استخدام الطرق المقلصة اللاخطية لتقدير مصفوفة التباين والتباين المشترك في البيانات عالية الابعاد مع تطبيق عملي

(الخصائص الكيميائية لتربة حوض النهر في محافظة واسط)

Using Nonlinear Shrinkage Estimation methods to Estimate the Covariance Matrix in the High Dimension Data

(The Chemical Properties of the Soil in the River Basin in the Wasit State)

م.م.أحمد مهدي صالح

كلية الادارة والاقتصاد/ جامعة واسط

المستخلص

عندما تكون ابعاد مصفوفة التباين والتباين المشترك كبيرة بالنسبة الى حجم العينة اي ان المصفوفة ذات ابعاد تكون قريبة الى حجم العينة او اكبر منها. ستكون هناك صعوبات في ايجاد تقدير جيد لها اذ ان اغلب المصفوفات بتلك الابعاد تعاني من صعوبة ايجاد المعكوس لهذه المصفوفات . لذلك فان طرق التقدير التقليدية مثل طريقة الامكان الاعظم او طريقة المربعات الصغرى ستعطي تقديرات متحيزة ويكون التقدير بعيدا عن قيمته الحقيقية في المجتمع. يهدف البحث الى التوسع في استخدام لمقدرات المقلصة لتقدير مصفوفة التباين والتباين المشترك في حالة استخدام عينات ذات ابعاد كبيرة .وهنا سيتم تقدير تلك المصفوفة باستخدام طريقتين والمقارنة فيما بينها بالاعتماد على اصغر مربعات خطاء. حيت تم استخدام مقدر الاوراكل (Oracle Estimator) كتقدير مقلص لمصفوفة التباين والتباين المشترك بالاضافة الى استخدام مقدر الاوراكل التقريبي (Approximate Oracle Estimator) والمقارنة بينهما حيث تم اجراء محاكاة لاحجام عينات مختلفة وبابعاد كبيرة وحساب اصغر مربعات خطاء عند ازدياد حجم العينة بالنسبة الى ابعاد مصفوفة التباين والتباين المشترك كذلك تم الحصول على بيانات عالية الابعاد حقيقية لبعض الخصائص الكيميائية لتربة حوض النهر في محافظة واسط واستخدامها كجزء من التطبيق العملي للبحث.

Abstract

*When the dimensions the covariance matrix are relatively large to the sample size or when the dimensions of the matrix are close to the sample size or larger than it . There will be difficulties in finding a good estimation for it. Most Matrices with high dimension suffer from the difficulty of finding the inverse of theme. Therefore the classical methods of estimation such least squares or maximum likelihood will give biased estimators and far from its true value. This search aims at expanding usage of shrinkage estimation to estimate the covariance matrix in the case of using samples with large dimensions.We will estimate the covariance matrix by using (**Oracle Estimator**) and (**Approximate Oracle Estimator)** and make comparison among them based on (MMSE) minimum mean square errors. Here we make a simulated experiment with high dimensions samples with multiple sizes and calculate MMSE as the increasing in sample size to the large dimension of covariance Matrix and we get real high*

dimension data represent the chemical properties of soil in the river basin in Wasit state and use it as a part of the practical side of this search.

1- مقدمة *Introduction*

تعود مسالة تقدير مصفوفة التباين والتباين المشترك الى عام 1960 حيث قدم الباحث Stein [15] افضل 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 [6] مقدر يدخل ضمن اسلوب بيز التجريبي (Empirical Bayes) وكذلك اشتق الباحثان Dey & Srinivasan [3] مقدر (Minimax) تحت دالة خسارة (Entropy) والتي قدمت بالاصل من Stein [8] .
وتمكن الباحثان Ledoit & Wolf [10] باقتراح مقدر جديد عندما يكون حجم العينة اصغر من عدد المتغيرات في العينة تحت الدراسة حيث يعمل هذا المقدر على تصغير متوسط مربعات الخطاء (MSE) أذ يعمل هذا المقدر جيدا في ظل الحجوم الصغيرة للعينات وبازدياد ابعاد مصفوفة التباين والتباين المشترك .
من الطبيعي ان تنصب الجهود الى ايجاد مقدر كفوء لمصفوفة التباين والتباين المشترك باستخدام طرائق غير تقليدية كطرائق التقدير المقلصة وغيرها في سبيل تطوير مقدر افضل لمصفوفة التباين والتباين المشترك .
يهدف البحث الى التوسع في استخدام لمقدرات المقلصة لتقدير مصفوفة التباين والتباين المشترك في حالة استخدام عينات ذات ابعاد كبيرة

2- مصفوفة التباين والتباين المشترك وأهمية تقديرها

أن تقدير مصفوفة التباين والتباين المشترك أو معكوسها يعتبر من أساسيات الكثير من التطبيقات الإحصائية لأنها تعكس العلاقة بين المتغيرات تحت الدراسة حيث أن مصفوفة التباين والتباين المشترك لمتجه المتغيرات العشوائية $\underline{X}$ هو [5]

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

فان مصفوفة التباين والتباين المشترك هي

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

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

ويرمز لها بالرمز Σ حيث تعد عملية تقديرها مهمة ويدخل مقدر $\left(\hat{\Sigma}\right)$ في الكثير من المواضيع الإحصائية ومنها .

2-1الانحدار ***Regression***

أن التداخل بين تقدير مصفوفة التباين والتباين المشترك وقيمة معلمات الانحدار قد لوحظ من أمد بعيد وهناك العديد من الأمثلة التطبيقية حول استخدام هذه المصفوفات في عمليات التنقية للتنبؤ والتقدير [2] . و يدخل تقدير مصفوفة التباين والتباين المشترك عن طريق المربعات الصغرى والإمكان الأعظم وكذلك بواسطة انحدار الحرف Ridge Regression في التنبؤ والاختبار لعدة معالم في الانحدار .

2-2 الأساليب القياسية لمتعدد المتغيرات ***Standard Multivariate Techniques***

عند استخدام أساليب الانحدار المتعدد بعد تقدير مصفوفة التباين والتباين المشترك أساسيا لأنة يمثل المرحلة الابتدائية (Initial Stage) في العديد من هذه التطبيقات مثل تحليل المركبات الأساسية Principal Component Analysis والتحليل المميز الخطي Linear Discernment Analysis فلابد من وجود تقدير كفؤ لمصفوفة التباين والتباين المشترك لتقليل اخطاء التصنيف [5] .Classification Errors

3- مجالات تطبيق البيانات عالية الابعاد واهم تحدياتها

تظهر البيانات عالية الابعاد في الكثير من مجالات العلوم الحديثة حيث ازداد ظهورها بازدياد التطور العلمي والتكنولوجي في السنوات الاخيرة ومنها

3-1 بيانات التقنية الاحيائية ***Biotechnology Data***

وهي من المجالات العلمية الحديثة وتظهر فية البيانات عالية الابعاد لتنوع بياناتها كالبيانات الخاصة بالعرق او التحليل العرقي DNA او بيانات الجينات والموروثات كتسلسل الاحماض الامينية في

الموروثات وتمتد الى سلاسل طويلة ومعقدة كذلك تظهر تلك البيانات في المجال الزراعي خصوصاً في مجال التعديل الوراثي للمحاصيل [12]

3-2 بيانات صور الاقمار الاصناعية ***Satellite Imagery Data***

تتطلب عملية تحليل الصور الواردة من الاقمار الصناعية التعامل مع قاعدة بيانات واسعة من الصور كصور التنبؤ بالمناخ او الصور الفلكية وغيرها حيث تكون هذة الصور عالية الدقة وعالية الابعاد باحجام هائلة تتطب كثيراً من الدقة والاحترافية للتعامل معها [12]

3-3 البيانات الكيميائية للمركبات ***Chemical Amides Data***

ان تنوع الحصائص الكيميائية للمركبات الكيميائية الصناعية منها والاحيائية خلق مجالا كبيرا لتحليل هذا النوع من البيانات وانعكس التنوع الكبير للكيمياء الطبيعية على نوعية تلك البيانات فوجد الباحثون المختصون بالمجال الصناعي و الزراعي انفسهم في مواجهة بيانات عالية الابعاد لتلك الخصائص وكون كل خاصية منها مهمه وفعاله مما يتطلب طرائق تحليل كفوءة لتلك البيانات .

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

4-3 مشاكل تقدير معالم الانحدار *Regression Parameters Estimating*

عندما تزداد اعداد المتغيرات في نموذج الانحدار العام فان هناك مشكلة وحدانية مصفوفة $(X'X)$ وكاالتالي 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 عدد كبير $(N < p)$ وبهذا تفقد شرط أن مصفوفة$(X'X)$ برتبة كاملة اذا ان تقدير معالم النموذج $\hat{\beta}$ باستخدام طريقة المربعات الصغرى الاعتيادية

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

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

لقد داب الباحثون لإيجاد طرائق جديدة لتقدير معالم نموذج الانحدار بوجود هذة المشكلة حيث استخدموا طرقاً مثل المربعات الصغرى الجزائية [7] Penalized least Squares

ومن اوائل الدراسات هي مقدرات الحرف Ridge Regression التي قدمت من قبل (Horrel & Kennard 1970) ولكن هذة 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 .

5-3 تحليل متعدد المتغيرات *Multivariate Analysis*

ان اغلب اساليب تحليل متعدد متغيرات تعتمد بشكل اساسي عل تحليل القيم الذاتية والموجهات الذاتية Eigen Analysis لمصفوفة التباين والتباين المشترك مثل التحليل القانوني Canonical Analysis وتحليل المركبات الرئيسية 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 حيث ظهرت طرائق تعاقبية Iterative Methods مثل طريقة كاوس سيدل Geuss-Seidel ولكن هذة الطرائق تضع قيوداً وشروطاً صعب التعامل معها بازدياد ابعاد مصفوفات التباين والتباين المشترك [7] .

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 طرائق تقدير مصفوفة التباين والتباين المشترك فالمقدر الافضل هو المقدر الذي يحقق اقل (MMSE) . حيث اقترح الباحثان Frost & Savarino استخدام القياس التربيعي للقياس التربيعي للمسافة من المصفوفة المقدرة والمصفوفة الحقيقية بالاستناد الى Frobenious Norm اذ انه في حالة كون المصفوفة مربعة متماثلة فان [4]

$$\|\boldsymbol{Z}\|_{\boldsymbol{F}}^{\boldsymbol{2}} = \boldsymbol{Trace}\,(\boldsymbol{Z})^{\boldsymbol{2}}$$

أي أن متوسط مربعات الخطاء سيكون

$$\boldsymbol{E}\left[\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\|_{\boldsymbol{F}}^{\boldsymbol{2}}\right] = \boldsymbol{Trace}\left(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right)^{\boldsymbol{2}} \qquad \boldsymbol{\ldots (1)}$$

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

فلتكن $\{X_i\}_{i=1}^{n}$تمثل p من المتجهات العشوائية بمتوسط (0) وتباين $\sum$ متماثلة التوزيع أي أنها متجهات كاوسية Gaussian Vectors المطلوب هنا أيجاد المقدر $\widehat{\sum}$ الذي يعمل على جعل MMSE المعرف بالمعادلة 1)) اقل ما يمكن من الصعوبة حساب $\widehat{\sum}$ بدون قيود اضافية وعندئذ سوف يجد الباحث نفسه مقيد بنوع من المقدرات الذي يجعله يستخدم طريقة التقليص Shrinkage كما أشار الباحثان Ledoit & Wolf [10]

أن المقدر الاعتيادي لمصفوفة التباين والتباين المشترك هو $\hat{S}$

$$\widehat{\boldsymbol{S}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i X_i'} \qquad \textbf{... (2)}$$

آذ أن هذا المقدر هو مقدر غير متحيز

$$\boldsymbol{E}(\widehat{\boldsymbol{S}}) = \boldsymbol{\Sigma}$$

ويمثل مقدر الإمكان الأعظم أيضا عندما (n>p) ولكنة ليس بالضرورة يحقق اصغر MSE بسبب التباين العادي وبسبب شروط ازدياد P لذلك من الناحية المنطقية أن يؤخذ مقدار أخر لتقدير $\sum$ وهو المقدر المعروف بالمعادلة (3)

$$\widehat{\boldsymbol{F}} = \frac{\boldsymbol{Tr}\,(\widehat{\boldsymbol{S}})}{\boldsymbol{p}} * \boldsymbol{I} \qquad \textbf{... (3)}$$

حيث أن I هي مصفوفة الوحدة بدرجة P وهذا المقدر يقلل التباين على حساب التحيز أذا يزداد تحيز هذا المقدر بزيادة P [1].

أن الحل المعقول للربط بين التحيز الصغير والتباين الصغير يتحقق من خلال التقليص بين $(\hat{S}\ ,\hat{F})$ مما ينتج عنه مقدر جديد يسمى مقدر مختلط (Mix Estimator) وغالباً ما يفضل تسميته Shrinkage ويعرف كالاتي[11] .

$$\widehat{\boldsymbol{\Sigma}} = (\mathbf{1} - \widehat{\boldsymbol{\rho}})\,\widehat{\boldsymbol{S}} + \widehat{\boldsymbol{\rho}}\,\widehat{\boldsymbol{F}} \qquad \textbf{... (4)}$$

وهذا المقدر يعتمد على معلمة الربط $(\hat{P})$ وهي معلمة واقعه بين الصفر والواحد ويمكن الإشارة إلى المقدر $\hat{F}$ بأنة صف التقليص Shrinkage Target سيتم تقدير معلمة الربط او معلمة الخلط اي تجعل MSE المعرفة بالمعادلة (1) اقل ما يمكن .

6- مقدر الاوراكل *The Oracle Estimator*

يعتبر مقدر الاوراكل من المقدرات المثلى اللاخطية لمصفوفة التباين والتباين المشترك عند ازدياد ابعاد تلك المصفوفة [11] اذ يعتمد هذا المقدر على استخدام المعامل الامثل غير العشوائي الذي يعمل على تصغير متوسط مربعات الخطاء اي ان $\hat{\Sigma}_0$ Oracle Estimator هو الحل للمعادلة التالية

$$\boldsymbol{Min}\ \left[\boldsymbol{E}\left\{\left\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}} - \widehat{\boldsymbol{\Sigma}}\right\|_{\boldsymbol{F}}^{\mathbf{2}}\right\}\right]$$

s.t

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}} = (\mathbf{1} - \widehat{\boldsymbol{\rho}})\,\widehat{\boldsymbol{S}} + \widehat{\boldsymbol{\rho}}\,\widehat{\boldsymbol{F}} \qquad \textbf{... (5)}$$

اذ ان $\hat{F}$, $\hat{S}$ كما في المعادلتين (2) (3) وان قيمة P المثلى يمكن الحصول عليها بتطبيق النظرية الاتية [10]

نظرية (1)

فليكن P من المتجهات والمعروفة $\{x_i\}_{i=1}^{n}$ وان $\hat{S}$ هو تقدير لمصفوفة التباين والتباين المشترك لهذه المتغيرات فإذا كانت ($(x_i)^n$) { متغيرات مستقلة ومتماثلة التوزيع توزيعاً طبيعياً فان حل المعادلة (5) هو:

$$\boldsymbol{\rho_0 = \frac{E\{Tr(\Sigma-\hat{S})(\hat{F}-\hat{S})\}}{E\{\|\hat{S}-\hat{F}\|_F^2\}}} \quad \textbf{... (6)}$$

أذا أن:

$$\boldsymbol{E\{Tr(\Sigma-\hat{S})(\hat{F}-\hat{S})\} = \frac{Tr(\Sigma)}{p}E\{Tr(\hat{S})\} - \frac{E\{Tr^2(\hat{S})\}}{p} - E\{Tr(\Sigma\hat{S})\} + E\{Tr(\hat{S}^2)\}} \quad \textbf{....(7)}$$

وأيضا:

$$\boldsymbol{E\{\|\hat{S}-\hat{F}\|_F^2\} = E\{Tr(\hat{S}^2)\} - 2E\{Tr(\hat{S}\hat{F})\} + E\{Tr(\hat{F}^2)\} = E\{Tr(\hat{S}^2)\} - \frac{E\{Tr^2(\hat{S})\}}{p}} \quad \textbf{... (8)}$$

وباستخدام التعاريف الاتية :

$$\boldsymbol{E\{Tr(\hat{S})\} = Tr(\Sigma)} \quad \textbf{... (9)}$$

$$\boldsymbol{E\{Tr(\hat{S}^2)\} = \frac{n+1}{n}Tr(\Sigma^2) + \frac{1}{n}Tr^2(\Sigma)} \quad \textbf{... (10)}$$

$$\boldsymbol{E\{Tr^2(\hat{S})\} = Tr^2(\Sigma) - \frac{2}{n}Tr(\Sigma^2)} \quad \textbf{.... (11)}$$

وعلية سيكون تقدير ρ_0 كالتالي

$$\boldsymbol{\rho_0 = \frac{(1-2/p)Tr(\Sigma^2)+Tr^2(\Sigma)}{(n+1-2/p)Tr(\Sigma^2)+(1-2/p)Tr^2(\Sigma)}} \quad \textbf{... (12)}$$

وتتحقق المعادلة (12) في ظل شروط التوزيع الطبيعي فقط آذ عندما يقترب توزيع المعاينة من التوزيع الطبيعي فان الصيغة (12) وتقرب من الصيغة (6)

7-مقدر الاوراكل التقريبي ***The Approximate Oracle Estimator AOE***

تم تركيز الجهود حول تطوير مقدر يقترب من مقدر الاوراكل أو يطوره و مقدر الاوراكل التقريبي هنا هو مقدر تكراري حيث يبدأ التكرار بعينة أولية تخمينية للمصفوفة $\boldsymbol{\Sigma}$ ومن ثم تحسين المقدر بصورة تكرارية بحيث يكون المقدر الأولي $\hat{\boldsymbol{\Sigma}}_0$ والذي يعطي حل $\hat{\boldsymbol{\Sigma}}_1$ وهكذا تستمر عملية تقريب حتى يتحقق التقارب

حيث ان :

$$\boldsymbol{\hat{\Sigma}_{AOE} = (1-\hat{\rho}_{AOE})\hat{S} + \hat{\rho}_{AOE}\hat{F}} \quad \textbf{... (13)}$$

آذ أن

$$\boldsymbol{\hat{\rho}_{j+1} = \frac{(1-2/p)Tr(\hat{\Sigma}_j\hat{S})+Tr^2(\hat{\Sigma}_j)}{(n+1-2/p)Tr(\hat{\Sigma}_j\hat{S})+(1-n/p)Tr^2(\hat{\Sigma}_j)}} \quad \textbf{... (14)}$$

$$\boldsymbol{\Sigma_{j+1} = (1-\hat{\rho}_{j+1})\hat{S} + \hat{\rho}_{j+1}\hat{F}} \quad \textbf{... (15)}$$

اذ ان هذا الاسلوب التكراري يؤدي الى تحسين المقدر ولكن يمكن الاستغناء عنه وذلك عن طريق النظرية التالية [11] .

نظرية (2)

لأي قيمة أولية تخمينية $\widehat{\boldsymbol{\rho}}_{\mathbf{0}}$ تقع بين الصفر والواحد فان الأسلوب التكراري المحدد بالمعادلتين (15) , (14) يقترب من المقدر التالي عندما $\boldsymbol{j \to 0}$.

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{AOE}} = (\mathbf{1} - \widehat{\boldsymbol{\rho}}^{*})\widehat{\boldsymbol{S}} + \widehat{\boldsymbol{\rho}}^{*}\widehat{\boldsymbol{F}} \qquad \textbf{... (16)}$$

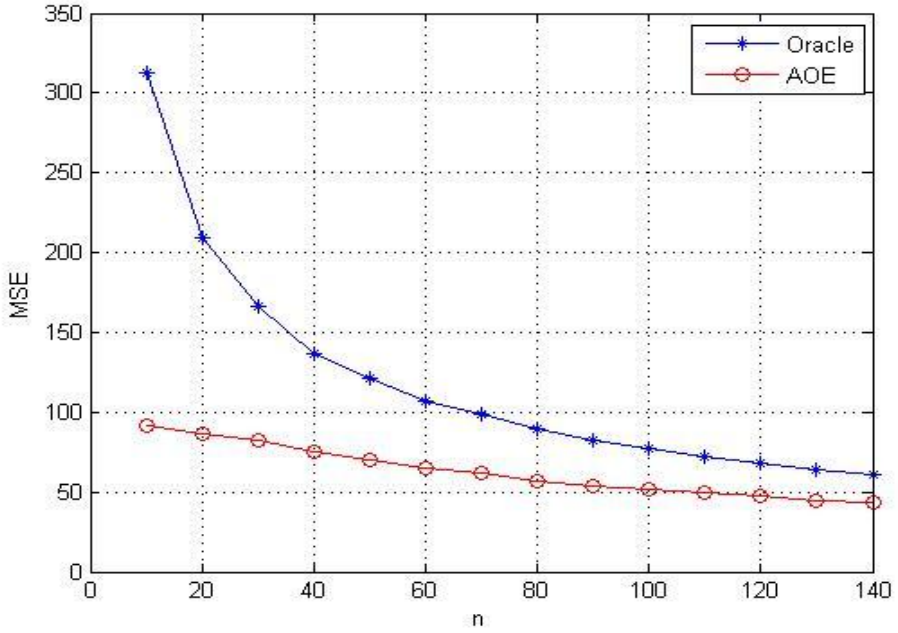
علما ان :

$$\widehat{\boldsymbol{\rho}}^{*} = \boldsymbol{min}\left(\frac{(\mathbf{1}-{}^{\mathbf{2}}/_{\boldsymbol{p}})\boldsymbol{Tr}(\widehat{\boldsymbol{S}}^{\mathbf{2}})+\boldsymbol{Tr}^{\mathbf{2}}(\widehat{\boldsymbol{S}})}{(\boldsymbol{n}+\mathbf{1}-{}^{\mathbf{2}}/_{\boldsymbol{p}})\left[\boldsymbol{Tr}(\widehat{\boldsymbol{S}}^{\mathbf{2}})-{}^{\boldsymbol{Tr}^{\mathbf{2}}(\widehat{\boldsymbol{S}})}/_{\boldsymbol{p}}\right]} \ , \mathbf{1} \right) \qquad \textbf{... (17)}$$

حيث ان المقدر بالمعادلة (16) هو مقدر الاوراكل التقريبي بمعلمة ربط كما هو في المعادلة (17)

8- المحاكاة ***Simulation***

تم توليد p =100 من المتغيرات التي تتوزع توزيعا طبيعيا بمتوسط صفر وتباين مساوي الى واحد حسب طريقة بوكس ميللر Box– Muller بأحجام عينات مختلفة (n= 10 , 20 , 30 ,...140) واحتسبت من تلك المتغيرات مصفوفة التباين والتباين المشترك للعينة وكررت التجربة بعدة تكرارات متنوعة (r = 1000 , 5000 , 10000) وتم احتساب مقدر الاوراكل حسب المعادلتين (6) , (5) وكذلك مقدر الاوراكل التقريبي كما في المعادلتين (16) , (17) وتم المقارنة بينهما بالاعتماد على MMSE كما هو في المعادلة (1) .



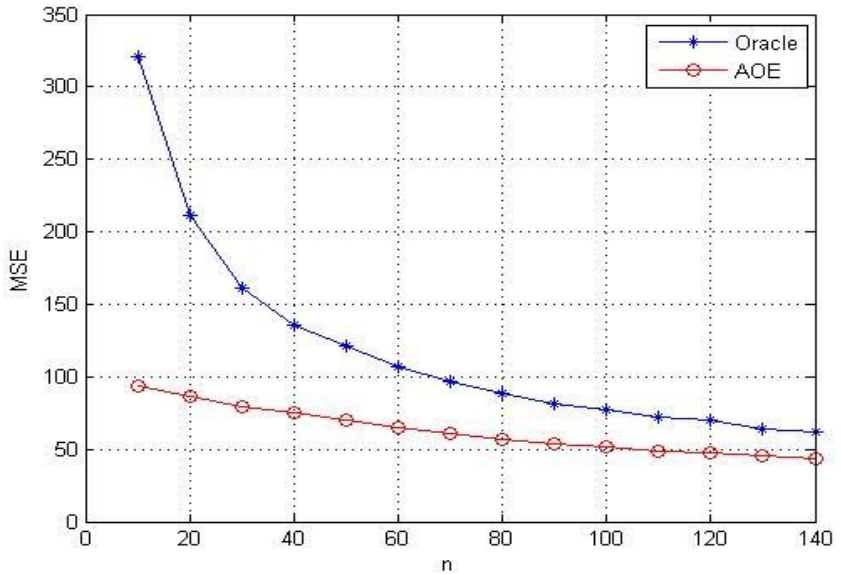
الشكل رقم (1)

مقدري Oracle وAOE عندما $p = 100 , r = 1000$

n	Oracle	AOE
10	312.2173	91.1616
20	209.4165	86.8229
30	166.3112	82.5460
40	136.3914	75.3670
50	120.8997	70.2532
60	106.4973	65.0109
70	98.2996	61.9502
80	89.2408	57.2698
90	82.5685	54.1869
100	77.6316	51.6624
110	72.4472	49.3338
120	68.5263	47.5158
130	63.9570	44.9679
140	61.0622	43.6546

جدول رقم (1)

ويمثل قيم MMSE لمقدري Oracle و AOE للشكل اعلاه

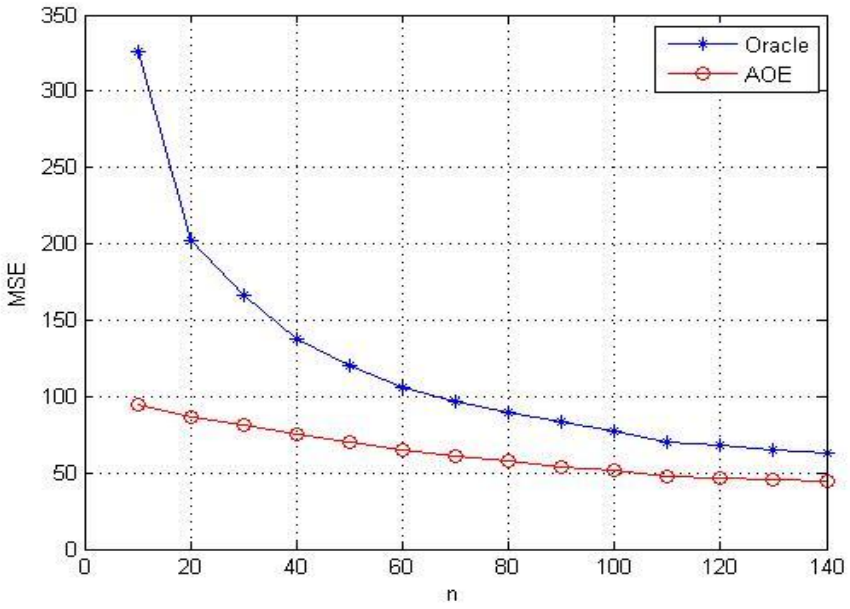


الشكل رقم (2)

مقدري Oracle وAOE عندما $p = 100 , r = 5000$

n	Oracle	AOE
10	320.3267	94.0545
20	211.1353	86.5023
30	161.5367	79.5672
40	135.8226	74.8482
50	121.1818	70.5334
60	107.0512	65.2931
70	96.4881	60.7108
80	88.9019	57.2399
90	81.6504	53.9329
100	76.9755	51.2332
110	72.3948	49.0166
120	69.7239	47.9203
130	64.2694	45.4824
140	61.5794	43.9396

جدول رقم (2)
ويمثل قيم MMSE لمقدري Oracle و AOE للشكل اعلاه



الشكل رقم (3)
مقدري Oracle وAOE عندما $p = 100 , r = 10000$

n	Oracle	AOE
10	326.2894	94.7494
20	201.9475	86.7880
30	166.6012	80.8861
40	138.0118	75.3607
50	120.6194	70.5447
60	106.0532	65.4387
70	97.0103	61.3666
80	89.8112	57.9252
90	83.0150	54.2015
100	77.3346	51.4727
110	70.4441	48.0827
120	67.7989	46.8488
130	64.7957	45.5870
140	62.5437	44.6037

جدول رقم (3)

ويمثل قيم MMSE لمقدري Oracle و AOE للشكل اعلاه

من خلال ملاحظة الشكل رقم (1) والجدول رقم (1) نلاحظ ان مقدر الاوراكل التقريبي AOE يكون افضل بشكل كبير من مقدر الاوراكل عندما يكون حجم العينة اقل من عدد المتغيرات فيها وتستمر افضلية مقدر الاوراكل التقريبي عندما يكون حجم العينة اكبر من عدد المتغيرات.
و من خلال ملاحظة الشكل رقم (2) والجدول رقم (2) بزيادة تكرار التجربة نلاحظ ان مقدر الاوراكل التقريبي AOE يكون افضل بشكل كبير من مقدر الاوراكل عندما يكون حجم العينة اقل من عدد المتغيرات فيها وتستمر افضلية مقدر الاوراكل التقريبي عندما يكون حجم العينة اكبر من عدد المتغيرات.
و من خلال ملاحظة الشكل رقم (3) والجدول رقم (3) و بزيادة تكرار التجربة نلاحظ ان مقدر الاوراكل التقريبي AOE يكون افضل بشكل كبير من مقدر الاوراكل عندما يكون حجم العينة اقل من عدد المتغيرات فيها أي بقاء الحال مشابه لما هو عليه وتستمر افضلية مقدر الاوراكل التقريبي عندما يكون حجم العينة اكبر من عدد المتغيرات.

9- البيانات الحقيقية

تم الحصول على عينة مكونة من 15 قطعة ارض حيث تم قياس 16 متغير لكل قطعة ارض تمثل الخصائص الكيميائية لتربة حوض النهر في محافظة واسط حيث كانت كالاتي :

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
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
10	16.32	13.40	2.36	0.02	0.97	7.97	7.89	1.05	3.87	14.96	7.34	1.44	6.87	24.35	28.10	23.70
11	13.28	8.13	2.49	0.01	1.05	6.26	3.76	0.77	3.34	14.35	7.63	1.60	1.45	21.64	33.95	25.10
12	9.68	4.37	2.44	0.05	1.01	3.82	2.17	1.25	2.49	15.21	7.29	1.16	2.56	23.50	30.50	25.55
13	18.98	5.41	2.64	0.03	0.99	14.64	4.22	3.08	4.34	27.59	7.75	1.40	4.49	36.60	0.24	22.85
14	15.89	4.46	2.53	0.02	0.71	14.36	7.93	2.32	3.38	30.02	7.51	1.14	2.68	33.70	1.23	21.20
15	18.42	5.91	2.05	0.02	0.95	10.70	9.27	4.17	4.28	28.69	7.52	1.00	3.80	29.18	2.30	23.20

جدول رقم (4)
الخصائص الكيميائية لتربة حوض النهر

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

جدول رقم (5)
توضيح المتغيرات

وتم احتساب مقدر الاوراكل حسب المعادلتين (6) , (5) وكذلك مقدر الاوراكل التقريبي كما في المعادلتين (16) , (17) وتم احتساب MMSE كما هو في المعادلة رقم (1) للبيانات اعلاه

	$\widehat{\rho}$	MMSE
Oracle	0.1682	1020.9
AOE	0.1911	1317.5

جدول رقم (6)
مقدري Oracle و OAE للبيانات الحقيقية

10- الاستنتاجات والتوصيات

من خلال ملاحظة الاشكال (3) , (2) , (1) تتضح افضلية مقدر الاوراكل التقريبي AOE خصوصا عندما يكون حجم العينة صغير جدا بالنسبة الى عدد المتغيرات ولكن المقدرين يقتربان من بعضهما عند ازدياد حجم العينة او اقتراب حجم العينة من عدد المتغيرات تحت افتراضات التوزيع الطبيعي .

اما بالنسبة للبيانات الحقيقية فاتضحت افضلية مقدر الاوراكل على مقدر الاوراكل التقريبي استنادا الى قيمة MMSE مع ملاحظة ان حجم العينة مساوي تقريبا لعدد المتغيرات.

يوصي الباحث باعتماد مقدر الاوراكل التقريبي عندما تكون احجام العينات صغيره نسبة الى اعداد المتغيرات.

كذلك يوصي الباحث باعتماد مقدر الاوراكل في حالة كون عدد المتغيرات صغير جدا بالنسبة الى حجم العينة .

كذلك يوصي الباحث بدراسة انواع اخرى لمقدرات مصفوفة التباين والتباين المشترك كالمقدرات الحصينة والمقدرات اللامعلمية .

المصادر

1- Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. Ann. Statist. 36 199–227.

2- Buhlmann, P. and Van De Geer.(2011).Statistics for High-dimensional Data Methods, Theory and Applications.Sprenger Press ..

3- Dey, D. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss, The Annals of Statistics, vol. 13, no. 4, pp. 1581-1591.

4- Frost, P. A. and Savarino, J. E.(1986). An empirical Bayes Approach to Portfolio selection. Journal of Finincial and Quantitave Analysis, 21: 293- 305

5- Fujikoshi, Y. , Ulyanov, V. , Shimizo, R.(2011). Multivariate Statistics : High-Dimensional and Large-Sample Approximations. Wiley Series in Probability and Statistics

6- Haff. L, (1980). Empirical Bayes estimation of the multivariate normal covariance matrix," The Annals of Statistics, vol. 8, no. 3, pp. 586-597.

7- Horel, A. E. & Kennard, R. W.(1970). Ridge regression; biased estimation for nonorthogonal problems. Technometrics 12, 55-82

8- James .W and Stein .C,(1956). Estimation with quadratic loss," in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, p. 361, University of California press,.

9- Johnstone, I. M. and Titterington, D. M. (2009). Statistical Challenges of high-dimensional data. Philosophical Transactions of The Royal Society. 367, 4237-4253

10- 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

11- Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics. Vol. 40, No. 2, 1024–1060

12- Pourahmadi, M. (2013). High-Dimensional Covariance Estimation: With High-Dimensional Data. Wiley Series in Probability and Statistics.

13- Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39the Annual Meeting IMS. Atlanta, Georgia