (2013).

Estimate the coefficients of penalty spline regression models using the (SOP) method.

Suhad Ahmed Abdullah/University of Baghdad/College of Administration and Economics/suhad.a@coadec.uobaghdad.edu.iq

Abstract

Nonparametric regression analysis techniques play a central role in statistical analysis, as the penalty spline regression method is considered one of the most currently used methods for smoothing data, as functions can be estimated directly from noisy data (which contain errors) or polluted (noisy data) instead of Relying on specific parameter models. The estimation method used to fit the penalty spline regression model is mostly based on least squares (OLS) methods, which are known to be sensitive to atypical (extreme) observations. In this research, penalty spline regression models (P-spline) will be estimated.) The generalized additive using the nested microarray separation method (SOP) proposed by the researcher (Rodríguez) and others in 2015, which takes extreme observations into account, where the estimation is based on the equivalence between (P-spline) and linear mixed models, and the parameters are estimated Variance and smoothing parameters based on the restricted maximum potential (REML) method. One of the most important conclusions reached is that there is no need to use numerical optimization methods, and the SOP method can be easily integrated into the estimation of generalized additive mixed models (GAMM) with independent random effects groups, in addition to the speed of applying the SOP method in implementing calculations .

Keywords : basis splines, penalty splines regression, nested matrix separation (SOP) method.

****
****
****