معالجة القيم المفقودة باستعمال طريقة تحليل المركبات الرئيسة (PCA) و خرائط التنظيم الذاتي(SOM)

أ.م .د. قتيبة نبيل نايف / كلية الادارة و الاقتصاد / جامعة بغداد
الباحث/ بشرى رحيم جاسم

تاريخ التقديم:2017/6/5
تاريخ القبول:2017/10/31

المستخلص :

ان فكرة انجاز بحث حول البيانات غير التامة جاءت من ظروف بلدنا العزيز وما تعرض له من ويلات الحروب حيث أدى ذلك الى فقدان الكثير من البيانات المهمة وفي جميع نواحي الحياة الاقتصادية والطبيعية والصحية والعلمية الصرفة ... الخ. كما ان أسباب الفقدان مختلفة ، منها ما يكون خارجاً عن ارادة المعنيين او تكون بإرادة المعنيين أي يكون مخططاً لذلك بسبب الكلفة او المخاطرة او بسبب عدم توافر الإمكانيات للمعاينة. وان معالجة البيانات المفقودة في هذا البحث تمت باستخدام طريقتي تحليل المركبات الرئيسة وتحليل المركبات الرئيسة الاحتمالية وذلك باستخدام المحاكاة، حيث تم اخذ متغيرات صحة الاطفال والمتغيرات التي تتاثر بها صحة الاطفال وهي الرضاعة وصحة الامهات ويحتوي متغير صحة الامهات على قيم مفقودة وتم معالجتها في برنامج (Matlab2015a) باستخدام طريقة تحليل المركبات الرئيسة وخرائط التنظيم الذاتيSOM حيث تم معالجة القيم المفقودة ومن ثم مقارنة الطرائق باستعمال جذر متوسط مربعات الخطأ وكانت افضل طريقة لمعالجة القيم المفقودة هي طريقة تحليل المركبات الرئيسة (PCA).

المصطلحات الرئيسية للبحث/ **مشكلة البيانات المفقودة، طرائق تقدير القيم المفقودة، تحليل المركبات الرئيسة (PCA) ، خرائط التنظيم الذاتيSOM.**

مجلة العلوم
الاقتصادية والإدارية
العدد 104 المجلد 24
الصفحات354ـ 373

*البحث مستل من رسالة ماجستير

المبحث الاول

1-1 المقدمة:

ان مشكلة البيانات المفقودة هي احدى مشاكل التحليل الاحصائي وهي مشكلة واقعية في الدراسات الإحصائية المختلفة ومن ضمن ذلك المسوحات والدراسات الميدانية وكذلك الدراسات الطبية كالاختبارات السريرية والدراسات الوبائية وغيرها الكثير. وبصورة عامة فأن اي مشاهدة مفقودة تشير الى ان هذه المشاهدة من المفترض ان تسجل بناءً على اسلوب المعاينة المخطط لها ولكن حدث الفشل في مشاهدتها. ولاسيما عندما يكون عدد القيم المفقودة كبير يؤثر في خصائص العينة المسحوبة من مجتمع الدراسة والذي يمكن ان يقود الى استنتاجات مضللة في تحليل البيانات . وفي الكثير من التحليلات الاحصائية وخاصة عند جمع البيانات نجد هناك مشكلة فقدان مشاهدات في بعض المتغيرات والتي يلجأ الباحثون الى تقديرها وفق طرائق خاصة للتقدير.

في هذا البحث قدمنا وسيلة جديدة لمعالجة القيم المفقودة في:

1- تحليل المركبات الرئيسية (Principal Component Analysis)(PCA).
2- تحليل المركبات الرئيسة الاحتمالية (probabilistic Principal Component Analysis) (PPCA)

وقد تم استخدام بيانات المسح الاقتصادي والاجتماعي في العراق MICS4 لعام 2011. وتم الاعتماد في بحثنا هذا على استبيان الاطفال والمرأة حيث يعتبر المسح العنقودي متعدد المؤشرات المصدر الرئيس للمعلومات عن وضع الاطفال حيث يوفر بيانات مؤشرات احصائية اساسية لقياس معدلات التنمية البشرية. ويعد المسح العنقودي متعدد المؤشرات اداة علمية ذات وجودة عالية لا يمكن الاستغناء عنها لتحديد وضع الاطفال والنساء.

2-1 مشكلة البحث:

يحدث فقدان البيانات بشكل واسع في المسوح نتيجة لعدم استجابة بعض المستجوبين لبعض الاسئلة، وايضا قد يكون سبب الفقدان بسبب عطل المعدات او تلف البيانات وتعتبر مشكلة فقد البيانات واقعا لابد من التعامل معه باسلوب علمي ممنهج بعيد عن الاخذ بالحلول السهلة (كالحذف) والتي قد لاتكون ملائمة لبعض الحالات.

فهذه المشكلة تقودنا الى التعريف بها والطرائق الممكنة لمعالجتها بحسب نوع وكمية القيم المفقودة ومن المهم ان نسال لماذا القيم مفقودة وهذا ممكن ان يساعدنا في ايجاد حل لهذه المشكلة وتتم معالجة القيم المفقودة باستعمال تحليل المركبات الرئيسية وتحليل المركبات الرئيسة الاحتمالية باستخدام خوارزمية التعويض (imputation algorithm) مع اجراء مقارنة بين اسلوبي التحليل.

3-1 هدف البحث:

يهدف البحث الى تقدير القيم المفقودة في تحليل المركبات الرئيسة وخرائط التنظيم الذاتي مع اجراء مقارنة بين اسلوبي التحليل وتطبيقها على بيانات المسوح ومقارنتها ومعرفة افضل طريقة لمعالجتها.

المبحث الثاني/الجانب النظري

2-1 المفاهيم الاساسية لخرائط التنظيم الذاتي

1- الطوبولوجيا(Topolgy): هي احد فروع علم الرياضيات الذي يهتم بدراسة تراكيب ومكونات وخصائص جميع الفضاءات المختلفة، بحيث تبقى هذه الخصائص متشابهة تحت عملية التشكيل المتصلة (Smoth deformations) دون ان تقوم بعملية تمزيق او يترك فتحات في الانتقال من احدهما الى الاخر وبالعكس ايضاً[p.p1:10].

2- الشبكات العصبية (Neural Net works): وهي من اهم مجالات الذكاء الاصطناعي حيث انها تعكس تطوراً ملموساً وهاماً في طريقة التفكير الانساني، وان فكرة الشبكات العصبية تدور حول محاكاة العقل البشري باستخدام الحاسب الالي. حيث ان عملية المحاكاة تتم عن طريق حل المشاكل التي تواجهه، وذلك باتباع عمليات التعلم الذاتي والتي تعتمد على الخبرات المختزنة في الشبكة التي تحقق افضل النتائج.[p.p1:12]

3- نموذج خليط جاوس(Gaussian Mixture Model): هو دالة كثافة احتمالية معلمية متمثلة بمجموع اوزان كثافات مركبات جاوس. ونموذج خليط جاوس(GGMS) عادة ما يستخدم كنموذج معلمي من التوزيع الاحتمالي للقياسات المستمرة. ونموذج خليط جاوس(GGM) يقدر المعلمات من بيانات التدريب باستخدام خوارزمية تعظيم التوقع EM المتكررة او تقدير الحد الاعلى اللاحق (MAP) من نموذج سابق مدرب جيداً. ونموذج خليط جاوس هو مجموع الاوزان لـM من مركبات كثافات جاوس تعطى بواسطة المعادلة الاتية:

$$P(Y/\lambda) = \sum_{i=1}^{N} w_i \; g(Y/\mu_i, \Sigma_i) \ldots (1.2)$$

حيث Y هي D من الابعاد المستمرة، قيم بيانات المتجه، w_i، i=1....N، هو خليط من الاوزان $g(Y/\mu_i, \Sigma_i)$، هي مركبات كثافات جاوس. كل مركبة كثافة هي Dمن متغيرات دالة(Gaussian) وكما موضح في النموذج الاتي:

$$g(Y/\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{D/2} \; /\Sigma_i/^{1/2}} \exp\left\{-\frac{1}{2}(Y-\mu_i)' \sum_i^{-1} (Y-\mu_i)\right\} \ldots\ldots (2.2)$$

حيث ان Σ_i هي مصفوفة التباين المشترك [p.p1:29].

2-2 مشكلة البيانات المفقودة Missing Data Problem

تنشأ البيانات المفقودة في المسوحات من اسباب عديدة ومختلفة ،فقد تكون نتيجة لعدم استجابة الاشخاص لبعض الاسئلة اوبعض الاشخاص لايمكن الاتصال بهم، وايضا قد يكون سبب الفقدان بسبب عطل المعدات اوتلف البيانات و لقد شهدت مشكلة البيانات المفقودة اهتماماً ملحوظاً في السنوات الاخيرة، ومع تطور السريع لأجهزة الحاسوب في معالجة العمليات اصبح تطوير طرائق تحليل البيانات المفقودة ممكن نظرياً وعلى الرغم من ذلك مازال الكثير منها بحاجة للتطوير ويعاني من مشاكل عديدة ، حيث تم استعمال طريقتي تحليل المركبات الرئيسة وتحليل المركبات الرئيسية الاحتمالية بهدف ايجاد افضل طريقة ملائمة لتقدير القيمة المفقودة. وفي هذا المبحث سوف يتم عرض انماط البيانات المفقودة والية البيانات المفقودة.

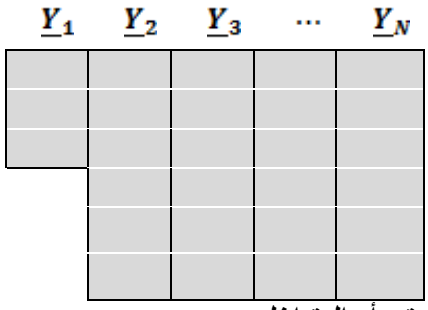
1.2.2 أنمـاط البيانات المفقودة: Patterns of Missing Data

ان الطرائق الإحصائية المستخدمة لمعالجة مشكلة فقدان البيانات تعتمد على نمط البيانات المفقودة. وعليه فأن أنماط البيانات المفقودة تقسـم على قسميـن الأولى منهـا يكون ضمن الأنماط الخاصة Special Patternsوالثانية ضمن النمـط العامGeneral Pattern.

1.1.2.2 النمـط الأول: Pattern of Univariate Missing Data

وهو ابسط حالة من حالات البيانات غير التامة والتي تكون فيها المتغيرات جميعها تامة المشاهدة عدا متغير واحد يعاني من فقدان في بياناته، و يعد هذا النوع من الأنماط الخاصة، والشكل (1.1) يوضح النمط الاول [p.p18:4] :

الشكل(1.1) نمط فقدان البيانات لاحد المتغيرات

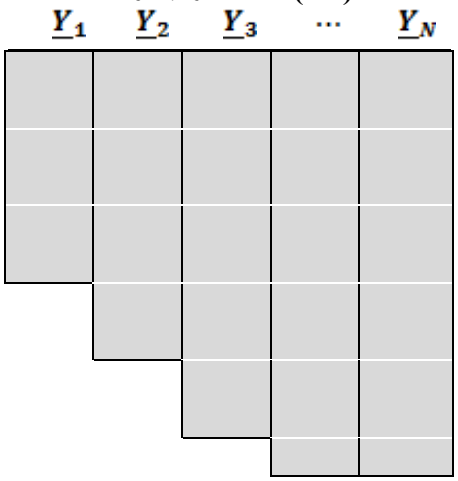


2.1.2.2 النمـط الثانـي : النمط المرتب أو المتداخل

Monotone or Nested Missing Data

في هذا النمط من البيانات غير التامة ترتب المتغيرات بحيث يكون المتغير الذي يحتـوي على اكبر عدد من الحالات البيانات هو أول المتغيرات ثم يأتي المتغير الذي يحتوي العدد الثاني اكبر عدد من الحالات المفقودة وهكذا بالنسبة لبقية المتغيرات . و من الممكن ترتيب المتغيرات بالاتجاه المعاكس كما موضح في الشكل(1.2) :[p.p234:1]

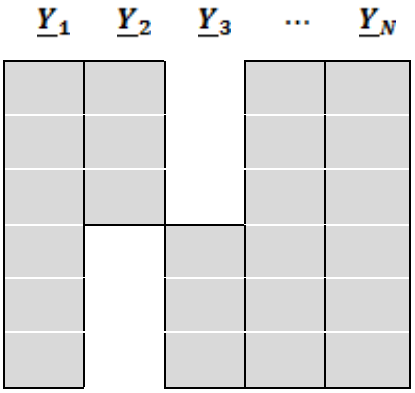
الشكل (1.2)النمط المرتب او المتداخل



3.1.2.2 النمـط الثالـث : Missing Data with Unidentified Parameters

ويسمى بنمط فقدان البيانات في حالة عدم تطابق المعالم ، ويحدث هذا النمط عند دمج او توليف عينتين، و يعد هـذا النمط أخر الأنماط الخاصة، والشكل(1.3) يبين هذا النمط من البيانات غير التامة[p.p19:4]:

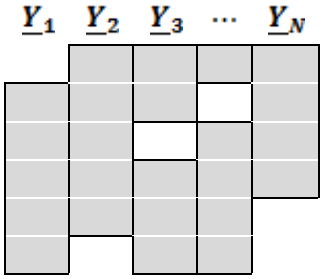
الشكل (1.3) نمط فقدان البيانات في حالة عدم تطابق المعالم



4.1.2.2 النمـط الرابـع : النمـط العـامGeneral Pattern

في هذا النمط فقدان البيانات هو فقدان عشوائي أي انه لا يوجد نمط معين لفقدان البيانات، والشكل (1.4) يبين هذا النمط من البيانات غير التامة[p.p234:1] .

الشكل (1.4) النمط العام للبيانات المفقودة



2.2.2.آلية البيانات المفقودة Missing Data Mechanisms

تختلف الطرائق الاحصائية الخاصة بتحليل البيانات غير التامة في فرضياتها حول الالية التي تؤدي الى فقدان البيانات .وان فهم هذه الالية وتحديد طبيعتها يساعد كثيرا في اختيار الطريقة المناسبة للتحليل بل يعد المدخل لتشخيص الطريقة التي تقترب نتائجها من الأمثلية للبيانات المدروسة.

وفي الادبيات الاحصائية تم تبني الاليات الاتية من انماط الفقدان:

1-فقدان البيانات تماما بشكل عشوائي

Missing Complete At Random(MCAR)

تفقد البيانات بشكل عشوائي تام (MCAR)اذا كان سبب الفقدان مستقلا عن القيمة المفقودة نفسها وعن قيم المتغيرات الاخرى في العينة.

2-فقدان البيانات بشكل عشوائي

Missing At Random (MAR)

تفقد البيانات بشكل عشوائي اذا كان سبب الفقدان له علاقة بقيم المتغيرات الأخرى فقط ومستقل عن القيمة المفقودة .

3-فقدان البيانات بشكل غير عشوائي

Missing Not At Random(Not MAR)

سبب الفقدان ناتج عن القيمة المفقودة نفسها ومستقل عن قيم المتغيرات الأخرى فالبيانات هنا لا تفقد بشكل عشوائي (Not MAR)عند تحليل هذا النوع من البيانات يجب اخذ آلية الفقدان بنظر الاعتبار . اما في حالة (MCAR) و(MAR) يمكن ان تهمل آلية التوزيع [p.p413:7].

ويمكن التعبير عن آلية الفقدان رياضيا وذلك من خلال التوزيع الخاص بها والمقترح من قبل Rubin (1976) والمتمثل بالتوزيع الشرطي ل(R/M)وبمعالم مجهولة هي φ.[p.p20:4]

P(M/R, φ)

حيث ان :

R:مصفوفة من درجة (n*P) وتدعى مصفوفة مؤشر البيانات المفقودة Indictor (Missing Data Matrix)وتمثل القيم المشاهدة والمفقودة لجميع المتغيرات .

M تمثل مصفوفة لها نفس درجة المصفوفة Rبحيث تكون عناصر المصفوفة mمعرفة كالاتي:

$$m_{ij} = \begin{cases} 1 & if\ r_{ij\,obs} \\ 0 & if r_{ij\,miss} \end{cases}$$

فأذا كان:

$$P(M_m/R,\varphi)=P(M_m/\varphi) \quad \text{for all } M_m \qquad \dots\dots(1.2)$$

حيث ان M_m تمثل مصفوفة القيم المفقودة.

Φ المعلمة المجهولة.

فأن البيانات تفقد تماما بشكل عشوائي (MCAR).

اما اذا كان :

$$P(M_m/R,\varphi)=P(M_m/R_0,\varphi) \quad \text{for all } M_m \qquad \dots\dots(2.2)$$

فأن البيانات تفقد بشكل عشوائي (MAR).

اما الحالة التي يعتمد بها التوزيع على القيمة المفقودة وبحسب الصيغة الاتية :

$$P(M_m/M_0R,\varphi)=P(M_m/R_m,\varphi) \quad \text{for all } M_m \qquad \dots\dots(3.2)$$

حيث ان R_0 تمثل مؤشر مصفوفة القيم المفقودة.

هنا يمكن القول أن البيانات لا تفقد بشكل عشوائي (Not MAR)عند تحليل هذا النوع من البيانات يجب اخذ توزيع آلية الفقدان بنظر العناية .

اما في حالة (MCAR) و (MAR)يمكن ان يهمل توزيع آلية الفقدان. [pp235-236:1]

3-2 طرائق تقدير القيم المفقودة

منذ بدء العمل بمعالجة مشكلة البيانات غير التامة في نماذج الإنحدار والباحثين يحاولون جاهدين حل المشكلة بإستخدام اسلوب التعويض عن كل قيمة مفقودة بقيمة تقديرية تعتمد في تقديرها على عدة طرائق منها طريقة اقرب مجاور او التعويض بالمتوسط او التعويضات المتعددة، واشار الباحثون الى ان تقدير القيم المفقودة يعني زيادة في حجم العينة قيد الدراسة ومن ثم يكون لهذه المعلومات تأثير رئيس في زيادة كفاءة مقدرات انموذج الإنحدار لكن السؤال الذي يطرح نفسه هو ماذا لو كانت هذه القيمة المقدرة غير ممثلة بشكل كفوء للقيمة المفقودة فسوف يؤدي ذلك الى زيادة المشكلة وتضليل المعلومات حول تلك المشاهدات بشكل اكبر هو اشبه مايكون بالمريض الذي يوصف له دواء غير مناسب مما يزيد في معاناته[2: p.p46]. وفي هذا المبحث طرائق تعويض القيم المفقودة يمكن ان تقسم الى:

1*-طرائق التعويض المفرد *single imputation methods

2*-طرائق التعويض المتعدد *multiple imputation methods

اذ قمنا باستخدام تحليل المركبات الرئيسية و تحليل المركبات الرئيسة الاحتمالية لمعالجة القيم المفقودة.

1.3.2 التقدير باستخدام المتوسط الحسابي *Mean Imputation*

تعتمد هذه الطريقة على استبدال البيانات المفقودة بقيمة المتوسط (***Mean***) المقدر باستخدام المشاهدات غير المفقودة وكما يأتي[p.p237: 1]:

$$\tilde{X} = \sum_{n_j} X_{obs} / n_j \qquad (1)$$

حيث n_j تمثل عدد القيم المشاهدة فعلا للمتغير X_j.

2.3.2 تقدير البيانات المفقودة باستخدام تحليل الانحدار

تستخدم هذه الطريقة لتقدير القيم التي سيتم تعويضها بدلاً من القيم المفقودة، وذلك من خلال تكوين مصفوفة الإرتباطات الاساسية للمتغيرات ، وكل متغير يتضمن قيماً مفقودة ، تتم معاملته على انه متغير تابع من خلال معادلة الإنحدار التي يتم تكوينها لكل فقرة تتضمن قيماً مفقودة، ثم تستخدم المعادلات الناتجة، في الحصول على تقديرات للقيم المفقودة لكل متغير، وبعد ذلك تتم عملية إدخال او تعويض هذه التقديرات في مجموعة البيانات الناقصة التي تتضمن قيماً مفقودة، والقيم المتنبأ بها من معادلة خط الانحدار ، يتم تعويضها بدلاً من القيم المفقودة بكل فقرة . وهكذا تتكرر هذه العملية لكل فقرة تتضمن قيما مفقودة. [p.p25: 5]

4.2 طريقة التعويضات المتعددة

Multiple Imputation Method(MI)

يتم في هذه الطريقة استبدال كل قيمة مفقودة بمتوسط مجموعة من القيم المختارة عشوائياً ، ولذلك ينظر اليها على انها تقدم قيما تعويضية باخطاء معيارية غير متحيزة في التحاليل الاحصائية، وهو ما يختلف عن طريقة حساب القيمة التعويضية الواحدة.[p.p26:3]

5.2 تحليل المركبات الرئيسة (Principal Component Analysis) (PCA)[p.p6:6]

المركبات الرئيسة تعد من اكثر طرائق التحليل العاملي دقة وشيوعاً في البحث ولهذه الطريقة مزايا عدة منها أنها تؤدي إلى تشبعات دقيقة وتؤدي إلى اقل قدر ممكن من البواقي كما أن المصفوفة الارتباطية تختزل إلى اقل عدد من العوامل المتعامدة (غير المرتبطة).وهي عبارة عن تراكيب خطية ***Linear Combination*** لمتغيرات عشوائية ذات خصائص معينة تدعى بالتباينات ***Variances*** وهي في النهاية متجهات مميزة ***Characteristic Vectors*** لمصفوفة التغاير. حيث ان التباينات المرافقة لهذه المركبات تتميز بخصائص احصائية. ففي التجارب الاحصائية تستخدم طريقة المركبات الرئيسة لإيجاد التراكيب الخطية بتباين كبير وفي كثير من الدراسات عندما تكون المتغيرات المتناولة في البحث كبيرة جداً فالطريقة الملائمة لتقليص هذا العدد الكبير من المتغيرات هو اهمال التراكيب الخطية التي لها تباينات صغيرة، ودراسة التراكيب التي لها تباينات كبيرة فقط. وتنطوي طريقة المركبات الرئيسية على تحويل مجموعة في المتغيرات المرتبطة خطياً $X's$ الى مجموعة جديدة من المركبات الرئيسية $V's$ على هيئة تراكيب خطية مشتقة من المتغيرات التوضيحية لتحل محلها، بحيث تكون مؤهلة لتفسير معظم التباين الكلي للقيم الاصلية.

وقبل ان نتطرق الى ماهية المركبات الرئيسة لابد لنا من التعريف بالجذور والمتجهات الذاتية (المميزة) ***Eigen Values and Eigen Vectors***، فلو فرضنا المصفوفة X درجتها q وحصلنا على متجه عمودي غير صفري $\underline{b}$ عدد عناصره q وقيمة غير متجه هي λ_i فأن:-

$$Xb_i = \lambda_i \, b_i \qquad \dots (1.2)$$

حيث قيم λi التي تحقق هذه المعادلة تسمى بالجذور المميزة للمصفوفة X، اما المتجهات $\underline{bi}$ التي تناظر هذه الجذور فتسمى بالمتجهات المميزة للمصفوفة X. من المعادلة (*1.2*):

$$[X - \lambda_i I][\underline{b}_i] = \underline{0} \qquad \dots (2.2)$$

حيث ان : I مصفوفة الوحدة.

فاذا كانت المصفوفة $[X-\lambda_i I]$ غير احادية، فيمكن ايجاد $\underline{b_i}$ بضرب المعادلة (1.2) ضرباً قبلياً في معكوسة هذه المصفوفة، وفي هذه الحالة تكون $\underline{b_i}$ متجهاً صفرياً وهو ما يتعارض مع كون $\underline{b_i}$ متجهاً غير صفري، لذلك فأن الشرط اللازم لإيجاد المتجه $\underline{b_i}$ ان تكون المصفوفة $[X-\lambda_i I]$ احادية، أي ان قيمة محددها يساوي صفراً.

$$|X - \lambda_i I| = 0 \quad \dots (3.2)$$

وتسمى المعادلة (3.2) بالمعادلة المميزة *Characteristic Equation* للمصفوفة X وبحلها يمكن ايجاد قيم λ_i، وباستخدام المعادلة (2.2) يمكن ايجاد المتجهات المميزة المناظرة لتلك الجذور بحيث تكون هذه المتجهات متعامدة *Orthogonal* فيما بينها.

و لو كان لدينا q من المتغيرات العشوائية $x_q x_1$, ... , بمتوسط مجتمع $\underline{\mu} = \underline{0}$ ومصفوفة التباين المشترك Σ أي ان $X \sim N_q(\underline{0}, \Sigma)$ ونفترض ان المصفوفة S تمثل تقديراً لمصفوفة التباين المشترك للمجتمع بدرجة حرية $n = N - 1$ وتكون متماثلة حقيقية معرفة موجبة $(q.d)$ او شبه موجبة $(q.s.d)$ فأهم خواص الجذور والمتجهات المميزة للمصفوفة S هي:

1. جميع الجذور المميزة للمصفوفة S تكون موجبة او غير سالبة وقد يكون بعضها متساوياً أي ان:

$$\lambda_1 > \lambda_2 > .... > \lambda_q > 0 \quad \dots (4.2)$$

2. نفترض ان المتجهات المميزة المناظرة للجذور المميزة هي $\underline{b_1}^*, \underline{b_2}^*, \dots, \underline{b_q}^*$ على الترتيب، وتكون المتجهات المميزة المتعامدة المعدلة *Normalised* للمصفوفة S هي $\underline{b1}\, b_q$, ..., $\underline{b2}$, بحيث ان:

$$\underline{b}'_i \, b'_j = \begin{cases} 0 & if \quad i \neq j \\ 1 & if \quad i = j \end{cases} \qquad i, j = 1, ..., q \quad \dots (5.2)$$

وان خاصية التعامد لها دور كبير في عملية تحليل المكونات الرئيسة المتعاقبة الى مجموع التباين الكلي للمتغيرات.

3. توجد المصفوفة T بحيث ان:

$$TST = D \quad \dots (6.2)$$

حيث ان:

D : مصفوفة قطرية عناصر القطر فيها تمثل الجذور المميزة للمصفوفة S.

T : مصفوفة متعامدة، اعمدتها تمثل المتجهات المميزة المتعامدة المعدلة المناظرة للجذور المميزة للمصفوفة X.

وبامعان النظر بالمعادلة (6.2) نستطيع ان نميز:

$$\underline{b}'_i \, S \, \underline{b}_j = \begin{cases} o & if \quad i \neq j \\ \lambda_i & if \quad i = j \end{cases} \qquad i, j = 1, ..., q \quad \dots (7.2)$$

المكونة الرئيسة الاولى V_1 للمتغيرات الاصلية $X's$ هي عبارة عن تركيب خطي حيث ان:

$$V_1 = b_{11}X_1 + b_{21}X_2 + ... + b_{q1}X_q = \underline{b}'_1 \underline{X} \quad \dots (8.2)$$

حيث ان $\underline{b_1}$ يمثل المتجه المميز المناظر للجذر المميز λ_1.

وطالما افترضنا ان $X \sim N_q(\underline{0}, S)$ فأن:

$$V_1 \sim Nq(\underline{0}, \underline{b}_1' S \, b_1) \quad \dots (9.2)$$

وطالما ان $\underline{b}_1' S \underline{b}_1 = \lambda_1$ نستطيع القول ان:

$$V_1 \sim N_q(\underline{0}, \lambda_1) \quad \ldots(10.2)$$

أي ان الجذر المميز الاكبر للمصفوفة S يستخدم لتقدير التباين الاعظم للمكونة الرئيسة الاولى، والمتجه المميز $\underline{b_1}$ يستخدم لتقدير المعاملات للمكونة الرئيسة الاولى. وبالطريقة نفسها تكون المكونة الرئيسة الثانية:

$$V_2 \sim N_q(\underline{0}, \lambda_2) \quad \ldots (11.2)$$

وبذلك نستطيع التعميم:

$$\underline{V} = \Gamma' X \quad \ldots (12.2)$$

اما عن التباين المشترك بين V_2 , V_1 فهو:

$$cov(V_1, V_2) = E(V_1, V_2) - E\,V_1\,E\,V_2 \quad \ldots (13.2)$$

$$= E\,\underline{\grave{a}_1 X \grave{X} a_2}$$

$$= \underline{\grave{b}_1 E(X\acute{X})\ \grave{b}_2}$$

$$= \underline{\grave{b}_1 S\, b_2}$$

$$= zero$$

أي ان الارتباط بين المكونة الرئيسة الاولى والمكونة الرئيسة الثانية يساوي صفرا.

ان اهمية الجذور المميزة هي انها تعد وسيلة لقياس الابعاد الموضوعة في تلك البيانات طالما ان:

$$\sum_{i=1}^{q} \lambda_i = tr(S) \quad \ldots (14.2)$$

$$\prod_{i=1}^{q} \lambda_i = |S| \quad \ldots (15.2)$$

لهذه الخاصية اهمية كبيرة في تفسير المكونات الرئيسة، حيث ان المكونة الرئيسة V_i تكون خطية ومستقلة عن المكونة الرئيسة V_j والجذور المميزة للمصفوفة S λ_j, λ_i تمثل التباينات للمكونات الرئيسة V_j, V_i على الترتيب، لذلك ومن المعادلة (14.2):

$$\sum_{i=1}^{q} \lambda_i = \sum_{i=1}^{q} Var(V_i) \quad \ldots (16.2)$$

وان الاهمية النسبية للمكونة الرئيسة (jth) في وصف النموذج تقاس نسبتها الى مجموع التباين، أي ان:

$$\frac{Var(V_i)}{\sum_{i=1}^{q} Var(V_i)} = \frac{\lambda_i}{tr(S)} = \frac{\lambda_i}{\sum_{i=1}^{q} \lambda_i} \quad \ldots (17.2)$$

في حالة كون المتغيرات المدروسة لها وحدات قياس مختلفة ففي هذه الحالة يصار الى تحويل المتغيرات الى الصيغة القياسية اولاً، وفي هذه الحالة فأن معاملات المركبات الرئيسة (b_{ij}) هي قيم المتجهات المميزة لمصفوفة الارتباط للمتغيرات القياسية. وطالما ان مصفوفة الارتباط متماثلة حقيقية موجبة ($q.d$) فأن جميع الجذور المميزة لها تكون موجبة، وكذلك:

$$\mathrm{tr}(R) = \sum_{i=1}^{q} \lambda_i^* = q \qquad \dots (18.2)$$

$$|R| = \prod_{i=1}^{q} \lambda_i^* \qquad \dots (19.2)$$

حيث ان λ_i^* , $i = 1, \dots, q$ الجذور المميزة للمصفوفة R، والاهمية النسبية للمكونة الرئيسية (V_i) في وصف النموذج تقاس بـ :

$$\frac{Var(V_i)}{\sum_{i=1}^{q} Var(V_i)} = \frac{\lambda_i^*}{\sum_{i=1}^{q} \lambda_i^*} = \frac{\lambda_i^*}{q} \qquad \dots (20.2)$$

واذا كانت بعض الجذور المميزة لمصفوفة الارتباط متساوية، فالمحاور المناظرة لهذه الجذور لها الطول نفسه، واذا كانت جميع الجذور المميزة متساوية فأن القطع الناقص Ellipsoid يصبح كروي[p.p26:6].

وان تحليل المركبات الرئيسة (Principal Component Analysis) هي مثال على تقنيات اختزال الابعاد حيث وظيفتها هو ايجاد رسم الخرائط ل D من ابعاد الفضاء الاصلي الى K من ابعاد الفضاء الجزئي حيث $K<D$. فضلا عن اختزال الأبعاد وضغط البيانات. حيث تحليل المكونات الرئيسة اقترحت من قبل العالم(JOLLIFFE عام 1986).

وعلى نطاق واسع تستخدم تحليل المركبات الرئيسة في تطبيقات اخرى مثل تصور البيانات ،معالجة الصور، التعرف على الانماط والتنبؤ بالسلاسل الزمنية.

وان الصيغ الأكثر شيوعا من تحليل المركبات الرئيسة (*Principal Component Analysis*) تعرف بصيغة تباين الحد الاعلى وتعرف *Principal Component Analysis* بانها الإسقاط المتعامد للبيانات على فضاء الأبعاد الخطية القليلة ، الفضاء الجزئي الرئيس، بحيث يتم تكبير تباين البيانات ويتم الحفاظ على الحد الاعلى من المعلومات (*information*) وهذ مقترح من قبل (*HOTOLLING*) عام *1995*) ويمكن ان نظهر ان الاسقاط الامثل ل *K* من ابعاد الفضاء الجزئي هو ان نختار *k* من المتجهات الذاتية $\{Wj\}$ $j=1,\dots,k$ ،من مصفوفة البيانات المشتركة:

$s = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$ المقابلة لاكبر القيم الذاتية *K* ،حيث $\bar{x} = N^{-1}\sum_n x_n$ $k,\lambda 1 \dots\dots \lambda$

,وان التحول الخطي من متجه المشاهدات x_n الى الفضاء الجزئي الرئيس يعرف بواسطة المتجهات الذاتية *k*ويكون الناتج كما في المعادله *(21.2)*

$$z_n = W^T (x_n - \mu) \qquad (21.2)$$

بحيث *Zn* تسمى Z-scors المعيارية هي مجموعة نقاط *Z* لمتجه المشاهدات و *K*عدد الاعمدة للمصفوفة *W* حيث *K* تؤدي الى المتجهات الذاتية لل*S* وμ هو متوسط المشاهدات . الخاصية المكملة ل *Principal* Component Analysis هو ايجاد التمثيل الخطي ل *k* من الأبعاد للبيانات مثل إعادة بناء الخطأ التربيعي للبيانات كما في المعادلة الاتية :

$$\hat{x}_n = W z_n + \mu \qquad (22.2)$$

تم تصغيرها[pp13:8] .

2-5-1 تحليل المركبات الرئيسة في تقدير القيم المفقودة *component Analysis With missing data*

تستخدم خوارزمية التعويض(imputation algorithm) في تحليل المركبات الرئيسة لتقدير القيم المفقودة، حيث انه في تحليل المركبات الرئيسة (الكلاسيكية) التقليدية (Principal Component Analysis) ليس هناك طريقة واضحة للتعامل مع القيم المفقودة ، وقام بوصفها كل من (Jolliffe، Iline، Riko، 2010) .وان خوارزمية التعويض هي عبارة عن اجراءات متكررة حيث تكون احد البدائل بين تعويض القيم المفقودة في البيانات، T_{mis} وتطبيق تحليل المركبات الرئيسة القياسية لمصفوفة البيانات الكاملة.

ويتم حساب القيم الاولية للعناصر المفقودة T_{mis} بحساب الوسط الحسابي الخاص بكل صف من صفوف المصفوفة T ،وكذلك يتم تحديث متوسط المشاهدات $\bar{t}$ في كل تكرار.ولخصت نتائج الخوارزمية مما ادى الى خوارزمية 3.1.4.2 [pp16:8].

2-5-2 خوارزمية التعويض لتحليل المركبات الرئيسة

The imputation algorithm for Principal Component Analysis يتم تحديد المعطى من بيانات T غير التامة مع العناصر المشاهدة T_{obs} والعناصر المفقودة T_{mis}.

1-تعريف العناصر المفقودة T_{mis} واعطاؤها قيم اولية وذلك بحساب المتوسط لكل صف من صفوف مصفوفة المشاهدات T_{obs}.

$$T_{mis} \leftarrow mean(T_{obs})$$

حيث T_{mis} العناصر المفقودة من المصفوفة T، T_{obs} العناصر المشاهدة للمصفوفة T.

وهذا ادى الى حساب البيانات التامة T_{imp}.

2-يتم تحديث متوسط المشاهدات μ لكل صف من صفوف المصفوفة T_{imp}.

$$\mu \leftarrow mean(T_{imp})$$

3-ايجاد الحل لـ p من المركبات الرئيسية للمصفوفة W باستخدام البيانات التامة.

4-يتم تحديث العناصر المفقودة كما يلي:

$$\mathrm{T_{mis}} \leftarrow \mathrm{WW^T}\left(\mathrm{T_{imp}} - \bar{t}\right) + \bar{t}$$

5-يتم التحقق من وجود تقارب اما T_{imp} او $Z=W^T(\ T_{imp}-\bar{t})$ اذا كان معيار التقارب ليس له عودة الى الخطوة 2. [pp17:8]

6.2 خريطة التنظيم الذاتي Self-Organizing Map (*SOM*)

هي احدى انواع الشبكات العصبونية الإصطناعية التي تعتمد على مبدأ التعليم غير المراقب وتستخدم في مجالات متعددة نذكر منها التصنيف وتقليل او تخفيض الأبعاد. وتسمى خارطة التنظيم الذاتي(*SOM*) بشبكة كوهنين العصبية نسبة الى الباحث الفنلندي البروفيسور (Teuvo kohonen) الذي صممها في عام (1982) حيث انه اشتهر بعمله ومساهماته في استخدام الحاسوب والشبكات العصبية الاصطناعية. وان الخلايا العصبية في خريطة التنظيم الذاتي(*SOM*) عادةً ماتسمى وحدات خريطة او نماذج ،وبالتالي يمكن ان ينظر اليها على انها عينات ممثلة للبيانات .و ان كل وحدة خريطة ترتبط مع متجه إشارة(*reference vector*) $\mathrm{W_i}$، ويتم رسم الخريطة لكل بيانات المتجه لوحدة الخريطة التي تمتلك متجه اشارة ويكون معظم تشابهها لبيانات المتجه نفسه .متجهات الاشارة $\mathrm{W_i}$ هي المعدلات الموزونة الموضعية للبيانات المرتبطة مع وحدة الخريطة التي تعطى في الاصل لفضاء البيانات .

حيث تستخدم خريطة التنظيم الذاتي(SOM) لصنع الأبعاد القليلة ، وعادة ثنائية الأبعاد وتمثيل وتصوير البيانات ذات الابعاد العالية. و تضع خريطة التنظيم الذاتي شرط الطوبولوجيا للمحافظة على رسم الخرائط من فضاء البيانات الاصلية الى وحدات الخريطة، وذلك لتشكيل شبكة ثنائية الابعاد وهذا يوفر الوسائل لتصوير البيانات على سطح مستو(on a plane)، حيث تم ملائمتها ورسم خرائط التنظيم الذاتي SOM من نوع البيانات الاعتباطية والتي تكون متبادلة وكذلك يتم تعريف المسافات الزوجية وهذا مقترح من قبل (kohonen and somervuo) عام 2002[pp18:8] .

1.6.2 خوارزمية خريطة التنظيم الذاتي

Self-Organizing Map(algorithm) (SOM)

تستخدم خريطة التنظيم الذاتي (SOM) مع متجه ثنائي الأبعاد من الوحدات ومن ثم خريطة التنظيم الذاتي (SOM) تحدد برسم فضاء البيانات المدخلة لمستوى ثنائي الأبعاد . حيث انه لكل وحدة خريطة i يكون لديها متجه اشارة (reference) معلمي (نموذج متجه) $w_i \in R^d$ حيث ان:

d:هو الابعاد للبيانات .

w_i : متجه اشارة يستخدم للإشارة الى وحدات ونموذج المتجهات الإبدالية.

. حيث $T \in R^{N*D}$ يكون تدريب(training) مصفوفة البيانات مع عينات *N* لـ q من الأبعاد. وان كل بيانات المتجه t_n يجب مقارنتها مع كل متجهات الاشارة w_i . ويتم استخدام المسافة الاقليدية(***Euclidean distance***) والوحدة مع اصغر مسافة اقليدية كما موضح في المعادلة(32.2):

$$w_{c(t_n)} = \arg\min_i \|t_n - w_i\| \quad (32.2)$$

وتشير المعادلة (32.2) اعلاه الى افضل وحدة مطابقة لمتجه المشاهدات t_n.

حيث ان التدريب(learning) يبدأ بواسطة تهيئة متجهات الاشارة(***reference vectors***) $w_i(\tau = 0)$

حيث 0= τ تشير الى متغير الوقت المتقطع(discrete-time) وتمثل المقياس الزمني للتدريب .

ويتم تحديث متجهات الاشارة باستخدام المعادلة الاتية:

$$w_i(\tau + 1) = w_i(\tau) + h_{ci}(\tau)\big(x(\tau) - w_i(\tau)\big) \qquad (33.2)$$

حيث $h_{ci}(\tau)$ هي الدالة المجاورة التي تُعرف على طول الشبكة لوحدات الخريطة .حيث تستخدم على نطاق واسع دالة جاوس المجاورة(Gaussian neighborhood function) كما في المعادلة الاتية:

$$h_{ci} = \alpha(\tau).\exp\left\{-\frac{\|r_c - r_i\|^2}{2}\right\} \qquad (34.2)$$

حيث $\|r_c - r_i\|^2$ هي المسافة بين افضل وحدة مطابقة r_c ووحدة i في *الصف*.

$0<\alpha(t)<1$ هو نسبة التعلم ذات القيمة العددية و $\sigma(\tau)$ معلمة القطر.

ومن الضروري للتقارب ان $h_{ci} \to 0$ عندما $\tau \to \infty$ وعادة $\alpha(\tau)$ و $\sigma(\tau)$ كلاهما يأخذ بالتناقص بصورة تدريجية(monotonically).

حيث تؤخذ الاوزان بعين الاعتبار اثناء تحديث متجهات الاشارة لكل البيانات او لمجموعة من البيانات في وقت واحد ووفقاً لذلك يتم تحديث قاعدة w_i كما موضح في المعادلة(35.2) :

$$w_i = \frac{\sum_n h_{ni} t_n}{\sum_j h_{ni}} \qquad (35.2)$$

حيث n : هي مؤشر لكل متجهات البيانات التي تمتلك افضل الوحدات المطابقة التي تقبل ب $h_{ni} > 0$.[pp61:9]

2.6.2 خريطة التنظيم الذاتي في تقدير القيم المفقودة

تم استخدام ***Self-Organizing Map with missing values***

خريطة التنظيم الذاتي لحساب القيمة المفقودة مع انواع كثيرة مثل بيانات المسح التي قام بها(Fessant and Midenet)عام 2002 و Wanyعام 2003،والبيانات الاقتصادية والاجتماعية التي قام بها (Cottrell and letrémy)عام 2007,وآخرون Gaubert عام1996.والبيانات الصناعية التي قام بها (Rustum and Adeloye)عام 2007وآخرون Sorjamaa واخرون ،عام 2009،Merlinوآخرون ،عام2010، والبيانات المناخية لSorjamaaعام*2010* .وان حساب القيم المفقودة في البيانات اعلاه يكون وفق ما اقترح من قبل كل من Cottrell and letrémy عام 2007 , حيث انه يتم حساب افضل الوحدات المطابقة لمتجهات البيانات مع القيم المفقودة كما موضح في المعادلة(36.2) التالية:

$$w_{C\,xn} = \arg\min_{i} \|t_n - w_i\| = \sum_{k\ s.t\ I_{nk}} (t_{nk} - w_{ik})^2 \qquad (36.2)$$

اي انه يتم حساب المسافة فقط باستخدام المركبات الموجودة في المتجه t_n. وكذلك يتم تجاهل القيم المفقودة اثناء تحديث متجهات الإشارة. وبعد التدريب، يتم التعويض عن القيم المفقودة وفقا لأفضل الوحدات المطابقة المناظرة لمتجهات البيانات، وهذا مقترح من قبل Cottrell وletrémy عام 2007 .[pp20:8]

المبحث الثالث/ الجانب التجريبي

3 -1 المقدمة (Introduction):

ان عملية المحاكاة هي تمثيل او تقليد للواقع الحقيقي باستخدام نماذج معينة، وكثيرا ما نجد في الواقع الحقيقي عمليات معقدة الفهم والتحليل، لذا من الأفضل ان نوصف هذه العمليات بصورة مشابهة للصورة الحقيقية بنماذج معينة، ففهم النموذج سيحقق قدرا من الادراك للعملية الاصلية او الواقع الحقيقي من خلال نماذج المحاكاة . هذا وان درجة المشابهة بين أي تجربة محاكاة والواقع الحقيقي تعتمد على مدى مطابقة او مشابهة نموذج المحاكاة للواقع الحقيقي. وقد اعتمد اسلوب المحاكاة(Simulation) لتطبيق الطرائق المدروسة في الجانب النظري لمحاكاة أكبر عدد ممكن من الحالات التي من الممكن أن تواجهنا في الواقع العملي وللحصول على نظرة اكثر شمولية. ونظرا للسرعة التي قد توفرها الحاسبات الالكترونية والبرامج من حزم جاهزة وبرمجية مما ادى بالباحثين الى اعتمادها في محاكاة نماذجهم بسبب الصعوبة في البرهان الرياضي والنظري وكذلك توفر الامكانية في مقارنة أكثر من طريقة في الوقت نفسه ومعرفة أي من الطرائق هي الأفضل.

فضلاً عما ذكر انفا فأن المحاكاة توفر مرونة في التعامل وسهولة في تكرار تجارب من الصعوبة تكرارها في الواقع العملي وكذلك وضع افتراضات من ناحية التباين وحجوم عينات مختلفة.

تم استعمال بعض الدوال الجاهزة والصيغ البرمجية في برنامج الـ (Matlab 2015a)في توليد البيانات وبناء نماذج المحاكاة لغرض المقارنة بين الطرائق باختلاف أحجام العينات والتباينات ولأكثر من نسبة للفقدان في قيم المتغير المستقل X_2.

2-3 الأنموذج المستعمل في المحاكاة Simulation Model

تم استخدام نموذج انحدار خطي متعدد

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

وبنسب فقدان (10%،20%،30%،40%).

حيث المتغير X_2 يحتوي على قيم مفقودة والخطوات الاتية تبين سير الخوارزمية لكل طريقة.

3-3 خطوات طريقة تحليل المركبات الرئيسة PCA:

1-حساب مصفوفة التباين والتباين المشترك ذات الابعاد $N \times N$ لمصفوفة أنحرافات المتغير X ذو الابعاد $N \times P$.

2-أستخراج المتجهات المميزة المرافقة للكل جذر مميز وحذف المتجهات التي ليس لها تأثير، عدد المتجهات الكلي $N \times N$ وبعد الحذف $N \times K$ ، إذ أن $N > K$.

3-حساب z ذات الابعاد $K \times P$ من المعادلة $(W'W)^{-1}W'(X-\mu)$ ،إذ أن μ تمثل الوسط الحسابي الخاص بكل صف من صفوف المصفوفة X.

4-حساب مصفوفة المتجهات المميزة W من جديد من المعادلة $XZ'(Z'Z)^{-1}$.

5-تكرار الخطوتين 3 و 4 لحين الوصول لحالة الاستقرارية.

6-حساب القيم المفقودة ولكل عمود من X من المعادلة $X_P = WZ_P + \mu$.

4-3 خطوات طريقة تحليل المركبات الرئيسة الاحتمالية PPCA:

- نفس الخطوات المتبعة في طريقة PCA لكن الفرق بدخول حد الخطأ العشوائي في الخطوة السادسة على تقدير القيمة المفقودة.

5-3 خطوات طريقة خريطة التنظيم الذاتي SOM:

1-فرض متجه اوزان اولي m يتم توليده من البيانات المدروسة وبالابعاد $N \times 1$.

2-حساب المسافات لكل قيمة من قيم m مع كل قيم المتغير المعتمد X (هنا يتم اخذ فقط المتجه الذي يحتوي على مشكلة الفقدان، أي متجه بطول N) وكما في المعادلة:

$$d(i) = \| m(i) - X \| = \sqrt{\sum_{j=1}^{N} (m(i) - X(j))^2}$$

1-يتم إختيار أقل قيمة من المسافات المستخرجة في الخطوة 2.

2-فرض قيم أولية لكل من نسبة التعلم $\eta(0)$ ومعلمة القطر $\sigma(0)$ وعدد تكرارات التجربة r_{max}=1000 .

$$\eta(0) = 0.01, \sigma(0) > 0$$

3-تحديث الأوزان وفق المعادلة:

$$m_{(i)}^{(r)} = m_{(i)}^{(r-1)} + \eta(r) h^{(r-1)}(i, i_n)(X(i) - m_{(in)}^{(r-1)})$$

إذ أن:

- $\eta(r) = \eta(0) e^{(-\frac{r}{r_{max}})}$.

- $\sigma(r) = \sigma(0)e^{(-\frac{rlog(\sigma(0))}{r_{max}})}$.
- $h(r) = e^{(-\frac{dw^2_{(i)}}{2\sigma^2_{(r)}})}$, $where\ dm(i) = \| m(i) - m(min) \|$.

4-يتم أختيار ادنى قيمة من المتجه m.

5-يتم اعادة الخطوتين 5 و 6 الى ان تصل عدد الخطوات الى r_{max}.

6-يتم اعتماد القيمة $m(i)$ بدل القيمة $X(i)$ المفقودة.

5-3-تحليل نتائج المحاكاة

متوسط مربعات الخطأ عند التعويض بدل المشاهدات المفقودة :

1-**حجم عينة** 3000

من الجدول المرقم (3-1) نلاحظ في حالة متوسط مربعات الخطأ لX انه في حالة فقدان البيانات بنسبة(10%،20%،30%،40%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين (1,2,3).

اما في حالة متوسط مربعات الخطأ لY نلاحظ ايضاً انه في حالة فقدان البيانات بنسبة(10%،20%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين (1,2,3).

اما عند النسبة(30%) افضل طريقة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين(1,2)اما عند التباين (3) فان افضل طريقة هي طريقة تحليل المركبات الرئيسيةPCA تليها طريقة خرائط التنظيم الذاتيSOM.

اما عند النسبة(40%) افضل طريقة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين(1,3)اما عند التباين (2) فان افضل طريقة هي طريقة تحليل المركبات الرئيسةPCA تليها طريقة خرائط التنظيم الذاتيSOM.

2-**حجم عينه** 4000

من نتائج الجدول المرقم (3-2) في حالة متوسط مربعات الخطأ لX نلاحظ انه في حالة فقدان البيانات بنسبة(10%،20%،30%،40%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين (1,2,3).

اما في حالة متوسط مربعات الخطأ لY نلاحظ انه في حالة فقدان البيانات بنسبة(10%،20%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين (1،2،3).

اماعند النسبة (20%) فأن افضل طريقة عند التباين (1) طريقة تحليل المركبات الرئيسة PCA PCA ومن ثم هي طريقة خرائط التنظيم الذاتيSOM ، اماعند التباين(2،3) فأن افضل طريقة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA .

اما عند النسبة (40%) فأن افضل طريقة في حالة التباين (1،3) هي طريقة خرائط التنظيم الذاتيSOM ومن ثم طريقة تحليل المركبات الرئيسة PCA ، اما في حالة التباين (2) فأن افضل طريقة هي طريقة تحليل المركبات الرئيسةPCA تليها طريقة خرائط التنظيم الذاتي SOM.

3-**حجم عينة** 5000

من نتائج الجدول المرقم (3-3) في حالة متوسط مربعات الخطأ لX نلاحظ انه في حالة فقدان البيانات بنسبة (10%،20%،30%،40%) ان افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM ومن ثم طريقة تحليل المركبات الرئيسة PCA عند التباين (1،2،3).

اما في حالة متوسط مربعات الخطأ ل Y نلاحظ انه في حالة فقدان البيانات بنسبة (10%،30%) ان افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM ومن ثم طريقة تحليل المركبات الرئيسية PCA عند التباين (1،2،3) .

اما عند النسبة 20% اظهرت النتائج ان افضل طريقة لمعالجة القيم هي طريقة خرائط التنظيم الذاتيSOM ومن ثم طريقة تحليل المركبات الرئيسة PCA عند التباين(1)، اما عند التباين (2،3) فأن افضل طريقة هي طريقة خرائط التنظيم الذاتيSOM ومن ثم طريقة تحليل المركبات الرئيسة PCA .

اما عند النسبة (40%) اظهرت النتائج ان افضل طريقة هي طريقة خرائط التنظيم الذاتيSOM ومن ثم طريقة تحليل المركبات الرئيسة PCA عند التباين (1،3)، اما عند التباين (2) فان افضل طريقة هي طريقة تحليل المركبات الرئيسة PCA تليها طريقة خرائط التنظيم الذاتي SOM.

الجدول رقم (3-1) يبين النتائج في حالة حجم العينة (n = 3000)

Sigma = 3		Sigma = 2		Sigma = 1			
SOM	PCA	SOM	PCA	SOM	PCA	الفقدان	المعيار
21.24477	26.58473	21.09917	26.36360	21.24401	26.38033	0.1	RMSEX
28.16595	28.16641	21.88156	21.88246	19.88270	19.88280		RMSEY
2.36730	2.36715	2.35502	2.35586	2.35080	2.35083		Bo
-0.00456	-0.00474	0.00301	0.00306	0.00084	0.00082		B1
0.00169	0.00190	-0.00008	-0.00066	0.00014	0.00013		B2
29.79074	35.76747	30.01972	36.09015	30.13491	35.75867	0.2	RMSEX
27.67983	27.68110	21.66396	21.66519	20.06740	20.06834		RMSEY
2.39569	2.39340	2.40591	2.40600	2.34318	2.34384		Bo
0.00194	0.00201	0.00263	0.00286	0.00107	0.00109		B1
-0.00151	-0.00008	-0.00088	-0.00112	0.00014	-0.00028		B2
36.71944	42.66059	36.76631	42.35509	36.73772	42.30932	0.3	RMSEX
27.68745	27.68696	21.82236	21.82245	20.37860	20.37903		RMSEY
2.41812	2.41663	2.40888	2.40805	2.35616	2.35590		Bo
-0.00328	-0.00287	-0.00476	-0.00490	0.00082	0.00116		B1
-0.00226	-0.00152	0.00008	0.00069	-0.00124	-0.00130		B2
42.38401	47.16254	42.53275	46.86629	42.38895	47.04983	0.4	RMSEX
28.14414	28.14417	21.57942	21.57918	20.01201	20.01242		RMSEY
2.39970	2.39927	2.41302	2.41196	2.36721	2.36736		Bo
0.00067	0.00009	0.00287	0.00196	-0.00067	-0.00045		B1
0.00054	0.00133	0.00094	0.00233	-0.00038	-0.00063		B2

الجدول رقم(3-2) يبين النتائج في حالة حجم العينة(n = 4000)

Sigma = 3		Sigma = 2		Sigma = 1			
SOM	PCA	SOM	PCA	SOM	PCA	الفقدان	المعيار
24.57884	30.69848	24.52659	30.59508	24.49072	30.36183		RMSEX
32.00823	32.00837	25.11177	25.11215	22.96421	22.96430		RMSEY
2.40522	2.40574	2.40821	2.40858	2.36458	2.36458	0.1	Bo
0.00597	0.00582	-0.00047	-0.00048	-0.00309	-0.00314		B1
0.00128	0.00107	0.00021	-0.00001	0.00054	0.00056		B2
34.57605	41.38934	34.73819	41.28692	34.66773	41.48269		RMSEX
31.93115	31.93130	25.08366	25.08367	22.88088	22.88030		RMSEY
2.40034	2.40071	2.39065	2.39024	2.36451	2.36331	0.2	Bo
0.00024	-0.00018	0.00361	0.00392	0.00028	0.00038		B1
0.00190	0.00196	-0.00149	-0.00144	-0.00122	-0.00050		B2
42.43562	48.97716	42.35003	48.92284	42.55899	48.74211		RMSEX
32.05088	32.05202	25.33456	25.33508	23.15350	23.15358		RMSEY
2.42730	2.42791	2.39841	2.40027	2.35741	2.35814	0.3	Bo
-0.00286	-0.00304	0.00034	0.00009	0.00110	0.00111		B1
0.00076	0.00051	0.00195	0.00088	0.00047	-0.00002		B2
48.83821	54.54682	49.08363	54.47800	48.84881	54.37116		RMSEX
32.61679	32.61734	25.08913	25.08839	22.84751	22.84795		RMSEY
2.41983	2.42074	2.40121	2.40213	2.36341	2.36289	0.4	Bo
-0.00010	-0.00076	-0.00178	-0.00192	0.00016	0.00090		B1
0.00174	0.00164	0.00082	0.00032	-0.00159	-0.00183		B2

الجدول رقم (3-3) يبين النتائج في حالة حجم العينة (n = 5000)

Sigma = 3		Sigma = 2		Sigma = 1			
SOM	PCA	SOM	PCA	SOM	PCA	الفقدان	المعيار
27.29256	34.26899	27.42381	33.87566	27.40306	34.20887		RMSEX
36.16747	36.16795	27.83088	27.83088	25.71035	25.71012		RMSEY
2.41382	2.41235	2.40860	2.40900	2.36273	2.36159	0.1	Bo
0.00185	0.00200	0.00207	0.00212	-0.00328	-0.00330		B1
-0.00232	-0.00144	-0.00023	-0.00052	-0.00048	0.00029		B2
38.73983	46.32366	38.60847	46.48456	38.66675	46.28463		RMSEX
35.89933	35.89943	28.09282	28.09294	25.95932	25.95980		RMSEY
2.41071	2.41158	2.41829	2.42015	2.36797	2.36794	0.2	Bo
-0.00223	-0.00212	-0.00113	-0.00106	-0.00257	-0.00264		B1
0.00001	-0.00063	0.00102	-0.00027	0.00026	0.00034		B2
47.29509	54.41269	47.42364	54.44474	47.43532	54.55176		RMSEX
36.08338	36.08542	27.66990	27.67030	25.66001	25.66034		RMSEY
2.40946	2.40794	2.40478	2.40413	2.36982	2.37012	0.3	Bo
-0.00060	-0.00022	-0.00074	-0.00098	-0.00276	-0.00322		B1
-0.00181	-0.00112	0.00012	0.00074	0.00132	0.00148		B2
54.66361	61.07873	54.71910	60.82617	54.77897	60.57682		RMSEX
36.42409	36.42350	28.41217	28.41302	26.30715	26.30779		RMSEY
2.40234	2.40239	2.40462	2.40652	2.35767	2.35682	0.4	Bo
0.00162	0.00278	0.00079	0.00042	0.00306	0.00327		B1
-0.00192	-0.00288	0.00197	0.00099	-0.00091	-0.00052		B2

المبحث الرابع/ الاستنتاجات والتوصيات

4-1الاستنتاجات:

بعد تنفيذ تجارب المحاكاة وما تم عرضه من النتائج والتحليل في الجانب التجريبي أستنتج الباحث مايأتي :-

4-1-1 في حال جذر متوسط مربعات الخطأ لـX

1-اظهرت النتائج ان في جميع تجارب المحاكاة وعند حجوم العينات (3000,4000,5000) ونسب الفقدان(10%,20%,30%,40%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتي تليها طريقة تحليل المركبات الرئيسة (PCA) عند التباين (1,2,3) .
2-اظهرت النتائج ان طريقة تحليل المركبات الرئيسة (PCA) اسوأ طريقة لتقدير القيم المفقودة في نتائج المحاكاة.

4-1-2 في حالة جذر متوسط مربعات الخطأ لـY

1- عند حجم عينة 3000

اظهرت نتائج المحاكاة انه في حالة فقدان البيانات بنسبة(10%،20%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتيSOM تليها طريقة تحليل المركبات الرئيسة PCA عند التباين (1,2,3).

اما عند النسبة(30%)افضل طريقة هي طريقة خرائط التنظيم الذاتي SOM ثم طريقة تحليل المركبات الرئيسة PCA عند التباين (2,1)اما عند التباين(3) فان افضل طريقة هي طريقة تحليل المركبات الرئيسية PCA وثم طريقة خرائط التنظيم الذاتي SOM.

2-عند حجم عينة 4000

اظهرت نتائج المحاكاة انه في حالة فقدان البيانات بنسبة(10%، 30%) افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتي تليها طريقة تحليل المركبات الرئيسية PCA عند التباين (1،2،3).

اماعند النسبة (20%) فأن افضل طريقة عند التباين (1) طريقة تحليل المركبات الرئيسة PCA تليها طريقة خرائط التنظيم الذاتي SOM، اماعند التباين(2،3) فأن افضل طريقة خرائط التنظيم الذاتي SOM تليها طريقة تحليل المركبات الرئيسة PCA .

اما عند النسبة (40%) فأن افضل طريقة في حالة التباين (1،3) هي طريقة خرائط التنظيم الذاتي SOM تليها طريقة تحليل المركبات الرئيسة PCA، اما في حالة التباين (2) فان افضل طريقة هي طريقة تحليل المركبات الرئيسة PCA تليها طريقة خرائط التنظيم الذاتي SOM.

3- عند حجم عينة 5000

اظهرت نتائج المحاكاة انه في حالة فقدان البيانات بنسبة(10%،30%)ان افضل طريقة لمعالجة القيم المفقودة هي طريقة خرائط التنظيم الذاتي SOM تليها طريقة تحليل المركبات الرئيسية PCA عند التباين (1،2،3) .

اما عند النسبة 20% اظهرت النتائج ان افضل طريقة لمعالجة القيم المفقودة طريقة تحليل المركبات الرئيسية PCA عند التباين(1)، اما عند التباين (2،3) فأن افضل طريقة هي طريقة خرائط التنظيم الذاتي SOM تليها طريق تحليل المركبات الرئيسية PCA .

اما عند النسبة(40%) اظهرت النتائج ان افضل طريقة هي طريقة خرائط التنظيم الذاتي SOM تليها طريقة تحليل المركبات الرئيسية PCA عند التباين(1,3)،اما عند التباين (2) فان افضل طريقة هي طريقة تحليل المركبات الرئيسة PCA تليها طريقة خرائط التنظيم الذاتي SOM .

4-2 التوصيات:

في دراسات قادمة نقترح استعمال ما يلي :-

1- اسلوب بيز في طريقة تحليل المركبات الرئيسية لمعالجة القيم المفقودة .

2- طريقة تحليل المركبات الرئيسية(PCA)لمعالجة القيم المفقودة لما تبديه من كفاءة في تقدير القيم المفقودة.

3- طريقة خرائط التنظيم الذاتي لمعالجة القيم المفقودة .

4- الخرائط المولدة الطوبوغرافية لمعالجة القيم المفقودة.

4-3 المصادر:

1- حسين، علي ناصر،2012،"تقدير القيم المفقودة لمتغير الاستجابة في نموذج الانحدار الخطي المتعدد"، مجلة العلوم الاقتصادية،العدد30،المجلد الثامن، قسم الاحصاء، كلية الادارة والاقتصاد ،جامعة بغداد.

2-حمزة، سعد كاظم،2009، مقارنة بعض الطرائق اللبية في تقدير نماذج الانحدار اللامعلمي بوجود بيانات تامة وغير تامة"، رسالة ماجستير علوم في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

3-الرحيل، راتب صايل الخضر والدراسة، رياض احمد صالح، "اثر طريقتي التعامل مع القيم المفقودة وطريقة تقدير القدرة على دقة تقدير معالم الفقرات والأفراد"، المجلة الدولية المتخصصة، العدد(6)، المجلد(3)، 2014،الاردن.

4-عبد الرزاق، علي صلاح،2015،"تقدير القيم المفقودة في انموذج البيانات المزدوجة مع تطبيق عملي"، رسالة ماجستير علوم في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

5-عز الدين ،مها و حافظ، ندى محمد ، 2006،" دليل تكوين المؤشرات المركبة "، الإدارة العامة لجودة البيانات ، مركز المعلومات ودعم اتخاذ القرار ، مجلس الوزراء ، العراق.

6-العلوي، لقاء علي محمد،2003،"مقارنة مقدرات التباين المشترك الحصينة في تحليل المركبات الرئيسية"، اطروحة دكتورا فلسفة في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

7 - كشمر، ختام سوادي، 2005 ،"الطرق المختلفة لتقدير القيم المفقودة لأغراض احتساب الرقم القياسي لأسعار المستهلك"، بحث دبلوم في الاحصاء، كلية الادارة والاقتصاد ، جامعة بغداد.

8 - Tapani Raiko,D.Sc. (Tech.) and Krista Laugus .D.Sc.(Tech.),2012,"Missing Value Imputation using SubSpace Methods With Application on Survey Data",Masters Thesis ,Department of information and computer science , school of science, alto university.

Processing of missing values in survey data using Principal Component Analysis and probabilistic Principal Component Analysis methods

Abstract:

The idea of carrying out research on incomplete data came from the circumstances of our dear country and the horrors of war, which resulted in the missing of many important data and in all aspects of economic, natural, health, scientific life, etc.,. The reasons for the missing are different, including what is outside the will of the concerned or be the will of the concerned, which is planned for that because of the cost or risk or because of the lack of possibilities for inspection. The missing data in this study were processed using Principal Component Analysis and self-organizing map methods using simulation. The variables of child health and variables affecting children's health were taken into account: breastfeeding and maternal health. The maternal health variable contained missing value and was processed in Matlab2015a using Methods Principal Component Analysis and probabilistic Principal Component Analysis of where the missing values were processed and then the methods were compared using the root of the mean error squares. The best method to processed the missing values Was the PCA method.

Key words: Missing Data Problem, Methods of estimating missing values, Principal Component Analysis,self-organizing Map.