AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

مقارنة لبعض طرائق تقدير مصفوفة التباين والتباين المشترك للبيانات الكبيرة مع تطبيق عملي

A Comparison of Some Estimation Methods for Var-Covariance matrix in Big Data with Application

جلال وحيد جاسم[1] أ.م .د. أحمد مهدي صالح[2]

Jalal Waheed Jassem Ahmed Mahdi Salih

amahdi@uowasit.edu.iq

كلّية الإدارة الاقتصاد / جامعة واسط

المستخلص

تمثّل عملية تقدير مصفوفة التباين والتباين المشترك حجر أساس في الكثير من التحليلات الإحصائية لمجالات علمية عدّة. وظهرت العديد من الطرائق التي تُعنى بتقدير مصفوفة التباين والتباين المشترك بالاعتماد على أسلوب الإمكان الأعظم، إذْ عُدّتْ تلك الطرائق بالطرائق الكلاسيكية، إذْ تزداد عملية تقدير مصفوفة التباين والتباين المشترك صعوبة بازدياد عدد المتغيّرات تحت الدراسة. في بحثنا هذا استعملنا طريقتين لتقدير مصفوفة التباين والتباين المشترك في ظلّ وجود بيانات كبيرة تحت الدراسة هما: مقدّر الحزمة BE، و كذلك المقدّر المستدقّ TE. وتَمّ هنا استعمال بيانات محاكاة، و كذلك بيانات حقيقية لمستويات المعيشة، تَمّ الحصول عليها من وزارة التخطيط العراقية. وتَمّ استعمال معيار النسبة المئوية للتحسّن في متوسّط الخسارة (PRIAL). و تَمّ التوصّل إلى أنّ مُقدّر الحزمة BE يعمل أفضل من المُقدّر المستدقّ TE تحت وجود بيانات كبيرة تحت الدراسة.

Abstract

The of the var-covariance matrix estimating is very important and represents a cornerstone in many statistical methods for several scientific fields, as many methods have emerged that are concerned with estimating the var-covariance matrix based on method of maximum likelihood. Nowadays these methods are classical methods and the process of estimating the covariance difficult with the increase in the number of variables under study. In this research, we used two

methods to estimate the var-covariance matrix in big data in tow methods . first is Banding Estimator BE and the Tapering Estimator TE. Simulation study were used, as well as real data for living standards obtained from The Iraqi Ministry of Planning, the percentage improvement in average loss (PRIAL) criterion was used, as it was concluded that the BE Banding Estimator works better than the TE Tapering Estimator under big data conditions.

1- المقدّمة

يُعدّ تقدير مصفوفة التباين والتباين المشترك أمراً ذا أهمّية في العديد من مشكلات الاستدلال الإحصائي مثل: تحليل المركبات الرئيسة، التصنيف عن طريق التحليل التمييزي، استدلال النماذج الرسومية وغيرها. ويكون الاهتمام أحيانا بتقدير البنية الذاتية لمصفوفة التباين والتباين المشترك وأحيانا في تقدير معكوسها. إنّ تقدير مصفوفة التباين والتباين المشترك من عينات البيانات الكبيرة هو مشكلة أساس في التحليل الإحصائي متعدّد المتغيّرات؛ لأنّه يتعامل مع البيانات بوصفها وحدات متعدّدة القياس أو متّجهات عشوائية. إنّ الطرائق التقليدية لتقدير مصفوفة التباين والتباين المشترك للبيانات الكبيرة ستعطي تقديرا متحيّزا، ويكون التقدير بعيدا عن قيمته الحقيقة. ومن هذه الطرائق: طريقة الإمكان الأعظم، طريقة انموذج العوامل وطريقة تقدير بييز وغيرها [1]. وقد دأب الباحثون على إيجاد طرائق تقدير مناسبة لمصفوفة التباين والتباين المشترك. وقَدّم Fan آخرون عام 2008 مُقدّرا مستدقّا لمصفوفة التباين والتباين المشترك. وكذلك قدّم Fran وآخرون عام 2020 مُقدّر غير خطّي لمصفوفة التباين والتباين المشترك.

وفي بحثنا هذا قمنا باستعراض مفعوم مصفوفة التباين والتباين المشترك و كذلك مفهوم البيانات الكبيرة و مشاكلها فضلاً عن طريقتي التقدير لمصفوفة التباين والتباين المشترك.

2-الهدف من البحث

1) تقدير مصفوفة التباين والتباين المشترك باستعمال بعض طرائق التقدير، وهما: مُقدّر الحزمة، والمُقدّر المستدقّ.
2) المقارنة بين الطريقتين باستعمال معيار النسبة المئوية للتحسّن في متوسّط الخسارة (PRIAL).

3- مصفوفة التباين والتباين المشترك

يرمز لمصفوفة التباين والتباين المشترك بـ(∑)، وهي من الدرجة (p×p) وتحتوي على تباين المتغيّرات العشوائية على عناصر القطر الرئيس، وتغايرات المتغيرات العشوائية على باقي عناصر المصفوفة. تتميّز هذه المصفوفة بكونها مصفوفة متماثلة ومعرفة غير سالبة، ومحيداتها الرئيسة جميعها وقيمها الذاتية غير سالبة [5] .

عند سحب أي عيّنة عشوائية ذات متغيّرات متعدّدة، تتكوّن من n من المتّجهات العشوائية، وكُلّ متّجه يتكوّن من p من المتغيّرات العشوائية. وعليه تكون مصفوفة المعلومات كالآتي [2] :

$$X = (X_1, X_2, \cdots, X_n)^{\square}$$

$$= \begin{pmatrix} X_1^{\square} \\ X_2^{\square} \\ \vdots \\ X_n^{\square} \end{pmatrix}$$

$$= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

إذْ إنْ متوسّط المتّجه العشوائي يكون كالآتي:

$$\underline{u}=E(\underline{X}) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ . \\ . \\ E(x_p) \end{bmatrix}$$

وتكون مصفوفة التباين والتباين المشترك كالآتي:

$$Cov(X) = E[(X-u)\,(X-u)^{\square}]$$

$$= \begin{bmatrix} var(x_1) & cov(x_1,x_2) & \cdots & cov(x_1,x_p) \\ cov(x_1,x_2) & var(x_2) & \cdots & cov(x_2,x_p) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_p,x_1) & cov(x_p,x_2) & \cdots & var(x_p) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} = \Sigma$$

إنّ البيانات الكبيرة قد تعمل على استحالة تحديد بنية التباين المشترك غير المعروفة. الأمر الذي يعيق تفسير النتائج. لذلك في الأدبيات تَمّ تقديم مناهج التنظيم لتقدير مصفوفة التباين والتباين المشترك الكبيرة ومعالجتها سواء أكان من الناحية النظرية أم التطبيقية. يقترح بعض الباحثين تقليص نموذج مصفوفة التباين والتباين المشترك نحو مصفوفة الوحدة، في حين يقترح البعض الآخر تطبيق تحويلات غير خطّية لعيّنة القيم الذاتية أو تنظيمها عن طريق تقسيم العيّنة. في حين يفكر بعض الدارسين في تقليص نموذج مصفوفة التباين والتباين المشترك، أيْ انكماش عناصر خارج القطر الرئيس تدريجيا نحو الصفر. في

الوقت نفسه هنالك نهج شائع يتمثّل في تشجيع التناثر، إمّا من خلال نهج الإمكان الأعظم الجزائي أو عن طريق تحديد نموذج مصفوفة التباين والتباين المشترك بأحد الطرائق المختلفة، وهي: (العتبة الصلبة، العتبة الناعمة، العتبة المعممة، والعتبة التكيفية). وهنا سنتطرق إلى طريقتين هما: مُقدّر الحزمة، والمُقدّر المستدقّ.

4- مفهوم البيانات الكبيرة

يتميّز العقدان المعاصران بتطوّر التكنولوجيا التي تولّد بيانات ومعلومات هائلة ، وتنقل وتخزن عبر شبكات واسعة ومعقّدة. و هذه التطوّرات والمعلومات أدّت إلى إحداث تغيير جذري في حياة اليوم، إذْ تستعمل غالبية القطاعات العامّة والخاصّة الأجهزة والإجراءات الرقمية لتزويد عملائها بخدمات عالية الجودة والموثوقية. نتج عن هذا الاستعمال الواسع النطاق حجمٌ لا يمكن تصوّره من البيانات التي يَتُمّ إنشاؤها ومعالجتها وتخزينها. إنّ أهمّية مثل هذه البيانات الكبيرة وحساسيتها قد حّولتها إلى هدف لا يقدّر بثمن لمخترقي الشبكات. [4]

شهدت قدرات جمع البيانات في السنوات الأخيرة نموّا هائلا بحيث أصبح تخزين البيانات المولّدة وتحليلها بشكل تدريجي أكثر صعوبة، وتأثّرت العديد من المجالات بالتوافر المتزايد للبيانات. تشير البيانات الكبيرة إلى مجموعات من البيانات الكبيرة جدا، والتي لا يمكن التعامل معها باستعمال الأجهزة التقليدية، وهي تمثّل مشكلة بالنسبة إلى البرامج مثل أنظمة قواعد البيانات والحزم الإحصائية وما إلى ذلك. [2]

تجلب البيانات الكبيرة فرصا جديدة للمجتمع الحديث وتحدّيات لعلماء البيانات. من ناحية كبيرة تحمّل البيانات الكبيرة وعودا كبيرة؛ لاكتشاف الأنماط الدقيقة وأوجه عدم التجانس للمجتمعات، التي لا يمكن اكتشافها مع البيانات صغيرة الحجم. ومن ناحية أخرى يقدّم الحجم الهائل للعيّنة والأبعاد العالية للبيانات تحديات حسابية وإحصائية فريدة منها: قابلية التوسّع والتخزين، الاختناق، تراكم الضوضاء، الارتباط الزائف وأخطاء القياس. [6].

ذكر العديد من المؤلّفين تعريفات للبيانات الكبيرة تبعا لأبعادها. مثل:

Schrock وآخرون عرّفوا البيانات الكبيرة بأنّها عبارة عن مزيج من الحجم، السرعة، التنوّع والصدق الذي يخلق فرصة للمؤسّسات لتحقيق مكاسب تنافسية في أسواق اليوم الرقمية [7] .

Chang and Gray عرّفا البيانات الكبيرة بأنّها مجموعة بيانات واسعة النطاق بشكل أساس في خصائص الحجم والسرعة و/ أو التنوع التي تتطلّب توسيع البنية للتخزين الفعّال والمعالجة والتحليل.

Boy and Crawford عرّفا البيانات الكبيرة بأنّها ظاهرة ثقافية، تكنولوجية وعلمية تستند إلى تفاعل التكنولوجيا والتحليل والأساطير.

Chen وآخرون عرّفوا البيانات الكبيرة بأنّها البيانات والتقنيات التحليلية في التطبيقات الكبيرة والمعقّدة لدرجة أنّها تتطلّب تقنيات تخزين وإدارة وتحليل وتصور للبيانات متقدّمة وفريدة من نوعها.[3]

5- مُقدّر الحزمة Banding Estimator BE

إذا كان هنالك ترتيب طبيعي لعناصر المتّجه X_i (على سبيل المثال إذا كان X_i يحتوي على متغيّرات مكانية أو زمنية)، فمن الطبيعي أنّ العناصر خارج القطر الرئيس في مصفوفة التباين والتباين المشترك Σ ، تكون متناقصة في المقدار كلّما ابتعدت عن القطر الرئيس. وتسمّى هذه المصفوفة بمصفوفة التباين والتباين المشترك ذات الحزم. تكون هذه العناصر صغيرة جدا بحيث يمكن تنظيم حزم تلك العناصر ومساواتها إلى الصفر. مع هذه الفكرة قدّم الباحثان Bickel and Levina (2008) فرضية أنّ مصفوفات التباين تتّبع الفئة الآتية **[9]** :

$$\mathcal{U}(\varepsilon_0, \alpha, C) = \{\Sigma: \max_j \sum_i \{|\sigma_{ij}|: |i-j| > k\} \leq Ck^{-\alpha} for\ all\ k > 0, and\ 0 < \varepsilon_0 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq {}^{1}/_{\varepsilon_0}\} \quad (1)$$

إذْ تمثّل $\mathcal{U}(\varepsilon_0, \alpha, C)$ دالّة الفئة و أنّ $\lambda_{min}(\Sigma)$ و $\lambda_{max}(\Sigma)$ أصغر قيمة وأكبر قيمة من القيم الذاتية لـ Σ. ثُمّ افترضا نطاق مصفوفة التباين والتباين المشترك للعيّنة $\boldsymbol{S} = (\boldsymbol{s_{ij}})$ ولأيّ قيمة لـ k بحيث إنّ $\mathbf{0} < k < p$ ليكون المُقدّر ذو النطاقات كما يأتي:

$$\hat{\Sigma}_k = B_k(S) = (\boldsymbol{s_{ij}} I(|i-j| \leq k)) \quad (2)$$

إنّ المُقدّر آنف الذكر يكون لبعض قيم k لذا وضع الباحثان النظرية الآتية [10]:

نظرية (1): بفرض أنّ مصفوفة التباين والتباين المشترك من فئة $\mathcal{U}(\varepsilon_0, \alpha, C)$ المعرفة في المعادلة رقم (2-43)، و X يتضمّن بيانات كاوسية أو بيانات ذات الذيل الخفيف فإذا كان

$k = k_n \asymp (n^{-1}\log(p))^{-1/(2(\alpha+1))}$ فإنّ:

$$\left\|\hat{\Sigma}_{k_n,p} - \Sigma_p\right\| = O_p\left(\left(\frac{\log(p)}{n}\right)^{\alpha/(2(\alpha+1))}\right) = \left\|\hat{\Sigma}^{-1}_{k_n,p} - \Sigma_p^{-1}\right\| \quad (3)$$

وبشكّل موّحد $\Sigma \in \mathcal{U}(\varepsilon_0, \alpha, C)$.

أمّا إذا كان $\gamma(\alpha) = (1 + \alpha + 2/\beta)^{-1}$ إذْ إنّ $k_n \asymp (n^{-1/2}p^{2/\beta})^{-\gamma(\alpha)}$ فإنّ:

$$\left\|B_{k_n}(\hat{\Sigma}) - \Sigma_p\right\| = O_p\left(\left(n^{-1/2}p^{2/\beta}\right)^{\alpha\gamma(\alpha)}\right) \quad (4)$$

إنّ معدّل k_n لا يزال بشكل مقارب هو الأمثل. ومن الملاحظ أنّ k_n المثلى لا تعتمد بشكل عام على n و p فقط وإنّما تعتمد على هيكلية النموذج المعتمدة أيضا والمُمَثّل عنها بـ α. إنّ النتائج المذكورة آنفا أعطتنا معدّل k_n الذي يضمن تقارب مُقدّر

الحزمة $\hat{\Sigma}_k$ ، لكّنها لا تقدّم توجيهات عملية كثيرة لاختيار قيمة k لمجموعة بيانات معيّنة. إنّ الطريقة القياسية لتحديد معلمة ضبط لتقليل المخاطرة.

$$R(k) = E\left\|\hat{\Sigma}_k - \Sigma\right\|_{(1,1)} \quad (5)$$

مع اختيار k الاوراكل المعطى بـ:

$$k_O = arg\ min_k\ R(k) \quad (6)$$

اقترح الباحثان مخطّطا لتقدير المخاطرة وk الاوراكل عن طريق تقسيم العيّنة الاصلية على عينتين عشوائيا واستعمال مصفوفة التباين والتباين المشترك لإحدى العينتين لاختيار أفضل k للعيّنة الأخرى. ليكون مُقدّر المخاطرة كالآتي:

$$\hat{R}(k) = \frac{1}{N}\sum_{v=1}^{N}\left\|B_k(\hat{\Sigma}_1^{(v)} - \hat{\Sigma}_2^{(v)}\right\|_{(1,1)} \quad (7)$$

إنّ مُقدّر k اختير كـ:

$$\hat{k} = arg\ min_k\ \hat{R}(k) \quad (8)$$

إذْ أن $\hat{\Sigma}_1^{(v)}$ و$\hat{\Sigma}_2^{(v)}$ مصفوفتي التباين والتباين المشترك للعينتين من القسم العشوائي $\boldsymbol{v}$.

N عدد الأقسام العشوائية.

إنّ k الاوراكل يعدّ أفضل اختيار من حيث الخسارة المتوقّعة. في حين يحاول $\hat{k}$ التكّيف مع البيانات المتوفّرة.

6- المُقدّر المستدقّ Tapering Estimator TE

تظهر في مصفوفة التباين والتباين المشترك ذات الحزم البسيطة مشكلة واحدة، وهي الافتقار إلى تأكيد خاصّية التعريف الموجب (positive definiteness). مع ذلك أشار الباحثان Furrer and Bengtsson [10] إلى أنّه يمكن المحافظة على خاصّية التعريف الموجب من خلال استدقاق مصفوفة التباين والتباين المشترك، وذلك عن طريق استبدال مُقدّر الحزمة $\hat{\Sigma}_p$ بـ $\hat{\Sigma}_p * R$. إذْ أنّ $*$ يشير إلى ضرب هادامار أو Schur matrix multiplication وإنّ $R = [r_{ij}]$ مصفوفة متماثلة معّرفة موجبة [10].

قَدّم الباحثون Chi. T, Zhang .C and Zhou .H إجراءات الاستدقاق لتقدير مصفوفة التباين والتباين المشترك [10]. متبعين Bickel and Levina (2008) إذْ عدّوا مصفوفة التباين والتباين المشترك تّتبع الفئة الآتية:

$$\mathbb{F}_{\alpha} = \mathbb{F}_{\alpha}(\boldsymbol{M_O}, \boldsymbol{M}) = \left\{\Sigma: \boldsymbol{max_j \sum_i}\left\{\left|\boldsymbol{\sigma_{ij}}\right|: \boldsymbol{|i-j| > k}\right\} \leq \boldsymbol{Mk^{-\alpha}\ for\ all\ k,\ and\ \lambda_{max}}(\Sigma) \leq \boldsymbol{M_O}\right\} \quad \textbf{(9)}$$

إذْ إنّ $M > 0\ , M_O > 0$ و $\alpha > 0$ مع ملاحظة أنّ λ_{min} (أصغر قيمة مميزة لأيّ مصفوفة تباين وتباين مشترك تّتبع الفئة $\mathcal{F}_\alpha$) يسمح لها أنْ تساوي الصفر، وهذا أكثر عمومية من الفرضية (1). إنّ المعلمة α في المعادلة رقم (9) التي تحّدد معدّل اضمحلال التباينات المشتركة σ_{ij} كلّما ابتعدت عن القطر الرئيس، يعتمد عليها بشكل مؤكّد معدّل التقارب الأمثل لتقدير مصفوفة التباين والتباين المشترك ضمن الفئة $\mathcal{F}_\alpha$. قدّم الباحثون مُقدّرا مستدقّ لمصفوفة التباين والتباين المشترك مبنيا على استدقاق مُقدّر الإمكان الأعظم، فعندما k عدد صحيح زوجي وإنّ $1 \le k \le p$ يكون المُقدّر المستدقّ كما يأتي:

$$\hat{\Sigma} = \hat{\Sigma}_k = \left(w_{ij}\sigma^*_{ij}\right)_{p\times p} \qquad (10)$$

إذْ ان σ^*_{ij} مدخلات مُقدّر الإمكان الأعظم Σ^*، وإنّ الاوزان تكون:

$$w_{ij} = k_h^{-1}\{(k - |i-j|)_+ - (k_h - |i-j|)_+\} \qquad (11)$$

إذْ إنّ $k_h = k/2$، $(x)_+ = max(x, 0)$ وإنّ الاوزان يمكن كتابتها كما يأتي:

$$w_{ij} = \begin{cases} 1, & when\ |i-j| \le k_h \\ 2 - \dfrac{|i-j|}{k_h}, & when\ k_h \le |i-j| \le k \\ 0, & otherwise. \end{cases}$$

إنّ هذا المُقدّر المستدقّ يختلف عن المُقدّر ذي النطاقات السابق، إذْ يمكن كتابته على أنّه مجموع مصفوفات عدّة للكتل الصغيرة على طول القطر الرئيس للمصفوفة. وتعرّف مصفوفات الكتل كما يأتي:

$$M_l^{*(m)} = \left(\sigma^*_{ij} l\{l \le i \le l+m, l \le j \le l+m\}\right)_{p\times p} \qquad (12)$$

إنّ:

$$S^{*(m)} = \sum_{l=1-m}^{p} M_l^{*(m)} \qquad (13)$$

لكلّ عدد صحيح $1 - m \le l \le p$ و $m \ge 1$.

وعليه يمكن كتابة المُقدّر المستدقّ من المعادلة رقم (10) بالصورة الآتية:

$$\hat{\Sigma}_k = k_h^{-1}\left(S^{*(k)} - S^{*(k_h)}\right) \qquad (14)$$

من الواضح أنّ المُقدّر المستدقّ $\hat{\Sigma}_k$ يعتمد على اختيار معامل الاستدقاق أو التناقص k. إذْ يعتمد الاختيار الأمثل لهذا المعامل على المعيار الذي تَمّ بموجبه قياس خطا التقدير فيما إذا كان معيار العامل *Operator norm* أو معيار فروبنيوس

Frobenius norm. وقد أثبت الباحثون Chi. T, Zhang .C and Zhou .H أنّ المُقدّر المستدقّ يصل إلى المعدّل الأمثل للتقارب بموجب المعيارين السابقين عند استعمال معامل الاستدقاق الأمثل الآتي:

$$k = n^{1/(2\alpha+1)} \qquad (15)$$

7- أسلوب المقارنة

كما سَيتُمّ المقارنة بين المُقدّرين باستعمال معيار النسبة المئوية للتحسّن في متوسّط الخسارة (PRIAL) وصيغته كالآتي [8] :

$$PRIAL(\hat{\Sigma}_a) = \frac{E\|\hat{S}-\Sigma\|^2 - E\|\hat{\Sigma}_a-\Sigma\|^2}{E\|\hat{S}-\Sigma\|^2}.100\% \qquad (16)$$

إذْ إنّ $\hat{S}$ مُقدّر مصفوفة التباين والتباين المشترك بأقلّ متوسّط لمربعات الخطأ (MSE)، وعادة ما يستعمل مُقدّر الإمكان الأعظم.

$\hat{\Sigma}_a$ مُقدّر مصفوفة التباين والتباين المشترك الطرائق السابقة والمراد المقارنة بينهما.

إنّ PRIAL هو مقياس متعلّق بالخطأ التربيعي المتوّقع لمُقدّر العيّنة $E\|\hat{S}-\Sigma\|^2$ فإذا كانت قيمة المقياس تساوي 100%، فهذا يعني أنّه لا يوجد خطأ في المُقدّر الجديد $E\|\hat{\Sigma}_a-\Sigma\|^2$. فعند المقارنة بين المُقدّرين (مُقدّر الحزمة، والمُقدّر المستدقّ) فإنّ أيّهما يمتلك PRIAL أكبر فإنّه الأفضل.

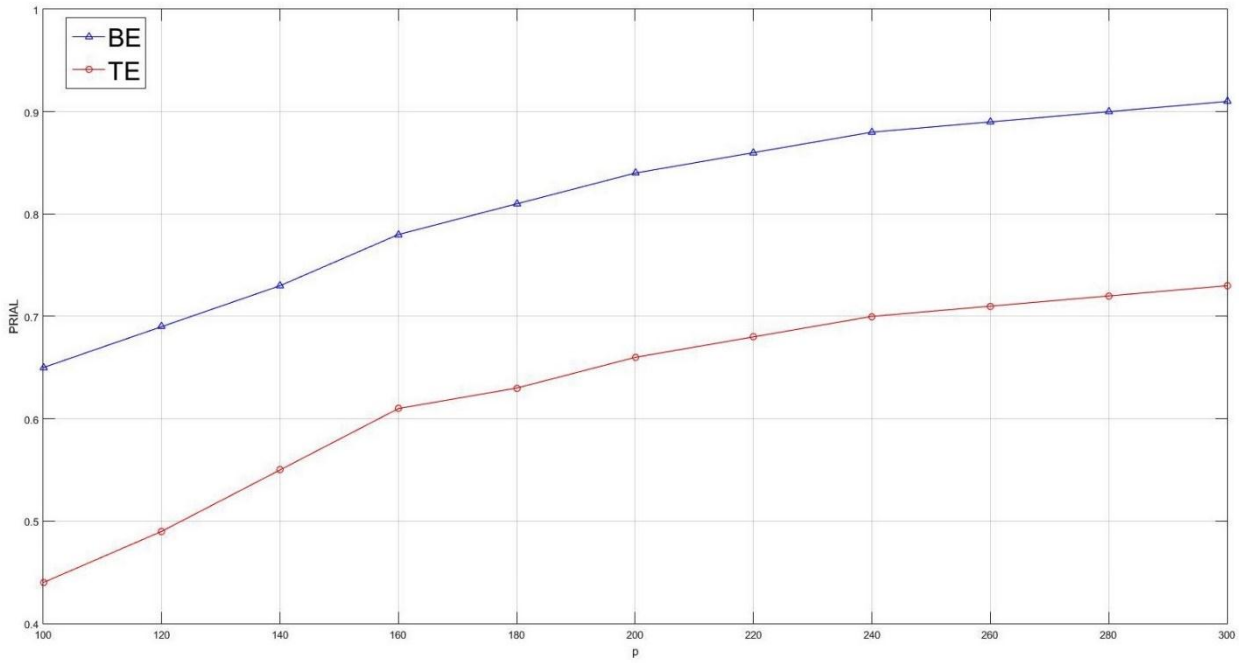
***8*- المحاكاة والبيانات الحقيقية**

تَمّ إجراء دراسة لمحاكاة بيانات كبيرة حقيقية تَمّ تجميعها من الجهاز المركزي للإحصاء، وتمثّل عيّنة بحجم 2000 مجموعة من الأسر في مختلف المحافظات العراقية. وتَمّ تسجيل 200 متغيّر تمثّل مؤشّرات المعيشة بشكل عام (الصحية والتعليمية والاقتصادية)، إذْ تَمّ في تكرار تجربة المحاكاة 1000 مرة تَمّ فيها تثبيت حجم العيّنة وتوليد أعداد مختلفة من المتغيّرات $p = 100, 120, 140, \ldots..300$ وتَمّت مقارنة فيما بينها بالاعتماد على معيار النسبة المئوية للتحسّن في متوسّط الخسارة (PRIAL) إذْ كانت النتائج كالآتي

الجدول رقم (1)

قيم PRIAL لطريقتي التقدير للبيانات المولّدة

p	BE	TE
100	0.65	0.44
120	0.69	0.49
140	0.73	0.55
160	0.78	0.61
180	0.81	0.63
200	0.84	0.66
220	0.86	0.68
240	0.88	0.70
260	0.89	0.71
280	0.90	0.72
300	0.91	0.73



الشكل رقم (1)

قيم PRIAL لطريقتي التقدير للبيانات المولّدة

من خلال ملاحظة الجدول رقم (1)، و كذلك الشكل رقم (1) نجد أنّ مُقدّر الحزمة **BE** يتفوّق على المُقدّر المستدقّ **TE** بشكل عام و لأحجام العيّنات كلّها، إذْ لم تتّضح أيّ أفضلية تذكر للمُقدّر المستدقّ.

أمّا بالنسبة للبيانات الحقيقية فكانت النتائج كالآتي:

الجدول رقم (2)

قيم PRIAL **لطريقتي التقدير للبيانات الحقيقية**

Method	PRIAL
BE	**0.81**
TE	**0.58**

أمّا بالنسبة للبيانات الحقيقية فتأكّدت أفضلية مُقدّر الحزمة **BE** على المُقدّر المستدقّ **TE** ؛ لكونه يمتلك قيمة أكبر للنسبة المئوية للتحسّن في متوسّط الخسارة **PRIAL**.

9- التوصيات:

- يوصي الباحثان باعتماد مُقدّر الحزمة لتقدير مصفوفة التباين والتباين المشترك بوجود بيانات كبيرة تحت الدراسة.
- يوصي الباحثان بالتوسّع في دراسة البيانات الكبيرة ودراسة أنواع أخرى من المُقدّرات لمصفوفة التباين والتباين المشترك كالمُقدّرات اللامعلمية والحصينة.
- يوصي الباحثان بدراسة البيانات الكبيرة بشكل عام وتأثيراتها على مختلف المعالم الأخرى كمعلمات الموقع والشكل، و كذلك تطبيقات البيانات الكبيرة في الانحدار والقياس الاقتصادي.

المصادر Refences:

1. Choo, K. & Dehghantanha, A. (2020). Handbook of big data privacy. Switzerland: Springer Nature Switzerland AG.
2. Anderson, A. & Semmelroth, D. (2015). Statistics for big data for dummies. New Jersey: John Wiley & Sons.
3. Fan, J. Han, F. & Liu, H. (2014). Challenges of big data analysis. National Science review, 10.1093/nsr/nwt032. 293-314.
4. Johnstone, I. & Titterington, M. (2009). Statistical challenges of high-dimensional data. Philosophical transaction of the royal society, 10.1098/rsta.2009.0159. 4237-4253.
5. Mahdi, A. & Ilhasme, M. (2019). Robust and nonlinear shrinkage estimates for high dimension covariance matrix under autocorrelation problem. ResearchGate, 337783398.
6. Farne, M.& Montanari, A. (2020). A large covariance matrix estimator under intermediate spikiness regimes. Journal of multivariate analysis 176 (2020) 104577. from http://www.elsevier.com/locate/jmva.
7. Fan, J., Fan, Y. & Lv, J. (2008). High dimensional covariances matrix estimation using Factor model. Journal of Econometrics, Volume 147, Issue 1. 186-197.
8. Touloumis, A. (2015). Nonparametric Stein-type Shrinkage Covariance Matrix Estimators in High-Dimensional Settings. Academia.edu. From http://www.academia.edu/10145166/Nonparametric_Stein_Type_Shrinkage_Covariance_Matrix_Estimators_in_High-Dimensional_Settings.
9. Bickel, P. & Levina, E. (2008). Regularized estimation of large covariance matrices. The annals of statistics, vol. 36, NO. 1. 199-227.
10. Chi, T., Zhang, C. & Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. The annals of statistics, vol. 38, NO. 4. 2118-2144.