# استعمال انحدار الاسقاطات المتلاحقة و الشبكات العصبية في تجاوز مشكلة البعدية

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

تاريخ التقديم:2017/6/2
تاريخ القبول:2017/10/25

## المستخلص

يهدف هذا البحث الى تجاوز مشكلة البعدية من خلال طرائق الانحدار اللامعلمي والتي تعمل على تقليل جذر متوسط الخطأ التربيعي (RMSE) ، أذ تم استعمال طريقة انحدار الاسقاطات المتلاحقة (PPR) ,والتي تعتبر احدى طرائق اختزال الابعاد التي تعمل على تجاوز مشكلة البعدية ( curse of dimensionality) ، وان طريقة (PPR) من التقنيات الاحصائية التي تهتم بأيجاد الاسقاطات الاكثر أهمية في البيانات المتعددة الابعاد ، ومع ايجاد كل اسقاط تتقلص البيانات بواسطة المركبات الخطية على طول الاسقاط ويتم تكرار العملية لايجاد اسقاطات جيدة لحين الحصول على افضل الاسقاطات والفكرة الاساسية لانحدار الاسقاطات المتلاحقة (PPR) هو نمذجة الانحدار المتعدد كمجموع للدوال غير الخطية للتراكيب الخطية للمتغيرات .

ومن اجل التخلص من مشكلة البعدية تم استعمال اسلوبين الاسلوب الاول طريقة انحدار الاسقاطات المتلاحقة (PPR ) المقترحة والاسلوب الثاني طريقة الشبكات العصبية ( NN ) المتمثلة ( بالانبعاث الخلفي للخطأ ) وهي من الطرائق المستخدمة في اختزال الابعاد ، وقد تم اجراء دراسة محاكاة للمقارنة بين الطرائق المستخدمة وتم التوصل من خلال تجارب المحاكاة الى استنتاجات بينت ان الطريقة (NN) في هذا البحث اعطت نتائج افضل مقارنة بطريقة ( PPR ) اعتمادا على معيار جذر متوسط مربعات الخطأ (RMSE).

**المصطلحات الرئيسة للبحث/** مشكلة البعدية ، انحدار الاسقاطات المتلاحقة ، الشبكات العصبية.

مجلة العلوم الاقتصادية والإدارية
العدد 104 المجلد 24
الصفحات 344ـ 353

* البحث مستل من أطروحة دكتوراه .

## 1. المقدمة :

نظراً للتطور الحاصل في العلوم التكنولوجيا والمعلوماتية في عصرنا الحديث الذي كان له الاثر الكبير في تطور باقية العلوم الطبية والطبيعية والانسانية ، ولقد انعكس هذا التطور التكنولوجي المعلوماتي بشكل واضح على علم الاحصاء وذلك لارتباطه الوثيق به ، فعندما يراد دراسة و تحليل بيانات الظواهر الاقتصادية و الطبية و الزراعية و المالية وغيرها ، يجب ان تتوفر المعرفة المسبقة لهذه الظواهر ، بمعنى اخر ، معرفة نوع بياناتها والتي غالباً ماتكون كمية ، ويتطلب ذلك بناء اأنموذج رياضي مناسب يمثل العلاقات السببية ( دالة سببية او سلوكية ) بين عواملها افضل تمثيل وهي ماتدعى مرحلة الوصف (description) ، لأعتماد التحليل المناسب والذي يمكننا بعد ذلك من أتخاذ العديد من القرارات بشأن أهم الدلالات والخصائص (characteristics) المتعلقة بتلك الظواهر، وتدعى تلك الدلالات المذكورة سابقاً بالمعلمات (parameters) .

فعند ثبات الظروف الاخرى المحيطة بالظاهرة و بمعرفة العلاقة السببية التي تربط المتغيرات العشوائية فان التحليل الاحصائي المناسب لها يدعى بالتحليل المعلمي ، حيث توفر المعلمات ملخص وجيز تنوب عن المشاهدات لتسهل الاستدلال الاحصائي ، اما في حالة عدم ثبات تلك الظروف أو عدم معرفة العلاقة التي تربط بين المتغيرات العشوائية من حيث كونها علاقة سببية او سلوكية فان التحليل المناسب يدعى بالتحليل اللامعلمي ، حيث لا تتوفر معلومات عن خصائص (أو دلالات) الظاهرة.

ومن أهم نماذج التحليل الاحصائي مايدعى 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 ،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 لتناسب 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 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 انحدار خوارزمية الاسقاطات المتلاحقة (projection pursuit regression) للتخلص من مشكله البعدية اذ تعمل على تحليل البيانات بعد ان يتم تقسيمها الى مجاميع او عناقيد او سطوح ويكون تحليلها بشكل منفصل وكل على حده ، وان احدى اهم مميزات (PP) انه يلائم مجموعات البيانات المتناثرة في فضاء عالي الابعاد والتخلص من مشاكل الانحدار المختلفة ومقارنتها مع طريقة الشبكات العصبية (Neural Networks) (وهي تعد من الطرائق العددية في التقدير) في تجاوز مشكلة البعدية بأستعمال المحاكاة بالاعتماد على معيار جذر متوسط مربعات الخطأ للحصول على افضل النتائج.

## 3. الجانب النظري :

### (3-1) أنموذج الانحدار اللامعلمي[1] (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 الباحثين وذلك لان البيانات الحقيقة لا تكون ذات مواصفات مثالية بشكل دائم ، حيث تم تمثيل الأنموذج بالصيغة الاتية :

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

حيث ان :

$y_i$ : يمثل متغير الاستجابة.

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

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

### (3-2) مشكلة التعدد الخطي [1] (Multicollinearity):

تظهر مشكلة الارتباط الخطي المتعدد في حالة وجود ارتباط خطي قوي بين متغيرين أو أكثر من المتغيرات التوضيحية (Explanatory Variable) المفسّرة ( X's) لتغير المتغير التابع (Y) (dependent Variable) ، بمعنى اخر: عدم وجود استقلالية بين المتغيرات التوضيحية ( Explanatory Variable) ، مما يؤدي لصعوبة عزل تأثير كل منها عن المتغير التابع (dependent Variable) ، وذلك يفقد معنوية معاملات الانحدار المحتسبة بأسلوب المربعات الصغرى ، وغالباً ماتظهر في تحليل السلاسل الزمنية نتيجة لتغير المتغيرات الاقتصادية معاً وتأثرها بعوامل اقتصادية متعددة خلال الفترة الزمنية المعنية.

كما تظهر هذه المشكلة في بعض الاحيان عندما يصغر حجم العينة بحيث يكون مقارباً لعدد المتغيرات المستقلة (n=K) ، وللكشف عن وجود هذه المشكلة تستخدم عدة طرائق واختبارات.

### (3-3) مشكلة الابعاد (البعدية) [8,4] (Curse of Dimensionality) :

ان مشكلة البعدية (curse of dimensionality) ليست مشكلة بيانات ذات الابعاد العالية وانما هي مشكلة مشتركة بين البيانات ذات البعدية العالية والخوارزمية الواجب تطبيقها اذا يصعب تطبيق الخوارزمية لمثل هذه البيانات لذلك عندما نواجه هذه المشكلة يجب علينا ان نجد حلا اما عن طريق معالجة البيانات الاصلية او عن طريق تطوير الخوارزمية.

وتنشأ مشكلة البعدية (curse of dimensionality) في حالة وجود ارتباطات خطية بين البيانات ذات الابعاد العالية، فعندما يراد تقدير كثافة البيانات فان تكامل مربع الخطأ يكون كبيراً جداً حتى اذا كان حجم العينة كبير جدا.

### (3-4) طريق انحدار الاسقاطات المتلاحقة [5,7] (Projection Pursuit Regression) :

اذ تعد طريقة (PPR) من التقنيات الاحصائية التي تهتم بأيجاد الاسقاطات الاكثر اهمية في البيانات المتعددة الابعاد ، ومع ايجاد كل اسقاط تتقلص البيانات بواسطة المركبة على طول الاسقاط ويتم تكرار العملية لايجاد اسقاطات جيدة لحين الحصول على افضل الاسقاطات، واهم ميزه لها انها من الاساليب القليلة التي تستطيع تجاوز مشكلة البعدية (curse of dimensionality) الناجمة عن فضاء الابعاد العالية .

، وان الفكرة الاساسية لانحدار الاسقاطات المتلاحقة هو لنمذجة سطح الانحدار كمجموع للدوال غير الخطية للتراكيب الخطية للمتغيرات والتي يعبر عنها بالدالة الاتية :

$$Y = \sum_{j=1}^{M} \vartheta_j(\beta_j' X) \quad \dots \quad (2-3)$$

حيث ان :

$Y$: يمثل متغير الاستجابة .

$\beta_j$: تمثل المعلمات الاولية .

**M : عدد الاسقاطات في ألنموذج.**
**$\vartheta_j$: تمثل دالة تمهيد محددة .**
**X: تمثل مصفوفة المتغيرات التوضيحية بدرجة (n*p) .**
**اذ ان :**
**P :تمثل عدد المتغيرات التوضيحية .**
**n : تمثل حجم العينة .**
**ان العديد من طرائق التحليل الاعتيادية (الكلاسيكية) لمتعدد المتغيرات تكون حالات خاصة من طريقة الاسقاطات المتلاحقة و من الامثلة عليها تحليل المركبات الرئيسة (PC) والتحليل المميّز (Discriminant Analysis).**

**(1-4-3)خوارزمية انحدار الاسقاطات المتلاحقة The PPR Algorithm**

**و لغرض تقدير دالة الاستجابة $f(x)$ للبيانات $\{(x_{1,1},.., x_{1,M}, y_1)\}.,\{(x_{n,1},.,x_{n,M}, y_n)\}$وكما يأتي:**

**1.تحديد** $r_i^{[0]} = y_i$ .

**2.لكل j=1,…, maximize.**

**حيث ان :**

**j: تمثل عدد الاسقاطات .**

$$R^2_{[j]} = 1 - \frac{\sum_{i=1}^{n}\left(r_i^{[j-1]} - \hat{\vartheta}_j(\hat{\beta}_{[j]}^T x_i)\right)^2}{\sum_{i=1}^{n}\left(r_i^{[j-1]}\right)^2} \quad \text{...} \quad (2\text{-}4)$$

**حيث ان :**

$r_i^{[j-1]}$ **: يمثل البواقي .**

$\hat{\beta}_{[j]} \in \Re^p\left(\|\beta_{[j]}\| = 1\right)$ **varying over the parameters ومن خلال اجراء تغيير على المعلمات**

**ودالة الانحدار أحادية المتغير** $\hat{\vartheta}_{[j]}$ **.**

**3.جعل** $r_i^{[j]} = r_i^{[j-1]} - \hat{\vartheta}_{[j]}(\hat{\beta}_{[j]}^T x_i)$ **ثم نستمر بإعادة الخطوة الثانية حتى يتم تقليل قيمة** $R^2_{[j]}$ **. وأن قيمة** $R^2_{[i]}$ **الصغيرة تعني أن** $\hat{\vartheta}_{[j]}(\hat{\beta}_{[j]}^T x_i)$ **هي تقريباً الدالة الصفرية وبذلك لا نستطيع الحصول على اتجاه آخر يكون مفيداً.**

**هذه الخوارزمية تؤدي إلى تقدير دالة الاستجابة عن طريق:**

$$\hat{f}_M(x) = \sum_{j=1}^{M} \hat{\vartheta}_{[j]}(\hat{\beta}_{[j]}^T x) \quad \text{...} \quad (2\text{-}5)$$

### (5-3) الإنتشار الخلفى للخطأ في الشبكات العصبية (Back in neural network Propagation of error [2,9]

**بعد عملية التقدير الاولية في الشبكة العصبية وحساب الخطأ الاولي ، يتم إجراء مقارنة بين القيم المحسوبة والقيم المرغوبة (حساب الخطأ) من خلال الفرق بين قيم تلك المخرجات وذلك من خلال معادلة الخطأ الاتية:**

$$E = ( d_i - Y_i ) \quad \dots \quad (2\text{-}6)$$

**اذن ان :**

**$d_i$ : تمثل الاخراج المرغوب فيه اي متغير الاستجابة .**

**$Y_i$ : تمثل قيمة المخرج من الشبكة اي متغير الاستجابة المقدر .**

**E : يمثل الخطأ الاولي المحسوب .**

**وبعد ذلك يتم تصحيح الوزن وتعديله من خلال عملية التعلم التى تتم على الشبكة من خلال طريقة الانبعاث الخلفي ، وتتلخص هذه الطريقة بالخطوات الاتية :**

**<u>الخطوة (1) :</u> البدء بالأوزان و Offsets ، إعطاء الأوزان وعقدة Offsets قيم عشوائية قليلة.**

**<u>الخطوة (2) :</u> تهيئة الإدخال ووصف الإخراجات المرغوب فيها ، تحضير القيم المستمرة لمتجه الإدخال $X_0, X_1, X_2, ..., X_{n-1}$ ووصف (تخصيص) الإخراجات المرغوب فيها $d_0, d_1, d_2, ..., d_{m-1}$ .**

**<u>الخطوة (3) :</u> حساب الإخراجات الحقيقية ، تم استعمال الدالة اللاخطية السيبية (Sigmoid Function) لحساب الإخراجات $Y_0, Y_1, Y_2, ...Y_{M-1}$ .**

**<u>الخطوة (4) :</u> تعديل الأوزان.**

**تبدأ الخوارزمية بتوليف عقد الإخراج (أي أيجاد أفضل قيمة للوزن المحسوب) وتعمل بصورة خلفية الى طبقة مخفية وتعدل الأوزان بواسطة :**

$$W_{ij}(t+1) = W_{ij}(t) + \eta\ \delta_i\ \bar{X}_i \qquad (2-7)$$

**حيث أن :**

**$W_{ij}(t)$ : الوزن في العقدة المخفية i أو من الإدخال الى العقدة j في الزمن t.**

**$\delta_i$ : مصطلح الخطأ للعقدة j .**

**$\eta$ : تمثل معلمة التعلم وتكون محصوره بين ( 0 ، 1 ) .**

**فإذا كانت j تمثل عقدة إخراج فأن :**

$$\delta_i = Y_i(1-Y_i)(d_i - Y_i) \qquad (2-8)$$

**حيث ان :**

**$d_i$ : تمثل الإخراج المرغوب فيها لعقدة j ، وتمثل $Y_i$ الإخراج الحقيقي.**

**أما إذا كانت j تمثل عقدة مخفية داخلية فأن :**

$$\delta_i = \bar{X}_i(1-\bar{X}_i)\sum_k \delta_k W_{ik} \qquad (2-9)$$

**حيث ان:**

**k : تمثل كل العقد في الطبقات الموجودة فوق العقدة j .**

تُعدل قيمة العتبة للعقد الداخلية بطريقة متشابهة بواسطة الافتراض بأنها عبارة عن أوزان مترابطة حيث ترتبط من الإدخالات ذات القيم الثابتة ولكي يكون التقارب (Convergence) سريعاً يضاف مصطلح معامل التعجيل (Momentum) ويرمز له بالرمز ( $\alpha$ ) علماً أنه يساعد في تغيير الأوزان بصورة منتظمة كالآتي :

$$W_{ij}(t+1) = W_{ij}(t) + \eta\,\delta_i\,\overline{X}_i + \alpha(W_{ij}(t) - W_{ij}(t-1)) \quad (2-10)$$

حيث أن :

$$0 < \alpha < 1$$

بشكل عام لايوجد معيار عام لكيفية اختيار معامل التعلم ومعامل التعجيل وتعتمد القيم المثالية على المشكلة التي يتم معالجتها وتقليلها .

## 4- الجانب التجريبي :

تم استعمال المحاكاة في هذا البحث لتوليد البيانات ومن ثم تطبيق طرائق الاختزال 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 التوضيحية :**

تم توليد خمسة من المتغيرات التوضيحية لكي تلائم واقع المشكلة تحت الدراسة حيث ان مشكلة البعدية تحدث في حالة وجود خمسة متغيرات توضيحية أو اكثر وقد تم استعمال ثلاث أحجام من العينات وهي n = 100 , 200 , 400 بواقع تكرار (1000) حيث توزعت أحجام المتغيرات التوضيحية على هذه الطرائق ، أذ تم الاستعانة ببرنامج (MATLAB R2012A)* في توليد البيانات حيث يعتبر من البرامج ذا قدرة الكبيرة في المجال البرمجي والرياضي ، اذ تم توليد المتغيرات التي تعاني من مشكلة البعدية عن طريق توليد مصفوفة ارتباط اولية تتحكم بمقدار الارتباط بين هذه المتغيرات ، أذ تم تحديد قيمة الارتباط P=0.5 والتي تكون هي المسببة لمكشلة التعدد الخطي في البيانات ومن ثم تعاني مشكلة البعدية وتتلخص خطوات المحاكاة كمايأتي :

اولا:- التوليد الاولي للبيانات :

تم التوليد الاولي للمتغيرات التوضيحية توليدا طبيعيا كما يلي

$$x \sim N(0 , 0.5) \ldots\ldots\ldots (2-10)$$

ثانيا :- توليد المتغيرات المتعددة التوضيحية :

تم توليد المتغيرت المتعددة بمتجهة الوسط الحسابي M و بمصفوفة تباين SIGMA بالاعتماد على التوليد الاولي لمتغيرات في ( أولاً ) اذ يتم حساب متوسط كل متغير والذي سيمثل متجه الوسط الحسابي M وحساب مصفوفة التباين والتباين المشترك للتوليد الاولي والذي سيمثل مصفوفة SIGMA أي :

$$X \sim MN(M , SIGMA) \ldots\ldots\ldots (2-11)$$

وبذلك يتم توليد متغيرات متعددة تعاني مشكلة التعدد الخطي .

ثالثا :- الكشف عن مشكلة التعدد الخطي [3] :

تم الكشف عن مشكلة التعدد الخطي من خلال هذه المتغيرات عن طريق معامل تضخم التباين اذ يتم اختبار جميع المتغيرات التوضيحية فيما اذا كان معامل التضخم لديها عالي والذي يسمى V.I.F ( variance inflation factor ) ولمعرفة فيما اذا كانت هنالك مشكلة تعدد خطي في هذه المتغيرات بأرتباطات عالية وبقيمة معامل تضخم (VIF>4 ) لهذه المتغيرات ، دلّ ذلك على وجود تعدد خطي بين المتغيرات .

**(4-1-2) توليد المتغير المعتمد [4]:**

مما تقدم في الفقرة (4-1-1) نلاحظ ان المتغيرات التي تم توليدها هي متغيرات ذات ارتباط عالي فيما بينها وبغية توليد المتغير المعتمد ( Y ) تم استعمال الأنموذج التالي وبالاعتماد على المتغيرات التوضحية المولدة في الفقرة (4-1-1) اذ تم اختيار اول متغيرين من المتغيرات التوضيحية والتي تكون مترابطة مع بقية المتغيرات التوضيحية وذلك لانها تعاني من مشكلة التعدد الخطي مع اضافة تباين المعلومات ($\sigma$) مع تشويش ($\delta$) وهذا هو الاسلوب المتبع في عملية توليد المتغير المعتمد بشكل عام ، اذ كان النموذج كما يلي :

$$y = \frac{X_1}{0.5+(X_2+1.5)^2} + (1+X_2)^2 + \sigma * \delta \quad \dots \quad (2\text{-}12)$$

يمثل تباين البيانات المولدة. $\sigma$:

$\delta$: تمثل تشويش يتبع التوزيع الطبيعي القياسي للمتغير المعتمد ( او متغير الاستجابة )(Y).

ويقصد بالتشويش هنا على انه قيم تضاف الى النموذج بغية تشتيت المتغير المعتمد واعطائه العشوائية التامة

**(4-1-3) توليد الاخطاء العشوائية :**

يتم توليد الاخطاء العشوائية على وفق التوزيع الطبيعي القياسي وعلى وفق الصيغة الاتية :

$\delta i \sim N(0,1)$ ... (2-13)

**(4-2) تطبيق الطرائق :**

تم تطبيق طريقتين في اختزال الابعاد وهي (PPR) المقترحة و طريقة(NN) اذ تم تقسيم عدد العقد في طبقة الشبكة العصبية الى (H) وهي على التوالي الى ( 5 ، 7 ، 10 ) عقدة 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 | w=10 |
| 1 | ppr | 100 | 1.0115 | | | 2.0666 | | | 4.2423 | | |
| | | 200 | 1.0501 | | | 1.8791 | | | 3.8064 | | |
| | | 400 | 1.0652 | | | 2.0099 | | | 3.7477 | | |
| 2 | NN | 100 | 0.3438 | 0.2603 | 0.4523 | 1.6109 | 1.8925 | 1.5208 | 2.9540 | 3.1632 | 2.9544 |
| | | 200 | 0.2826 | 0.4130 | 0.3791 | 1.5287 | 1.7164 | 1.9154 | 3.1334 | 2.9223 | 3.1082 |
| | | 400 | 0.3009 | 0.3039 | 0.3031 | 1.4062 | 1.5848 | 1.4337 | 2.9489 | 2.8656 | 3.0582 |

نلاحظ من خلال نتائج جدول رقم ( 1 ) ان طريقة (PPR) تمتلك اقل (RMSE ) عند حجم مشاهدات (100) عند الارتباط ( $\sigma^2 = 0.1$ ) و كذلك تمتلك اقل ( RMSE ) عند حجم مشاهدات (200) عند الارتباط ( $\sigma^2 = 0.1$ ) و ايضا تمتلك اقل ( RMSE ) عند حجم مشاهدات (400 ) عند الارتباط ( $\sigma^2 = 0.1$ ).

كذلك نلاحظ من خلال نتائج جدول رقم (1) الى ان طريقـــة (NN) تمتـــلك اقل ( RMSE ) عند الارتباط ( $\sigma^2 = 0.1$) عند حجم المشاهدات (100) عندما ( H=7) وكذلك عند حجم المشاهدات (200) عند ( H=10 ) وعند حجم المشاهدات (400 ) عند ( H=5 )، وفي حالة الارتباط ( $\sigma^2 = 0.5$ ) تمتـــلك اقل (RMSE ) عند حجم مشاهدات (100) عند (H=10) و كذلك تمتـــلك اقـــل ( RMSE )عند حجم المشاهدات (200) عند ( H=5 ) وايضا عند حجم المشاهدات ( 400 ) تمتـــلك اقـــل (RMSE ) عند (H=5) ، وفي حالة الارتباط ( $\sigma^2 = 1$) تمتـــلك اقـــل (RMSE )عند حجم المشاهدات ( 100 ) عند (H=5) و كذلك عند حجم المشاهدات ( 200) فانها تمتـــلك اقـــل ( RMSE ) ايضا عند (H=7) وايضا تمتـــلك اقـــل ( RMSE ) عند حجم المشاهدات ( 400 ) عند ( H=7 ).

جدول رقم (2)

جدول يبين افضلية المقدرات بحسب معيار جذر متوسط مربعات الخطأ

| RMSE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | method | $\sigma^2 = 0.1$ | | | $\sigma^2 = 0.5$ | | | $\sigma^2 = 1$ | | |
| | | 100 | 200 | 400 | 100 | 200 | 400 | 100 | 200 | 400 |
| 1 | ppr | 1.0115 ثانيا | 1.0501 ثانيا | 1.0652 ثانيا | 2.0666 ثانيا | 1.8791 ثانيا | 2.0099 ثانيا | 4.2423 ثانيا | 3.8064 ثانيا | 3.7477 ثانيا |
| 2 | NN | 0.2603 اولا | 0.2826 اولا | 0.3009 اولا | 1.5208 اولا | 1.5287 اولا | 1.4062 اولا | 2.9540 اولا | 2.9223 اولا | 2.8656 اولا |

تشير نتائج جدول رقم ( 2 ) الى افضلية طريقة (NN) عند التباينات ( 0.1 و 0.5 و 1 ) ولجميع احجام العينـــات ( 100 ، 200 ، 400 ) وذلك لآنها تمتـــلك اقـــل (RMSE) .

## 5- الاستنتاجات والتوصيات :

### (5-1) الاستنتاجات :

في ضوء تحليل تجارب المحاكاة ، تم التوصل للاستنتاجات الآتية:

1. أن طريقة ( NN ) هي الافضل من خلال نتائج معيار المقارنة إذا أعطت أقل (RMSE) بعد تطبيقها على جميع التباينات واحجام العينات .
2. لقيمة معلمة الضبط (H) أهمية قصوى في التحكم بدقة المقدرات، وتحقق الأمثلية في حالة أختيارها مساوية لعدد المتغيرات التوضيحية (H= P).
3. أن نتائج طريقة ( PPR ) كانت قريبة جدا من نتائج طريقة ( NN ) مما يعطي قوة لهذه الطريقة كون طريقة ( NN ) من الطرائق العددية وهذا ما بدى واضحا في تحليل النتائج .

### (5-2) التوصيات :

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

1. نوصي باستعمال طريقة (NN ) وطريقة ( PPR ) في تقدير النماذج عندما يكون فيها عدد المتغيرات التوضيحية كبيراً وذلك لان زيادة عدد المتغيرات التوضيحية يؤدي غالبا الى حدوث مشكلة البُعدية ( Curse of Dimensionality ) ومن ثم حدوث بقية المشاكل كعدم التجانس والتعدد الخطي والارتباط الذاتي .
2. نوصي باستعمال طريقة (PPR) عندما يكون حجم العينة صغير وبوجود مشكلة البعدية ( curse of dimensionality).

**3. نوصي بمحاولة تقدير قيمة ضبط المعلمة (H) (tuning parameter ) وذلك لأهميتها في حساب ( NN).**
**4. نوصي بأستعمال خوازميات عددية حديثة وذلك لأهميتها في ايجاد الاسقاطات في عملية تجاوز مشكلة البعدية في طريقة ( PPR ) .**

## 6- المصادر :

### المصادر العربية :

**1. الحسناوي ، اموري هادي – القيسي ، باسم شليبه " القياسي الاقتصادي المتقدم –النظرية والتطبيق " المكتبة الوطنية ، دار الكتب والوثائق ببغداد 552 ،(2002).**
**2.رضا ، صباح منفي " تحديد هوية المتكلم باستعمال الشبكات العصبية " اطروحة دكتوراه ، جامعة بغداد ،(2004).**
**3.يوسف ، حنين مراد " مقارنة بين طرائق تقدير انحدار الحرف العامة في معالجة مشكلة التعدد الخطي شبة التام مع تطبيق عملي " ، رسالة ماجستير ، جامعة بغداد ، (2014) .**

### المصادر الاجنبية :

**4.Francis Bach ;(2017) "Breaking the Curse of Dimensionality with convex Neural Networks" JMLR ,PP.(1-53).**
**5.Jerome H. Friedman , Werne Stuetzie ; (1981) " Projection Pursuit Regression ",JASA,Vol.76 , No.376 , PP. (817-823) .**
**6.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).**
**7.Nathan Intrator ; (1993) " Combining Exploratory Projection Pursuit Regression with application to Neural Networks" , MIT , Vol.5 , No.3 , PP.(443-455).**
**8.Swati kaur, S. M. Ghosh ; (2016) "A survey on dimension reduction techniques for classification of multidimensional data", IJSTE ,Vol.2 , Issue.12 , PP.(31-37).**
**9.Wei Wang , Yan Huang , Yizhou Wang and Liang Wang ; (2014) " Generalized Autoencoder A.Neural Network Framework for Dimensionality Reduction " CVPR , PP.(490-497).**

## Use projection pursuit regression and neural network to overcome curse of dimensionality

**Abstract**

**This research aim to overcome the problem of dimensionality by using the methods of non-linear regression, which reduces the root of the average square error (RMSE), and is called the method of projection pursuit regression (PPR), which is one of the methods for reducing dimensions that work to overcome the problem of dimensionality (curse of dimensionality), The (PPR) method is a statistical technique that deals with finding the most important projections in multi-dimensional data , and With each finding projection , the data is reduced by linear compounds overall the projection. The process repeated to produce good projections until the best projections are obtained. The main idea of the PPR is to model the multiple regression as a sum of the nonlinear functions of the linear structures of the variables.**

**Two approaches were used to solve the problem curse of dimensionality : the first approach is proposed projection pursuit regression method (PPR) and The second approach is the method of neural networks (NN) representing by (Back Propagation of error) which is one of the methods used in reducing dimensions . A simulated study was conducted to compare the methods used. The simulations were based on findings that showed that the method (NN) in this study gave better results than the (PPR) based on RMSE.**

**Key words:-** **curse of dimensionality ; projection pursuit regression ; neural networks.**