تطوير مقدر مقلص لتقدير مصفوفة التباين والتباين المشترك ذات الابعاد الكبيرة باستخدام الدوال المثلثية الزائدية (الخصائص الكيميائية لتربة حوض دجلة في محافظة واسط)

م . م . أحمد مهدي صالح

كلية الادارة والاقتصاد

جامعة واسط

الملخص

عندما تكون ابعاد مصفوفة التباين والتباين المشترك كبيرة بالنسبة الى حجم العينة اي ان المصفوفة ذات ابعاد تكون قريبة الى حجم العينة او أكبر منها. ستكون هناك صعوبات في ايجاد تقدير جيد لها اذ ان اغلب المصفوفات بتلك الابعاد ستعاني من صعوبة ايجاد المعكوس لهذه المصفوفات. لذلك فان طرق التقدير التقليدية مثل طريقة الامكان الاعظم ستعطي تقديرات متحيزة ويكون التقدير بعيدا عن قيمته الحقيقية. يهدف البحث الى التوسع في استعمال لمقدرات المقلصة لتقدير مصفوفة التباين والتباين المشترك في حالة استعمال عينات ذات ابعاد كبيرة. وهنا سيتم تقدير تلك المصفوفة باستعمال ثلاث طرق والمقارنة فيما بينها بالاعتماد على أصغر مربعات خطاء. تم استعمال مقدر الإمكان الأعظم *MLE* وكذلك مقدر غير خطي وهو مقدر الاوراكل *Oracle Estimator* وكذلك مقدر مقلص خطي وهو *Lediot and Wolf Estimator LW* والمقدر المقترح من لدن الباحث واجرينا محاكاة لأحجام عينات مختلفة وبأبعاد كبيرة وحساب أصغر مربعات خطاء عند ازدياد حجم العينة بالنسبة الى ابعاد مصفوفة التباين والتباين المشترك

وتوصل الباحث الى ان المقدر المقترح يكون الافضل عندما يكون حجم العينة صغيراً بالنسبة الى عدد المتغيرات فيها وكذلك ان مقدر الاوراكل يعمل جيدا عندما يكون حجم العينة صغيرا نوعا ما الى عدد المتغيرات فيها بينما لم تتوضح اية افضلية لمقدر الإمكان الأعظم وكذلك مقدر *LW*.

Abstract

When the dimensions of the covariance matrix are relatively large compared with the sample size ; or when the dimensions of the matrix are close to the sample size or larger, There will be difficulties in finding a good estimation for it. Most Matrices with high dimension suffer from the difficulty of finding their inverse. Therefore, the classical methods of estimation such as maximum likelihood will give biased estimators and far from their true value. This research aims at expanding usage of shrinkage

estimation to estimate the covariance matrix in the case of using samples with large dimensions.The covariance matrix will be estimated by using three methods. The Maximum Likelihood estimator *MLE* and the nonlinear shrinkage estimator *Oracle,* and the linear shrinkage estimator *Lediot and Wolf (LW)* and the *Suggested Estimator* and make comparison among them based on (MMSE) minimum mean square errors. Here, a simulated experiment with high dimensions samples was made with multiple sizes and calculated MMSE as the increasing in sample size to the large dimension of covariance matrix.

As conclusions, the *Suggested Estimator* is perfect when the sample size is very small compared with the number of variables in it. Moreover, the *Oracle* estimator is working well when the sample size is fairly small to the number of variables in it while it has not cleared that the maximum likelihood estimator *MLE* and the *(LW)* have any goodness.

1–مقدمة

ينصب اهتمام البحث عن ايجاد تقدير جيد وكفوء لمصفوفة التباين والتباين المشترك (Σ) كونها تدخل في العديد من التطبيقات الاحصائية المهمة وعنما تزداد ابعاد تلك المصفوفة تزداد صعوبة ايجاد التقدير لها بالطرق الكلاسيكية لذلك كان من الضروري استعمال طرق اخرى غير تقليدية كالمقدرات المقلصة.

تعود مسالة تقدير مصفوفة التباين والتباين المشترك الى عام 1960 قدم الباحث Stein [16] أفضل تمثيل حصل عليه من مقدر مقلص لتباين العينة. وبعد ذلك اقترحت العديد من المقدرات المقلصة ولها مقاييس اداء مختلفة فعلى سبيل المثال قدم الباحث Haff [7] مقدر يدخل ضمن اسلوب بيز الجزئي(Empirical Bayes) وكذلك اشتق الباحثان Dey & Srinivasan [3] مقدر(Minimax) تحت دالة خسارة (Entropy) والتي قدمت بالأصل من Stein .وتمكن الباحثان Ledoit & Wolf [12] باقتراح مقدر جديد عندما يكون حجم العينة أصغر من عدد المتغيرات في العينة تحت الدراسة إذ يعمل هذا المقدر على تصغير لأصغر متوسط مربعات الخطاء (MMSE) إذ يعمل هذا المقدر جيدا في ظل الحجوم الصغيرة للعينات وبازدياد ابعاد مصفوفة التباين والتباين المشترك .

2– مصفوفة التباين والتباين المشترك وأهمية تقديرها

Var-Covariance Matrix and its estimation

أن تقدير مصفوفة التباين والتباين المشترك أو معكوسها يعد من أساسيات الكثير من التطبيقات الإحصائية لأنها تعكس العلاقة بين المتغيرات تحت الدراسة إذ أن مصفوفة التباين والتباين المشترك لمتجه المتغيرات العشوائية $\underline{X}$ هو [6] .

$$\underline{X} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_p \end{bmatrix}$$

فان مصفوفة التباين والتباين المشترك هي

$$cov(\underline{X}) = E\left[\left(\underline{X} - \underline{\mu}\right)\left(\underline{X} - \underline{\mu}\right)'\right]$$

$$= \begin{bmatrix} var(x_1) & cov(x_1, x_2) & cov(x_1, x_3) & \cdots & cov(x_1, x_p) \\ cov(x_2, x_1) & var(x_2) & cov(x_2, x_3) & \cdots & cov(x_2, x_p) \\ cov(x_3, x_1) & cov(x_3, x_2) & var(x_3) & \cdots & cov(x_3, x_p) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ cov(x_p, x_1) & cov(x_p, x_2) & cov(x_p, x_3) & \cdots & var(x_p) \end{bmatrix}$$

ويرمز لها بالرمز Σ إذ تعد عملية تقديرها مهمة ويدخل مقدر $(\widehat{\Sigma})$ في الكثير من الموضوعات الإحصائية.

3- التحديات التي تفرضها الابعاد الكبيرة *High Dimensions Challenges*

لقد زاد الاهتمام في السنوات الأخيرة بمسالة تحليل البيانات عالية الأبعاد إذ فرض التطور الكبير الذي شهدته الكثير من القطاعات كالاتصالات والتقنيات الحياتية تحدياً كبيراً للباحثين الإحصائيين للتعامل مع المتغيرات ومحاولة التعرف على طبيعة هذه البيانات وتحليلها واختبار النماذج الملائمة لتمثيلها وتقدير معالم تلك النماذج ولكن مشكلة الأبعاد الكبيرة للبيانات وضعت الباحثين في مواجه المشاكل التي يتسبب بها ومنها.

3-1 مشاكل تقدير معالم الانحدار *Regression Parameters Estimating problems*

عندما تزداد اعداد المتغيرات في نموذج الانحدار العام فان هناك مشكلة وحدانية مصفوفة $(X'X)$ وكلآتي فليكن

$$Y = X\beta$$

نموذج انحدار عام بمتجه معالم β وحجم $(p \times 1)$ ومصفوفة المتغيرات X بحجم $(N \times p)$ اذ بازدياد عدد المتغيرات ينتفي شرط $(N > p)$ ويكون P عدداً كبيراً $(N < p)$ وبهذا تفقد شرط أن مصفوفة$(X'X)$ برتبة كاملة إذا ان تقدير معالم النموذج $\widehat{\beta}$ باستعمال طريقة المربعات الصغرى الاعتيادية

$$\widehat{B} = (\acute{X}X)^{-1}\acute{X}Y$$

وفي هذه الحالة لا يكون لمصفوفة $(X'X)$ معكوس لان محددها سيساوي صفراً

لقد دأب الباحثون لإيجاد طرائق جديدة لتقدير معالم نموذج الانحدار بوجود هذه المشكلة إذ استخدموا طرائق مثل المربعات الصغرى الجزائية [8] Penalized least Squares ومن اوائل الدراسات هي مقدرات الحرف Ridge Regression التي قدمت من لدن (Horrel & Kennard 1970) ولكن هذه التقديرات تتسم بالتحيز كما يزداد التحيز بازدياد عدد المتغيرات [10]. ومن ناحية اخرى تفترض طريقة المربعات الصغرى فروضاً منها ان يكون المتغير (x) قابل للقياس (Measurable) بدون اخطاء وتزداد صعوبة توفير او تحقق هذا الفرض بازدياد اعداد المتغيرات.

3-2 تحليل متعدد المتغيرات *Multivariate Analysis*

ان اغلب اساليب تحليل متعدد متغيرات تعتمد بشكل اساسي على تحليل القيم الذاتية والموجهات الذاتية Eigen analysis لمصفوفة التباين والتباين المشترك مثل التحليل القانوني Canonical Analysis وتحليل المركبات الرئيسة Principal Component Analysis وتحليل تباين متعدد متغيرات MANOVA وبازدياد عدد المتغيرات اي بازدياد ابعاد مصفوفة التباين والتباين المشترك تزداد صعوبة الحصول على القيم الذاتية Eigen Values والموجهات الذاتية Eigen Vectors [10] وحاول الباحثون ايجاد القيم الذاتية عندما تكون ابعاد المصفوفة كبيرة إذ ظهرت طرائق تعاقبية Iterative Methods مثل طريقة كاوس سيدل Geuss-Seidel ولكن هذه الطرائق تضع قيوداً وشروطاً صعب التعامل معها بازدياد ابعاد مصفوفات التباين والتباين المشترك.

4- اصغر متوسط مربعات خطاء *Minimum Mean Square Error (MMSE)*

قبل الدخول الى انواع مقدرات مصفوفة التباين والتباين المشترك لابد من عرض بسيط لمفهوم (MMSE) لكونه يمثل اسلوب مقارنة بين مختلف طرق تقدير مصفوفة التباين والتباين المشترك فالمقدر الافضل هو المقدر الذي يحقق اقل (MMSE) . إذ اقترح الباحثان Frost & Savarino استعمال القياس التربيعي للقياس التربيعي للمسافة من المصفوفة المقدرة والمصفوفة الحقيقية بالاستناد الى Frobenius Norm اذ انه في حالة كون المصفوفة مربعة متماثلة فان [5]

$$\|Z\|_F^2 = Trace\ (Z)^2$$

أي أن متوسط مربعات الخطاء سيكون

$$E\left[\left\|\Sigma - \widehat{\Sigma}\right\|_F^2\right] = Trace\left(\Sigma - \widehat{\Sigma}\right)^2 \quad \dots (1)$$

يكون MMSE اسلوباً جيداً للمقارنة بين مختلف مقدرات مصفوفة التباين والتباين المشترك

5–طريقة التقدير المقلصة لمصفوفة التباين والتباين المشترك

Shrinkage Estimator for Covariance Matrix

تقوم فكرة المقدر المقلص Shrinkage Estimator على تحسين المقدرات الاعتيادية كمقدر الإمكان الأعظم من خلال إيجاد مقدر يجمع بين خصائص التحيز القليل وأقل مربعات الخطأ إذ ان المقدرات الاعتيادية لمصفوفة التباين والتباين المشترك تكون ضعيفة عند ازدياد ابعاد هذه المصفوفة [2]. إذ يكون عادة المقدر المقلص لمصفوفة التباين والتباين المشترك بالصيغة الآتية.

$$\hat{\Sigma} = (1-\rho)S + \rho F \qquad \dots (2)$$

إذ ان ρ هو معامل التقليص Shrinkage Coefficient والذي تكون قيمته $1 > \rho > 0$ و المصفوفة F تعرف بهدف التقليص Shrinkage Target وتكون موجبة Positive Definite . وقد تنوعت أساليب اختيار مصفوفتي S, F والتي تستخدم في تقدير $\hat{\Sigma}$ ومن الأساليب الأكثر شيوعا هي اختيار S على انها مقدر الإمكان الأعظم.

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)' \qquad \dots (3)$$

ويمثل مقدر الإمكان الأعظم أيضا عندما $(n \geq p)$ ولكنه ليس بالضرورة يحقق أصغر MMSE بسبب التباين العادي وبسبب شروط ازدياد p أذ ان $E(S) = \Sigma$. اما بالنسبة الى هدف التقليص Shrinkage Target المتمثلة بالمصفوفة F قام بعض الباحثين باختيار F على انها مصفوفة الوحدة من الدرجة p أي ان [1] , [4] .

$$F = I_{(p \times p)}$$

وكذلك فضل البعض الاخر الباحثين اختيار F بالصيغة الآتية [2],[12].

$$F = \frac{Tr(S)}{p} . I_{(p \times p)} \qquad \dots . (4)$$

إذ أن I هي مصفوفة الوحدة بدرجة p وهذا المقدر يقلل التباين على حساب التحيز إذ يزداد تحيز هذا المقدر بزيادة p . ويمثل المقدر المقلص هو الحل المعقول للربط بين التحيز الصغير والتباين الصغير يتحقق من خلال التقليص بين (S , F) وهذا المقدر يعتمد على ρ معامل التقليص Shrinkage Coefficient وهي معلمة سيتم تقديرها أي إيجاد $\hat{\rho}$ بالشكل الذي يجعل MMSE المعرفة بالمعادلة (1) اقل ما يمكن أي ان المقدر المقلص هو حل للمعادلة الآتية [1].

$$\min_{\rho} E\left\{\left\|\hat{\Sigma} - \Sigma\right\|_F^2\right\}$$

$$s.t\ \hat{\Sigma} = (1-\rho)S + \rho F \qquad \dots (5)$$

6- مقدر LW ***Ledoit & Wolf Estimator***

عندما يكون من الصعب الحصول على تقدير جيد لمصفوفة التباين والتباين المشترك لا يعتمد على المصفوفة غير المعلومة ($\sum$) وعند المعرفة القليلة عن توزيع العينة او عدم معرفته اقترح الباحثان Ledoit Wolf [2] المقدر الخطي التالي

$$\hat{\rho}_{LW} = \frac{\sum_{i=1}^{n}\left\|X_i X_i' - \hat{S}\right\|_F^2}{n^2[Tr(\hat{S}^2) - Tr^2(\hat{s})/p]} \qquad \dots (6)$$

اذ ان

$$n\,, p \longrightarrow \infty \ ,\ \ {}^{n}\!/_{p} \longrightarrow c \ ,\ \ 0 < c < \infty$$

وعندئذ يعرف مقدر (LW) للمصفوفة Σ بالمقدر $\hat{\Sigma}_{LW}$

$$\hat{\Sigma}_{LW} = (1 - \hat{\rho}_{LW})\,\hat{S} + \hat{\rho}_{LW}\,\hat{F} \qquad \dots (7)$$

وكذلك بين الباحثان Ledoit Wolf بان القيمة المثلى لـ $\hat{P}$ غالباً ماتقع بين الواحد والصفر وبعد ذلك قام الباحثان بتحسين المقدر من خلال

$$\hat{\rho}_{LW} = \min(\hat{\rho}_{LW}\,, 1)$$

7-مقدر الاوراكل ***The Oracle Estimator***

يعد مقدر الاوراكل من المقدرات المثلى اللاخطية لمصفوفة التباين والتباين المشترك عند ازدياد ابعاد تلك المصفوفة [13] اذ يعتمد هذا المقدر على استعمال المعامل الامثل غير العشوائي الذي يعمل على تصغير متوسط مربعات الخطاء. أي ان $\hat{\Sigma}_0$ Oracle Estimator مقدر الاوراكل لمصفوفة التباين والتباين المشترك هو الحل للمعادلة الآتية [2].

$$\min_{\rho}\ E\left\{\left\|\hat{\Sigma}_0 - \Sigma\right\|_F^2\right\}$$

$$\text{s.t. } \hat{\Sigma}_0 = (1-\rho)S + \rho F$$

اذ ان $\widehat{\Sigma}_O$ هو مقدر الاوراكل لمصفوفة التباين والتباين المشترك و تكون قيمة S كما في المعادلة رقم (3) وتكون قيمة F كما هي في المعادلة رقم (4) ويمكن الحصول على قيمة معامل التقليص عن طريق تعويض قيمة $\widehat{\Sigma}_O$ في الدالة ينتج لدينا

$$E\left\{\left\|\widehat{\Sigma}_0-\Sigma\right\|_F^2\right\}=E\{\|(1-\rho)S+\rho F-\Sigma\|_F^2\}$$

$$=E\{\|(S-\Sigma)-\rho(S-F)\|_F^2\}$$

$$=E\{\|S-\Sigma\|_F^2\}-2\rho E\{\langle(S-\Sigma),(S-F)\rangle\}+\rho^2E\{\|S-F\|_F^2\} \quad \dots (8)$$

وباشتقاق المعادلة رقم(8) بالنسبة الى ρ ومساواتها بالصفر ينتج لدينا.

$$2\rho E\{\|S-F\|_F^2\}-2E\{\langle(S-\Sigma),(S-F)\rangle\}=0$$

$$\widehat{\rho}_0=\frac{E\{\langle(S-\Sigma),(S-F)\rangle\}}{E\{\|S-F\|_F^2\}}$$

$$\widehat{\rho}_0=\frac{E\{Tr((S-\Sigma)(S-F))\}}{E\{Tr(S-F)^2\}} \quad \dots (9)$$

بالنسبة الى البسط

$$E\left\{Tr\left((S-\Sigma)(S-F)\right)\right\}=$$

$$E\{Tr(S^2)\}-\frac{E\{Tr^2(S)\}}{p}-E\{Tr(\Sigma S)\}+\frac{Tr(\Sigma)}{p}E\{Tr(S)\} \quad \dots (10)$$

اما المقام فان.

$$E\{Tr(S-F)^2\}=E\{Tr(S^2)\}-2E\{Tr(SF)\}+E\{Tr(F^2)\}$$

$$=E\{Tr(S^2)\}-\frac{E\{Tr^2(\widehat{S})\}}{p} \quad \dots (11)$$

وباستعمال نتائج التوقعات الآتية

$$\dots (12) \quad E\{Tr(S)\}=Tr(\Sigma)$$

$$E\{Tr(S^2)\}=\frac{n+1}{n}Tr(\Sigma^2)+\frac{1}{n}Tr^2(\Sigma) \quad \dots (13)$$

$$\dots (14)\ E\{Tr^2(S)\}=Tr^2(\Sigma)+\frac{2}{n}Tr(\Sigma^2)$$

سيكون

$$\widehat{\rho}_O = \frac{(1-2/p)\ Tr(\Sigma^2)+Tr^2(\Sigma)}{(n+1-2/p)\ Tr(\Sigma^2)+(1-2/p)Tr^2(\Sigma)} \qquad \dots (15)$$

وعليه فأن مقدر الاوراكل لمصفوفة التباين والتباين المشترك سيكون بالصورة الآتية.

$$\widehat{\Sigma}_O = (1-\widehat{\rho}_O)S + \widehat{\rho}_O F \qquad \dots (16)$$

ومن الممكن التعبير عن المقدرين بدلالة قياسات العينة عن طريق المعادلتين الآتيتيين[4] [15].

$$a_1 = \frac{1}{p} Tr(\Sigma) \qquad \dots (17)$$

$$a_2 = \frac{1}{p}\ Tr(\Sigma^2) \qquad \dots (18)$$

إذ يكون التوقع بالصورة الآتية

$$\widehat{a}_1 = \frac{1}{p} Tr(\widehat{S}) \qquad \dots (19)$$

$$\widehat{a}_2 = \frac{n^2}{(n-1)(n+2)p}\left[Tr(\widehat{S}^2) - \frac{1}{n} Tr^2(\widehat{S})\right] \qquad \dots (20)$$

فينتج لدينا.

$$\widehat{\rho}_O = \frac{(1-2/p)\ pa_2+p^2a_1^2}{(n+1-2/p)pa_2\ +(1-2/p)p^2a_1^2} \qquad \dots. (21)$$

ويمكن التعويض عن قيم a_2 , $_1$ بالتقديرات لها كما هو في المعادلتين (20),(19) ليكون مقدر معامل التقليص كالآتي.

$$\widehat{\rho}_{O2} = \frac{(1-2/p)\ p\widehat{a}_2+p^2\widehat{a}_1^2}{(n+1-2/p)p\widehat{a}_2\ +(1-2/p)p^2\widehat{a}_1^2} \qquad \dots (22)$$

اما مقدر الاوراكل لمصفوفة التباين والتباين المشترك يكون بالصيغة الآتية [2]

$$\widehat{\Sigma}_{O2} = (1-\widehat{\rho}_{O2})S + \widehat{\rho}_{O2} F \qquad \dots (23)$$

8- المقدر المقترح لتقدير مصفوفة التباين والتباين المشترك

Suggested Estimator for Covariance Matrix

تم تركيز الجهود لإيجاد مقدر يقترب من مقدر الاوراكل أو يحسنه واقترح الباحث هنا استعمال الدوال المثلثية إذ تم استعمال دالة الجيب الزائدية Hyperbolic Sine Function إذ يبدا بقيمة أولى تخمينية مقترحة لمعامل التقليص

Shrinkage Coefficient ρ أي ان معامل التقليص هنا سيكون $\sinh\rho$ ابتداءاً ومنه سيتم الحصول على قيمة معامل التقليص الذي يجعل أصغر متوسط مربعات خطاء MMSE المعرف بالمعادلة (1) اقل ما يمكن إذ أن قيمة معامل التقليص هو الحل للدالة المقيدة الآتية [1]

$$\min_{p} E\left\{\left\|\widehat{\Sigma}_S - \Sigma\right\|_F^2\right\}$$

$$s.t\ \ \widehat{\Sigma}_S = (1-\sinh(\rho))S + \sinh(\rho)\, I \qquad \dots (24)$$

إذ أن $\widehat{\Sigma}$ هي المقدر المقترح لمصفوفة التباين والتباين المشترك. اذ أختار الباحث قيمة S كما هي في المعادلة (3) وقيمة Fكمصفوفة وحدة $I_{p\times p}$ وبتعويض قيمة القيد داخل الدالة في المعادلة رقم (24) ينتج لدينا.

$$E\left\{\left\|\widehat{\Sigma}_S - \Sigma\right\|_F^2\right\} = E\{\|(1-\sinh(\rho))S + \sinh(\rho)\, I - \Sigma\|_F^2\}$$

$$= E\{\|(S-\Sigma) - \sinh(\rho)\,(S-I)\|_F^2\}$$

$$= E\{\|S-\Sigma\|_F^2\} - 2\sinh(\rho)\, E\{\langle (S-\Sigma),(S-I)\rangle\} + (\sinh(\rho))^2 E\{\|S-I\|_F^2\} \qquad \dots (25)$$

وباشتقاق المعادلة رقم (25) بالنسبة الىρ ومساواتها بالصفر ينتج لدينا.

$$2\sinh(\rho)\cosh(\rho)\, E\{\|S-I\|_F^2\} - 2\cosh(\rho)\, E\{\langle (S-\Sigma),(S-I)\rangle\} = 0$$

$$\sinh(\rho) = \frac{E\{\langle (S-\Sigma),(S-I)\rangle\}}{E\{\|S-I\|_F^2\}}$$

وعليه فأن معامل التقليص سيكون

$$\widehat{\rho} = \sinh^{-1}\left[\frac{E\{Tr((S-\Sigma)(S-I))\}}{E\{Tr(S-I)^2\}}\right] \qquad \dots (26)$$

ولكون قيم معامل التقليص Shrinkage Coefficient هي قيم واقعة بين الصفر والواحد الصحيح اقترح الباحث استعمال دالة معكوس الجيب الزائدية بالصيغة الآتية [17].

$$\sinh^{-1} x = \text{lin}\left[x + \sqrt{x^2+1}\right] \qquad \dots (27)$$

وبتبسيط المقدار داخل القوس في المعادلة (26) وبالاعتماد على نتائج التوقع المعادلات

(13),(13),(12) ينتج لدينا الآتي بالنسبة الى البسط.

$$E\left\{Tr\left((S-\Sigma)(S-I)\right)\right\} =$$

$$E\{Tr(S^2)\} - E\{Tr(S)\} - E\{Tr(\Sigma S)\} + Tr(\Sigma)$$

$$= \frac{n+1}{n} Tr(\Sigma^2) + \frac{1}{n} Tr^2(\Sigma) - Tr(\Sigma^2)$$

$$= \frac{1}{n} Tr(\Sigma^2) + \frac{1}{n} Tr^2(\Sigma) \qquad \dots (28)$$

أما بالنسبة الى المقام فأن

$$E\{Tr(S-I)^2\} = E\{Tr(S^2)\} - 2E\{Tr(S)\} + p$$

$$= \frac{n+1}{n} Tr(\Sigma^2) + \frac{1}{n} Tr^2(\Sigma) - 2Tr(\Sigma) + p \qquad \dots (29)$$

وعليه سيكون مقدر معامل التقليص بالصورة الآتية.

$$\hat{\rho}_S = \sinh^{-1}\left[\frac{Tr(\Sigma^2)+Tr^2(\Sigma)}{(n+1)\, Tr(\Sigma^2)+Tr^2(\Sigma)-2nTr(\Sigma)+np}\right] \qquad \dots (30)$$

وعليه فأن المقدر المقترح لمصفوفة التباين والتباين المشترك يكون بالصيغة الآتية.

$$\hat{\Sigma}_S = (1-\hat{\rho}_S)S + \hat{\rho}_S I \qquad \dots (31)$$

ويكون المقدرين أعلاه دالة لمصفوفة التباين والتباين المشترك ومن الممكن التعبير عنهما بدلالة قياسات العينة عن طريق المعادلتين (18),(17) فيمكن التعبير عن المقدر $\hat{\rho}_S$ في المعادلة رقم (30) بالصورة الآتية.

$$\hat{\rho}_S = \sinh^{-1}\left[\frac{pa_2+p^2a_1^2}{(n+1)\, pa_2+p^2a_1^2-2npa_1+np}\right]$$

وبالعودة الى المعادلات (20),(19) يمكن التعبير عن مقدر معامل التقليص بدلالة قياسات العينة بالصورة الآتية .

$$\hat{\rho}_{S2} = \sinh^{-1}\left[\frac{\hat{a}_2+p\hat{a}_1^2}{(n+1)\, \hat{a}_2+p\hat{a}_1^2-2n\hat{a}_1+n}\right] \qquad \dots (32)$$

وعليه سيكون المقدر المقترح لمصفوفة التباين والتباين المشترك بدلالة قياسات العينة بالصيغة الآتية.

$$\hat{\Sigma}_{S2} = (1-\hat{\rho}_{S2})S + \hat{\rho}_{S2} I \qquad \dots (33)$$

9- المحاكاة *Simulation* وهنا تكون مصفوفة التباين والتباين المشترك Σ هي ناتج لعملية الانحدار الذاتي من الدرجة الأولى Autoregressive Process AR(1) عملية الانحدار الذاتي لكاوس هي نوع من انواع العمليات العشوائية التي تصف التغيير الذي يطرأ على المتغير في مدة زمنية محددة إذ تفترض هذا النموذج ان قيمة المتغير الحالي تعتمد بشكل خطي على قيمته السابقة وتكون بالصيغة الآتية [9].

$X_k = \rho X_{k-1} + \varepsilon_k$ وهي معادلة فروق ... (34)

تصف علاقة X_k بقيمته السابقة إذ ان .... $k = 0, \pm 1, \pm 2$ و ε_k هو حد الخطاء العشوائي بمتوسط صفر وتباين σ_ε^2 و ρ هي معلمة النموذج وتمثل مؤشر الارتباط الذاتي بين قيم X_k.

تم هنا ادخال مصفوفة التباين والتباين المشترك على انها تباين عملية الانحدار الذاتي من الدرجة الاولى لكاوس والتي تكون بالصيغة [2],[11].

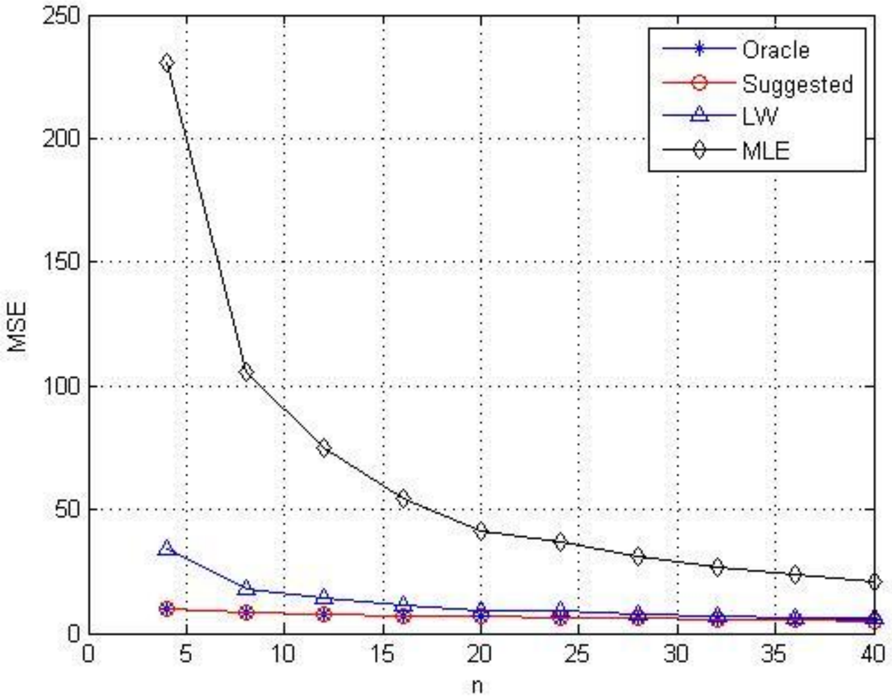
$$\Sigma_{ij} = \rho^{|i-j|} \quad \dots (35)$$

اذ ان $0 < \rho < 1$ إذ تضمن عملية الانحدار لكاوس الذاتي توافر الخصائص الرئيسة لمصفوفة التباين والتباين المشترك [11]. وهنا اقترح الباحث استعمال قيم وهي كالآتي $r = 0.45 , 0.7 , 0.9$ وتوليد عدد من المتغيرات (p= 20) تتوزع توزيعاً طبيعياً قياسيا و اختيار عدة احجام للعينة من صغيرة نسبيا الى عدد المتغيرات وصولا الى حجم عينة اكبر من عدد المتغيرات $n = 4 , 8, 12, \dots, 40$. و تم حساب قيم المقدرات الاربع مقدر الاوراكل Oracle estimator حسب المعادلة رقم (16) والمقدر المقترح Suggested Estimator حسب المعادلة رقم (31) ومقدر LW حسب المعادلة (7) ومقدر MLE حسب المعادلة (3) . والمقارنة فيما بينها بالاعتماد على أصغر متوسط مربعات خطاء MMSE وحسب المعادلة رقم (1) وباستعمال برنامج MATLAB [14] تم تكرار التجربة 5000 مرة وكانت النتائج كالآتي.

جدول(1)

قيم متوسط مربعات الخطاءMMSE للطرق الثلاثة عندما $\rho = 0.45$

	Minimum Mean Squares Error , ρ = 0.45				
	n	Oracle	Suggested	LW	MLE
Sample Size	4	9.6896	9.9309	39.4803	230.8511
	8	8.3685	8.4376	18.8109	105.8724
	12	7.7273	7.7573	17.1753	74.6462
	16	7.1119	7.0874	15.1335	54.6369
	20	6.7951	6.7045	14.1898	41.5309
	24	6.3501	6.2752	13.4646	37.0942
	28	6.0287	5.9661	12.5233	31.4039
	32	5.6145	5.5545	11.5393	26.7718
	36	5.3391	5.2847	10.9186	23.6439
	40	5.1358	5.0896	10.4289	21.0283



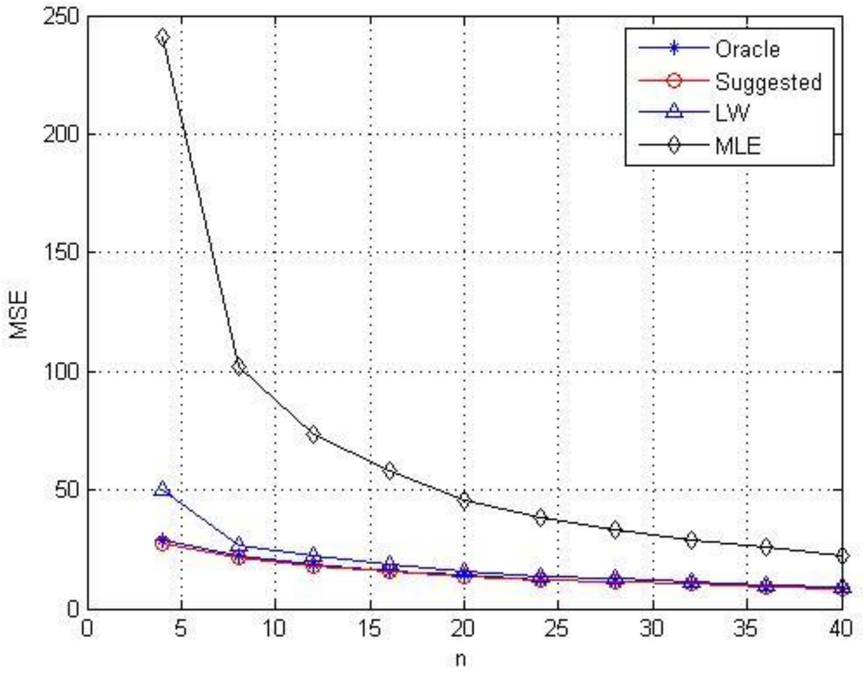
شكل (1)

متوسط مربعات الخطاءMMSE للطرق الثلاثة عندما $\rho = 0.45$

جدول(2)

قيم متوسط مربعات الخطاءMMSE للطرق الثلاثة عندما $\rho = 0.7$

	Minimum Mean Squares Error , $\rho = 0.7$				
Sample Size	*n*	*Oracle*	*Suggested*	*LW*	*MLE*
	4	28.7709	27.4005	74.9033	240.9809
	8	22.3739	21.6146	46.9168	101.9495
	12	18.5076	18.1046	39.4267	73.1855
	16	15.8522	15.6357	34.0649	58.0118
	20	13.9987	13.8640	29.3432	45.5564
	24	12.2085	12.1329	25.2860	38.4333
	28	11.3740	11.2696	24.1668	33.1798
	32	10.4072	10.3045	21.7295	29.1589
	36	8.8799	8.8124	18.4682	25.9067
	40	8.6943	8.6239	17.9318	22.0129



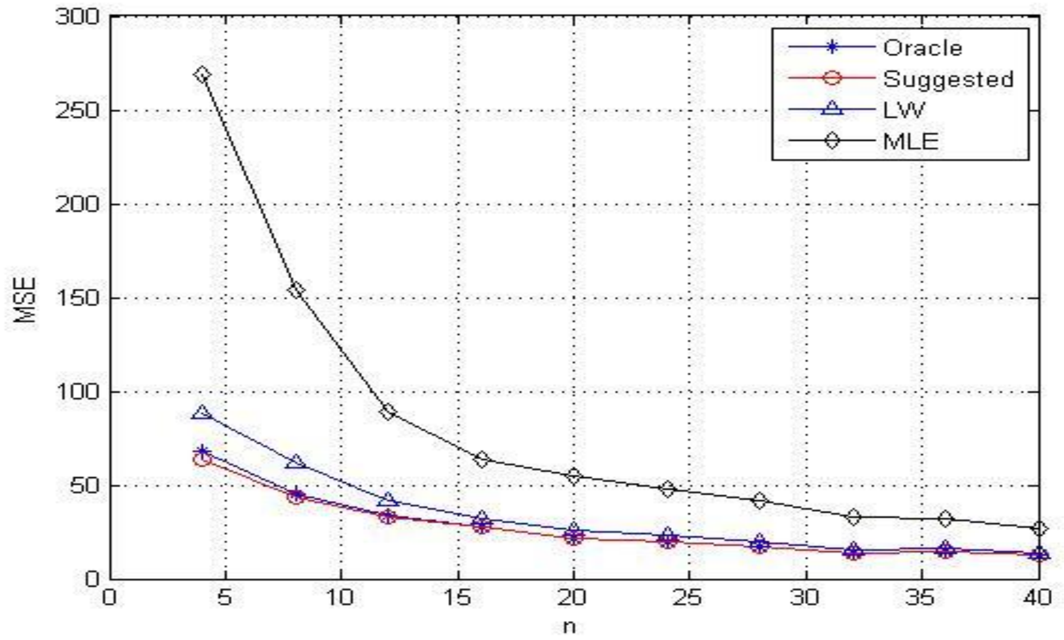
شكل (2)

متوسط مربعات الخطاءMMSE للمقدرات الثلاثة عندما $\rho = 0.7$

جدول(3)

قيم متوسط مربعات الخطاءMMSE للطرق الثلاثة عندما $\rho = 0.9$

	Minimum Mean Squares Error , $\rho = 0.9$				
Sample Size	*n*	*Oracle*	*Suggested*	*LW*	*MLE*
	4	68.2274	63.9288	155.9729	273.4953
	8	44.9663	43.7992	107.3286	158.5127
	12	33.8325	33.1453	77.5245	93.8462
	16	27.8659	27.4026	60.0569	68.5306
	20	21.5229	21.2861	45.6633	59.6152
	24	19.7495	19.5946	42.6959	53.0937
	28	17.4379	17.3240	37.0592	46.8846
	32	13.8516	13.7677	28.7248	35.8192
	36	14.1960	14.1112	29.9547	34.8338
	40	12.7084	12.6291	26.4099	29.5188



شكل (3)

متوسط مربعات الخطاءMMSE للمقدرات الثلاثة عندما $\rho = 0.9$

من جدول رقم (1) و شكل رقم (1) $\rho = 0.45$ عندما يكون الارتباط الذاتي قليلاً نجد عندما يكون حجم العينة صغير بالنسبة الى عدد المتغيرات $n = 4, 8, 12, 16$ نلاحظ ان مقدر الاوراكل يكون افضل من بقية المقدرات يليه المقدر المقترح ونلاحظ عدم وجود اية افضلية بالنسبة الى مقدري LW, MLE . وعندما يكون حجم العينة قريبا من عدد المتغيرات $n = 20, 24, 28$ نلاحظ افضلية بسيطة للمقدر المقترح على مقدر الاوراكل وكذلك ابتعاد مقدري LW, MLE عن بقية المقدرات. وعندما يكون حجم العينة اكبر من عدد المتغيرات $n = 32, 36, 40$ نجد استمرار افضلية المقدر المقترح على بقية المقدرات مع بقاء مقدري LW, MLE بدون أي افضلية تذكر.

من جدول رقم (2) و شكل رقم (2) $\rho = 0.7$ عندما يكون الارتباط الذاتي قوي نسبيا نجد عندما يكون حجم العينة صغير بالنسبة الى عدد المتغيرات $n = 4, 8, 12, 16$ نلاحظ ان المقدر المقترح يكون افضل من بقية المقدرات يليه مقدر الاوراكل ونلاحظ عدم وجود اية افضلية بالنسبة الى مقدري LW, MLE . وعندما يكون حجم

العينة قريبا من عدد المتغيرات $n = 20, 24 , 28$ نلاحظ بقاء افضلية للمقدر المقترح على مقدر الاوراكل وكذلك اقتراب مقدر LW من بقية المقدرات مع بقاء مقدر MLE بعيدا عن بقية المقدرات. وعندما يكون حجم العينة اكبر من عدد المتغيرات $n = 32, 36 , 40$ نجد استمرار الحال على ما هو عليه.

من جدول رقم (3) و شكل رقم (3) $\rho = 0.9$ عندما يكون الارتباط الذاتي عال جدا نجد انه عندما يكون حجم العينة صغيراً بالنسبة الى عدد المتغيرات $n = 4, 8 , 12 , 16$ نلاحظ ان المقدر المقترح يكون افضل من بقية المقدرات يليه مقدر الاوراكل ونلاحظ عدم وجود اية افضلية بالنسبة الى مقدري LW, MLE . وعندما يكون حجم العينة قريبا من عدد المتغيرات $n = 20 , 24 , 28$ نلاحظ افضلية بسيطة للمقدر المقترح على مقدر الاوراكل وكذلك ابتعاد مقدري LW, MLE عن بقية المقدرات. وعندما يكون حجم العينة اكبر من عدد المتغيرات $n = 32 , 36 , 40$ نجد استمرار افضلية المقدر المقترح على بقية المقدرات مع بقاء مقدري LW, MLE بدون أي افضلية تذكر.

9- البيانات الحقيقية *Real Data*

تم الحصول على عينة مكونة من 9 قطعة ارض إذ تم قياس 16 متغيراً لكل قطعة ارض تمثل الخصائص الكيميائية لتربة حوض النهر في محافظة واسط إذ كانت كلآتي:

جدول (4)

الخصائص الكيميائية لتربة حوض النهر

no	ESB	SO4	Hco3	p	ESR	Ca	Mg	K	Na	Cl	PH	OM	EC	Lime	Gypsum	IC
1	11.74	2.50	3.49	0.03	0.70	10.23	1.50	1.52	2.41	30.50	7.74	1.35	2.37	22.35	23.20	20.60
2	12.67	2.65	3.25	0.02	0.79	8.28	1.49	1.47	2.50	29.50	7.50	1.45	1.99	23.15	26.45	19.67
3	9.56	2.21	3.76	0.03	0.77	3.97	2.17	2.42	1.93	29.87	7.45	1.12	1.67	23.12	28.10	20.20
4	11.15	2.47	2.34	0.14	0.84	4.86	1.38	1.12	2.42	29.88	7.83	1.36	1.52	21.40	23.99	21.55
5	12.64	6.94	3.25	0.02	0.99	6.96	2.11	1.78	2.99	21.75	7.81	1.57	2.05	24.10	30.95	23.65
6	10.05	5.60	2.37	0.02	0.47	7.50	2.13	2.24	2.33	22.15	7.44	1.58	1.74	23.41	33.45	23.10

جدول (5)

7	6.94	7.24	3.43	0.02	0.62	3.90	1.92	1.34	1.51	21.45	7.65	1.40	1.83	22.19	29.30	21.55
8	13.60	7.53	4.30	0.03	1.02	7.23	2.27	1.39	3.15	18.10	7.95	1.67	3.15	25.45	33.55	23.15
9	17.15	15.95	2.23	0.02	0.89	13.40	10.53	0.69	4.37	14.05	7.42	1.39	9.64	23.15	30.95	25.40

توضيح المتغيرات

ESP	النفاذية المحسوسة
So4	ايون الكبريتات
Hco3	ايون الهيدروكاربونات
P	ايون الفسفور
ESR	صفوه الطين الخالصة
Ca	ايون الكالسيوم
Mg	ايون المغنيسيوم
K	ايون البوتاسيوم
Na	ايون الصوديوم
Cl	ايون الكلور
PH	دليل الحامضية
OM	المواد العضوية
EC	الاملاح المذابة
Gypsum	الجبس
Lime	الكلس
IC	النفاذية المتبادلة

تم حساب قيم المقدرات بدلالة قياسات العينة مقدر الاوراكل Oracle estimator حسب المعادلة رقم (23) والمقدر المقترح Suggested Estimator حسب المعادلة رقم (33) ومقدر LW حسب المعادلة (7). ومن ثم احتساب أصغر متوسط مربعات خطاء MMSE وحسب المعادلة رقم (1).

جدول (6)

MMSEللمقدرات الثلاث للبيانات الحقيقية

	MMSE
Oracle	102.9

Suggested	101.5
LW	178.5

من ملاحظة جدول رقم (6)

نجد ان أصغر متوسط لمربعات الخطاء يكون باقل قيمة عند المقدر المقترح اذ يتفوق المقدر المقترح بمقدار قليل على المقدر اللاخطي المتمثل بمقدر Oracle اما بالنسبة الى المقدر الخطي LW فنجد انه يمتلك اصغر متوسط مربعات خطاء بقيمة عالية تبتعد عن بقية المتغيرات.

وتعد النتائج التي حصل عليها الباحث قريبة جدا من نتائج المحاكاة.

10-الاستنتاجات *Conclusions*

مما سبق تم استنتاج ما ياتي

1-يكون مقدر الاوراكل أفضل من بقية المقدرات في حالة كون العينة صغيرة جدا نسبة الى عدد المتغيرات وعندما يكون الارتباط الذاتي قليل.

2-ويكون المقدر المقترح أفضل من بقية المقدرات بوجود ارتباط ذاتي عال بمختلف احجام العينات نسبة الى عدد المتغيرات في العينة.

3-عدم وجود اية افضلية لمقدري MLE , LW مقارنة بمقدر الاوراكل والمقدر المقترح .

10- التوصيات *Recommendations*

يوصي الباحث بما يأتي

1-اعتماد المقدر المقترح عندما يكون حجم العينة صغيراً وعندما تعاني البيانات من وجود ارتباط ذاتي عالي.

2-التوسع في استعمال الدوال اللاخطية لتطوير المقدرات المقلصة لمصفوفة التباين والتباين المشترك وكذلك واقتراح مقدرات جديدة اخرى للتقدير في ظل ظهور البيانات عالية الابعاد في مختلف مجالات الحياة.

المصادر

[1]. CHEN, Y., WIESEL, A. and HEREO, A. O. (2011). Robust shrinkage estimation of high dimensional covariance matrices. *IEEE Transactions on Signal Processing,* vol. 59, issue. 9.

[2]. CHEN, Y. A., WIESEL, A. and ELDAR, A. O. (2010). Shrinkage Algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing,* vol. 58, issue 10.

[3]. DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stien's loss. *The Annals of Statistics,* vol. 13, No. 4, pp. 1581-1591.

[4]. FISHER, T. J. and SUN, X. (2011). Improved Stein-type shrinkage estimator for the high-dimensional multivariate normal covariance matrix. *Comp. Statist. Data Analysis,* 55, 1909-1918.

[5]. FROST, P. A. and SAVARINO, J. E. (1986). An empirical Bayes Approach to Portfolio selection. *Journal of Financial and Quantitate Analysis,* 21: 293- 305

[6]. FUJIKOSHI, Y., ULYANOV, V. and SHIMIZU, R. (2011). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* Wiley Series in Probability and Statistics.

[7]. HAFF, L. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix," *The Annals of Statistics,* vol. 8, no. 3, pp. 586-597.

[8]. HOREL, A. E. and KENNARD, R. W. (1970). Ridge regression; biased estimation for nonorthogonal problems. *Technometrics,* vol. 12, pp. 55-82.

[9]. HORVATH, Z. and JOHNSTONE, R. (2000). AR(1) Time series process. *Econometrics 7590,* Utah University Press.

[10]. JOHNSTONE, I. M. and TITTERINGTON, D. M. (2009). Statistical Challenges of high-dimensional data. *Philosophical Transactions of the Royal Society.* 367, pp. 4237-4253.

[11]. KINCAID, C. (2005). Guidelines for selection the covariance structure in mixed model analysis. Proceedings of the 30th annual SAS, Paper 198-30.

[12]. LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large- dimensional covariance matrices, Journal of Multivariate Analysis, vol. 88, no.2, pp. 365-411.

[13]. LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of*

Statistics. Vol. 40, No. 2, pp. 1024–1060.

[14]. MARTINEZ, W. L. and MARTINEZ, A. R. (2002). *Computational Statistics Handbook with MATLAB*, Chapman & Hall/CRC Press.

[15]. SRIVASTAVA, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.* 35(2), 251-272.

[16]. STIEN, C., EFRON, B. and MORRIS, C. (1972). Improving the usual estimator of a normal covariance matrix, *National Science Foundation Grant, Technical Report,* No. 37.

[17]. ZWILLINGER, D. (2012). *Standard Mathematical Tables and Formulae.* 32ed. CRC Press. Taylor & Francis Groups. NW. USA