ISSN (1681- 6870)

استعمال خوارزمية توقع التعظيم الشرطي (ECM) لتقدير القيم المفقودة والمقارنة بين طريقة (MLE) والخوارزمية الجينية (GA) في تقدير معلمات التوزيع الطبيعي الملتوي متعدد المتغيرات (MSN)[1]

أ.د قتيبة نبيل نايف
dr.qutaiba.n@gmail.com
لينا نضال شوكت
lola1990@yahoo.com
كلية الادارة والاقتصاد - جامعة بغداد، بغداد - العراق

المستخلص

إن تقدير المعلمات الإحصائية لبيانات متعددة المتغيرات تؤدي إلى إهدار المعلومات إذا ما تم اهمال القيم المفقودة ، وبالتالي فأن ذلك يؤدي الى تقديرات غير دقيقة، لذا يجب تقدير القيم المفقودة بإحدى طرق التقدير الإحصائية ، للحصول على نتائج دقيقة وبالتالي الحصول على تقديرات معلمات جيدة .

يهدف البحث الى تقدير القيم المفقودة لدالة التوزيع الطبيعي الملتوي المتعدد المتغيرات (MSN) باستعمال خوارزمية توقع التعظيم الشرطي Expectation Conditional Maximization (ECM) . ومن ثم يتم ايجاد مقدرات المعلمات للبيانات بعد تقدير القيم المفقودة عن طريق مقدرات الإمكان الأعظم ($Maximum\ likelihood\ Estimation$ (MLE)) باستعمال خوارزمية نيوتن رافسون (Newton Raphson Algorithm)، و استعمال الخوارزمية الجينية (Genetic Algorithm (GA)). و باستعمال أسلوب المحاكاة من خلال إيجاد متوسط مربعات الخطأ (MSE) للدالة لمعرفة افضل طريقة للتقدير من خلال المقارنة بين الطريقتين و بأحجام عينة مختلفة (n= 400 , 600 800)، وأثبتت الخوارزمية الجينية (GA) المعتمدة على خوارزمية توقع التعظيم الشرطي (ECM) لتقدير القيم المفقودة، كفاءتها و تفوقها على طريقة *(MLE)* من حيث النتائج التي تم التوصل اليها.

الكلمات المفتاحية: التوزيع الطبيعي الملتوي متعدد المتغيرات (MSN)، خوارزمية توقع التعظيم الشرطي (ECM)، مقدرات الإمكان الأعظم (MLE)، خوارزمية نيوتن رافسون، الخوارزمية الجينية (GA)، متوسط مربعات الخطأ (MSE).

Using (ECM) Algorithm to Estimate the Missing Values and Make Comparison between (MLE) and (GA) Algorithm for Estimating Parameters of Multivariate Skew Normal Distribution (MSN)

Prof. Dr. Qutaiba N. Nayef
dr.qutaiba.n@gmail.com
Lina N. Shawkat
lola1990@yahoo.com
College of Administration and Economics - University of Baghdad, Baghdad - Iraq
Received 10/8/2020 **Accepted 4/10/2020**

Abstract: *The estimation of statistical parameters for multivariate data leads to waste in the information if the missing values are neglected, which will subsequently lead to inaccurate estimates. Therefore, the incomplete data must be estimated using one of the statistical estimation methods to obtain accurate results and thus obtaining good estimates for the parameters*.

The aim of this paper is to estimate the missing values for the multivariate skew normal distribution function using the Expectation Conditional Maximization (ECM) algorithm. After estimating the missing values, the parameters are estimated using Maximum Likelihood Estimation (MLE) with the Newton-Raphson algorithm, as well as using the Genetic Algorithm (GA). Using simulation, the

---

[1] البحث مستل من رسالة ماجستير.

ISSN (1681- 6870)

Mean Squared Error (MSE) was calculated to find out which method is the best for estimation by comparing the two methods using different sample sizes (400, 600, and 800). The (GA) that is based on the (ECM) algorithm to estimate the missing values proved to be better and more efficient than the (MLE) method in terms of the results.

Keywords: Multivariate skew normal distribution (MSN), Expectation Conditional Maximization algorithm (ECM), Maximum Likelihood Estimation (MLE), Newton Raphson algorithm, Genetic Algorithm (GA), Mean Squared Error (MSE).

1. المقدمة

يعد التوزيع الطبيعي الملتوي من عائلة التوزيعات التي يتضمنها التوزيع الطبيعي، الا ان التوزيع الطبيعي الملتوي يحتوي على معلمة إضافية لتنظيم الالتواء، ويعرف الالتواء هو عدم تناسق في التوزيع الإحصائي، اذ يظهر فيه المنحنى مشوهاً أو مائلاً إما إلى اليسار أو إلى اليمين، يمكن تحديد الانحراف لتحديد مدى اختلاف التوزيع عن التوزيع الطبيعي. و يظهر في الرسم البياني التوزيع الطبيعي، " كمنحنى على شكل جرس" كلاسيكي ومتماثل، المتوسط، أو المعدل، والمنوال، أو الحد الأقصى للنقطة على المنحنى، متساويان.

وتعد التوزيعات الملتوية مهمة ومفيدة في مجالات عديدة منها صياغة الأسهم المالية وعائدات البورصة، أذ أن نسبة العائدات المتوقعة على الموجودات المالية والتي تكون معرضة لمخاطر كثيرة مثل صكوك التأمين، وراس المال، وحقوق البيع والشراء لأسهم او لسلع معينة بأسعار معينة خلال مدد العقد وغيرها، عادة ما يفترض انها تتوزع توزيع طبيعي الا انها تكون عرضة للتكتل اما بالاتجاه السالب او بالاتجاه الموجب، اذ ان التكتل بالاتجاه السالب يقود الى نماذج التواء سالبة، اما التكتل بالاتجاه الموجب يقود الى نماذج التواء موجبة.

ان ظهور مشكلة البيانات المفقودة في توزيع متعدد المتغيرات الطبيعي الملتوي (MSN) هي من المشكلات الشائعة عند جمع البيانات و تحليلها، وتعني فقدان جزء من بيانات العينة ، كأن يكون فقدان البيانات مثلاً في تجربة صناعية، أي تكون بعض النتائج مفقودة بسبب الأعطال الميكانيكية غير المرتبطة بالعملية التجريبية، او مثلاً في استطلاع للرأي، قد لا يتمكن بعض الأفراد من التعبير عن تفضيل مرشح واحد على آخر، او على سبيل المثال، قد يرفض المجيبون في استطلاع للأسرة الإبلاغ عن الدخل او يرفضون الإجابة عن بعض الأسئلة الموجهة اليهم، او عن طريق تلف جزء من البيانات او فقدانها، ان البيانات المفقودة تعد من المشكلات الكبيرة التي تواجه الباحث، وان الأساليب الإحصائية المستعملة لتحليل البيانات، تفترض وجود معلومات تامة عن جميع المتغيّرات المستخدمة في التحليل، وعدمُ معالجتها بشكلٍ مناسب قد يُسبّب للباحث بعض المشكلات منها عدم تقدير التباين بشكل صحيح، أو الحصول على نتائج متحيّزة، لذا فمن من الواجب تقدير البيانات المفقودة باستعمال بعض الطرق الإحصائية. و من هنا سيتم تقدير البيانات المفقودة باستعمال خوارزمية توقع التعظيم الشرطي (ECM).

في هذا البحث سيتم التطرق الى اهم الأنماط والاليات للبيانات المفقودة، وطرق تقدير القيم المفقودة ، التوزيع الطبيعي الملتوي متعدد المتغيرات (MSN) والذي تم تقديمه لأول مرة من قبل [3] $(Azzalini\ and\ Dalla\ Valle)$ عام(1996) ، والذي يعد من التوزيعات المهمة ذو الثلاث معلمات وهي: معلمة الموقع (ξ)، ومعلمة القياس ($\boldsymbol{\Sigma}$)، ومعلمة الالتواء (معلمة الشكل)($\boldsymbol{\Lambda}$).

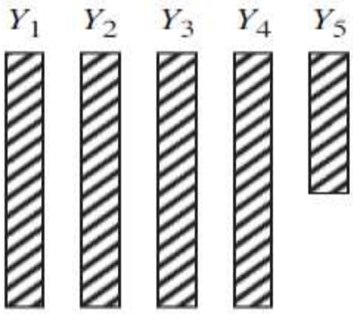
2. أنماط البيانات المفقودة Patterns of data Missingness

توجد أنماط مختلفة للبيانات المفقودة منها النمط العام (General Pattern)و الانماط الخاصة (Special Patterns)، كما توجد طرائق إحصائية مناسبة للأنماط الخاصة من البيانات المفقودة والتي يمكن ترتيبها بشكل محدد ، كما وان معرفة هذه الأنماط تساعد الباحث لمعرفة الطريقة الإحصائية المناسبة والتي تلائم تقدير معلمات الدالة او الانموذج . وعادة يتم اللجوء الى الطرائق الإحصائية التقدير ذات الأنماط الخاصة لأنها سهلة التطبيق وذات خطوات واضحة و متسلسلة بعيدة عن التعقيد ، عكس الطرائق الإحصائية للنمط العام من البيانات المفقودة والتي تكون معقدة وصعبة. ما يجعل العديد من الباحثين يميلون الى ترتيب البيانات وفق نمط محدد لتجنب استعمال الطرائق الحسابية المعقدة. وعليه فإن أنماط البيانات المفقودة يمكن ان تقسم الى قسمين هي الأنماط الخاصة و النمط العام ، ويمكن تعريف كل نمط منها على حدة وكالاتي : [1]

1) **النمط الأول: نمط فقدان البيانات لمتغير واحد: Pattern of Univariate Missing Data**

يعتبر من ابسط حالات البيانات المفقودة وهو من الأنماط الخاصة والتي تكون فيها جميع المتغيرات تامة المشاهدة فيما عدا متغير واحد يحتوي على قيم مفقودة في قسم من مشاهداته .وكما في الشكل التالي: [2]

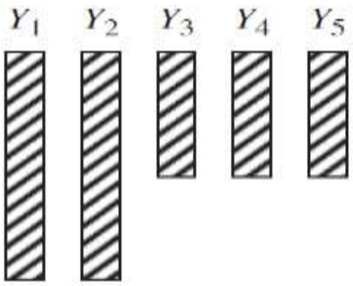
ISSN (1681- 6870)



شكل (1): نمط الفقدان لمتغير واحد

2) النمط الثاني: متعدد المتغيرات ذات النمطين :Multivariate with two Patterns

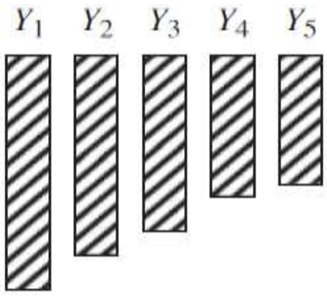
يعد هذا النمط من الأنماط الخاصة في البيانات ذات المتغيرات المتعددة والتي تكون فيه بعض المتغيرات تامة المشاهدة اما البعض الاخر فتكون مفقودة ومتساوية في الفقدان. وكما مبين في الشكل ادناه [7]



شكل (2): متعدد المتغيرات ذات النمطين

3) النمط الثالث: النمط المرتب او المتداخل Monotone or Nested Missing Data

يعد هذا النمط من الأنماط الخاصة التي يتم ترتيب البيانات فيه تصاعدياً او تنازلياً وفقاً لعدد القيم المفقودة، ويتكون هذا النمط عادة عند اختيار عينة لحساب عدد من المتغيرات التوضيحية ومن ثم سحب عينة جزئية لحساب عدد من المتغيرات التوضيحية الأخرى. وكما مبين في الشكل ادناه: [1]

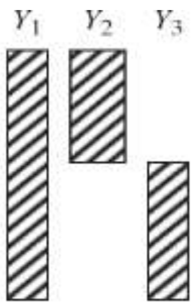


شكل (3): النمط المتداخل او المرتب

4) النمط الرابع: نمط البيانات المفقودة في حالة عدم تطابق المعلمات Missing Data with Unidentified Parameters

يدعى بنمط البيانات المفقودة في حالة عدم تطابق المعلمات ويعتبر هذا النمط هو اخر الأنماط الخاصة ، ويتكون في حالة مشاهدات متغيرين Y_2 و Y_3 غير مسجلة في مشاهدات واحدة أي ان أي مشاهدة في Y_2 يقابلها مشاهدة مفقودة في المتغير Y_3 ، ويتكون هذا النمط عند دمج او توليف عينتين، ويمكن ملاحظته في الشكل التالي (4-2) : [2]

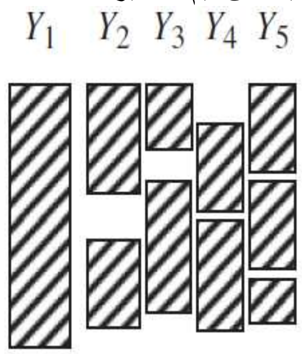
ISSN (1681- 6870)



شكل (4) : نمط الفقدان للبيانات في حالة عدم تطابق المعلمات

5) النمط الخامس: النمط العام General Pattern

ويكون نمط فقدان البيانات عشوائياً لأي قيمة من قيم المتغيرات تحت الدراسة. الشكل التالي (5-2) يوضح هذا النمط. [1]



شكل (5): النمط العام لفقدان البيانات

3. آلية فقدان البيانات Missing Data Mechanism

ان الطرائق الخاصة بتحليل البيانات التي تحوي قيماً مفقودة تختلف في فرضياتها حول الالية التي تؤدي الى فقدان البيانات . وان فهم الالية وتحديد طبيعتها مهم جدا لاختيار الطريقة المناسبة للتحليل، يمكن تقسيم آلية الفقدان كما يلي: [5]

1) الفقدان العشوائي التام للبيانات Missing Completely At Random (MCAR)

ويحدث الفقدان العشوائي التام للبيانات (MCAR)عندما يكون سبب الفقدان مستقلا عن القيمة المفقودة نفسها و مستقلا عن قيم المتغيرات الأخرى. وتكون هذه الحالة نادرة في الفقدان.

2) الفقدان العشوائي للبيانات Missing At Random (MAR)

يحدث الفقدان العشوائي للبيانات (MAR) عندما يكون سبب الفقدان مستقل عن القيمة المفقودة نفسها و يمكن ان تكون مرتبطة او لها علاقة بقيم متغير اخر، وتكون هذه الحالة شائعة ويسهل التعامل معها.

3) الفقدان الغير عشوائي للبيانات (Missing Not At Random (MNAR))

يحدث الفقدان غير العشوائي للبيانات (MNAR) عندما يكون سبب الفقدان ناتجاً عن القيمة المفقودة نفسها أي ان (القيمة المفقودة ترتبط بالقيم الأخرى لنفس المتغير)، ويصعب التعامل مع هذه الحالة.

ويمكن ان يتم التعبير عن الية الفقدان بصيغة رياضية عن طريق التوزيع الخاص بها والتي تم اقتراحها عام (1976) م من قبل (Rubin) والمتمثل بالتوزيع الشرطي لــ(X|R) و بمعلمات غير معلومة هي: (θ) [2]

$$P(R|\,X,\theta)$$

اذ ان :

X: مصفوفة البيانات الحقيقية ذات رتبه $(n\ x\ p)$

R:: مصفوفة ثنائية تأخذ القيم (0,1) مناظرة للمصفوفة X وتدعى بمصفوفة مؤشر البيانات المفقودة $(Missing\ data\ indicatormatrix)$

وان:

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

ISSN (1681- 6870)

ولنفرض اذا كان :

$$P(R \mid X, \theta) = P(R \mid \theta) \qquad for\ all\ X^m \tag{1}$$

فتفقد البيانات بصورة عشوائية تامة $(MCAR)$
من الصيغة (1) أعلاه يتضح ان التوزيع لا يعتمد على القيم المشاهدة X^o ولا على القيم المفقودة X^m.
اما اذا كان

$$P(R \mid X, \theta) = P(R \mid X^o, \theta) \qquad for\ all\ X^m \tag{2}$$

فتفقد البيانات عشوائياً(MAR)
يتضح من أعلاه بان التوزيع يعتمد على القيم المشاهدة X^o ولكنه لم يعتمد على القيم المفقودة X^m.
واذا كان:

$$P(R \mid X, \theta) = P(R \mid X^m, \theta) \qquad for\ all\ X^m \tag{3}$$

فان البيانات لا تفقد بصورة عشوائية (MNAR)
من الصيغة (3) يمكن ان نلاحظ ان التوزيع يعتمد على القيم المفقودة X^m.
يجب اخذ توزيع الية الفقدان في الاعتبار عند تحليل هذا النوع من البيانات، كما ويمكن ان يهمل توزيع آلية الفقدان في حالة (MAR) و (MCAR).

4. التوزيع الطبيعي الملتوي متعدد المتغيرات The multivariate skew normal Distribution (MSN)

ليكن المتجه العشوائي X يتبع توزيع (MSN) لـ p من المتغيرات [9] مع متجه الموقع $\xi \in \mathbb{R}^p$ (متجه لـ p من الابعاد وكل بعد يمثل مجموعة الاعداد الحقيقية)، Σ هي مصفوفة قياس التباين والتباين المشترك بالأبعاد $p \times p$، وان $\Lambda = Diag(\lambda)$ هي مصفوفة التواء، وان $\lambda = (\lambda_1, \dots, \lambda_p)'$ تمثل متجه معلمة الالتواء، إذا كانت دالة الكثافة الاحتمالية لها (pdf) هي:[6]

$$f(x \mid \xi, \Sigma, \Lambda) = 2^p \phi_p(x \mid \xi, \Omega) \Phi_p(\Lambda\, \Omega^{-1}(x - \xi) \mid \Delta) \tag{4}$$

ξ : يمثل متجه معلمة الموقع من درجة$(p \times 1)$ أي ان $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)'$.

وان:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

σ_{11} ، σ_{22} ، .... ،σ_{pp}: تمثل التباينات
σ_{12} ، σ_{13} ، ... ،σ_{1p} ،σ_{2p} ...: تمثل التباينات المشتركة

$$\Omega = \Sigma + \Lambda^2$$

$$\Delta = (I_p + \Lambda\, \Sigma^{-1} \Lambda)^{-1} = I_p - \Lambda\, \Omega^{-1} \Lambda$$

وان $\phi_p(\,.\mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ هي دالة الكثافة الاحتمالية $(\,pdf)$ للتوزيع الطبيعي $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ، وان $(\Phi_p\,(\,.\mid \boldsymbol{\Sigma}))$ هي دالة الكثافة التجميعية (cdf) للتوزيع الطبيعي $N_p(0\,, \boldsymbol{\Sigma})$. كما ويمكن كتابة $\boldsymbol{X} \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ للدلالة على ان $\boldsymbol{X}$ يملك دالة كثافة، معادلة (4) .

5. نماذج (MSN) مع معلومات مفقودة (MSN) models with missing information

لنفرض ان $X = (X_1, \dots, X_n)$ هي عينة عشوائية ذات حجم n حيث ان كل X_j مأخوذة من $SN_p(\xi, \Sigma, \Lambda)$ حيث ان $n > p$ ولتحليل نماذج (MSN) لمجموعات من البيانات ذات أنماط عامة مفقودة بمعنى ان المشاهدات لم تشاهد بشكل كامل .

ISSN (1681- 6870)

ولوضع معادلات التقدير لبيانات متعددة المتغيرات والتي تسمح بفقدان البيانات ، وعلى هذا الأساس نقوم بتقسيم $X_{j(p\times1)}$ الى مكونين $\left(X_j^o, X_j^m\right)$ ، حيث ان $X^o_{j\left(p_j^o\times1\right)}$ هو المكون المشاهد، وان $X^m_{j((p-p_j^o)\times1)}$ هو المكون المفقود ، إضافة الى ذلك يوجد لدينا مصفوفتين في الدراسة (M_j, O_j) ، وبذلك تتوافق مع X_j بحيث ان $X_j^o = O_jX_j$ ، و $X_j^m = M_jX_j$ على التوالي ، أي ان (O_j) مصفوفة تعطينا القيم المشاهدة في (X)، و مصفوفة M_j تعطينا القيم المفقودة في (X) .
وبدقة اكبر فأن $\boldsymbol{O}_{\boldsymbol{j}(p_j^o\times p)}$ و $\boldsymbol{M}_{j((p-p_j^o)\times p)}$ هي مصفوفات جزئية مستخرجة من صفوف المصفوفة I_p المتوافقة مع مواقع الصف لـ$\boldsymbol{X}_j^o$ و $\boldsymbol{X}_j^m$ في $\boldsymbol{X}_j$ ، على التوالي. عندما $\boldsymbol{X}_j = \boldsymbol{X}_j^o$ ، $\boldsymbol{O}_j = \boldsymbol{I}_p$، وان M_jمصفوفة صماء (صفرية)، ومن خصائص هاتين المصفوفتين :[6]

a) $X_j = O_j' \, X_j^o + M_j' \, X_j^m$ (5)

b) $O_j' \, O_j + M_j' M_j = I_p$ (6)

6. خوارزمية توقع التعظيم الشرطي Expectation Conditional Maximization (ECM) Algorithm

تعد خوارزمية (EM) هي من الأدوات التكرارية الشائعة لتقديرات الإمكان الأعظم (ML) للنماذج ذات البيانات المفقودة [4] ، إضافة الى امتلاكها بعض الخصائص المرغوبة مثل ثبات وتيرة او نمط التقارب و بساطة تطبيقها. مع ذلك فان خوارزمية (EM) تفقد بعض من مميزاتها عندما تصبح الخطوة M مستعصية تحليلياً او بمعنى اخر يصبح تحليلها صعباً .
خوارزمية (ECM) المقترحة من قبل [8] (Meng & Rubin) عام (1993)هي تحديث بسيط لخوارزمية (EM) والتي يتم فيها استبدال خطوة التعظيم M بسلسلة من خطوات التعظيم الشرطي CM البسيطة حسابياً . ، نقوم باستعمال خوارزمية ECM لايجاد مقدرات ML للمعلمات.
لسهولة الترميز لنفرض $\mathrm{X^o} = (\mathrm{X_1^o}, \dots, \mathrm{X_n^o})$ و $\mathrm{X^m} = (\mathrm{X_1^m}, \dots, \mathrm{X_n^m})$، هي الأجزاء المشاهدة والمفقودة على التوالي من البيانات التجريبية ، ولتكن $\tau = (\tau_1, \dots, \tau_n)$ تمثل جميع المتغيرات الكامنة (Latent variables).
دالة الإمكان اللوغاريتمية لجميع البيانات لـθ الخاصة بدالة MSN، بعد استبعاد الحدود الجمعية الثابتة هي [6]

$$\ell_c(\theta|X^O, X^m, \tau) = -\frac{1}{2}\sum_{j=1}^{n}\left\{log|\Sigma| + \left(X_j - \xi - \Lambda\tau_j\right)'\Sigma^{-1}\left(X_j - \xi - \Lambda\tau_j\right) + \tau_j'\tau_j\right\} \tag{7}$$

$\boldsymbol{\theta}$: تمثل كافة المعلمات المجهولة أي ان :

$\theta = (\xi, \Sigma , \Lambda)$

في خطوة E – (خطوة التوقع) من خوارزمية ECM يتم حساب دالة Q – و التي تمثل التوقع الشرطي لدالة الإمكان اللوغاريتمية للبيانات كافة من معادلة (7) مع العلم بالبيانات المشاهدة $\mathrm{X^o}$ و التقدير الحالي$(\hat{\theta}^{(k)})$ ، أي ان $\hat{\theta}^{(k)} = (\hat{\xi}^{(k)}, \hat{\Sigma}^{(k)}, \hat{\Lambda}^{(k)})$.
ان الجزء $-\frac{1}{2}\mathrm{E}(\tau_j'\tau_j \,|\mathrm{X}_j^o, \hat{\theta}^{(k)})$ يمكن أن يهمل لأنه لا يحتوي أي من المعلمات، و بذلك نحصل على:

$$Q\left(\theta\middle|\hat{\theta}^{(k)}\right) = \frac{1}{2}\sum_{j=1}^{n}\left\{\log|\Sigma^{-1}| - tr(\Sigma^{-1}R_j^{(k)}(\xi, \Lambda))\right\} \tag{8}$$

اذ ان:

$$R_j^{(k)}(\xi, \Lambda) = E\left(\left(X_j - \xi - \Lambda\,\tau_j\right)\left(X_j - \xi - \Lambda\tau_j\right)'|X_j^o, \hat{\theta}^{(k)}\right)$$

$$= \left(\left(I_p - \hat{\Sigma}^{(k)}\hat{S}_j^{oo(k)}\right)\hat{\Lambda}^{(k)} - \Lambda\right)\left(\hat{\Psi}_j^{(k)} - \hat{\eta}_j^{(k)}\hat{\eta}_j^{(k)'}\right)\left(\left(I_p - \hat{\Sigma}^{(k)}\hat{S}_j^{oo(k)}\right)\hat{\Lambda}^{(k)} - \Lambda\right)' + \left(I_p - \hat{\Sigma}^{(k)}\hat{S}_j^{oo(k)}\right)\hat{\Sigma}^{(k)} + \left(\hat{X}_j^{(k)} - \xi - \Lambda\hat{\eta}_j^{(k)}\right)\left(\hat{X}_j^{(k)} - \xi - \Lambda\hat{\eta}_j^{(k)}\right)' \tag{9}$$

و ان:

ISSN (1681- 6870)

$$\hat{S}_j^{oo(k)} = o_j'(o_j\hat{\Sigma}^{(k)}o_j')^{-1}o_j$$

حيث ان المعالجتين $\hat{\eta}_j^{(k)}$ و $\hat{\Psi}_j^{(k)}$ تعرف كالاتي:

$$\hat{\eta}_j^{(k)} = E\left(\tau_j \mid X_j^o, \hat{\theta}^{(k)}\right) \quad , \quad \hat{\Psi}_j^{(k)} = E\left(\tau_j\tau_j' \mid X_j^o, \hat{\theta}^{(k)}\right) \tag{10}$$

والتي يمكن تقييمها باستعمال (نظرية 1) و (نتيجة 2) والواردة في [6]، فأن التنبؤ بـ X_j في التكرار k يعطى كالاتي:

$$\hat{X}_j^{(k)} = E\left(X_j \mid X_j^o, \hat{\theta}^{(k)}\right) = \hat{\Sigma}^{(k)}\hat{S}_j^{oo(k)}X_j + \left(I_p - \hat{\Sigma}^{(k)}\hat{S}_j^{oo(k)}\right)(\hat{\xi}^{(k)} + \hat{\Lambda}^{(k)}\hat{\eta}_j^{(k)}) \tag{11}$$

باختصار، يتم تنفيذ خوارزمية ECM على النحو التالي:

- **الخطوة E:**

حساب $\hat{X}_j^{(k)}, \hat{\Psi}_j^{(k)}, \hat{\eta}_j^{(k)}$ حيث ان $(j = 1, \dots, n)$ باستعمال المعادلتين **(10)** و **(11)**

- **خطوات CM :**

✓ **الخطوة الأولى:** تحديث قيمة $(\hat{\xi}^{(k)})$ بتعظيم (8) بالاعتماد على ξ مما ينتج:

$$\hat{\xi}^{(k+1)} = \frac{1}{n}\left(\sum_{j=1}^{n}\hat{X}_j^{(k)} - \hat{\Lambda}^{(k)}\sum_{j=1}^{n}\hat{\eta}_j^{(k)}\right)$$

✓ **الخطوة الثانية:** تحديث قيمة $(\hat{\Sigma}^{(k)})$ بتعظيم (8) بالاعتماد على (Σ) مما ينتج:

$$\hat{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\hat{R}^{(k)}$$

حيث ان $(\hat{\boldsymbol{R}}_j^{(k)})$ هي $(\boldsymbol{R}_j^{(k)}(\xi, \boldsymbol{\Lambda}))$ في معادلة (9) مع (ξ) و $(\boldsymbol{\Lambda})$ التي تستبدل بـ$\hat{\xi}^{(k+1)}$ و $\hat{\boldsymbol{\Lambda}}^{(k+1)}$على التوالي.

✓ **الخطوة الثالثة:** تحديث قيمة $(\hat{\Lambda}^{(k)})$ بتعظيم (8) بالاعتماد على Λ مما ينتج:

$$\hat{\Lambda}^{(k+1)} = Diag\left\{\left(\hat{\Sigma}^{(k+1)^{-1}} \odot \sum_{j=1}^{n}\hat{\Psi}_j^{(k)}\right)^{-1}\left(\hat{\Sigma}^{(k+1)^{-1}} \odot \sum_{j=1}^{n}\hat{\Upsilon}_j^{(k+1)}\right)1_p\right\}$$

حيث ان :

$$\hat{\Upsilon}_j^{(k+1)} = \left(\hat{\Psi}_j^{(k)} - \hat{\eta}_j^{(k)}\eta_j^{(k)'}\right)\hat{\Lambda}^{(k)}\left(I_p - \hat{\Sigma}^{(k)}\hat{S}_j^{00(k)}\right) + \hat{\eta}_j^{(k)}(\hat{X}_j^{(k)} - \hat{\xi}^{(k+1)})'$$

$\odot$: يمثل حاصل الضرب هادامارد **[11]** $(Hadamard\ product)$ ، لمصفوفتين لهما نفس الابعاد.

الآن يتم استعمال طريقة بسيطة للحصول على قيم ابتدائية منطقية للمعلمات ، ولتكن $\hat{\boldsymbol{\theta}}^{(0)} = (\hat{\boldsymbol{\xi}}^{(0)}, \hat{\boldsymbol{\Sigma}}^{(0)}, \hat{\boldsymbol{\lambda}}^{(0)})$. وتكون خطوات الطريقة كما يلي: [6]

1. لمجموعة بيانات مشاهدة جزئياً ( ) X^o، ببساطة تقوم بملىء القيم المفقودة بقيمة الوسط الحسابي للقيم المشاهدة التي تتوافق مع المتغير . نرمز لهذه البيانات المخصصة بـ (X^{IM}) .
2. حساب العينة لمتجه الوسط الحسابي $(\bar{\boldsymbol{x}})$ ، وحساب العينة لمصفوفة التباين والتباين المشترك لـ $(\boldsymbol{X}^{IM})$. ويرمز لها بـ $(\boldsymbol{S})$، اذ ان $\boldsymbol{S} = [s_{ij}]$.

ISSN (1681- 6870)

3. توليد ارقام عشوائية (u) من التوزيع المنتظم $Uniform\ (0,1)$ تستعمل في استخراج قيم أولية للمعلمات ، وبعد ذلك يتم وضع ما يلي :

$$\hat{\Sigma}^{(0)} = S + (u - 1)\, Diag(S)$$

$$\hat{\lambda}_i^{(0)} = (\pm)\sqrt{(1-\mathrm{u})S_{ii}/\,(1-2/\pi)} \qquad ,i = 1,\ldots,p$$

$$\hat{\xi}^{(0)} = \bar{x} - \sqrt{2/\pi}\,\hat{\lambda}^{(0)}$$

حيث إن إشارة $\hat{\lambda}_i^{(0)}$ تعتمد على إشارة التواء العينة للمتغير رقم i .
إن خطوة $E -$ و خطوات $CM -$ تتغير بشكل متكرر حتى يتحقق شرط تقارب مناسب، فعلى سبيل المثال ان الفرق في القيم المتتالية لدالة الإمكان اللوغاريتمية هي اقل من القيمة المسموح بها ، و لتقييم الثبات للتقديرات الناتجة ينصح باستعمال قيم ابتدائية مختلفة عند تطبيق الخوارزمية .الحل الأمثل الشامل يمكن الحصول عليه من خلال مقارنة قيم الإمكان اللوغاريتمي للتقارب، ان تقديرات (ML) الناتجة يرمز لها $\hat{\boldsymbol{\theta}}^{(0)} = (\hat{\boldsymbol{\xi}}^{(0)}, \hat{\boldsymbol{\Sigma}}^{(0)}, \hat{\boldsymbol{\lambda}}^{(0)})$ من خلال (نتيجة 1) الواردة في [6] يمكن توقع المكون المفقود من $\boldsymbol{X}_j^m$ خلال الصيغة:

$$\hat{X}_j^m = M_j\left(\hat{\xi} + \hat{\Lambda}\hat{\eta}_j + \hat{\Sigma}\hat{S}_j^{oo}\left(X_j - \hat{\xi} - \hat{\Lambda}\hat{\eta}_j\right)\right)$$

حيث ان $(\hat{\boldsymbol{\eta}}_j)$ و $(\hat{\boldsymbol{\Psi}}_j)$ هي كل من $(\hat{\boldsymbol{\eta}}_j^{(k)})$ و $(\hat{\boldsymbol{\Psi}}_j^{(k)})$ في معادلة (2.10) ، و حل محله $(\hat{\theta})$ و $(\boldsymbol{S}_j^{oo})$ اذ ان :

$$S_j^{oo} = O_j'(O_j\Sigma O_j')^{-1}O_j$$

7. طريقة تقدير الإمكان الأعظم Maximum Likelihood estimation Method (MLE)

تعتبر طريقة تقدير الامكان الاعظم (MLE) من الطرق المهمة لإيجاد مقدرات المعلمات للبيانات التامة بعد تقدير القيم المفقودة كما تطرقنا اليه سابقاً وسنتطرق في هذا البحث لتقدير معلمات انموذج متعدد المتغيرات للتوزيع الطبيعي الملتوي (MSN) وبأخذ اللوغاريتم كما يلي:

$$\ell_c(\theta \mid X,\tau) = \sum_{j=1}^{n} \log f\left(X_j \mid \xi, \Sigma, \Lambda\right)$$

اي ان:

$$\ell_c\,(\boldsymbol{\theta} \mid \boldsymbol{X},\boldsymbol{\tau}) = -\frac{1}{2}\sum_{j=1}^{n} \log|\boldsymbol{\Sigma}| + \left(\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda\tau_j}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda\tau_j}\right) + \boldsymbol{\tau_j'}\,\boldsymbol{\tau_j} \qquad (12)$$

ويتم استعمال خوارزمية نيوتن رافسون لإيجاد افضل تقدير تقريبي للجذور لدالة القيمة الحقيقية.

8. خوارزمية نيوتن-رافسون Newton Raphson algorithm (NRA)

تعرف خوارزمية نيوتن رافسون (NRA) بطريقة نيوتن "*Newton's Method*" والتي تم تطويرها بعد السير إسحاق نيوتن عام 1669 . وفي وقت لاحق قام جوزيف رافسون بنشر وصف مبسط لهذه الطريقة المسماة "*Newton Raphson*" عام (1690) .
ان تقنية (NR) هي تقنية حتمية لإيجاد افضل تقدير تقريبي للصفر (او جذور) لدالة القيمة الحقيقية. ان فكرة الطريقة هي على النحو التالي: [13]
تبدأ الطريقة بتخمين اولي معقول يكون قريب الى قيمة الجذر الحقيقي. يتم تقريب دالة الفائدة باستعمال الخط المماس. بعد ذلك يتم حساب تقاطع (X) لهذا الخط المماس. هذه القيمة سوف تكون في العادة افضل تقريب لجذر الدالة من التخمين الأصلي. يتم تكرار الطريقة حتى يتم استيفاء معايير التقارب المحددة مسبقاً.
ولغرض فهم افضل لخطوات خوارزمية (NR) ، لنفترض ان المتجه (U) يعرف كالاتي:

$$U = (U_1, U_2, U_3)$$

حيث ان:

ISSN (1681- 6870)

$$U_1 = \left[\frac{\partial lnL}{\partial \xi_1}, \frac{\partial lnL}{\partial \xi_2}, \frac{\partial lnL}{\partial \xi_3}, \ldots . \frac{\partial lnL}{\partial \xi_p} \right]' = \frac{\partial lnL}{\partial \boldsymbol{\xi}}$$

$$U_2 = \left[\frac{\partial lnL}{\partial \sigma_{11}}, \frac{\partial lnL}{\partial \sigma_{12}}, \frac{\partial lnL}{\partial \sigma_{22}}, \ldots . \frac{\partial lnL}{\partial \sigma_{pp}} \right]' = \frac{\partial lnL}{Vech\ \partial \Sigma}$$

$$U_3 = \left[\frac{\partial lnL}{\partial \lambda_1}, \frac{\partial lnL}{\partial \lambda_2}, \frac{\partial lnL}{\partial \lambda_3}, \ldots . \frac{\partial lnL}{\partial \lambda_p} \right]' = \frac{\partial lnL}{\partial \Lambda}$$

$$Z = \begin{bmatrix} \frac{\partial^2 lnL}{\partial \xi^2} & \frac{\partial^2 lnL}{\partial \xi \partial \Sigma} & \frac{\partial^2 lnL}{\partial \xi \partial \Lambda} \\ \frac{\partial^2 lnL}{\partial \Sigma \partial \xi} & \frac{\partial^2 lnL}{\partial \Sigma^2} & \frac{\partial^2 lnL}{\partial \Sigma \partial \Lambda} \\ \frac{\partial^2 lnL}{\partial \Lambda \partial \xi} & \frac{\partial^2 lnL}{\partial \Lambda \partial \Sigma} & \frac{\partial^2 lnL}{\partial \lambda^2} \end{bmatrix} \quad (13)$$

أذ ان (Z) هي مصفوفة الميل التي تم الحصول عليها من شروط المشتقات الثانية الجزئية.

- **خطوات خوارزمية نيوتن رافسون NR**

✓ **الخطوة الأولى:** إعطاء قيم أولية لمعلمات التوزيع $\boldsymbol{\xi}$، $\boldsymbol{\Sigma}$، $\boldsymbol{\Lambda}$

$$\theta = (\xi^{(0)}, \Sigma^{(0)}, \Lambda^{(0)})$$

✓ **الخطوة الثانية:** حساب المتجه $U = U(\xi^{(k)}, \Sigma^{(k)}, \lambda^{(k)})$
والذي تمثل مكوناته:

$$U\left(\xi^{(k)}\right) = \frac{\partial lnL}{\partial \xi} \quad , \quad U\left(\Sigma^{(k)}\right) = \frac{\partial lnL}{Vech\ \partial \boldsymbol{\Sigma}} \quad , \quad U\left(\Lambda^{(k)}\right) = \frac{\partial lnL}{\partial \boldsymbol{\Lambda}}$$

و حساب المصفوفة $Z = Z\left(\xi^{(k)}, \Sigma^{(k)}, \Lambda^{(k)}\right)$ والتي يتم حسابها وفق **(13)** اما (k) فتمثل عدد التكرارات $k = 0,1, \ldots .$

✓ **الخطوة الثالثة:** تحديث قيم المعلمات وفق المعادلة:

$$\begin{bmatrix} \xi^{(k+1)} \\ \Sigma^{(k+1)} \\ \Lambda^{(k+1)} \end{bmatrix} = \begin{bmatrix} \xi^{(k)} \\ \Sigma^{(k)} \\ \Lambda^{(k)} \end{bmatrix} - Z^{-1}\left(\xi^{(k)}, \Sigma^{(k)}, \Lambda^{(k)}\right) U\left(\xi^{(k)}, \Sigma^{(k)}, \Lambda^{(k)}\right) \quad (14)$$

✓ **الخطوة الرابعة:** تستمر التكرارات حتى $\| \xi^{(k+1)} - \xi^{(k)},\ \Sigma^{(k+1)} - \Sigma^{(k)}, \Lambda^{(k+1)} - \Lambda^{(k)} \| < c$
حيث ان c هو ثابت صغير محدد سابقاً.

9. الخوارزمية الجينية: "Genetic Algorithm (GA)"

الخوارزمية الجينية (GA) هي في الأساس من الأفكار التطورية للانتقاء الطبيعي وعلم الوراثة، وهي طريقة بحث عشوائية (تصادفية) مفيدة و فعّالة، لمستوحاة من نظرية داروين لبقاء التطور للأصلح، و من الشائع بالطبيعة أنه في منافسة يبحث فيها الأفراد عن الموارد، يهيمن الأفراد الأكثر براعة على من هم الأضعف، الحوسبة التطورية اليوم تعتبر الخوارزمية الجينية واحدة من الأجزاء المهمة من بين طرق البحث العشوائية المستخدمة لحل مشاكل التحسين، تمثل (GA) بنية ذكية سهلة التنفيذ.

لأي مشكلة معينة تعمل(GA) لحلها عن طريق عمليات محاكاة استخدام الطبيعة، مثل الاختيار، العبور، طفرة وقبول، لتطوير حل جيد لهذه المشكلة. تبدأ الخوارزمية الجينية (GA) بمجتمع مختار بشكل عشوائي يتكون من الحلول المُحتملة و تنتهي بحلٍ أمثل عن طريق تحديثات تكرارية معينة و ذلك من خلال تقليد آليات التطور البيولوجي. في (GA)، المفردات (الحلول) الأنسب تسيطر أو تهيمن على الأضعف باستعمال هذه الآليات.[10]

ISSN (1681- 6870)

✓ **عوامل الخوارزمية الجينية (GA):**

تستخدم الخوارزمية الجينية (GA) العوامل الجينية للحفاظ على التنوع الجيني. وان لمن المهم الحفاظ على التنوع الجيني أو التباين لعملية التطور. العوامل الوراثية هي نفسها المستوحاة من التركيب الوراثي الطبيعي، فيما يلي العوامل المستخدمة في الخوارزميات الجينية: [10]

أ. التكاثر / الاختيار: عادةً ما يكون العامل الأول المطبق على السكان عبارة عن نسخة طبق الأصل. يتم اختيار الكروموسومات من السكان ليكونوا ألابوين (الصنفين) لخطوة التزاوج وإنتاج الذرية، وفقا لنظرية داروين البقاء للأصلح، أي ان الأفضل هو من يبقى على قيد الحياة ويقوم بإنشاء سلالة جديدة. يُطلق على عامل التكاثر أيضاً بعامل الاختيار لأنه في الأساس يعمل على استخراج مجموعة جينات فرعية من السكان الحاليين استنادًا إلى بعض معايير الجودة أو التعريف. ان دالة التطابق (اللياقة) (fitness function) هي قياس الجودة التي يمكنها تحديد أفضل مجموعة فرعية للجينات، وان كل جينة تحتوي على معنى معين .ويتم الحصول عليها بتحويل دالة الهدف (Objective function) الى دالة مناسبة للحل في الخوارزمية.

ب. التزاوج / إعادة التركيب: يطلق على هذا العامل الوراثي بالتزاوج لأنه يتزاوج (يجمع) بين صنفين (كروموسومات) لإنتاج سلالة جديدة (كروموسوم). وان الأساليب الأكثر استخداما لاختيار تزاوج الأصناف هي: اختيار الرتبة (Rank selection) ، اختيار بولتزمان (Boltzmann selection) ، اختيار حالة ثابتة (Steady state selection)، اختيار الدورة (Tournament selection). فكرة التزاوج هي أنه بعد جمع كروموسومات أي من الصنفين (الابوين) التي تم اختيارها بناءً على دالة معينة ، السلالات الناتجة (الكروموسومات) سوف تكون مجدية كما تكون مستمدة كنتيجة لأفضل خصائص الابوين. و وفقًا لاحتمال التزاوج المعرف من قِبل المستخدم، يتم ذلك أثناء مرحلة التطور.

ج. الطفرة: تظهر الطفرة أثناء مرحلة التطور حيث يحدد المستخدم احتمال الطفرة، عادة ما يتم تعيين (ضبط) هذا الاحتمال إلى قيمة منخفضة إلى حد ما، مثل (0.01) هو الخيار الأول الجيد. وان الطفرة هي العامل الوراثي المستخدم للحفاظ على التنوع الوراثي من جيل واحد من السكان من الكروموسومات إلى الجيل التالي.

10. خطوات الخوارزمية الجينية (GA):

تتلخص خطوات الخوارزمية الجينية كالاتي: [13]

✓ الخطوة 1: تعريف معايير التقارب، دالة الهدف، فضاء البحث (فترات الحلول المُحتملة) و معلمات (GA) الإبتدائية (مثل حجم المجتمع (N)، رقم النخبة (EN)، إحتمالية التحوّل (MP)، إحتمالية العبور (CP) و نسبة الإختيار (SR)).

✓ الخطوة 2: توليد المجتمع الابتدائي المتكون من (N) من الكروموسومات من فضاء البحث عن طريق استراتيجية التهيئة. المجتمع الابتدائي يكتب بالشكل $\left\{\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_N^{(0)}\right\}$ حيث $\theta = \{\xi, \Sigma, \Lambda\}$ في هذه الدراسة. الخوارزمية الجينية (GA) تتعامل مع مجتمع من الحلول الممكنة. كل حل مُمثل من خلال كروموسوم و دالة مطابقة (fitness function) الكروموسومات في المجتمع يتم تمثيلها بـ(θ_j)، $j = 1, \dots, N$ ، سيتم استعمال استراتيجية التوليد العشوائي.

✓ الخطوة 3: يتم حساب قيمة المطابقة (اللياقة) لكل كروموسوم في المجتمع عند أي تكرار k، $\ln L(\theta_j^{(k)})$.

✓ الخطوة 4: عند نسبة اختيار تم تحديدها سابقاً، الحلول (المفردات) التي تمتلك أضعف قيم لدالة المطابقة (اللياقة)، و في ظل تقييم المفردات عند تنفيذ الخطوة السابقة، يتم استبدالها بمفردات جديدة مولدة بشكل عشوائي. إضافةً الى ذلك ، عدد معين (EN) من المفردات، و التي تمتلك أفضل قيم مطابقة (لياقة)، يتم قبولها على إنها مفردات نخبة و التي يتم نقلها بدون أي تحديث للجيل الجديد.

✓ الخطوة 5: بتنفيذ طريقة عجلة الروليت (مبنية على أساس مبدأ إن هنالك فرصة أكبر للاختيار إن كان هناك مطابقة (لياقة) أفضل) على إنها طريقة اختيار تناسبية، يتم اختيار مفردتين مرشحتين كأبوين من المفردات، عدا المفردات النخبة.

✓ الخطوة 6: يتم تنفيذ عمليات العبور و التحوّل ، كآليات اضطراب، لترشيح مفردات طبقاً لاحتمالات (CP) و (MP). يتم تنفيذ عبور الأبوين للحصول على مفردات ذرية جديدة و تحويل مفردات جديدة. و بذلك يتم الحصول على الجيل رقم (k+1) و الذي يرمز له $\left\{\theta_1^{(k+1)}, \theta_2^{(k+1)}, \dots, \theta_N^{(k+1)}\right\}$.

✓ الخطوة 7: أخيراً، جعل k = k+1 و الاستمرار بالتكرارات مع خطوة تقييم المطابقة حتى تتحقق معايير التقارب. عندما يتوقف التطور, الحل الذي يمتلك أفضل قيمة لياقة عند المجتمع الأخير هو الحل الأفضل. القيم لأفضل حل و التي يرمز لها $\{\hat{\xi}, \hat{\Sigma}, \hat{\Lambda}\}$ يطلق عليها تقديرات المعلمات.

الجانب التجريبي

11. مفهوم المحاكاة: Concept of simulation

يعد التحليل باستعمال المحاكاة امتدادا طبيعيا للأساليب التحليلية وبصورة منطقية، اذ يعد اسلوب المحاكاة من الاساليب الرصينة باعتبارها اسلوب للاختبار قبل تطبيق التجربة على البيانات الحقيقية، اذ يتم اللجوء اليها لبعض الحالات التي لا يمكن

ISSN (1681- 6870)

تمثيلها رياضياً ولأسباب عديدة منها تعقيد صياغة المسألة المدروسة او قد تكون المسألة ذات طبيعة عشوائية او بسبب التفاعلات اللازمة لوصفها وصفاً دقيقاً، ولجميع الحالات التي يصعب صياغتها رياضياً. حيث تعرف المحاكاة بانها عملية تمثيل سلوك الظاهرة الحقيقية قيد الدراسة بشكل يكون اقرب الى الواقع، و تعد الوسائل المهمة لحل المشكلات (Problem Solving Techniques) ، كما يمكن القول بأنها الوسيلة الوحيدة والأخيرة لحل أي مشكلة اذا ما استعصى حلها بالطرق العددية (Numerical Methods) او الطرق التحليلية (Analytic Methods)، كما وتعتمد المحاكاة على طرق إعادة المعاينة (Resampling Methods) وتوليد متغيرات و ارقام عشوائية لها صفات معينة.

12. خطوات تجربة المحاكاة: Steps of the simulation experiment

بالاعتماد على دالة توزيع متعدد المتغيرات الطبيعي الملتوي (MSN) الذي تم ذكرها في الجانب النظري من المعادلة (4)، اذ تم استعمال متغيرين (X_1 و X_2) مما يجعل الدالة تتوزع التوزيع الطبيعي الملتوي ثنائي المتغيرات (Bivariate Skew Normal). $X \sim SN_2(\xi, \Sigma, \Lambda)$ ، لملائمة البيانات ، حيث ان $\xi = (\xi_1, \xi_2)'$ ، $\sigma = vech(\Sigma) = (\sigma_{11}, \sigma_{12}, \sigma_{22})'$ ، $\lambda = diag(\Lambda) = (\lambda_1, \lambda_2)'$.

سوف يتم وصف مراحل تجربة المحاكاة ومن خلال برنامج R من خلال الخطوات التالية:

✓ **المرحلة الاولى:**

في هذه الخطوة تم اختيار القيم الافتراضية لمعلمات الدالة حيث تعتبر هذه الخطوة من اهم الخطوات التي سيتم الاعتماد عليها بشكل اساسي في الخطوات اللاحقة وكما موضح في الجدول التالي:

جدول (1): يبين المعلمات الافتراضية لمعلمات دالة التوزيع الطبيعي الملتوي ثنائي المتغيرات

function	ξ_1	ξ_2	σ_{11}	σ_{21}	σ_{22}	λ_1	λ_2
1	8.0	5.0	3.0	0.0	1.0	2.0	1.0
2	9.0	6.0	4.0	0.0	2.0	3.0	2.0
3	10.0	7.0	5.0	0.0	3.0	4.0	3.0

وقد تم ايضاً اختيار احد اهم العوامل المؤثرة وهي حجوم العينات المختلفة، فقد تم اخذ حجوم عينات هي (400 ، 600 ، 800) لتنفيذ تجارب المحاكاة ، وتم تكرار التجربة (500) مرة.

✓ **المرحلة الثانية:**

في هذه المرحلة يتم توليد المتغيرات وفق توزيع ثنائي المتغيرات الطبيعي الملتوي (BSN) باستعمال الدوال الجاهزة وكما يلي:

$$[x_1 \quad x_2]' \sim BSN(L, Sigma, Lambda)$$

✓ **المرحلة الثالثة:**

يتم اخذ نسبتين للفقدان هي 12% و 20% وان الفقدان يكون حسب آلية الفقدان العشوائي (MAR) التي تم ذكرها في الجانب النظري في المعادلة (22.) للمتغير (X_2)، والتي تم توليدها وفق الصيغة التالية: [12]

➢ لتوليد نسبة فقدان 12%

$$p(y, x_1) = p(R = 1|Y = y, X_1 = x_1) = \pi_i$$
$$= 1/(1 + exp(-ln(3) - 0.3(y - \bar{y}) - 0.2(x_1 - \bar{x}_1)$$

أي ان الفقدان اعتمد على القيم المشاهدة للمتغير ولم يعتمد على القيمة المفقودة أي تكون القيمة المفقودة مستقلة عن أي قيم أخرى في البيانات.

➢ لتوليد نسبة فقدان 20% في المتغير X_2 [12]:

$$p(y, x_1) = p(R = 1|Y = y, \boldsymbol{X}_1 = x_1) = \pi_i$$
$$= 1/(1 + exp(-ln(5) - 0.2(y - \bar{y}) - 0.2(x_1 - \bar{x}_1)$$

هنا تم اخذ المتغير (Y) كعامل مساعد مرتبط بالمتغير المراد توليد الفقدان فيه. فمثلا لدينا بيانات ثلاثة متغيرات هي (X_2،X_1 ،Y) ، وهناك ارتباط منطقي لـ (Y) بالمتغيرين الاخرين وكانت بعض البيانات مفقودة في المتغير (X_2) وحسب الالية اعلاه نستخدم (X_1) والـ(Y) كما هو، بعد ذلك يتم توليد قيم ثنائية (0 ، 1) باستخدام توزيع برنولي $Ber(\pi_i)$ وبعد ذلك نفقد القيم التي تقابل 0.

ISSN (1681- 6870)

✓ **المرحلة الرابعة:**

يتم تقدير القيم المفقودة وفق طريقة توقع التعظيم الشرطي Expected Conditional Maximization (ECM)

✓ **المرحلة الخامسة:**

يتم تقدير معلمات دالة توزيع الطبيعي الملتوي ثنائي المتغيرات ، ووفق طرائق التي تم ذكرها والتي هي: طريقة تقدير الإمكان الأعظم (MLE) ، و الخوارزمية الجينية (GA).

✓ **المرحلة السادسة:**

يتم في هذه المرحلة استعمال معيار متوسط مربعات الخطأ (MSE) للأنموذج لأغراض المقارنة بين طرائق التقدير ومعرفة افضل طريقة، وكما موضح بالصيغة التالية:

$$\mathrm{MSE} = \left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{Y} - Y\right]^2\right\}$$

سيتم عرض وتحليل نتائج تجربة المحاكاة حسب القيم الافتراضية لمعلمات دالة التوزيع الطبيعي الملتوي ثنائي المتغيرات بعد تقدير القيم المفقودة لنسب فقدان (12%) و (20%) وكما موضح ادناه:

جدول (2): يبين تقدير المعلمات للدالة الأولى بنسب فقدان 12% ولكافة الحجوم.

	Estimators / Method	$\xi_1 = 8$	$\xi_2 = 5$	$\sigma_{11} = 3$	$\sigma_{12} = 0$	$\sigma_{22} = 1$	$\lambda_1 = 2$	$\lambda_2 = 1$
N=400	MLE	8.0441790	5.2526765	2.8314646	-0.3351498	1.0629585	2.0206628	0.8894530
	GA	8.8512536	5.0355592	2.4011825	-0.4991624	1.3915561	0.4804912	0.8607022
N=600	MLE	7.8693511	5.4455999	3.3255249	-0.6370371	0.9172098	2.1059666	0.6124373
	GA	8.1678361	5.2807286	2.3827517	-0.5208845	0.9916469	1.9204198	0.7053029
N=800	MLE	8.0013733	5.2031714	3.0510936	-0.3223887	1.0360373	2.1771370	0.8744561
	GA	8.3110395	5.0110056	2.3057102	-0.2000053	1.0851872	1.6209340	0.9076372

جدول (3): يبين تقدير المعلمات للدالة الأولى بنسب فقدان 20% ولكافة الحجوم

	Estimators / Method	$\xi_1 = 8$	$\xi_2 = 5$	$\sigma_1 = 3$	$\sigma_{12} = 0$	$\sigma_2 = 1$	$\lambda_1 = 2$	$\lambda_2 = 1$
N=400	MLE	9.0517242	6.3225301	3.5442830	-0.3765316	1.8371492	2.3503017	1.5206025
	GA	9.1892460	6.2267584	3.1090116	-0.3251702	2.0084950	1.9214479	1.5802468
N=600	MLE	8.8417782	6.7031423	4.1521574	-0.9952034	1.7710069	2.6766105	1.0087986
	GA	9.0635829	6.4753784	3.1960073	-0.8638147	2.1208739	2.2263872	1.1684219
N=800	MLE	8.7943481	6.4204488	4.3433131	-0.5707985	1.7172992	3.7114204	1.5782344
	GA	9.0211951	6.2436377	3.2210470	-0.4688219	2.0006601	2.8216746	1.6014685

جدول (4): يبين تقدير المعلمات للدالة الثانية بنسب فقدان 12% ولكافة الحجوم

	Estimators / Method	$\xi_1 = 9$	$\xi_2 = 6$	$\sigma_1 = 4$	$\sigma_{12} = 0$	$\sigma_2 = 2$	$\lambda_1 = 3$	$\lambda_2 = 2$
N=400	MLE	8.9949590	6.3730277	4.2134981	-0.6120338	1.8087059	2.8449368	1.4286862
	GA	9.6427144	5.6364878	3.2356237	-0.4128737	3.3770229	1.6182936	2.1336063
N=600	MLE	9.0524209	6.2040533	4.3284431	-0.6610244	1.9182807	2.8794026	1.5509901
	GA	9.8068608	5.4674930	3.3955071	-0.7056098	3.4395166	1.5070209	2.5153355
N=800	MLE	8.9493776	6.2140330	4.1582952	-0.6431441	2.1906869	3.4550733	2.3018609
	GA	9.0293573	6.1962990	3.1312098	-0.5865183	2.4056046	3.0295645	2.0472562

ISSN (1681- 6870)

جدول (5): يبين تقدير المعلمات للدالة الثانية بنسب فقدان 20% ولكافة الحجوم

N	Estimators / Method	$\xi_1 = 9$	$\xi_2 = 6$	$\sigma_1 = 4$	$\sigma_{12} = 0$	$\sigma_2 = 2$	$\lambda_1 = 3$	$\lambda_2 = 2$
N=400	MLE	8.9537540	6.1011527	4.1424262	-0.4140544	2.1852189	3.7113977	2.2142177
	GA	9.1389378	5.9888316	3.2152232	-0.3463338	2.3136926	2.9847227	2.0307128
N=600	MLE	8.595463	6.720478	5.351359	-1.085898	1.931245	3.791096	1.345967
	GA	8.9284994	6.4489723	3.5320571	-0.9064192	2.4300682	3.1540896	1.4803259
N=800	MLE	8.9149455	6.2655659	4.0400540	-0.4303733	1.8425273	2.6978843	1.4047647
	GA	9.0806569	6.1526523	3.2084572	-0.3932574	2.4006160	1.9886561	0.7767995

جدول (6): يبين تقدير المعلمات للدالة الثالثة بنسب فقدان 12% ولكافة الحجوم

N	Estimators / Method	$\xi_1 = 10$	$\xi_2 = 7$	$\sigma_1 = 5$	$\sigma_{12} = 0$	$\sigma_2 = 3$	$\lambda_1 = 4$	$\lambda_2 = 3$
N=400	MLE	9.9880589	7.2435910	5.2767775	-0.6599032	3.2612750	3.8741777	2.4735168
	GA	10.4490655	6.7591275	3.8995925	-0.5193386	5.5300774	1.8419337	2.4701777
N=600	MLE	9.9307731	7.7313203	5.3558665	-0.9416804	2.8487402	2.9724942	1.3497154
	GA	10.2456402	7.1861074	4.2770381	-0.6962784	4.0474749	2.4079542	1.6730566
N=800	MLE	10.0936380	6.9977551	4.3432149	-0.4461771	3.6298138	4.4482180	3.5536604
	GA	10.0118345	7.0307000	4.0217322	-0.4191516	3.9582871	3.8623788	2.9543135

جدول (7): يبين تقدير المعلمات للدالة الثالثة بنسب فقدان 20%ولكافة الحجوم

N	Estimators / Method	$\xi_1 = 10$	$\xi_2 = 7$	$\sigma_1 = 5$	$\sigma_{12} = 0$	$\sigma_2 = 3$	$\lambda_1 = 4$	$\lambda_2 = 3$
N=400	MLE	10.1420961	6.8504839	5.2389760	-0.5590355	3.8477954	3.9410341	3.0531280
	GA	10.2829169	6.7339789	4.2365709	-0.5019436	4.3005739	2.6975835	2.3706762
N=600	MLE	10.0672107	7.0748642	4.8588906	-0.2662033	3.1125129	3.4274634	2.4844980
	GA	10.3643388	6.6905371	3.9256183	-0.1766377	4.8144253	2.0029790	2.3697741
N=800	MLE	10.2759258	6.7436216	4.1283472	-0.3824393	4.0660372	3.6702256	3.7682956
	GA	10.3492189	6.7800959	4.0470181	-0.3572998	3.9783603	2.7704005	3.0234769

جدول (8): يمثل متوسط مربعات الخطأ (MSE) للدالة

Model	Missing Rate	N	MLE	GA_	Best
1	12%	400	0.005019439	0.0005241275	GA
		600	0.0002547981	0.000362525	MLE
		800	0.0002759635	0.000314884	MLE
	Missing Rate	N	MLE_	GA_	Best
	20%	400	0.008693741	0.0001835651	GA
		600	0.004671554	0.0001537649	GA
		800	0.003342661	0.0001476836	GA
Model	Missing Rate	N	MLE	GA	Best
2	12%	400	0.00988455	0.0002620756	GA
		600	0.007153127	0.0001866411	GA
		800	0.00058510884	0.0001056634	GA
	Missing Rate	N	MLE	GA _	Best
	20%	400	0.009645271	0.0005963493	GA
		600	0.0071077278	0.0002074227	GA
		800	0.0046906757	0.0001138706	GA
Model	Missing Rate	N	MLE	GA	Best
3	12%	400	0.006174455	0.0006173906	GA
		600	0.0051357847	0.0004087494	GA
		800	0.004749128	0.00005017162	GA
	Missing Rate	N	MLE	GA	Best
	20%	400	0.0064797482	0.0005515852	GA
		600	0.0043856181	0.00007204327	GA
		800	0.0003705512	0.00004897988	GA

ISSN (1681- 6870)

13. الاستنتاجات : (Conclusion)

يمكن الملاحظة من نتائج المحاكاة بأنه وعلى الرغم من تفاوت قيم التحيز للمعلمات المقدرة ومن الطريقتين للأنموذج الواحد ولكل من النماذج الثلاثة ولكل حجوم العينات وبنسبتي الفقدان الافتراضيتين الا ان مقدرات طريقة الخوارزمية الجينية سجلت تحيزاً اقل في قيم المتغير X_1 (والذي لا يعاني من الفقدان) مقارنة مع X_2 وعند مقارنة متوسط مربعات الخطأ لنفس التجارب أعلاه يمكن ملاحظة بأن الأفضلية كانت للخوارزمية الجينية و لأغلب التجارب ماعدا نسبة الفقدان 12% وللحجمين 600 و 800 لذا وبشكل عام كان أداء الخوارزمية الجينية افضل من أداء طريقة الإمكان الأعظم.

المصادر

[1] Abdul-Razak, Ali Salah, 2015, " Estimation of Missing Data in Panel Data Model With Practical Application", M. Sc. Thesis in Statistics, College of Administration and Economics, University of Baghdad.

[2] Al,kazaz, Qutaiba Nabeel, 2007, "A comparison of Robust Bayesian Approaches with other Methods for Estimating Parameters of Multiple Linear Regression Model with missing Data", a Dissertation of Doctor Philosophy in Statistics. College of Administration and Economics, University of Baghdad.

[3] Azzalini, A., & Valle, A. D., (1996). "The multivariate skew-normal distribution", Biometrika, 83(4), 715-726.

[4] Dempster, A. P., Laird, N. M., & Rubin, D. B., (1977), "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

[5] Hussain, Inaam Aboud, 2010, "Incomplete Data Analysis Multiple Regression Models using Algorithms EM, ECM, ECME With Practical Application", M. Sc. Thesis in Statistics, College of Administration and Economics, University of Baghdad.

[6] Lin, T. I., Ho, H. J., & Chen, C. L., (2009), "Analysis of multivariate skew normal models with incomplete data", Journal of Multivariate Analysis, 100(10), 2337-2351.

[7] Little, R. J., & Rubin, D. B. (2020). Statistical analysis with missing data, John Wiley & Sons.

[8] Meng, X. L., & Rubin, D. B., (1993), "Maximum likelihood estimation via the ECM algorithm: A general framework", Biometrika, 80(2), 267-278.

[9] Sahu, S. K., Dey, D. K., & Branco, M. D., (2003), "A new class of multivariate skew distributions with applications to Bayesian regression models", Canadian Journal of Statistics, Vol. 31, No.(2) ,129-150.

[10] Shahzad, W., Rehman, Q., & Ahmed, E., (2017), "Missing data imputation using genetic algorithm for supervised learning", International Journal of Advanced Computer Science and Applications (IJACSA).

[11] Styan, G. P., (1973), "Hadamard products and multivariate statistical analysis" , Linear algebra and its applications, 6, 217-240.

[12] Wang, Q. H., (2009), "Statistical estimation in partial linear models with covariate data missing at random" , Annals of the Institute of Statistical Mathematics, 61(1), 47-84.

[13] Yalçınkaya, A., Şenoğlu, B., & Yolcu, U., (2018), "Maximum likelihood estimation for the parameters of skew normal distribution using genetic algorithm", Swarm and Evolutionary Computation, 38, 1-28.