مقارنة الانحدار الشرائحي المعكوس مع المركبات الرئيسة في اختزال البيانات ذات الابعاد العالية بأستعمال المحاكاة

أ . م . د. عمر عبد المحسن علي/ كلية الادارة والاقتصاد / جامعة بغداد
الباحث/ زينة ابراهيم حسن

تاريخ التقديم:2017/6/6
تاريخ القبول:2017/7/9

المستخلص

يدرس هذا البحث طرائق اختزال الابعاد التي تعمل على تجاوز مشكلة البعدية عندما تفشل الطرائق التقليدية في ايجاد تقدير جيد للمعلمات، لذلك يتوجب التعامل مع هذه المشكلة بشكل مباشر. ومن اجل ذلك، يجب التخلص من هذه المشكلة لذا تم استعمال اسلوبين لحل مشكلة البيانات ذات الابعاد العالية الاسلوب الاول طريقة الانحدار الشرائحي المعكوس (SIR) والتي تعتبر طريقة غير كلاسيكية وكذلك طريقة (WSIR) المقترحة والاسلوب الثاني طريقة المركبات الرئيسة (PCA) وهي الطريقة العامة المستخدمة في اختزال الابعاد ، ان عمل طريقة انحدار الشرائحي المعكوس (SIR) و طريقة المركبات الرئيسة (PCA) يقوم على عمل توليفات خطية مختزلة من مجموعة جزئية من المتغيرات التوضيحية الأصلية والتي قد تعاني من مشكلة عدم التجانس ومن مشكلة التعدد الخطي بين معظم المتغيرات التوضيحية ، وستقوم هذه التوليفات الجديدة المتمثلة بالمركبات الخطية الناتجة من الطريقتين بإختزال أكثر عدد من المتغيرات التوضيحية للوصول الى بُعد جديد واحد او اكثر يسمى بالبعد الفعّال . وسيتم استعمال معيار جذر متوسط مربعات الخطأ للمقارنة بين الاسلوبين لبيان افضلية الطرائق ، وقد تم اجراء دراسة محاكاة للمقارنة بين الطرائق المستعملة وقد بينت نتائج المحاكاة ان طريقة weight standard Sir المقترحة هي الافضل .

المصطلحات الرئيسية للبحث/ اختزال الابعاد ، الانحدار الشرائحي المعكوس ، المركبات الرئيسية.

* البحث مستل من أطروحة دكتوراه.

1. المقدمة :

لدراسة و تحليل بيانات الظواهر الاقتصادية و الطبية و الزراعية و المالية وغيرها، يتوجب ان تتوفر المعرفة المسبقة لهذه الظواهر ، بمعنى اخر ، معرفة نوع بياناتها والتي غالباً ماتكون كمية ، ويتطلب ذلك بناء أنموذج رياضي مناسب والذي يمثل العلاقات السببية (دالة سببية او سلوكية) بين عواملها افضل تمثيل وهي ماتدعى مرحلة الوصف (description) ، لأعتماد التحليل المناسب والذي يمكن بعد ذلك من أتخاذ العديد من القرارات بشأن أهم الدلالات والخصائص (characteristics) المتعلقة بتلك الظواهر، وتدعى تلك الدلالات المذكورة سابقاً بالمعلمات (parameters) . ان من أهم نماذج التحليل الاحصائي هو ماتدعى بتحليل نماذج الانحدار و يوجد منهجين مختلفين لتناول هذه النماذج ، ولكل منهج أو أسلوب توجد شروط أوقيود . فالاسلوب الاول هو: اسلوب الانحدار المعلمي الذي يفترض ان تكون العينة متأتية من مجتمع محدد من احدى التوزيعات الاحتمالية المعروفة ويتم تقدير المعلمات باستعمال طرائق مختلفة ، وبسبب سهولة العمليات الاستدلالية وقوة الاختبارات المعلمية ظلت الاساليب المعلمية مهيمنة على تحليل الانحدار لحقبة من الزمن ، لكن قد يؤدي الافتراض الخاطئ للتوزيع المعلمي الى استنتاجات خاطئة و تقديرات غير متسّقة و كذلك لانها لاتناسب البيانات المعقّدة ، ولهذه الاسباب يلجأ الباحثون الى الاسلوب الثاني وهو الاسلوب اللامعلمي أو الشبه المعلمي لتحليل البيانات وكذلك للبيانات المعقدة ولتقييم شرعية الأنموذج المعلمي المفترض وبالعكس ، وقد تم تطوير هذه الاساليب الأخيرة لتتناسب دراسة الانحدار المتعدد والتي سيفرزعنها مشكلة جديدة تدعى بمشكلة البُعدية او الأبعاد (curse of dimensionality) بسبب تزاحم البيانات في الفضاءات الممثلة لها مع محدودية المتغيرات التي تمثلها ، عندها ستفشل الطرائق التقليدية في ايجاد تقدير جيد للمعلمات، لذلك يتوجب التعامل مع هذه المشكلة بشكل مباشر.

وبسبب ذلك، يتوجب التخلص من هذه المشكلة (أو على الأقل التخفيف منها) عن طريق ايجاد حلول مناسبة لها من خلال استعمال بعض الاساليب التي تؤدي الى الحصول على نتائج دقيقة وغالبا ما يتم استعمال الاساليب التي تعمل على دمج (أو ضغط) المتغيرات دون خسارة اية معلومات من البيانات وهذا ما يُدعى باختزال الابعاد (Dimensionality Reduction: RD). ان الهدف المشترك لجميع هذه الاساليب المستعملة هو اختزال ابعاد البيانات (أو ضغطها) ، في حين تتم المحافظة على محتوى المعلومات الكامنة فيها مهما كانت طرائق تحليلها واستخلاص النتائج منها.

2. هدف البحث :

يهدف البحث الى استعمال اسلوب الانحدار الشرائحي المعكوس (sliced inverse regression: SIR) لأختزال الأبعاد والتخلص من مشكله البعدية وتحويل الأنموذج المستعمل من أنموذج متعدد الى أنموذج ابسط وتحسين اداء العمل الاستدلالي والتخلص من مشاكل الانحدار المختلفة ومقارنته مع الطريقة الكلاسيكية (PCA) بأستعمال المحاكاة بالاعتماد على معيار جذر متوسط مربعات الخطأ للحصول على افضل النتائج .

3. الجانب النظري :

(1-3) أنموذج الانحدار المعلمي (Parametric Regression Model (PRM)):

ان أنموذج (PRM) يدرس العلاقة بين متغيرين او عدة متغيرات X's مستقلة (Independent Variable) أو توضيحية (Explanatory Variable) ، والتى يُعتقد أنها تؤثر فى المتغير المعتمد (Y) (dependent Variable) أو متغير الاستجابة (Response Variable) ، وان هذا الأنموذج يتطلب توفر شروط معينة في البيانات مثل: معرفة توزيع المشاهدات وتوزيع الأخطاء ، أو المعرفة المسبقة بالصيغة الدالّية التي تتحكم بالعلاقة السببية بين المتغيرات، ويتم تمثيل الأنموذج بالصيغة الآتية:

$$f(X_i,\beta) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_z X_{iz} + \epsilon_i \qquad \ldots \qquad (3-1)$$

إذ ان:

$f(X_i,\beta)$: تمثل دالة خطية للمتغيرات المستقلة.

X_{n*z} : تمثل مصفوفة المتغيرات التوضيحية .

n : تمثل حجم المشاهدات.

Z : تمثل عدد معلمات المتغيرات التوضيحية (β) ، حيث ان ($i= 1, \ldots, z$).

β : تمثل معلمات غير معروفة والتي يتم تقديرها بأحدى طرائق التقدير المعلمي.

ϵ_i : يمثل الخطاء العشوائي وبافتراض $E(\epsilon) = 0 , var(\epsilon) = \sigma^2 I$.

(2-3)أنموذج الانحدار اللامعلمي (Nonparametric Regression Model(NPRM)):

ان أنموذج (NPRM) يتمتع هذا الأنموذج بمرونة عالية، اذ لا يتطلب توفر الشروط كما في أنموذج الانحدار المعلمي مما جعل أنموذج الانحدار اللامعلمي مرغوبا لدى اكثر الباحثين وذلك لان البيانات الحقيقة لا تكون ذات مواصفات مثالية بشكل دائم ، حيث تم تمثيل الأنموذج بالصيغة الاتية :

$$y_i = m(\chi_i) + \epsilon_i \quad \ldots \quad (3-2)$$

حيث ان :

y_i : يمثل متغير الاستجابة.

$m(\chi_i)$: تمثل دالة تمهيد مناسبة، وهي لا تحتوي على معلمات ويتم تقديرها باحدى الطرائق اللامعلمية.

ϵ_i : يمثل الخطاء العشوائي ، بمتوسط $E(\epsilon) = 0$ ، وتباين $var(\epsilon) = \sigma^2$.

(3-3) اختزال الابعاد (dimension reduction)[7] :

في مجال الانحدار المعلمي او اللامعلمي عندما يكون عدد المشاهدات او البيانات كبير جدا فاننا بحاجة لاستعمال أساليب معينة للتعامل معها، ولتسهيل التعامل مع البيانات ذات الابعاد العالية نعمل على اختزال الابعاد، حيث يوجد أسلوبين للاختزال، الأسلوب الاول: وهو يعمل على اهمال المتغيرات المستقلة التي لا تؤثر على متغير الاستجابة بالتالي حذفها من الأنموذج وجعل معاملها مساوٍ للصفر (β=0)، ويتم بذلك تقليل عدد المتغيرات، وهذا ليس ممكن عندما تكون العلاقات بين المتغيرات غير معروفة أصلاً. أما الأسلوب الثاني: فهو لا يحذف أي من المتغيرات بشكل قاطع بل يحاول الإبقاء على أكبر قدر ممكن من المعلومات لغرض الاستفادة منها فيعمد الى استعمالها ضمن مركبة خطية على البيانات الاصلية، بمعنى آخر، سيتم استبدال البيانات الأصلية بتراكيب خطية من المتغيرات الأصلية، وان الطريقة التي يتم بها اختيار هذه التراكيب الخطية تستند الى طريقة تخفيض او تقليص الابعاد المستعملة.

(4-3) طريقة الانحدار الشرائحي المعكوس[4,5,6,9] (Sliced Invers Regression (SIR)):

ان الاساس في طريقة الانحدار الشرائحي المعكوس (SIR) هو عكس العلاقة السببية في تحليل الانحدار التقليدي (الكلاسيكي) لدراسة العلاقة بين متغير الاستجابة (response variable) (Y) والمتغيرات التوضيحية(explanatory variable) (X) والمتمثلة بـ E(Y/X) والتي تمثل انحدار المتغير الخارج (output) (Y) مقابل العديد من المتغيرات الداخلة (X) الى انحدار(E(X/Y ، اي جعل المتغير (Y) يمثل المتغير المستقل والمتغير (X) يمثل المتغير المعتمد.

وتعمل هذه الطريقة على تجزئة (decomposition) الأنموذج الى شرائح متعددة بحسب قيم (Y) ثم القيام بعمليات احصائية مختلفة لكل شريحة، اذ يعمل على دمج (composition) معلومات كل الشرائح والحصول على الجذور الكامنة والتي يتم اختيار الأكبر من بينها لتمثل الاتجاهات الفعالة (e.d.r) لـ (SIR) ونقصد هنا بالـ (e.d.r) هو المتجه الناتج من عملية الاختزال الذي يمثل الشكل الجديد للبيانات ، ويكون تشتتها متناسب بالنسبة لتشتت متغيرات (X) الاصلية، و لايمكن ان يكون منحنى الانحدار العكسي مستقيماً، وان هذا الانحناء يؤدي دوراً مهماً في ايجاد اتجاهات (e.d.r). اما في حالة كون المنحنى الانحدار العكسي مستقيماً، فعندها قد لا نتمكن من ايجاد اكثر من اتجاه واحد.

تستند طريقة الانحدار الشرائحي المعكوس (SIR) الى إيجاد تقديرات اتجاهات فعالة تكون بمثابة معلمات (B_k's) تحول فيه البيانات إلى الشكل المختزل و تحل محل البيانات الأصلية وذلك لسهولة التعامل معها، و هي بدورها تعتبر معالجة لمشكلة البعدية، أن الأنموذج الذي تعتمد عليه طريقة الانحدار الشرائحي المعكوس (SIR) يكون أشبه بأنموذج الانحدار اللامعلمي والشبه المعلمي حيث يمثل بالصيغة الاتية:

$$y = f(\beta'_1 X_1, \beta'_2 X_2, \cdots, \beta'_k X_p, \epsilon_i) \quad \ldots \quad (3-3)$$

حيث أن:
f : تمثل دالة غير معلومة .
B_k: تمثل معلمات الاختزال ، وان عدد المعلمات الشكل المختزل (K) حيث ان K=1 , …, m .
X: مصفوفة المعلومات من درجة (nxp).
n : يمثل حجم الشاهدات وان (i= 1, … , n) .
P :عدد المتغيرات التوضيحية .
ϵ_i: يمثل متجه الخطأ العشوائي والمستقل عن (X) .

تقوم طريقة الانحدار الشرائحي المعكوس على مجموعة تحويلات وطرائق حسابية تتلخص في الخوارزمية الاساسية (SIR) :

(3-4-1) الخوارزمية الاساسية (SIR):

تتلخص هذه الخوارزمية بالخطوات الاتية:

1-ادخال المتغيرات التوضيحية (X) والمتغير المعتمد (او متغير الاستجابة) (Y) حيث أن كل صف في الجدول يمثل مشاهدة .

2-ترتيب المتغيرات التوضيحية (X) والمتغير المعتمد (او متغير الاستجابة) (Y) ترتيب تصاعدياً تبعاً للمتغير المعتمد (Y) ، ثم بعد ذلك يتم حساب الوسط الحسابي العام ومصفوفة التباين والتباين المشترك العامة وكما يأتي :

$$\bar{X}_i = \frac{\sum_{j=1}^{N} X_{ij}}{N} \quad \ldots \quad (3-4)$$

$$\hat{\Sigma}_x = \frac{\sum_i^n (X_i - \bar{X})(X_i - \bar{X})'}{N} \quad \ldots \quad (3-5)$$

3-يتم تقسيم المصفوفة افقياً الى H ثم يتم حساب الوسط الحسابي لكل شريحة وكما يأتي:

$$\bar{X}_h = \frac{\sum_{i=1}^{nh} X_{hi}}{n_h} \quad \ldots \quad (3\text{-}6)$$

حيث ان :
$\bar{X}_h$: يمثل متوسط الشريحة (h) .
n_h : يمثل حجم الشريحة (h).

4-يتم حساب مصفوفة التباين والتباين والتباين المشترك للمصفوفة متوسطات المتغيرات التوضيحية $\Sigma_{\bar{x}}$ وأن حساب المصفوفة يكون وفق الصيغة الآتية:

$$\hat{\Sigma}_{\bar{X}_h} = \frac{\sum_{h=1}^{H} N_h (\bar{x}_h - \bar{X})(\bar{x}_h - \bar{X})'}{N} \quad \ldots \quad (3\text{-}7)$$

حيث أن :

$$\bar{X} = \sum_{i=1}^{N} \frac{X_i}{N}$$

5- يتم حساب القيم والمتجهات الكامنة للمصفوفة ($\Sigma_{\bar{x}}$) والتي سنسميها (Σ_η) حيث يتم اختيار العمود من المصفوفة الناتجة من تكوّن المتجهات الكامنة الذي سيدعى (η_k) وأن (k) يمثل عدد أعمدة مصفوفة المتجهات الكامنة المقابل للقيمة الكامنة التي تكون اكبر او تساوي (0.5).

6- يتم إيجاد مصفوفة $\Sigma_x^{-1/2}$ وذلك من خلال الصيغة الاتية:

$$\Sigma_x^{-1/2} = D\Sigma_\eta^{-1/2}\, D' \quad \ldots \quad (3\text{-}8)$$

حيث أن:
D : يمثل مصفوفة المتجه الذاتي.
Σ_η : يمثل مصفوفة القيم الكامنة القطرية.

7- يتم حساب $\widehat{\beta}_K$ من خلال :

$$\widehat{\beta}_k=\Sigma_x^{-1/2} * \eta_k \qquad \dots \qquad (3\text{-}9)$$

حيث أن :

$\widehat{\beta}_K$: يمثل متجه (SIR) الجديدة .

8- حساب $E(X|y)$ بواسطة طريقة (OLS)(Ordinary Least Square) .

(2-4-3) الخوارزمية (WSIR)المقترحة (Proposed weighted Slices Inverse Regression):

تتلخص هذه الخوارزمية بالخطوات الاتية :

1- ادخال المتغيرات التوضيحية (X) والمتغير المعتمد (Y) اذ أن كل صف في الجدول يمثل مشاهدة .

2- ترتيب المتغيرات التوضيحية X والمتغير المعتمد Y تصاعديا تبعاً للمتغير المعتمد (Y) ثم بعد ذلك يتم حساب الوسط الحسابي العام ومصفوفة التباين والتباين المشترك العامة وكما يأتي:

$$\bar{X} = \frac{\sum_{j=1}^{N} X_{ij}}{N}$$

$$\widehat{\Sigma}_x = \frac{\sum_i^n (X_i - \bar{X})(X_i - \bar{X})'}{N}$$

3- يتم تقسيم المصفوفة افقيا الى H من الشرائح وثم يتم حساب الوسط الحسابي لكل شريحة كما في معادلة $(2-6)$.

4- يتم تحويل المصفوفة المتغيرات التوضيحية X الى الصيغة القياسية وفق الصيغة الاتية :

$$\widetilde{X}_i = \widehat{\Sigma}_X^{-\frac{1}{2}}(X_i - \bar{X}) \qquad \dots \qquad (3-10)$$

حيث ان :

$\widetilde{X}_i$: يمثل المصفوفة القياسية .

5- يتم حساب $\widehat{Wh}$ والذي يمثل وزن كل شريحة من خلال الصيغة الاتية :

$$\widehat{W}_h=\frac{\widehat{m}_h}{M} \qquad \dots \qquad (3\text{ - }11)$$

حيث ان :

$\widehat{m}_h$: يمثل الوسط الحسابي لكل شريحة.

M : يمثل الوسط الحسابي العام للشرائح .

6- يتم حساب متوسط كل سلايس وفق الصيغة الاتية :

$$\widehat{m}_h = \frac{\sum_{i=1}^{n_h} \widetilde{x}_{ni}}{n\widehat{p}_h} \qquad \dots \qquad (3\text{ - }12)$$

حيث ان :

$$\widehat{p}_h= \frac{nh}{N} \qquad \dots \qquad (3\text{ - }13)$$

7- يتم حساب مصفوفة التباين والتباين المشترك الموزونة وفق الصيغة الاتية :

$$\widehat{V} = \sum_{h=1}^{H} \widehat{Wh}\widehat{m}_h\widehat{m}_h' \qquad \dots \qquad (3\text{ - }14)$$

8- يتم حساب القيم الكامنة والمتجهات الكامنة لمصفوفة ($\widehat{V}$) والتي يرمز لها (Σ_η) حيث يتم اختيار العمود من مصفوفة المتجهات الكامنة والذي سنسميه (η_k) وأن (k) يمثل عدد أعمدة مصفوفة المتجهات الكامنة المقابل للقيمة الكامنة التي تكون أكبر أو تساوي (0.5).

9- يتم إيجاد مصفوفة $\Sigma_x^{-1/2}$ وذلك من خلال الصيغة الاتية :

$$\Sigma_x^{-1/2}= D\Sigma_\eta^{-1/2} D' \qquad \dots \qquad (3\text{ - }15)$$

حيث أن :

D : يمثل مصفوفة المتجه الذاتي.

Σ_η : يمثل مصفوفة القيم الكامنة القطرية.

10- يتم حساب $\widehat{\beta_K}$ من خلال :

$$\widehat{\beta}_k=\Sigma_x^{-1/2}*\eta_k \quad \dots \quad (3-16)$$

حيث أن :

$\widehat{\beta_K}$: يمثل متجه (SIR) الجديدة .

11- حساب $E(X|y)$ بواسطة طريقة (OLS)(Ordinary Least Square) .

(5-3) تحليل المركبات الرئيسة[1,2,8] (Principal Components Analysis(PCA)

يستند هذا التحليل على مبدأ تقليص عدد من المتغيرات التوضيحية (X's) الى عدد أقل من المركبات الخطية والتي تكون مستقله عن بعضها بعضا والتي ستدعى "المركبات الرئيسة" وان اهم ميزة في هذه الطريقة هي عدم خسارة معلومات من البيانات . وستكون المركبات الرئيسة عبارة عن توليفات خطية من هذه (X's) وللحصول على نتائج دقيقة وذات اهمية يجب ان تكون هذا المركبات الخطية متجانسة من حيث القياس ، ومن هذه ال (X) سيكون علينا إستعمال القيم المعيارية لتحويل مصفوفة (X) الى ذات الدرجة (Z) الى مصفوفة القيم المعيارية كما يأتي :

$$Z_{ij}=\frac{X_{ij}-\bar{X}_j}{\sqrt{S_{ij}}} \quad \dots \quad (3-17)$$

حيث ان :

$$S_{ij}=\frac{\sum_i^n(X_i-\bar{X})2}{N} \quad \dots \quad (3-18)$$

وتتلخص خوارزمية المركبات الرئيسية بالاتي :

1- ادخال المتغيرات التوضيحية (X) والمتغير المعتمد (او متغير الاستجابة) (Y) اذ أن كل صف في الجدول يمثل مشاهدة .

2- - يتم تحويل المصفوفة المتغيرات التوضيحية (X) الى الصيغة القياسية (Z).

3- يتم حساب القيم الكامنة والمتجهات الكامنة لمصفوفة($x'x$) والتي سنسميها($\wedge$) حيث يتم اختيار العمود من مصفوفة المتجهات الكامنة والذي يرمز له (V) وأن (k) يمثل عدد أعمدة مصفوفة المتجهات الكامنة المقابل للقيمة الكامنة التي تكون أكبر أو تساوي (0.5) والتي يرمز له (r).

حيث ان :

r : يمثل عدد القيم الكامنة التي تكون اكبر من (0.5).

V : هي مصفوفة متعامدة من درجة (k*k) اي :

$$V'V=VV'=I \quad \dots \quad (3-19)$$

وان :

$$V'x'xV=\wedge \quad \dots \quad (3-20)$$

4- تقسيم مصفوفة (V) الى $[V_r : V_{k-r}]$ و كذلك مصفوفة ($\wedge$) الى $\begin{bmatrix} \wedge r & 0 \\ 0 & \wedge k-r \end{bmatrix}$

5- حساب $\widehat{\beta}_{P.C}$ من خلال الصيغة التالية :

$$\widehat{\beta}_{P.C}=V_r*\wedge r^{-1}*V'_r*x'y \quad \dots \quad (3-21)$$

6 – حساب $\hat{y}$ من خلال الصيغة الاتية:

$$\hat{y}=X*\widehat{\beta}_{P.C} \quad \dots \quad (3-23)$$

(3-6) معيار المقارنة **Comparison Criterion**

معيار (RMSE) ويمثل جذر متوسط مربعات الخطأ (Root Of Mean Squared Error) ويحسب من خلال الصيغة

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}(\,yi - \widehat{yi}\,)^2/\,n} \qquad \ldots \qquad (3\text{-}24)$$

حيث تتم المفاضلة وفق هذا المعيار بأختيار اقل (RMSE) .

4- الجانب التجريبي :

تم استعمال المحاكاة في هذا البحث لتوليد البيانات ومن ثم تطبيق طرائق الاختزال عليها عليها واختيار افضل طريقة من خلال معيار(RMSE)، إذ تعد المحاكاة (Simulation) تقليد للبيانات الاصلية تحت البحث اذ تقوم بتوظيف أو تكوين نماذج تظهر فيها عدد كبير من الحالات الافتراضية لتكون نتائج التحليل أكثر شمولية وتعميماً.

(4-1) خطوات اجراء المحاكاة :

(4-1-1) توليد المتغيرات المستقلة :

تم توليد مصفوفة المعلومات (X) بواقع (5) متغيرات توضيحية وهي X_1(X_2,X_3 ,X_4 ,X_5) بحسب التوزيع الطبيعي بوسط حسابي يساوي صفر و تباين على التوالي (0.1 ، 0.5 ، 1) وباحجام مشاهدات على التوالي (100 ، 200، 400) بأستخدام برنامج (MATLAB).

(4-1-2) توليد المتغير المعتمد [3] :

تم استعمال الأنموذج التالي في توليد المتغير المعتمد (Y) وبالاعتماد على متغيرات مصفوفة المعلومات اذ تم اختيار اول متغيرين من مصفوفة المعلومات مع اضافة تباين المعلومات (σ) مع تشويش (δ) اذ كان :

$$Y = \frac{X_1}{0.5+(X_2+1.5)^2} + (1+X_2)^2 + \sigma * \delta \qquad \ldots \qquad (4\text{-}25)$$

حيث ان :

σ: يمثل تباين البيانات المولدة .

δ: تمثل تشويش للمتغير المعتمد (متغير الاستجابة) (Y).

(4-1-3) توليد الاخطاء العشوائية: Type equation here.

يتم توليد الاخطاء العشوائية المتمثلة بالتشويش (δ) على المتغير المعتمد (Y) على وفق التوزيع الطبيعي القياسي وفق الصيغة التالية :

$\delta \sim N(0,1)$

(4-2) تطبيق الطرائق :

تم تطبيق ثلاث طرائق في اختزال الابعاد وهي (SIR) الاساسية و طريقة (WSIR) المقترحة وطريقة (P.C) الكلاسيكية اذ تم تقسيم عدد الشرائح (H) بالنسبة الى (SIR) و (WSIR) على التوالي الى (5 ، 7 ، 10) شريحة بغية الحصول على افضل النتائج اذ تعد (H) بمثابة معلمة ضبط (tuning parameter).

جدول رقم (1)
يبين مقارنة طرائق الاختزال حسب توليد عدد المشاهدات وبتباينات مختلفة

RMSE											
	methods		$\sigma^2 = 0.1$			$\sigma^2 = 0.5$			$\sigma^2 = 1$		
		n	H=5	H=7	H=10	H=5	H=7	H=10	H=5	H=7	H=10
1	basic Sir	100	0.9822	0.9876	0.9667	0.7916	0.9529	0.8734	0.7610	0.6783	0.7849
		200	0.9693	0.9670	0.2603	0.9011	0.8546	0.8601	0.8630	0.8660	0.8848
		400	0.9806	0.9818	0.9883	0.8234	0.8314	0.8577	0.8684	0.8638	0.8486
2	weight standard Sir	100	1.0548	1.0470	1.0344	0.8617	0.8223	0.7696	0.6602	0.7294	0.9180
		200	0.9489	0.9847	0.9598	0.8183	0.8226	0.7802	0.8150	0.8421	0.8031
		400	0.9493	0.9689	0.9514	0.7178	0.7525	0.8876	0.7761	0.7938	0.8337
3	PC	100	1.0051			1.9193			3.1241		
		200	1.0231			2.0275			4.1535		
		400	1.0525			1.8640			3.9234		

تشير نتائج جدول رقم (1) الى ان نتائج طريقة (Basic SIR) عند ($\sigma^2 = 0.1$) وعند حجم مشاهدات (100) هو اقل (RMSE) عند (H=10) وعند (200) مشاهدة ايضا عند (H=10) وعند حجم مشاهدات (400) عند (H=5) ، وفي حالة ($\sigma^2 = 0.5$) وعند حجم مشاهدات (100) هو اقل (RMSE) عند (H=5) وعند (200) مشاهدة ان اقل (RMSE) ايضا عند (H=7) وعند حجم مشاهدات (400) ان اقل (RMSE) عند (H=5) ، وفي حالة ($\sigma^2 = 1$) وعند حجم مشاهدات (100) هو اقل (RMSE) عند (H=7) وعند (200) مشاهدة ان اقل (RMSE) ايضا عند (H=5) وعند حجم مشاهدات (400) ان اقل (RMSE) عند (H=10).

تشير نتائج جدول رقم (1) الى ان نتائج طريقة (weight standard Sir)(SIR) عند ($\sigma^2 = 0.1$) وعند حجم مشاهدات (100) هو اقل (RMSE)عند (H=10) وعند (200) مشاهدة ان اقل(RMSE) عند (H=5) وعند حجم مشاهدات (400) عند (H=5) ، وفي حالة ($\sigma^2 = 0.5$) وعند حجم مشاهدات (100) هو اقل(RMSE) عند (H=10) وعند (200) مشاهدة ان اقل (RMSE) ايضا عند (H=10) وعند حجم مشاهدات (400) ان اقل (RMSE) عند (H=5)، وفي حالة ($\sigma^2 = 1$) وعند حجم مشاهدات (100) هو اقل(RMSE) عند (H=5) وعند (200) مشاهدة ان اقل (RMSE) ايضا عند (H=10) وعند حجم مشاهدات (400) ان اقل (RMSE) عند (H=5).

تشير نتائج جدول رقم (1) الى ان نتائج طريقة (Principal Components) وعند حجم مشاهدات (100) هو اقل (RMSE)عند ($\sigma^2 = 0.1$) وعند حجم مشاهدات (200) ان اقل (RMSE) ($\sigma^2 = 0.1$) وعند حجم مشاهدات (400) ان اقل (RMSE) عند ($\sigma^2 = 0.1$) .

جدول رقم (2)
يبين ترتيب طرائق الاختزال

	methods	n	$\sigma^2 = 0.1$			$\sigma^2 = 0.5$			$\sigma^2 = 1$		
			RMSE								
		n	100	200	400	100	200	400	100	200	400
1	basic Sir	2	0.9667 اولا	0.2603 اولا	0.9806 ثانيا	0.7916 ثانيا	0.8546 ثانيا	0.8234 ثانيا	0.6783 ثانيا	0.8630 ثانيا	0.8486 ثانيا
2	weight standard Sir	1	1.0344 ثالثا	0.9489 ثانيا	0.9493 اولا	0.7696 اولا	0.7802 اولا	0.7178 اولا	0.6602 اولا	0.8031 اولا	0.7761 اولا
3	PC	3	1.0051 ثانيا	1.0231 ثالثا	1.0525 ثالثا	1.9193 ثالثا	2.0275 ثالثا	1.8640 ثالثا	3.1241 ثالثا	4.1535 ثالثا	3.9234 ثالثا

تشير نتائج جدول رقم (2) الى افضلية طريقة (weight standard SIR)(WSIR) عند التباينات (0.5 و 1) وفي جميع احجام العينات (100 ، 200 ، 400) وذلك لآنها تمتلك اقل (RMSE) يليها طريقة (basic SIR) وقد جائت اخيرا طريقة (Principal Components) اذ انها كانت تملك اعلى (RMSE).

5- الاستنتاجات والتوصيات :

(5-1) الاستنتاجات :

في ضوء تحليل تجارب المحاكاة ، تم التوصل للاستنتاجات الآتية:

1. عند توليد البيانات بثلاث تباينات مختلفة و لجميع احجام المشاهدات ولخمسة متغيرات توضيحية، لوحظ أن طريقة (weight standard Sir) (WSIR) حققت نتائج كفوءة بدرجة عالية إذا أعطى أقل (RMSE)، ماعدا عند تباين (0.1) عند حجم مشاهدات (100 ، 200) يحث يلاحظ ان الاختلافات قليلة بين المتغيرات وهذا يعطي مؤشر الى انه في حالة عدم اعطاء اقل (RMSE) من هذه الطريقة على ان البيانات ربما لا تعاني من مشاكل .

2. نلاحظ عند تغيير قيمة ضبط المعلمة (H) (tuning parameter) له تأثير كبير في جعل (RMSE) اقل ما يمكن في طريقتي (Basic SIR) وطريقة (RMSE) حيث نلاحظ تفاوت النتائج عند تغيير قيمة (H) ويلاحظ ايضا ان قيمة اقل RMSE كانت غالبا ماتكون عند (H= 5) و (H=10) اي عندما تكون قيمة (H) مساوية لعدد المتغيرات التوضيحية او ضعف عددها .

3. نلاحظ نجاح طريقة (Principal Components) عندما تكون البيانات قريبة من الصفر قيمة التباين (0.1) عند حجم مشاهدات(100) ولكن نلاحظ فشلها عندما يكبر حجم المشاهدات وايضا عندما ترتفع قيم التباينات وهذا واضح من تحليل النتائج .

4. نلاحظ ان طريقة (Basic SIR) كانت نتائجها مقبولة الى حد ما مما يجعلها تتفوق على الطريقة الكلاسيكية في اختزال الابعاد .

5. نلاحظ من خلال النتائج اعلاه الى امكانية استعمال طريقة (Basic SIR) و (weight standard Sir) (WSIR) كطرائق كشف للبيانات التي تحتوي على مشاكل اذ كلما اعطت نتائج غير مرضية دل ذلك على خلو البيانات من المشاكل والعكس صحيح .

(2-5) **التوصيات** :

بناءً على ما تم التوصل إليه من استنتاجات، فيما يأتي بعض التوصيات :

1. نوصي باستعمال طريقة (weight standard Sir) في تقدير النماذج عندما يكون فيها عدد المتغيرات التوضيحية كبير وذلك لان زيادة عدد المتغيرات التوضيحية يؤدي غالبا الى حدوث مشكلة البُعدية (Curse of Dimensionality) ومن ثم حدوث بقية المشكلات كعدم التجانس والتعدد الخطي والارتباط الذاتي .

2. نوصي بمحاولة تقدير قيمة ضبط المعلمة (H) (tuning parameter) وذلك لأهميتها في حساب (basic Sir) و (weight standard Sir) وذلك للحصول على نتائج اكثر دقة .

6- **المصادر** :

المصادر العربية :

1. القيسي ، عمر عبدالمحسن – المهنا ، فراس احمد " حول تقليص تقدير المركبات الرئيسة مع التطبيق "،المجلة العراقية للعلوم الاحصائية ،العدد (14)،الصفحة (371-384)،(2010).

2. النعيمي ، اسوان محمد "معالجة البيانات التامة وتقديرها بطريقة انحدار المركبات الرئيسية"،كلية علوم الحاسبات والرياضيات، الصفحة (6-7) ،جامعة الموصل ،(2009).

المصادر الاجنبية :

3. Kenji Fukumizu ,Francis R.Bach and Michael I.Jordan; (2009) "Kernel Dimension Reduction in Regression" Annals of statistics ,Vol.(37) ,No.(4) , PP.(1871-1905).

4. Ker-chau li. ; (1991) " sliced inverse regression for dimension regression" JASA, Vol. 86 , No.414 , PP.316-327.

5. Ni, Liqiang; Cook, R. Dennis; and Tsai, Chih-Ling; (2005); "Note on shrinkage sliced inverse regression"; Biometrika, Vol. (92), No. (1), PP. (242-247).

6. Shevlyakova, Maya and Morgenthaler, Stephan; (2014); "Sliced inverse regression for survival data"; Stat Papers at Springer-Verlag Berlin Heidelberg; Vol. (55), PP.(209-220).

7. Swatikaur, S. M. Ghosh ; (2016) "A survey on dimension reduction techniques for classification of multidimensional data", IJSTE ,Vol.2 , Issue.12 , PP.(31-37).

8. Tomoyuki Akita ;(2011) "Estimation on inverse regression using principle components of covariance matrix of sliced data", Hiloshima Mathematical journal , Vol.41, No.1, PP.(41-53) .

9. ZHU, Li-ping; and YU, Zhou; (2007); "On spline approximation of sliced inverse regression"; Sciences in China Series A: Mathematics; Vol. (50), No. (9), PP. (1289-1302).

Comparison of Slice inverse regression with the principal components in reducing high-dimensions data by using simulation

Abstract

This research aims to study the methods of reduction of dimensions that overcome the problem curse of dimensionality when traditional methods fail to provide a good estimation of the parameters So this problem must be dealt with directly . Two methods were used to solve the problem of high dimensional data, The first method is the non-classical method Slice inverse regression (SIR) method and the proposed weight standard Sir (WSIR) method and principal components (PCA) which is the general method used in reducing dimensions, (SIR) and (PCA) is based on the work of linear combinations of a subset of the original explanatory variables, which may suffer from the problem of heterogeneity and the problem of linear multiplicity between most explanatory variables. These new combinations of linear compounds resulting from the two methods will reduce the number of explanatory variables to reach a new dimension one or more which called the effective dimension. The mean root of the error squares will be used to compare the two methods to show the preference of methods and a simulation study was conducted to compare the methods used. Simulation results showed that the proposed weight standard Sir method is the best.

Key words:- dimensions reduction , Slice inverse regression, principal components.