

مقارنة بعض الطرق المستخدمة في تحديد عدد المركبات الأساسية التي يتم الاحتفاظ بها

منال خليفة عبدالله حسن

قسم الإحصاء/ كلية العلوم/ جامعة غريان

Email: manal.hasan@gu.edu.ly

الملخص:

تم فحص أداء خمس طرق لتقدير عدد المركبات الأساسية التي يجب الاحتفاظ بها (طريقة Parallel Analysis ، Auer-Gervini ، طريقة Broken-Stick، عامل التسارع، واختبار بارتلت حيث تم استخدام ثلاث مستويات معنوية له (0.05، 0.01، 0.005)). تم إنتاج أربع مصفوفات ارتباط باستخدام 9 و 18 متغيراً، وتقسيمها إلى مجموعات لها أعداد مختلفة من المتغيرات لتوليد ثلاث مركبات أساسية مهمة. في هذه الدراسة تم استخدام حجم العينة 72 و 180 في حالة عدد المتغيرات 9 متغيراً، بينما استخدم حجم العينة 144 و 360 عندما كان عدد المتغيرات 18 متغيراً. بصفة عامة فإن طريقة Parallel Analysis كانت الأفضل في تقدير عدد المركبات الأساسية مقارنة بباقي الطرق. أيضاً بينت هذه الدراسة أن أداء الطرق كان أفضل عندما كان عدد المتغيرات وحجم العينة كبيرين عدا طريقة بارتلت فقد كانت الأكثر تغيراً والأكثر حساسية اتجاه عدد المتغيرات والزيادة في حجم العينة.

الكلمات المفتاحية: التقدير، عدد المركبات الأساسية، عامل التسارع، اختبار بارتلت

Comparison of some methods for determining the number of principle components to retain

Manal Khalifa Hasan

Abstract

The performance of five methods for determining the number of components to retain (Auer-Gervini, Parallel Analysis, Broken-Stick, acceleration factor, and Bartlett's test where three significant levels were used for it (0.05, 0.01, 0.005)). Four sample correlation matrices were generated with 9 and 18 variables and divided into three groups having different numbers of variables to generate three non-trivial components. In general, the Parallel Analysis method was the best in estimating the number of principle components compared to other rules. This study also showed that the performance of the methods was better when the number of variables and the sample size were large, except the Bartlett method, which was the most variable and most sensitive towards the number of variables and the increase in the sample size.

Key words: estimation, number of principles components, acceleration factor, Bartlett's test

مقدمة

يعد تحليل المركبات الأساسية أحد أكثر طرق تحليل متعدد المتغيرات استخداماً في مختلف المجالات العلمية. إن تحليل المركبات الأساسية يقوم بتحويل عدداً من المتغيرات (والتي من الممكن أن تكون مرتبطة) إلى عدد أصغر من المتغيرات الغير مرتبطة تسمى مركبات أساسية، حيث يمثل المركب الأساسي الأول أكبر قدر من التباين في البيانات قدر الإمكان، ويمثل كل مركب أساسي يليه أكبر قدر ممكن من التغير المتبقي، وبالتالي فإن الهدف الأساسي من تحليل المركبات الأساسية هو تقليل أبعاد مجموعة البيانات (عدد المتغيرات).

إن المشكلة الشائعة في استخدام تحليل المركبات الأساسية هو تحديد عدد المركبات الأساسية التي يجب الاحتفاظ بها، [1] والتي مازالت تحتفظ بالمعلومات الأساسية.

تم اقتراح العديد من الطرق المختلفة في تحديد عدد المركبات الأساسية التي يجب الاحتفاظ بها في التحليل [2][3][4] [5][6][7]. العديد من الدراسات حققت في دقة طرق تحديد عدد المركبات الأساسية من خلال استخدام أسلوب المحاكاة سواء من خلال دراسة طريقة بشكل فردي، أو من خلال مقارنة عدة طرق. [1-4][6-12] بصفة عامة فإنه في العديد من الدراسات تم استخدام مصفوفات ارتباط معينة بحيث كان عدد المركبات الأساسية معلوم، ثم بعدها تم توليد البيانات وتطبيق تحليل المركبات الأساسية. [1-4][6-7] كما وضحت بعض الدراسات بأن الطرق المقترحة في تحديد عدد المركبات الأساسية لا تعطي نفس العدد للمركبات التي يجب الإبقاء عليها في التحليل. [13][14]

ورغم كل هذه الجهود، إلا أننا نرى بأن هناك حاجة للمزيد من الدراسات لمقارنة طرق تحديد عدد المركبات الأساسية وأثر بعض العوامل على تقدير العدد الذي يجب الاحتفاظ به في التحليل. فمثلاً نجد أن قليل من الدراسات استعرضت طريقة AG من خلال مقارنتها بطرق أخرى، كذلك بعض الدراسات تستخدم في بيانات فعلية لتقارن طرق تحديد عدد المركبات الأساسية والغير معلومة.

لذلك هذه الدراسة هدفت إلى فحص خمس طرق لتحديد عدد المركبات الأساسية، والمقارنة بينهم من خلال أثر عدد المتغيرات، حجم العينة، ودرجة تشبع المتغيرات في المركبات الأساسية على تقدير عدد المركبات الأساسية.

طرق تحديد عدد المركبات الأساسية

هناك العديد من الأساليب المستخدمة في تحديد عدد المركبات الأساسية التي يجب الاحتفاظ بها في التحليل، وسنتناول خمسة أساليب والتي هي:

1- اختبار بارتلت

اقترح Bartlett سنة 1950م اختبار إحصائي اختبر فيه الفرضية بأن $p-r$ من القيم الذاتية متساوية. تكون بداية الاختبار بالمركبة الأساسية الأولى، ويتم استبعاد كل مركبة من الاختبار إلى حين الفشل في رفض الفرضية الصفرية، وعدد المركبات الأساسية التي يجب الاحتفاظ بها هي المركبات الأساسية التي تم استبعادها قبل رفض الفرضية الصفرية. هذا الاختبار يتوزع تقريباً توزيع مربع كاي، وهو اختبار يتأثر بحجم العينة، حيث توصلت دراسة [1] إلى أن قوة معنوية الاختبار تزداد بازدياد حجم العينة، وبالتالي القليل من الفروق الجوهرية بين القيم الذاتية تكون معنوية. تم استخدام اختبار بارتلت على حسب توصيات الدراسة [12] بثلاث مستويات معنوية وهم 0.05، 0.01، 0.005، وستتم الإشارة إلى كل منهم على التوالي اختبار بارتلت1، اختبار بارتلت2، و اختبار بارتلت3.

2- طريقة Broken-stick

اقترح هذه الطريقة Frontier سنة 1976، وهي تعتمد على القيم الذاتية للبيانات العشوائية، وهي تتميز ببساطة الحساب، والتقييم الدقيق لعدد المركبات الأساسية مقارنة بباقي الطرق. [15][10] حدد جاكسون سنة 1993 نموذج Broken-stick ليكون نهجاً ثابتاً لتحديد العدد المناسب من المركبات الأساسية التي يمكن استخدامها للتفسير، وذكر أن هذه الطريقة هي إحدى أفضل الطرق لاختبار عدد المركبات الأساسية. [3] يفترض هذا النموذج أن التباين الكلي يتناسب بين المركبات، وأن توزيع القيمة الذاتية المتوقع يجب أن يتبع توزيع Broken-stick، بمعنى أنه إذا تم تقسيم التباين الكلي (مجموع القيم الذاتية) بشكل عشوائي على مختلف المركبات الأساسية، فإن التوزيع المتوقع للقيم الذاتية سوف يتبع توزيع broken-stick. القيم الذاتية الملاحظة قابلة للتفسير إذا زادت عن القيم الذاتية الناتجة من طريقة BS. وللتوضيح، نفترض أنه تم كسر عصا إلى p قطعة بعد أن تم تحديد عدد $1 - p$ نقطة كسر، وتم ترتيب أطوال القطع تنازلياً، فإن الطول المتوقع b_k لأطول قطعة k يمكن حسابه من المعادلة التالية:

$$b_k = \sum_{i=k}^p \frac{1}{i}$$

رغم أن هناك جدول لتوزيع Broken-stick، إلا أنه يمكن استخدام المعادلة السابقة لحسابه. إن التناقض في قيمة b_k يمكن استخدامه كقيمة حرجة لاقتراح كم عدد المركبات الأساسية التي يمكن إرجاعها والتي تحتوي على معلومات مفيدة؟ وعلى حسب هذه القاعدة فإنه يتم استرجاع عدد المركبات الأساسية التي تكون فيها القيمة الذاتية أعلى من القيم الذاتية لـ b_k . [9][10] [16][17] وسيتم استخدام الرمز BS للإشارة لهذه الطريقة.

3- طريقة Auer-Gervini

اقترح هذه الطريقة كلاً من Auer و Gervini، وهي تعتمد في تحديدها لعدد المركبات الأساسية على نموذج بيزي Bayesian model الذي يؤكد أن متجه التباينات المفسرة (القيم الذاتية) يجب أن تكون على الشكل التالي:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > \lambda_{d+1}, \lambda_{d+1} = \lambda_{d+2} = \dots = \lambda_n < \lambda_n$$

من أجل إيجاد البعد الحقيقي d . [18] حيث إنه يتم افتراض مجموعة من التوزيعات الأولية prior distributions على $d \in \{1, 2, \dots, n\}$ ، والتي تتضائل بشكل أسّي، مع التحكم في معدل الاضمحلال بواسطة المعلمة θ . لكل قيمة θ يتم اختيار قيمة d والتي تحقق:

$$\hat{d}(\theta) = \operatorname{argmax}_{0 \leq d \leq n-1} \{ \hat{F}(d) + \theta(n-1-d) \}, \quad \hat{F}(d) = (n-d) \log \left(\frac{\hat{G}_d}{\hat{A}_d} \right)$$

حيث \hat{G}_d و \hat{A}_d هما المتوسط الهندسي والحسابي لـ $\lambda_{d+1} \dots \lambda_n$ على التوالي.

إنّ الناتج من $\hat{d}(\theta)$ هو عبارة عن دالة الخطوة step function، وعن طريق الرسم البياني يتم اختيار أعلى خطوة، والتي تمثل عدد المركبات الأساسية التي يجب الاحتفاظ بها. [18] إلا أننا في هذه الدراسة اعتمدنا على حزمة (PCDimension) في حزمة R. [19]، حيث إنه يتم استخدام عدة خوارزميات، يتم من خلالها حساب عدد المركبات الأساسية بشكل تلقائي بناءً على طريقة Auer-Gervini. وستتم الإشارة لهذه الطريقة باستخدام الاختصار AG.

4- طريقة عامل التسارع Acceleration factor

تعتمد هذه الطريقة على طريقة منحنى الصخرة (Scee plot)، والتي يتم فيها رسم القيم الذاتية لمصفوفة الارتباط في المحور الكارتيزي، حيث يمثل المحور الرأسي القيم الذاتية، والمحور الأفقي عدد المركبات الأساسية، [11]. يتم البحث على النقطة والتي يتغير فيها انحناء المنحنى بشكل واضح. هذه النقطة تسمى بالنقطة الفاصلة، حيث إنه تنتهي عندها القيم الذاتية الكبيرة، وتبدأ القيم الذاتية الصغيرة. طريقة عامل التسارع تأخذ في الاعتبار عدد المركبات الأساسية التي تكون قيمها الذاتية أعلى من النقطة الفاصلة، والتي في نفس الوقت تكون أكبر من الواحد الصحيح. عامل التسارع هو المشتقة الثانية لمنحنى الصخرة:

$$f''(i) = \frac{f(i+h) - 2f(i) - f(i-h)}{h^2}$$

حيث $f''(i) \cong f(i+1) - 2f(i) + f(i-1)$ ، ومنها يتم تحديد قيمة i التي تحقق التالي:

$$\max_i \{\lambda_{i+1} - 2\lambda_i + \lambda_{i-1}\}$$

وبالتالي فإن عدد المركبات الأساسية يكون مساوياً لـ $i - 1$ ، بشرط أن القيمة الذاتية λ_{i-1} تكون أكبر من الواحد الصحيح. [11][20]

5- طريقة Parallel Analysis

اقترح هذه الطريقة Horn سنة 1965، وهي تعتمد على توليد عدد من المصفوفات العشوائية ذات أبعاد مقابلة للبيانات الأصلية، على سبيل المثال إذا كانت البيانات الفعلية حجمها 200 وعدد متغيراتها 9، فإنه سيتم توليد مجموعة من المصفوفات العشوائية ذات بعد (8×200) . [21] يتم إيجاد القيم الذاتية لمصفوفة الارتباط للبيانات الأصلية، وحساب القيم الذاتية لكل مصفوفات الارتباط العشوائية التي تم توليدها. وبعد ذلك يتم المقارنة بين القيم الذاتية للبيانات الأصلية والقيم الذاتية للبيانات العشوائية. [8][11][21]

في الوصف الأساسي لطريقة PA فإنه يتم استخدام متوسط القيم الذاتية لمصفوفات الارتباط العشوائية كقيمة أساس للمقارنة، حيث إن القيمة الذاتية للمركبة الأساسية للبيانات الفعلية التي تكون قيمتها الذاتية أعلى من متوسط القيم الذاتية للمركبة الأساسية من البيانات العشوائية يتم الاحتفاظ بها. [22] أما بعد ذلك تم الاعتماد على المئين 95 لتوزيع القيم الذاتية للبيانات العشوائية. [29] [21] [24] يتم تحديد المئين 95 لمجموعة القيم الذاتية الأولى لمصفوفات الارتباط العشوائية ولتكن $\hat{\lambda}_1$ ، ولتكن $\hat{\lambda}_2$ تمثل المئين 95 لمجموعة القيم الذاتية الثانية لمصفوفات الارتباط العشوائية، وهكذا حتى المئين 95 لمجموعة القيم الذاتية p لمصفوفات الارتباط العشوائية ويرمز له بالرمز $\hat{\lambda}_p$ ، وكل مركبة أساسية q يتم الاحتفاظ بها في التحليل إذا كانت $\lambda_q > \hat{\lambda}_q$. [25] [26] [27] ولقد استخدم الترميز PA للإشارة لهذه الطريقة في هذه الدراسة.

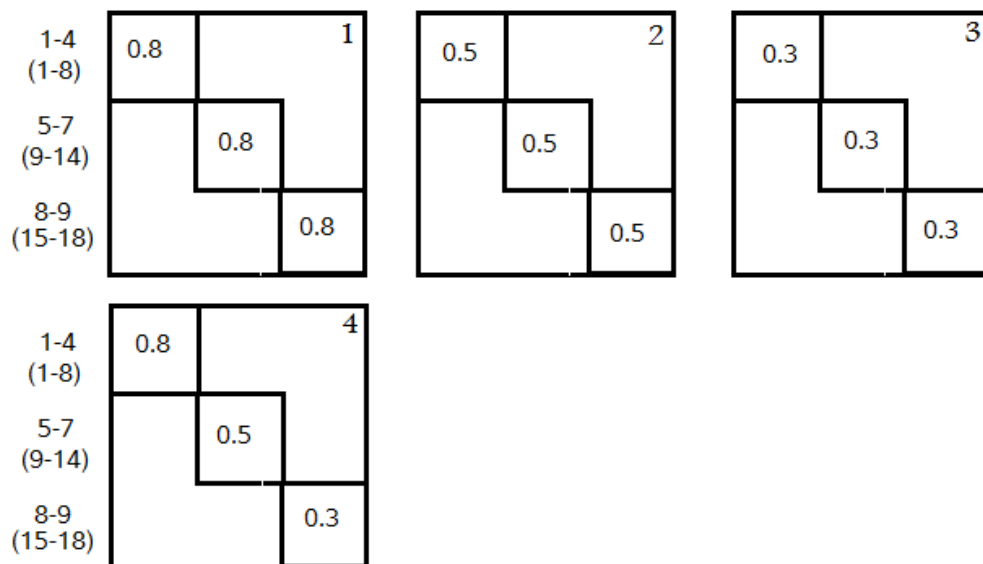
طريقة المحاكاة

تم إجراء سلسلة من عمليات محاكاة مونت كارلو. وتم استخدام نفس التصميم الذي اقترحه Perest-Neto في دراسته سنة 2005، [10] وهو نظير التصميم الذي أُستخدِم لتحديد عدد المركبات الأساسية في دراسات Zwick و Velicer سنة 1986، [8] ودراسة Jackson سنة 1993. [3]

سيتم توضيح التصميم من خلال الخطوات التالية:

استخدم في هذا التصميم حالتين لعدد المتغيرات وهما 9 و 18 متغيراً، وتم تقسيمها إلى ثلاث مجموعات لها أعداد مختلفة من المتغيرات، بحيث أن كل مصفوفة تتكون من ثلاث مركبات أساسية، وكل مجموعة من المتغيرات لها درجة تشيع معينة. تم انتاج مصفوفات ارتباط تختلف من واحدة إلى أخرى من حيث

درجة الارتباط داخل المجموعات (0.8، 0.5، 0.3)، وبين المجموعات كانت درجة الارتباط صفر سواء في حالة عدد 9 متغيرات أو 18 متغيراً، والشكل رقم (1) يعطي تفصيلاً للمصفوفات التي استخدمت. حيث إن كل مصفوفة تتكون من ثلاث مركبات أساسية



شكل (1): تمثيل بياني لمصفوفات الارتباط المستخدمة في الدراسة، فمثلا المصفوفة (4) في حالة عدد المتغيرات 9 تعني أن درجة الارتباط بين المتغيرات 1-4 ، 4-5 ، 5-7 ، و 8-9 هي 0.8، 0.5، و 0.3 على التوالي.

أيضا لدراسة أثر حجم العينة تم استخدام حجمين، ففي حالة عدد المتغيرات 9 كان حجم العينة الأولى 72 والثانية 180. أما في حالة عدد 9 متغيرات، فكان حجم العينة الأولى 144 والثانية 360.

توليد البيانات:

تم استخدام حزمة R النسخة (4.0.2) لإجراء عمليات المحاكاة، والتحليل الإحصائي في هذه الدراسة، وكانت الخطوات على النحو التالي:

(1) توليد بيانات تتبع التوزيع الطبيعي متعدد المتغيرات مع مصفوفة ارتباط محددة وحجم عينة معين. (2) تطبيق تحليل المركبات الأساسية على البيانات في كل حالة.

(3) تطبيق طرق تحديد عدد المركبات الأساسية الداخلة في هذه الدراسة وتسجيل عدد المركبات الأساسية التي يتم تقديرها في كل مرة.

(4) تكرار الخطوات الثلاثة الأولى 500 مرة.

(5) تكرار الخطوات السابقة لجميع المصفوفات، في حالة عدد المتغيرات 9 و 18، والمستويات الأربعة لحجم العينة. تم تقييم دقة طرق تحديد عدد المركبات الأساسية من خلال معيارين هما نسبة التقدير الصحيح، ومتوسط الفرق بين العدد المقدر والعدد الصحيح للمركبات الأساسية.

النتائج والمناقشة

كما أشرنا سابقاً فإنه تم استخدام خمس طرق في هذه الدراسة لتقدير عدد المركبات الأساسية من خلال توليد بيانات تتبع التوزيع الطبيعي باستخدام مصفوفات ارتباط تختلف من حيث درجة تشبع المتغيرات في المركبات الأساسية، وباستخدام عدد متغيرات وأحجام عينة مختلفة. تحت هذه القيود تم إيجاد متوسط الفرق بين القيمة التقديرية لعدد المركبات الأساسية بالطرق المستخدمة في هذه الدراسة، وعدد المركبات الأساسية الفعلي لمجتمع الدراسة وهو ثلاث مركبات أساسية، حيث إن الفرق السالب يشير إلى تقدير أقل من القيمة الفعلية Underestimation، والفرق الموجب يشير إلى تقدير أكبر من القيمة الفعلية Overestimation، أما الصفر فهو يشير إلى أن التقدير كان صحيحاً. هذه النتائج تم عرضها في الجدولين (1) و (2)، حيث الجدول (1) يعرض متوسط الفرق بين القيمة التقديرية والفعلية لعدد المركبات الأساسية لجميع الطرق في حالة حجم العينة 72 و 180، وعدد المتغيرات 9، بينما الجدول (2) في حالة حجم العينة 144 و 360 وعدد المتغيرات 18.

جدول(1): متوسط الفرق بين القيمة التقديرية من كل طريقة والعدد الفعلي لعدد المركبات الأساسية لجميع مصفوفات الارتباط لحجم العينة 72، و 180 وعدد المتغيرات 9.

Matrix	n	Bar1*	Bar2*	Bar3*	BS	AG	Ac*	PA
1	72	0.186	0	0	-0.228	0	-0.044	0
1	180	0.164	0.002	0.002	-0.092	0	0	0
2	72	0.44	-0.244	-0.29	-2.724	-0.242	-0.442	-0.14
2	180	0.064	0	0	-2.978	-0.01	0	-0.008
3	72	-0.694	-0.432	-1.08	-2.988	-1.336	-1.008	-0.962
3	180	-0.638	-0.722	-0.048	-2.994	-0.528	-0.43	-0.144
4	72	1.94	1.808	2.71	-1.472	-1.004	-1.602	-0.318
4	180	2.956	2.93	3.046	-1.348	-1.062	-1.806	-0.156

* تشير Ac إلى طريقة عامل التسارع، أما Bar1، Bar2، Bar3 إلى اختبار بارتل 1، 2، و 3 على التوالي، بينما n تشير إلى حجم العينة.

من خلال الجدول (1) يتضح أن جميع الطرق تعطي تقدير أقل من العدد الفعلي للمركبات الأساسية في هذه الدراسة عدا طريقة اختبار بارتل، حيث إنه عند المستويات المعنوية الثلاث يتذبذب تقدير عدد المركبات الأساسية بين التقدير الأقل والتقدير الأعلى من القيمة الفعلية لعدد المركبات الأساسية.

اختبار بارتل 1 أعطى تقدير أعلى من عدد المركبات الفعلي في كل الحالات عدا عند درجة تشبع 0.3 كان التقدير أقل من القيمة الفعلية، بينما كان هناك توافق في نتائج اختبار بارتل 2، واختبار بارتل 3،

فكان متوسط التقدير لعدد المركبات الأساسية 0 عند درجة التشبع 0.8 وحجم العينة 72، وعند درجة التشبع 0.5 وحجم العينة 180، وهذا يشير إلى أن في هذين الحالتين تم استرجاع عدد المركبات الأساسية بشكل صحيح في جميع المرات. أما عند درجة التشبع 0.5 وحجم العينة 72، وعند درجة التشبع 0.3 وحجم العينة 72 و 180 فالتقدير كان أقل من العدد الفعلي لعدد المركبات الأساسية، بينما في باقي الحالات كان التقدير لعدد المركبات الأساسية مبالغ فيه، أي تقدير أعلى من عدد المركبات الأساسية الفعلي.

جدول(2): متوسط الفرق بين القيمة التقديرية من كل طريقة والعدد الفعلي لعدد المركبات الأساسية لجميع مصفوفات الارتباط لحجم العينة 144، و 360 وعدد المتغيرات 18

Matrix	n	Bar1	Bar2	Bar3	Broken	Auer	Accel	Parr
1	144	0.008	0.006	0.144	0.002	0	0	0
1	360	0.018	0.01	0.166	0	0	0	0
2	144	0	0	0.054	-0.092	0	-0.012	0
2	360	0	0	0.072	-0.012	0	0	0
3	144	-0.272	-0.348	-0.01	-2.916	0.014	-0.208	0
3	360	0	0	0.012	-2.994	0	0	0
4	144	7.052	6.95	7.76	-0.868	-0.132	-1.776	-0.002
4	360	8	7.996	8.076	-0.94	-0.028	-1.918	0

أما نتائج التقدير لعدد المركبات الأساسية عند حجم العينة 144 و 360 فهي تشير إلى أن في حالة المصفوفة (1) كان متوسط فرق التقدير لعدد المركبات الأساسية إما صفر أو أعلى بقليل من القيمة الفعلية باستخدام جميع الطرق قيد الدراسة. أما في باقي درجات التشبع فكان متوسط فرق التقدير لعدد المركبات الأساسية إما صفر أو أقل من القيمة الفعلية باستخدام الطرق PA، BS، وعامل التسارع، أما طريقتي AG واختبار بارتلت فكانت القيمة تتذبذب بين تقدير أقل وأكبر من القيمة الفعلية لعدد المركبات الأساسية. وضحت النتائج أن درجة التشبع كان لها تأثير على نتائج كل الطرق المستخدمة في هذه الدراسة، حيث لم يتجاوز متوسط انحراف القيمة التقديرية من كل طريقة عن المعيار المستخدم في هذه الدراسة عن 0.23 عندما كان درجة التشبع عالية، بينما زادت قيمة متوسط الانحراف بانخفاض درجة التشبع.

جدول(3): نسبة التقدير الصحيح لعدد المركبات الأساسية لجميع مصفوفات الارتباط لحجم العينة 72، و 180 وعدد المتغيرات 9.

Matrix	n	Bar1	Bar2	Bar3	BS	AG	Ac	PA
1	72	88.4	100	100	84.2	100	97.8	100
1	180	88.8	99.8	99.8	95	100	100	100
2	72	93.8	75.6	71.2	2.2	83.4	72.4	86
2	180	95	100	100	0.2	99.2	100	99.6
3	72	1	0.6	17.6	0	26	3	4.6
3	180	36.6	30.6	84.2	0.2	61.8	71.4	85.6
4	72	6	8.6	0.6	10.2	37.6	9.2	68.4
4	180	0	0	0	10.8	41.4	6.4	84.4
1	144	99.4	99.4	89.2	99.8	100	100	100
1	360	98.8	99.4	88.8	100	100	100	100

2	144	100	100	96	91	100	99.4	100
2	360	100	100	94	98.8	100	100	100
3	144	72.8	65.2	94.2	0	98.2	86.4	100
3	360	100	100	98.8	0	100	100	100
4	144	0	0	0	15.4	90	6.8	99.8
4	360	0	0	0	6.2	97.4	3.2	100

النتائج الموضحة في الجدول (3) تعطي نسبة التقدير بشكل صحيح لطرق تحديد عدد المركبات الأساسية في هذه الدراسة. طريقة PA كانت الأفضل مقارنة بالطرق الأخرى وتليها طريقة AG ؛ حيث إن هذه الطرق الثلاث نجحت في تحديد عدد المركبات الأساسية بشكل صحيح بنسبة 89%، 84%، و 82% على التوالي في جميع السيناريوهات. نلاحظ من النتائج أن حجم العينة، ودرجة التشبع لهما أثر على التقدير، لاحظ أن أداء PA أصبح أفضل في حالة حجم العينة 144، و 360 وعدد المتغيرات 18، فقد تم ارجاع عدد المركبات الأساسية بشكل صحيح بنسبة 99%.

النتائج تشير إلى انخفاض في نسبة التقدير الصحيح بانخفاض درجة تشبع المتغيرات في المركبة الأساسية، فكانت نسبة التقدير الصحيح لعدد المركبات الأساسية تقريبا 5% باستخدام طريقة PA عندما كانت درجة التشبع 0.3 وحجم العينة 72، بينما طريقة AG بلغت 26%. كما توضح النتائج بأن نسبة التقدير بشكل صحيح زادت عندما زاد حجم العينة إلى 180 باستخدام كل الطرق عدا طريقة BS . حيث إن طريقة BS لم تتمكن من اختزال عدد المركبات الأساسية بشكل جيد إلا عندما كانت درجة التشبع 0.8 في العينتين ذات الحجم 72 و 180، فكانت نسبة التقدير الصحيح لعدد المركبات الأساسية 84% عند حجم العينة 72، وزادت إلى 95% في حجم العينة 180. أما في باقي الحالات فكان نسبة التقدير بشكل صحيح صغيرة.

أوضحت النتائج أن التقدير لعدد المركبات الأساسية صحيحاً بنسبة 100% باستخدام طريقة عامل التسارع في حالة المصفوفة (1) والمصفوفة (2)، وحجم العينة 180، إلا أنه انخفض بشكل ملحوظ في حالة المصفوفة (4)، أما في حجم العينة 72 فإنه يمكن ملاحظة انخفاض نسبة التقدير الصحيح بانخفاض درجة التشبع.

تبين النتائج أن نسبة التقدير الصحيح باستخدام اختبار بارتل عند مستويات المعنوية الثلاث المستخدمة في هذه الدراسة كانت وعلى التوالي 61%، 61% و 64%. بصفة عامة نسبة التقدير بطريقة بارتل لعدد المركبات الأساسية بشكل صحيح كان الأعلى في حجم العينة 180 مقارنة بحجم العينة 72، وكذلك اختبار بارتل 2، واختبار بارتل 3 كانا أفضل مقارنة باختبار بارتل 1 في حالة المصفوفة (1) والمصفوفة (2)، وحجم العينة 72 و 180، إلا أن هذه النسبة انخفضت في المستويات المعنوية الثلاث عند درجة التشبع 0.3 وانعدمت في حالة المصفوفة (4).

نتائج الجدول (3) توضح أن هناك تحسن في تقدير عدد المركبات الأساسية لجميع الطرق. الجدير بالذكر أن نسبة التقدير الصحيح لعدد المركبات الأساسية باستخدام كل الطرق عدا طريقة AG قد ارتفع عندما كانت درجة التشبع 0.3 عند حجم العينة 144، وعدد المتغيرات 18 مقارنة بحجم العينة 180 وعدد المتغيرات 9، وهذا يدل على أثر عدد المتغيرات على التقدير. أما طريقة AG فقد فشلت تماماً في تقدير عدد المركبات الأساسية. كذلك نلاحظ أنه عندما كان حجم العينة 360 ودرجة التشبع 0.3 كان التقدير صحيحاً بنسبة 100% باستخدام كل الطرق عدا طريقة بارتل 3 و BS.

بصفة عامة فإن طريقة PA أفضل في تقدير عدد المركبات الأساسية مقارنة بباقي الطرق قيد الدراسة. جميع الطرق كان أداؤها أفضل عندما كانت درجة التشبع عالية (المصفوفة 1)، بينما كان الأداء ضعيف عندما كانت درجة التشبع ضعيفة. أيضاً بيّنت هذه الدراسة أن أداء الطرق كان أفضل عندما كان عدد المتغيرات، وحجم العينة كبيرين عدا طريقة بارتل فقد فشلت تماماً في تحديد العدد الصحيح للمركبات الأساسية عند درجة الارتباط 0.8، 0.5، 0.3 داخل المجموعات.

أداء اختبار بارتل كان الأكثر تغيراً بين الطرق المستخدمة في هذه الدراسة، بالإضافة إلى ذلك فقد كان أكثر تأثراً اتجاه عدد المتغيرات، وكذلك درجة تشبع المتغيرات في المركبات الأساسية أكثر من تأثير حجم العينة. هذه النتيجة تختلف عن نتيجة دراسة استخدمت بيانات فعلية، وبيانات مولدة تتبع التوزيع الطبيعي، أشارت بأن اختبار بارتل يحدد عدد المركبات بشكل صحيح إذا كان العدد الفعلي للمركبات 3، [3] وهو عكس ما توصلت له هذه الدراسة. كذلك أشارت نفس الدراسة بأن طريقة بارتل قد تعطي تقدير مرتفع جداً لعدد المركبات الأساسية قد يصل إلى 11 مركبة أساسية رغم أن العدد الفعلي هو واحد فقط، [3] وهو يتطابق مع نتائج هذه الدراسة. المستويات الثلاثة من اختبار بارتل احتفظت بعدد كبير من المركبات الأساسية عند حجم العينة الكبير، وهذا أدى إلى زيادة المبالغة في التقدير. هذه النتيجة تتفق مع نتائج الدراسة [28] واختلفت مع نتيجة الدراسة [8] التي وجدت أن اختبار بارتل إلى حد ما أكثر دقة مع العينات الكبيرة مقارنة بالعينات الصغيرة. لذلك على الرغم من أن الاختبار يعمل بشكل صحيح في بعض السيناريوهات في هذه الدراسة، فقد لا يكون مناسباً للتطبيق بشكل عام، وهذا ما وصلت له بعض الدراسات الأخرى. [3] [8] رغم أنه تم استخدام مستويات معنوية مختلفة في اختبار بارتل والذي أدى إلى اختزال عدد مركبات أساسية أقل أو أكبر من عدد المركبات الأساسية الفعلي، فإن دقة نتائج هذه الطريقة لم تتأثر بشكل إيجابي ملحوظ عن طريق ضبط مستويات المعنوية مع حجم العينة.

نتائج طريقة BS كانت جيدة عندما كانت درجة تشبع المتغيرات في المركبات الأساسية مرتفعة، بينما تأثرت هذه الطريقة بشكل سلبي بانخفاض درجة التشبع حتى عندما كان حجم العينة وعدد المتغيرات كبيرين، وهذا بديهي لأن هذه الطريقة لا تعتمد على حجم العينة. [6] هذه الطريقة قد أعطت

تقدير أقل لعدد المركبات الأساسية من المعيار المستخدم في هذه الدراسة تحت جميع الشروط عدا في حالة المصفوفة (1). نتائج هذه الدراسة تتفق وتختلف عن ما توصلت إليه دراسة استخدمت درجتين من الارتباط (0.8 و 0.3)، وتوصلت إلى أن طريقة BS أعطت في الغالب تقديراً صحيحاً، ولم تعط تقديراً مبالغ فيه تحت الشروط المستخدمة، وأنه الأفضل مقارنة بالطرق المستخدمة. [3] أيضاً هناك دراسات قد توصلت إلى أن طريقة BS تميل إلى تقدير عدد مركبات أساسية أقل من العدد الفعلي. [3-4][29] ومع ذلك، دراسة تدعي أن طريقة BS قد حددت بدقة العدد الصحيح للمركبات الأساسية في ثلاث من المصفوفات الأربعة المستخدمة، بينما أعطت الطريقة تقديراً أقل في الأخرى، وذكرت هذه الدراسة أن هذه الطريقة تعتبر من أكثر الطرق دقة مقارنة بالطرق الأخرى قيد الدراسة. [3] دراسة تقترح أن الطريقة التي تعطي تقدير عدد المركبات الأساسية أقل من العدد الفعلي طريقة مقبولة للاستخدام طالما أن استخدام تحليل المركبات الأساسية في المقام الأول للأغراض الوصفية وليس التنبؤية. [4]

انخفاض درجة التشبع كان لها الأثر السلبي على طريقة عامل التسارع عندما كان عدد المتغيرات صغير في هذه الدراسة. النتائج توضح أن نسبة التقدير الصحيح ارتفعت باستخدام طريقة عامل التسارع بزيادة حجم العينة، هذه النتيجة تتوافق مع نتائج دراسة بينت أن كلما زاد حجم العينة كلما كانت طريقة عامل التسارع أدق في التقدير. [11] أيضاً نفس الدراسة أشارت إلى أنه طريقة عامل التسارع تعطي تقدير أقل من العدد الفعلي للمركبات الأساسية في حال كان حجم العينة ليس كبيراً، هذه النتيجة أيضاً مكافئة لما توصلت له هذه الدراسة، فقد لوحظ أنه في حالة حجم العينة 7 كانت نسبة التقدير الصحيحة صغيرة جداً في حالة المصفوفة (3)، ولكن بصفة عامة فإن نتائج الدراسة بينت دقة هذه الطريقة في عدة حالات. أدت الزيادات في حجم العينة وتشبع المكونات وعدد المتغيرات لكل مركبة أساسية إلى تحسين أداء طريقة PA. بالإضافة إلى أن هذه الطريقة لم تعط تقدير أعلى من المعيار المستخدم في هذه الدراسة تحت كل الشروط. هذه النتيجة عكس نتائج الدراسة [8]، فقد توصلت إلى أن طريقة PA تعطي ميول نحو التقدير الأعلى وأن أكثر من 50% من الأخطاء في التقدير باستخدام هذه الطريقة يكون تقدير مبالغ فيه. دقة طريقة PA في هذه الدراسة متوافقة مع نتائج دراسات أخرى [8] و [30].

طريقة PA تعتبر بديلاً ممتازاً للطرق الأخرى المستخدمة بشكل شائع لنفس الغرض مثل اختبار Scree أو قاعدة Kaiser، ومع ذلك فإن هذه الطريقة غير معروفة جيداً بين الباحثين، ويرجع ذلك لعدم توفره كخيار تحليل في الحزم الإحصائية شائعة الاستخدام. [5] إن العيب الرئيسي في طريقة PA هو الحاجة إلى توليد العديد من المصفوفات العشوائية، ولكن في النهاية تبقى الطريقة التي توصي بها عدة

دراسات نظراً لدقة نتائجها. [5-8]

الخاتمة:

توصلت هذه الدراسة إلى أن طريقة PA أفضل في تقدير عدد المركبات الأساسية مقارنة بباقي الطرق قيد الدراسة. جميع الطرق كان أداءها أفضل عندما كانت درجة التشبع عالية (المصفوفة 1)، بينما كان الأداء ضعيف عندما كانت درجة التشبع ضعيفة. أيضا بينت هذه الدراسة أن أداء الطرق كان أفضل عندما كان عدد المتغيرات، وحجم العينة كبيرين عدا طريقة بارتلت كانت الأكثر تغيراً. وبشكل عام، نوصي باستخدام أكثر من قاعدة واحدة لتحديد عدد المركبات الأساسية، على سبيل المثال طريقة PA وطريقة AG لتحديد عدد المكونات التي يجب الاحتفاظ بها للاستخدام في أي تحليل. إن استخدام طريقتين من شأنه أن يضيف قوة إلى قرار تحديد عدد المركبات الأساسية التي يجب الاحتفاظ بها. إن الاستخدام الروتيني لتحليل المركبات الأساسية، وطرق تحديد عدد المركبات الأساسية الأخرى لمستخدمي التقنيات متعددة المتغيرات سيسمح بمزيد من الثقة في النتائج.

المراجع:

- [1] Gorsuch, R. L. (1973). Using Bartlett's significance test to determine the number of factors to extract. *Educational and Psychological Measurement*, 33(2), 361-364.
- [2] Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate behavioral research*, 17(2), 253-269.
- [3] Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8), 2204-2214.
- [4] Ferré, L. (1995). Selection of components in principal component analysis: a comparison of methods. *Computational Statistics & Data Analysis*, 19(6), 669-682.
- [5] Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical assessment, research, and evaluation*, 12(1), 2.
- [6] Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, 70(6), 885-901.
- [7] Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1), 99-106.
- [8] Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3), 432.
- [9] Caron, P. O. (2016). A Monte Carlo examination of the broken-stick distribution to identify components to retain in principal component analysis. *Journal of Statistical Computation and Simulation*, 86(12), 2405-2410.
- [10] Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4), 974-997.
- [11] Sheytanova, T. (2015). A Monte Carlo Study Comparing Three Methods for Determining the Number of Principal Components and Factors.
- [12] Horn, J. L., & Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research*, 14(3), 283-300.

- [13] Cattell, R. B., & Vogelmann, S. (1977). A Comprehensive Trial Of The Scree And Kg Criteria For Determining The Number Of Factors. *Multivariate behavioral research*, 12(3), 289–325.
- [14] Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, 33(1), 37–72.
- [15] Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology direct*, 2(1), 1-21.
- [16] Wang, M., Kornblau, S. M., & Coombes, K. R. (2018). Decomposing the apoptosis pathway into biologically interpretable principal components. *Cancer informatics*, 17, 1176935118771082.
- [17] Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2003). Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology*, 84(9), 2347-2363.
- [18] Auer, P., & Gervini, D. (2008). Choosing principal components: a new graphical method based on Bayesian model selection. *Communications in Statistics—Simulation and Computation*®, 37(5), 962-977.
- [19] Coombes, K. R., Wang, M., & Coombes, M. K. R. (2019). Package ‘PCDimension’. Website: <https://cran.r-project.org/web/packages/PCDimension/PCDimension.pdf>.
- [20] Raïche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J. G. (2013). Non-graphical solutions for Cattell’s scree test. *Methodology*.
- [21] Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, 55(3), 377-393.
- [22] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185.
- [23] Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement*, 53(3), 585-596.
- [24] Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement*, 58(4), 541-568.
- [25] Saccenti, E., & Camacho, J. (2015). Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, 149, 99-116.
- [26] O’connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer’s MAP test. *Behavior research methods, instruments, & computers*, 32(3), 396-402.
- [27] Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1), 99-106.
- [28] Crawford, C. B. (1975). Determining the number of interpretable factors. *Psychological Bulletin*, 82(2), 226.
- [29] Bartkowiak A: How to reveal the dimensionality of the data? *Applied Stochastic Models and Data Analysis* 1991, 55-64.
- [30] Humphreys, L. G., & Montanelli Jr, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10(2), 193-205.