المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

التحقيق النظري للنماذج في خوارزمية تخفيض الأبعاد متعددة العوامل المعممة للأنماط الظاهرية الترتيبية

محمد إبراهيم عثمان **د. زيد طارق صالح الخالدي**

قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

معلومات النشر

تاريخ المقالة:
تم استلامه في24 ايلول 2023
تم القبول في 12 تشرين الثاني 2023
متاح على الإنترنت في 1 كانون الاول 2023

الكلمات الدالة:
خوارزمية تخفيض الأبعاد، الانحدار اللوجستي الترتيبي، التداخلات الجينية، الأنماط الظاهرية.

المراسلة:
محمد إبراهيم عثمان
mohammed.ibrahem.es208@gmail.com

الخلاصة

تشير الدراسات السريرية إلى العلاقة الوثيقة بين بعض الأمراض ووجود تداخلات محددة بين العوامل الجينية وكما هو الحال في كثير من الدراسات، فإن كشف التداخلات الجينية ذات التأثير الكبير على ظهور الأمراض الوراثية يحتاج إلى تحليلات احصائية مستفيضة. وبسبب الحجم الهائل للبيانات الجينية في الجنس البشري، فكان لا بد من تطوير طرق إحصائية مكيفة للتعامل مع البيانات الابعاد العالية. تعد خوارزمية تخفيض الأبعاد متعددة العوامل Multifactor Dimensionality Reduction (MDR) أحد الخوارزميات اللامعلمية الرائدة في هذا المجال. حيث تعمل الخوارزمية على تخفيض أبعاد البيانات الجينية للحصول على أهم تداخل ذات تأثير مباشر على زيادة احتمالية ظهور الأمراض الوراثية. وتعتمد الخوارزمية في تكوينها على مجموعة من الإجراءات اللامعلمية لتشخيص التداخل الجيني الأعلى تأثيراً على متغيرات الاستجابة الثنائية حصراً. وكأي طريقة إحصائية، فإن هذه الخوارزمية لا تخلو من نقاط الضعف والمحددات التطبيقية، لذا كان لا بد من تطوير الخوارزمية لتجاوز المعوقات. أحد نقاط الضعف في هذه الخوارزمية هي عدم إمكانية الخوارزمية من التعامل مع البيانات التي تحتوي على متغير استجابة من النوع الترتيبي. طور بعض الباحثين تعميماً لخوارزمية تخفيض الأبعاد متعددة العوامل لتمكينها من العامل مع البيانات الترتيبي. مع ذلك فإن الخوارزمية المعممة أكثر تعقيداً من الخوارزمية الأصلية. لذلك اقترحنا تطوير الخوارزمية المعممة تطويراً بسيطاً وذلك بتوظيف الانحدار اللوجستي الترتيبي في تصنيف الافراد في العينة، مع الإبقاء على جميع خطوات الخوارزمية الأصلية دون تغيير. ومن ناحية أخرى، فإن خوارزمية MDR تعتمد اسلوباً لا معلمياً للتحقق من معنوية التداخلات المرشحة في الخوارزمية. وينبني هذا الاجراء اللامعلمي على فكرة الاختبارات التباديلية، وهو يستهلك وقتاً زمنياً طويلاً جداً مقارنة بالإجراءات المعلمية المعتمدة على الأساليب النظرية. اقترح بعض الباحثين استخدام توزيع القيمة المتطرفة المعمم للتحقق من المعنوية الإحصائية للتداخلات المرشحة، لكن لم يرد استخدام هذا الأسلوب الا مع المتغيرات المعتمدة المستمرة والثنائية. تم في هذا البحث توظيف الاسلوب النظري المعتمد على توزيع القيمة المتطرفة المعمم بدلا من الاختبارات التباديلية المعتمدة في الخوارزمية وذلك عندما يكون متغير الاستجابة من النوع الترتيبي.

DOI: 10.33899/IQJOSS.2023.0181255

مقدمة

تعد الأنماط الظاهرية phenotypes انعكاساً للأنماط الجينية genotypes التي يحملها الانسان، مثل لون العينين، لون الشعر، شكل القدم، وغيرها من الصفات التي يمكن ملاحظتها بشكل ظاهري. ونعد الامراض الوراثية أحد الأنماط الظاهرية للتعابير الجينية التي يحملها الانسان، خصوصاً على مستوى التداخلات الجينية. حيث تعتبر القابلية على الإصابة ببعض الأمراض مرتبطة بشكلٍ كبير بالعوامل الجينية متعددة المواقع multilocus genetic factors على مستوى التأثيرات الرئيسة و/أو تأثيرات التداخلات (Pattin, White et al. 2009). تم استخدام العديد من الأساليب الإحصائية المعلمية لنمذجة العلاقة بين القابلية للإصابة بالمرض والعوامل الجينية. تتبني غالبية هذه الطرق على مفهوم النمذجة الخطية المعممة Generalized linear modeling (Gola, Mahachie John et al. 2016). ومع ذلك، نظراً للأبعاد العالية للبيانات الجينية و/أو حجم العينة الصغير نسبياً، حيث يُقدَّر عدد العوامل الجينية ثنائية الأليل في الجنس البشري بـ 84.7 مليون عامل جيني (Chauhan et. al, 2022) ، فقد لا تكون هذه الطرق فعالة في العمل في ظل هذه الظروف. ولرؤية ذلك، فإن مقدرات المربعات الصغرى الاعتيادية ordinary least squares (OLS) لمتجه معاملات الانحدار الخطي المتعدد $\boldsymbol{\beta}$ يمكن حسابه وفقًا للمعادلة رقم 1:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y} \qquad (1)$$

حيث أن $\boldsymbol{Y}$ هو متجه متغير الاستجابة ذات طول n، $\boldsymbol{X}$ تمثل مصفوفة المتغيرات التوضيحية ذات ابعاد $n \times p$، $\widehat{\boldsymbol{\beta}}$ هو متجه مقدرات OLS ذات طول p، n يمثل عدد المشاهدات، و p يمثل عدد المعلمات في أنموذج الانحدار المقدر.

عندما نواجه مشكلة كون عدد المعلمات في أنموذج الانحدار أكبر من عدد المشاهدات $p > n$، فإنَّ رتبة المصفوفة $\boldsymbol{X^T X}$ تكون مساوية لـ n على الأكثر. مما يعني أنَّ هناك مشكلة تعدد العلاقة الخطية multicollinearity في البيانات. في مثل هذه الحالة، لا يوجد معكوس للمصفوفة $\boldsymbol{X^T X}$ ، وهذا يعني أنَّ طريقة OLS غير قابلة للتطبيق. حتى في حالة استخدام معكوس معمم generalized inverse للمصفوفة $\boldsymbol{X^T X}$ لحساب $\widehat{\boldsymbol{\beta}}$ وفقاً للمعادلة 1، فإنه لن يكون هناك مقدر فريد بسبب كون أنَّ المعكوس المعمم غير فريد. علماً أنه لو تم استخدام معكوس معمم، فإنَّ تفسير معاملات الانحدار للمتغيرات التوضيحية المترابطة لن يكون دقيقاً (Sofroniou and Hutcheson 1999). في الدراسات الجينية، يمكن أن يصبح عدد المعلمات في الانموذج p كبيراً جداً، خصوصاً عندما يتم تضمين التداخلات من الرتب العليا بين العوامل الجينية في النمذجة. على سبيل المثال، فإنَّ الانموذج متعدد الحدود من الدرجة الثانية لبيانات فيها عشرة عوامل جينية ممكن أن تحتوي على 56 معلمة. فبالإضافة إلى معلمة المقطع، يكون لدينا المعلمات الخاصة بالتأثيرات الرئيسة والبالغ 10 معلمات، وجميع معلمات حدود التداخلات الثنائية في الأنموذج والبالغ 45 (توافيق 2 من 10). ويزداد عدد المعلمات في الأنموذج بشكل متسارع جداً في حال زيادة عدد العوامل المؤثرة في الأنموذج، وكذا الحال عند تضمين تداخلات من الدرجات العليا. علماً أن الدراسات الجينية غالبا ما تحتوي على عدد كبير من المتغيرات التوضيحية (العوامل الجينية)، ويكون التركيز فيها على التداخلات بين العوامل الجينية. عليه فإن ظهور خروقات في فروض التحليل للطرق المعلمية يكون وارداً جداً. لذلك تم تطوير بدائل لا معلمية للتغلب على صعوبات استخدام الطرق المعلمية.

خوارزمية تخفيض الأبعاد متعددة العوامل Multifactor-Dimensionality Reduction (MDR)

إحدى الخوارزميات اللامعلمية الرائدة في مجال تحليل ونمذجة البيانات الجينية هي خوارزمية تخفيض الأبعاد متعددة العوامل (Multifactor-Dimensionality Reduction (MDR، التي قدمها في الأصل ريتشي وآخرون (Ritchie, Hahn et al. 2001) ، حيث تم استخدامها على نطاق واسع لوصف العلاقة بين القابلية على الإصابة بالأمراض وتداخلات العوامل الجينية متعددة المواقع في دراسات المرضى والأصحاء case-control studies. علماً أن خوارزمية MDR تم تطويرها بالاعتماد على مفهوم طريقة التجزئة التوافقية combinatorial partitioning التي وصفها نيلسون وآخرون (Nelson, Kardia et al. 2001).

الهدف الرئيس لخوارزمية MDR هو تحديد التفاعل الجيني متعدد المواقع multilocus genetic interaction ذي المعنوية الأعلى إحصائياً مقارنةً مع باقي التداخلات. ويتم اختيار التداخل الأهم عن طريق تخفيض عدد المؤثرات الجينية إلى عامل واحد فقط، وذلك من خلال تصنيف كل التداخلات متعددة المواقع المتوفرة لدينا إلى تداخل ذي خطورة عالية أو خطورة منخفضة وفقاً لمعيار معين. كذلك يتم استخدام التحقيق التقاطعي cross-validation لتقييم صحة التداخل المقترح من أي درجة d لـ $d = 2,3, \dots, N-1$، حيث أن N هو عدد العوامل الكلية في البيانات. علاوةً على ذلك، يتم التحقق من أهمية التفاعل النهائي المرشح باستخدام اختبار التباديل permutation testing والذي يعتمد على توليد التوزيع التجريبي لإحصاءة المختبر الاحصائي، ومن ثم استخدام التوزيع التجريبي للوقوف على مستوى المعنوية الإحصائية للأنموذج المقترح.

يتمثل أحد أوجه القصور الشائعة في خوارزمية MDR في أنها قابلة للتطبيق على البيانات المتزنة فقط، أي أن عدد المرضى والأصحاء يكون متساوياً في مجموعة البيانات، وذلك بسبب اعتماد الخوارزمية لعتبة مساوية لواحد. ضعف كبير آخر في خوارزمية MDR هو استخدام عتبة ثابتة لتصنيف الأفراد إلى

مجموعات عالية المخاطر وأخرى منخفضة المخاطر. بغض النظر عن فوائد استخدام عتبة ثابتة، لأنها تقلل من العبء الحسابي، فقد يؤدي ذلك إلى خسارة هائلة في قوة الاختبار (Hua, Zhang et al. 2010).

ويلاحظ أيضا أن خوارزمية MDR تتعامل مع البيانات التي يكون فيها متغير الاستجابة من النوع الثنائي (مصاب، غير مصاب)، وهو ما يعد أحد نقاط الضعف الرئيسة في هذه الخوارزمية. حيث لا يمكن للخوارزمية أن تتعامل مع البيانات التي تحتوي على متغير استجابة مستمر (مثل ضغط الدم، تركيز الدهون في الدم)، أو متغير فئوي من النوع الترتيبي أو اسمي متعدد الفئات (مثل علاقة مستويات مرض سرطان الثدي بالعوامل الوراثية). ومن المآخذ الأخرى على هذه الخوارزمية هو العبء الحسابي الكبير الذي تتطلبه الخوارزمية لتحليل البيانات والتحقق من معنوية النماذج، وتحديدا الجزء الخاص باختبار التباديل. حيث يتطلب تنفيذ اختبار التباديل وقتا كبيرا باستخدام الحاسوب قد يتعدى عدة ساعات أو عدة أيام. عليه، فقد تم إجراء تعديلات وحلول بديلة لمعالجة المشاكل التي تواجهها خوارزمية MDR للتغلب على بعض المعوقات ونقاط الضعف التي تواجهها الخوارزمية.

تم التركيز في هذه البحث على تعديل خوارزمية MDR لتتعامل مع البيانات التي تحتوي على متغير استجابة من النوع الترتيبي. وكذلك تكييف الجزء الخاص بالتحقق من معنوية النماذج معلمياً وذلك بالاعتماد على الأساليب الإحصائية النظرية لغرض تخفيض الوقت المستغرق في تنفيذ الخوارزمية برمجياً.

ويمكن تلخيص طريقة MDR في الخطوات الآتية:

1. تحديد العوامل الجينية N و/أو العوامل البيئية في الدراسة.
2. استعراض التوزيع التكراري frequency distribution للبيانات في فضاء ذات بعد d لكل تفاعل مُفترض من الدرجة d، $d = 2, \dots, N-1$. أي أن بيانات أي تفاعل ثنائي الدرجة يتم تمثيلها باستخدام جداول تكرارية ثنائية الأبعاد 2-way contingency table. وبالمثل، يتم استخدام جداول تكرارية ثلاثية الأبعاد 3-way contingency cube (أو ثلاثة جداول تكرارية ثنائية الاتجاه) لتمثيل البيانات لأي تفاعل من الدرجة الثالثة، وهكذا. يتم تحديد أبعاد هذه الجداول المتقاطعة بعدد المستويات في كل عامل. على سبيل المثال، يتم تمثيل التوزيع التكراري للتفاعل cross tabulations بين عاملين لكل منهما ثلاثة مستويات باستخدام جدول تكراري ذي ابعاد 3×3. تحتوي كل خلية في الجدول التكراري على أعداد المصابين وغير المصابين بالمرض، حيث تمثل كل خلية في الجدول أحد التداخلات متعددة المواقع بين العوامل الجينية المدروسة. يتم حساب نسبة المرضى إلى الأصحاء لكل خلية، ثم تقارن هذه النسب مع العتبة threshold المحددة مسبقاً لتحديد فيما إذا كانت الخلية (التداخل الجيني) ذات خطورة عالية أم منخفضة على ظهور المرض. حيث يتم تقرير فيما إذا كان الأفراد في كل خلية معرضين لخطر الإصابة بالمرض عندما تكون نسبة المرضى إلى الأصحاء تتجاوز أو تساوي العتبة المحددة. بخلاف ذلك، فإن الأفراد الحاملين للتداخل الجيني المتعدد المواقع يُصنفون على أنهم أقل عرضة للإصابة بالمرض عندما تكون نسبة المرضى إلى الأصحاء أقل من العتبة. علما أن العتبة المستخدمة في خوارزمية MDR هي واحد (Ritchie, Hahn et al. 2001). الهدف من عملية التصنيف هو تقليل أبعاد البيانات إلى متغير تنبؤ ثنائي واحد فقط.
3. يتم اختيار نموذج مقترح (تفاعل) للرتبة d باعتباره الانموذج الذي يحتوي على أصغر خطأ تصنيف classification error (CE) لكل انموذج من الدرجة d ممكن بناؤه. للحصول على CE لكل انموذج، يتم تسجيل العدد الإجمالي للأفراد الذين تم تصنيفهم بشكل خاطئ (المرضى الذين تم تصنيفهم على أنهم منخفضو المخاطر، الأصحاء الذين تم تصنيفهم على أنهم مرتفعو المخاطر) لكل انموذج. عادةً ما يُطلق على التصنيف الخاطئ للمرضى بـ false negative error (FN)، في حين يُطلق على عدد الأصحاء المصنفين خطأً على أنهم مصابون بـ false positive error (FP).
4. بعد ذلك، ومن أجل تقييم صحة الأنموذج الذي تم اختياره في الخطوة الثالثة، يتم استخدام التحقيق التقاطعي بعشرة طويات (10-fold cross-validation CV) لكل تداخل من الدرجة d. لتنفيذ إجراء CV، يتم تقسيم البيانات عشوائياً إلى عشر مجاميع متساوية الحجم تقريباً، بحيث يكون لكل مجموعة العدد نفسه المحدد من المصابين وغير المصابين لتحقيق حالة التوازن بين عدد المرضى والأصحاء في العينة. في كل مرة، يتم استبعاد مجموعة واحدة كمجموعة بيانات اختبار، بينما تعتبر المجموعات التسعة المتبقية مجموعة بيانات تدريب. في وقت لاحق، يتم إجراء تصنيف البيانات وإجراءات اختيار الأنموذج الموضحة في الخطوتين الثانية والثالثة على مجموعة بيانات التدريب. بعد ذلك، يتم تصنيف الأفراد الذين ينتمون إلى مجموعة بيانات الاختبار (المجموعة المستبعدة) وحسب انتماءاتهم الجينية وفقاً للتنبؤ الثنائي الذي تم الحصول عليه من تنفيذ الخطوتين الثانية والثالثة على مجموعة بيانات التدريب. يتم تنفيذ إجراء الاستبعاد على كل مجموعة من المجموعات العشر، ويتم حساب أخطاء التصنيف CEs لكل انموذج من الدرجة d ممكن بناؤه باستخدام بيانات مجموعات التدريب. وبطريقة مماثلة لحساب CE، يتم حساب خطأ التنبؤ prediction error (PE) لكل مجموعة مستبعدة في الطيات العشرة. بمعنى آخر، فإن خطأ التنبؤ PE هو عدد الأفراد المصنفين بشكل خاطئ في بيانات مجموعة الاختبار. وللتخلص من التأثيرات المحتملة للتقسيم العشوائي للبيانات الى عشرة مجاميع، يتم تكرار جميع إجراءات الـ CV بالكامل عدة مرات (على سبيل المثال خمس مرات). حيث يتم إجراء تقسيم عشوائي جديد للبيانات الى عشر مجموعات متساوية الحجم في كل تكرار. ثم يتم حساب متوسط أخطاء التصنيف $\overline{CE}$

لكل انموذج من الدرجة d. بعد ذلك، يتم تحديد النماذج التي تصغر قيمة $\overline{CE}$ لكل درجة من درجات التداخلات الممكنة $d = 2,3, \dots, N-1$. أخيراً، الأنموذج الذي يمثل بشكل أفضل العلاقة بين التفاعل الوراثي متعدد المواقع وقابلية الإصابة بالمرض بين جميع النماذج المُختارة هو الأنموذج الذي يحتوي على الحد الأدنى لمتوسط أخطاء التنبؤ $\overline{PE}$، حيث يتم حساب $\overline{PE}$ بطريقة مماثلة لـ $\overline{CE}$ لكن باستخدام بيانات مجموعة الاختبار. يتم استخدام اتساق التحقيق التقاطعي Cross-validation consistency (CVC) لتقييم صلاحية الأنموذج المحدد. أي أن خوارزمية MDR تحسب عدد المرات التي يتم فيها تحديد كل أنموذج معين من جميع الطيات العشرة. يتم حساب متوسط اتساق التحقيق التقاطعي النهائي $\overline{CVC}$ لكل أنموذج مقترح بناءً على نتائج جميع التكرارات.

5. أخيرًا، للتحقق من معنوية التداخل المرشح، يتم استخدام اختبار التباديل مع 1000 مجموعة من البيانات المبدلة permuted data sets. حيث يتم خلط مشاهدات متغير الاستجابة (مصاب وغير مصاب) بشكل عشوائي بينما يتم الاحتفاظ بمعلومات المتغيرات التوضيحية كما هي. ثم يتم تطبيق خوارزمية MDR بالكامل على كل مجموعة من مجموعات البيانات المبدلة للحصول على انموذج جديد مع الـ $1000, \dots, 2,1 = i; \overline{CVC}_i$ الخاص به. وبذلك سيكون لدينا 1000 قيمة من قيم $\overline{CVC}_i$ تستخدم لبناء توزيع العدم التجريبي لـ $\overline{CVC}$ الناتجة من البيانات الأصلية. لفحص الأهمية الإحصائية للأنموذج المقترح، تتم مقارنة $\overline{CVC}$ المحسوبة من مجموعة البيانات الأصلية بالتوزيع التجريبي لـ $\overline{CVC}$ الناتج من 1000 اختبار التباديل. حيث أن:

$$p - value = \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{I}_{\{\overline{CVC} < \overline{CVC}_i\}} \qquad (2)$$

حيث أن $\mathbb{I}_{\{.\}}$ تمثل دالة الدليل indicator function. يُعدُّ الانموذج المقترح ذا دلالة إحصائية إذا كانت $p - value \leq 0.05$.

التعديلات المقترحة على خوارزمية MDR

نظراً لعدم تعامل خوارزمية MDR مع البيانات التي يكون فيها متغير الاستجابة من النوع الترتيبي، فقد تم اقتراح تعديل الخوارزمية لتتعامل مع هكذا نوع من المتغيرات وذلك من خلال تعديل الجزء الخاص بتصنيف خلايا الجداول التوافقية للتداخلات من الدرجات $d = 2,3, \dots, N-1$ إلى عوامل الخطورة. حيث أن التصنيف في خوارزمية MDR ينحصر بين (مصاب، غير مصاب). وتحسب بعدها نسبة المرضى إلى الاصحاء لتقرير فيما إذا كان التداخل الجيني ذا خطورة عالية أم لا، ومنها يتم حساب خطأي التصنيف والتقدير CE and PE. التعديل المقترح سيعتمد على توظيف الانحدار اللوجستي الترتيبي لتصنيف الخلايا إلى فئات تماثل فئات المتغير المعتمد. فلو فرضنا أن المتغير المعتمد له ثلاث فئات مختلفة وقابلة للترتيب، مثل مستويات الكولسترول في الدم (طبيعي، فوق الطبيعي، مرتفع)، فعندئذٍ يمكن استخدام الانحدار اللوجستي الترتيبي لتصنيف الخلايا وفقا لإحدى هذه الفئات الثلاثة. وبعد ذلك يتم حساب خطأ التصنيف وخطأ التقدير بنفس الطريقة السابقة. وتبقى آلية اختيار التداخلات ذات التأثير العالي على تطور المرض كما هي عليه في الخوارزمية الأصلية. وتكون صيغة الانحدار اللوجستي الترتيبي كالآتي:

$$p(y \leq k) = \frac{\exp(\alpha_k + \beta x_i)}{1 + \exp(\alpha_k + \beta x_i)} \qquad (3)$$

حيث أن x_i تمثل قيمة التداخل بين العوامل الجينية وحسب درجة التداخل، وتمثل β معامل الانحدار الخاص بالمتغير x_i، فيما تمثل α_k معلمة المقطع للفئة k من فئات المتغير المعتمد Y. وعلى سبيل المثال، التداخل الأليلي من الدرجة الثانية يتضمن على تسع تداخلات متعددة المواقع كما موضح في الجدول رقم 1 أدناه:

الجدول رقم 1: يمثل القيم التداخلات الأليلية متعددة المواقع من الدرجة الثانية بين عاملين وراثيين

		Factor B		
		BB	Bb	Bb
Factor A	AA	AABB=1	AABb=2	AAbb=3
	Aa	AaBB=4	AaBb=5	Aabb=6
	aa	aaBB=7	aaBb=8	aabb=9

والتداخلات الأليلية هي التداخلات التي تحدث بين عاملين جينيين ثنائيي الألائل. ويتم تمثيل بيانات التداخل من الدرجة الثانية أعلاه بمتغير وهمي وحسب التداخلات المتعلقة بتطور المرض. ثم يحسب خطأ التصنيف CE وخطأ التقدير PE ومعامل الاتساق CVC كما هو الحال عليه في خوارزمية تخفيض الأبعاد الأصلية.

وتم أيضاً اقتراح تعديل اسلوب التحقق من معنوية التداخلات المرشحة في خوارزمية MDR ليعتمد على التوزيعات النظرية بدلا من اختبار التباديل. حيث تكمن الفكرة في توظيف توزيع القيمة العظمى المعمم GEVD لحساب قيمة الـ $p-value$ للاحصاءة المحسوبة التي هي $\overline{CVC}$. ويعود السبب في اختيار توزيع القيمة العظمى المعمم إلى أن النماذج المرشحة من خوارزمية MDR يتم اختيارها بحيث تعظم قيمة $\overline{CVC}$. حيث يتم توليد توزيع العدم null distribution للإحصاءة عن طريق عدد محدود جداً من اختبارات التباديل (50 تباديل على الاكثر). والذي بدوره سيؤدي إلى تخفيض زمن التنفيذ بشكل كبير جدا مقارنة بأسلوب التباديل الذي يعتمد على توليد 1000 عينة تباديلية. ويتم تقدير معلمات توزيع القيمة العظمى المعمم من خلال العينة ذات الحجم 50، وتوظف لحساب قيمة الـ $p-value$ للتداخل المرشح. حيث يتم اشتقاق التوزيع النظري بالاعتماد على قيم $\overline{CVC}_i$, for $i = 1,2,\dots,50$ المولدة من اختبار التباديل في خوارزمية MDR المعدلة. حيث يتم تقدير معلمات توزيع GEV وتطبيق الدالة التراكمية لحساب الـ $p-value$ لمعامل الاتساق للتداخل المرشح، وكالآتي:

$$p(X > x) = \int_{x}^{\infty} \mathrm{e}^{-\left(1+\hat{\xi}\left(\frac{u-\hat{\mu}}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\xi}}}\right)} du \qquad (4)$$

حيث أن X تمثل المتغير الذي يتبع توزيع GEV والذي يمثل معدل معامل الاتساق $\overline{CVC}$ بالنسبة لنا، وأن x تمثل القيمة المشاهدة لـ $\overline{CVC}$ والخاصة بالتداخل المرشح من خوارزمية MDR، وأن $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ يمثلون مقدرات الإمكان الأعظم لمعلمات توزيع GEV، حيث أن μ تمثل معلمة الموقع، σ تمثل معلمة القياس، وأن ξ تمثل معلمة الشكل. ويعدُّ التداخل الجيني المقترح من خوارزمية MDR معنويا إذا كانت $p-value \leq 0.05$.

وبذلك يمكن سرد خطوات الخوارزمية المعدلة كالآتي:

1. تحديد العوامل الجينية N و/أو العوامل البيئية في الدراسة.
2. استعراض التوزيع التكراري frequency distribution للبيانات في جداول توافقية ذات بعد d لكل تفاعل مُفترض من الدرجة d، $d = 2,\dots,N-1$. تحتوي كل خلية في الجدول التوافقي على أعداد الأشخاص الحاملين لكل فئة من فئات متغير الاستجابة، حيث تمثل كل خلية في الجدول أحد التداخلات متعددة المواقع بين العوامل الجينية المدروسة. يتم استخدام الانحدار اللوجستي الترتيبي لإعادة تصنيف الخلايا ضمن كل تداخل من الدرجة d، وذلك لغرض حساب خطأ التصنيف CE لكل تداخل.
3. يتم اختيار أفضل انموذج (تداخل) من كل رتبة d بوصفه الانموذج الذي يحتوي على أصغر خطأ تصنيف classification error (CE). للحصول على CE لكل انموذج، يتم تسجيل العدد الإجمالي للأفراد الذين تم تصنيفهم بشكل خاطئ لكل انموذج، حيث يعتبر التصنيف غير صحيح إذا تم تصنيف الأشخاص بشكل مخالف للفئة التي ينتمون اليها في الواقع. ونلاحظ أنه لا يوجد أخطاء من نوع False Positive وFalse Negative في حالتنا هذه، كون أن الفئات قد تمثل جميعها أشخاصا مصابين، مثل درجات الإصابة بأحد أنواع الأورام الخبيثة.
4. بعد ذلك، ومن أجل تقييم صحة الأنموذج الذي تم اختياره في الخطوة الثالثة، يتم استخدام التحقيق التقاطعي بعشرة طويات (10-fold cross-validation CV) لكل تداخل من الدرجة d. لتنفيذ إجراء CV، يتم تقسيم البيانات عشوائياً إلى عشر مجاميع متساوية الحجم تقريباً. في كل مرة، يتم استبعاد مجموعة واحدة كمجموعة بيانات اختبار، بينما تعتبر المجموعات التسعة المتبقية مجموعة بيانات تدريب. في وقت لاحق، يتم إجراء تصنيف البيانات وإجراءات اختيار الأنموذج الموضحة في الخطوتين الثانية والثالثة على مجموعة بيانات التدريب. بعد ذلك، يتم تصنيف الأفراد الذين ينتمون إلى مجموعة بيانات الاختبار (المجموعة المستبعدة) وحسب انتماءاتهم الجينية وفقاً للتصنيف الذي تم الحصول عليه من تنفيذ الخطوتين الثانية والثالثة على مجموعة بيانات التدريب. يتم تنفيذ إجراء الاستبعاد على كل مجموعة من المجموعات العشر، ويتم حساب أخطاء التصنيف CEs لكل انموذج من الدرجة d ممكن بناؤه باستخدام بيانات مجموعات التدريب. وبطريقة مماثلة لحساب CE، يتم حساب خطأ التنبؤ prediction error (PE) لكل مجموعة مستبعدة في الطيات العشرة. بمعنى آخر، فإن خطأ التنبؤ PE هو عدد الأفراد المصنفين بشكل خاطئ في بيانات مجموعة الاختبار. وللتخلص من التأثيرات المحتملة للتقسيم العشوائي للبيانات الى عشر مجاميع، يتم تكرار جميع إجراءات الـ CV بالكامل خمس مرات. حيث يتم إجراء تقسيم عشوائي جديد للبيانات إلى عشر مجموعات متساوية الحجم في كل تكرار. ثم يتم حساب متوسط أخطاء التصنيف $\overline{CE}$ لكل انموذج من الدرجة d. بعد ذلك، يتم تحديد النماذج التي تصغر قيمة $\overline{CE}$ لكل درجة من درجات التداخلات الممكنة $d = 2,3,\dots,N-1$. أخيراً، الأنموذج الذي يمثل بشكل أفضل العلاقة بين التفاعل الوراثي متعدد المواقع وقابلية الإصابة بالمرض بين جميع النماذج المُختارة هو الأنموذج الذي يحتوي على الحد الأدنى

لمتوسط أخطاء التنبؤ $\overline{PE}$، حيث يتم حساب $\overline{PE}$ بطريقة مماثلة لـ $\overline{CE}$ لكن باستخدام بيانات مجموعة الاختبار. يتم استخدام اتساق التحقيق التقاطعي Cross-validation consistency (CVC) لتقييم صلاحية الأنموذج المحدد. أي أن خوارزمية MDR تحسب عدد المرات التي يتم فيها تحديد كل أنموذج معين من جميع الطيات العشرة. يتم حساب متوسط اتساق التحقيق التقاطعي النهائي $\overline{CVC}$ لكل أنموذج مقترح بناءً على نتائج جميع التكرارات.

5. أخيرًا، للتحقق من معنوية الأنموذج المرشح، يتم استخدام 50 عينة من البيانات المبدلة permuted data sets. حيث يتم خلط مشاهدات متغير الاستجابة (فئات متغير الاستجابة) بشكل عشوائي بينما يتم الاحتفاظ بمعلومات المتغيرات التوضيحية كما هي. ثم يتم تطبيق الخطوات السابقة بالكامل على كل مجموعة من مجموعات البيانات المبدلة للحصول على انموذج جديد مع الـ $\overline{CVC_i}; i = 1,2, \dots ,50$ الخاص به. وبذلك سيكون لدينا 50 قيمة من قيم $\overline{CVC_i}$ تستخدم لتقدير معلمات توزيع العدم لإحصاءة $\overline{CVC}$ الناتجة من البيانات الأصلية وذلك وفقا لتوزيع GEV. وللوقوف على المعنوية الإحصائية للأنموذج المقترح، يتم حساب قيمة الـ p-value كالآتي:

$$p - value = \int_{\overline{CVC}}^{\infty} \mathrm{e}^{-\left(1+\hat{\xi}\left(\frac{u-\hat{\mu}}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\xi}}}\right)} du \qquad (5)$$

ويعدُّ الانموذج المقترح ذا دلالة إحصائية إذا كانت $p - value \leq 0.05$.

دراسة محاكاة Simulation Study

لغرض اختبار فاعلية الخوارزمية وفقا للتعديلات والمقترحة في الجانب النظري في تشخيصها للتداخلات ذات التأثير العالي على الإصابة بالمرض، تم تطبيق الخوارزمية المعدلة على بيانات جينية مولدة وفقا لسيناريوهات مختلفة سيرد ذكرها لاحقاً. ولضيق الوقت وقلة الإمكانات الحاسوبية المتاحة، فقد تم الاقتصار في جميع المجموعات المولدة على توليد المعلومات الجينية لستة عوامل جينية فقط (A, B, C, D, E, F). وبذلك يكون عدد التداخلات من الدرجة الثانية الممكنة هو 15 تداخلاً. بالمقابل، يكون عدد تداخلات العوامل الجينية من الدرجة الثالثة هو 20 تداخلاً، وكما موضح في الجدول رقم 2:

الجدول رقم 2: التداخلات الجينية الممكنة في دراسة المحاكاة

التداخلات من الدرجة الثانية	التداخلات من الدرجة الثالثة
AB, AC, AD, AE, AF	ABC, ABD, ABE, ABF
BC, BD, BE, BF	ACD, ACE, ACF
CD, CE, CF	ADE, ADF
DE, DF	AEF
EF	BCD, BDE, BDF
	BDE, BDF
	BEF
	CDE, CDF
	CEF
	DEF

وعُدَّ وجود أليلين فقط لكل عامل جيني، أليل سائد يرمز له بحرف الكبير، وأليل متنحي يرمز له بحرف صغير. وعُدَّ التوزيع النظري للألائل المولدة خاضعا لقانون هاردي- وينبيرغ وفقا للاحتمالات $p(A) = p(a) = 0.5$

وبذلك تكون الاحتمالية النظرية المعتمدة في التوليد لظهور التداخل الأليلي السائد (Major Allele Combination) هي:

$$p(\text{Major Allele}) = p(AA) + p(Aa) + p(aA)$$
$$= 0.25 + 0.25 + 0.25 = 0.75$$

بينما تكون احتمالية ظهور التداخل الأليلي المتنحي (Minor Allele Combination) هي:

$$p(\text{Minor Allele}) = p(aa) = 0.25$$

حيث تم تطبيق الاحتمالات أعلاه على جميع العوامل الجينية المولدة.

وتم أولا توليد مجتمع بحجم مئة ألف مشاهدة، ثم قمنا بسحب عينة عشوائية واحدة فقط بحجم 500 مشاهدة. وتم تكرار هذا الإجراء عشرين مرة، أي أن مجموع العينات بحجم 500 التي تم سحبها عشوائياً هو 20 عينة، كل عينة مسحوبة من مجتمع مولد بشكل منفصل عن باقي المجتمعات. وتم توليد عشرة من هذه المجتمعات بحيث يكون التداخل الجيني المسبب لزيادة احتمالية تطور المرض من الدرجة الثانية، أي $(AB, AC, \dots, EF)$. فيما تم توليد المجتمعات العشرة

المتبقية بحيث يكون التداخل الجيني المرتبط بتفاقم حالة المرض من الدرجة الثالثة، أي $(ABC, ABD, \dots, DEF)$. وتم إعادة عملية التوليد اعلاه لتوليد عشرة مجموعات جديدة من البيانات بحجم 1000 مشاهدة مرتبطة بتداخل محدد من الدرجة الثانية، وعشر مجموعات أخرى بحجم 1000 مرتبطة بتداخل محدد من الدرجة الثالثة. وبالطريقة نفسها تم توليد 20 عينة بحجم 2000 مشاهدة. أي أن مجموع العينات المولدة بلغ 60 عينة في المجمل. علماً أن الدراسات الجينية تحتاج إلى عدد كبير نسبياً من المشاهدات لضمان ظهور النمط الحقيقي للأمراض الوراثية السارية في المجتمع.

ولربط مستويات الإصابة بالمرض (المتغير المعتمد) بأحد التداخلات الجينية، تم توليد متغير الاستجابة الترتيبي بثلاث فئات مرتبة (1,2,3) وفقاً لأنموذج انحدار لوجستي مفترض بحيث يكون فيه أحد التداخلات الجينية المدرجة في الجدول رقم 2 معنوياً بمعامل انحدار محدد مسبقاً، بينما تهمل بقية المتغيرات التوضيحية في الانموذج، أي اعتبار أن معاملات الانحدار لباقي المتغيرات والتداخلات تساوي صفراً. مثلاً في التداخل من الدرجة الثانية بين العاملين AB نفترض أن $\beta_{AB} = 5$ وباقي المعاملات أصفار. أو في التداخل من الدرجة الثالثة بين العوامل ABC نفترض أن $\beta_{ABC} = 4$ وباقي المعاملات أصفار، وهكذا. وتمت إضافة حد عشوائي يتوزع وفقاً للتوزيع الطبيعي بمتوسط وانحراف معياري محددين وبمعلمة انحدار تساوي واحد، وذلك لإضفاء السمة العشوائية على عملية توليد المتغير المعتمد. وتم أيضاً تحديد قيم لمعلمتي المقطع α_1, α_2. وأخيراً، تم تحديد التداخلات متعددة المواقع multilocus genetic interaction المرتبطة في تطور المرض وحسب درجة التداخل. ولتوضيح ذلك، افرض أن التداخل المستهدف هو بين العوامل الثلاثة ABC. وبما أن التداخل بين العوامل الثلاثة يمكن تمثيله بجدول تقاطعي ثلاثي الأبعاد (ثلاثة جداول توافقية كما موضح في الجدول رقم 3). فقد تم ربط التداخلات متعددة المواقع المؤشرة باللون الأحمر مع تطور المرض وذلك لضمان معنوية التداخل من الدرجة الثالثة بين العوامل الثلاثة.

الجدول رقم 3: يوضح التداخلات متعددة المواقع بين العوامل الثلاثة ABC

	BB	Bb	bb		BB	Bb	bb		BB	Bb	bb	
AA	$AABBCC$	$AABbCC$	$AAbbCC$		$AABBCc$	$AABbCc$	$AAbbCc$		$AABBcc$	$AABbcc$	$AAbbcc$	
Aa	$AaBBCC$	$AaBbCC$	$AabbCC$		$AaBBCc$	$AaBbCc$	$AabbCc$		$AaBBcc$	$AaBbcc$	$Aabbcc$	
aa	$aaBBCC$	$aaBbCC$	$aabbCC$		$aaBBCc$	$aaBbCc$	$aabbCc$		$aaBBcc$	$aaBbcc$	$aabbcc$	
	CC				Cc				cc			

بعد ذلك نطبق المعادلة الآتية:

$$p(Y_i \le k) = \frac{e^{(\alpha_k + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})}}{1 + e^{(\alpha_k + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})}} = \frac{e^{(\alpha_k + \boldsymbol{X_i \beta})}}{1 + e^{(\alpha_k + \boldsymbol{X_i \beta})}} \quad (6)$$

على المتغيرات التوضيحية المولدة في المجتمع والحد العشوائي المولد وفقاً للتوزيع الطبيعي وباستخدام المعلمات المحددة مسبقاً نحصل على الاحتمالات المقابلة لكل فئة من فئات متغير الاستجابة. ومن هذه الاحتمالات يمكن تحديد قيمة متغير الاستجابة لكل فرد من أفراد المجتمع. وبذلك تكون بيانات المجتمع جاهزة لسحب العينة العشوائية بالحجم المطلوب.

وبعد إتمام توليد البيانات وسحب العينة المطلوبة، يتم تنفيذ جميع خطوات الخوارزمية المعدلة والموضحة في الفصل الثاني من هذه الرسالة. حيث يتم استخدام بيانات كل عينة لبناء انموذج انحدار لوجستي يشمل جميع التداخلات الممكنة والموضحة في الجدول رقم 2 أعلاه بغض النظر عن معنويتها. وتم بناء الانموذج باستخدام الدالة polr من مكتبة MASS في برنامج R (Venables & Ripley, 2002). ثم يتم إعادة تصنيف الأفراد في كل عينة باستخدام الانموذج المقدر ومنها يتم حساب خطأ التصنيف $\overline{CE}$ وخطأ التقدير $\overline{PE}$ واتساق التحقق المتقاطع $\overline{CVC}$. وأخيراً يتم التحقق من معنوية التداخل المرشح بإعادة تنفيذ الخوارزمية على بيانات تبادلية بعدد 50 لكل عينة، ومنها نحسب $\overline{CVC}$ لكل تداخل مرشح من كل عينة تبادلية. حيث تم حساب مقدرات الإمكان الأعظم لمعلمات توزيع القيمة العظمى المعمم GEVD للإحصاءة $\overline{CVC}$ من كل عينة من خلال العينة التبادلية وذلك باستخدام الدالة egevd من المكتبة EnvStats في برنامج R (Millard SP, 2013). ومن التوزيع المقدر يتم حساب الـ $p - value$ للتداخل المرشح وفقا للمعادلة رقم 4 وذلك من خلال الدالة pgevd من المكتبة EnvStats في برنامج R.

يذكر أنه تم إجراء جميع إجراءات المحاكاة الموضحة في أعلاه باستخدام برنامج R (R Core Team (2023)) في بيئة نظام Windows 11 باستخدام حاسبة مدعمة بمعالج Intel Core i7-8550U.

نتائج المحاكاة

عندما $n = 500$ والتداخل من الدرجة الثانية

تم استخدام متجه المعلمات الآتي لتوليد مشاهدات المتغير المعتمد في جميع العينات العشرة:

$$[\alpha_1 \quad \alpha_2 \quad \beta_{ij} \quad \beta_e] = [-2 \quad 1 \quad 5 \quad 1]$$

حيث أن α_1 و α_2 تمثلان معلمتي المقطع لفئتي المتغير المعتمد الأولى والثانية، وأن β_{ij} تمثل معامل الانحدار للمتغير الذي يمثل التداخل بين العاملين الجينيين i,j، وأن β_e تمثل معامل الحد العشوائي الذي تمت اضافته لضمان عشوائية عملية التوليد.

وتم توليد الحد العشوائي من التوزيع الطبيعي بمتوسط صفري وانحراف معياري يساوي 3. ويوضح الجدول رقم 4 أدناه نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 500 والتداخل المسبب للمرض من الدرجة الثانية.

الجدول رقم 4: نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 500 والتداخل المسبب للمرض من الدرجة الثانية

Sample	True Model	Proposed Model	$\overline{CVC}$	$p-value$	Time (min)
1	AB	AB	7.02	0.000	18.268
2	CF	CF	7.64	0.000	18.647
3	BE	BE	8.38	0.000	18.824
4	AD	AD	7.74	0.000	17.786
5	DF	DF	2.52	0.607	18.095
6	EF	ED	4.03	0.205	17.671
7	CE	CE	7.81	0.000	18.667
8	BC	BC	6.99	0.000	18.598
9	AE	AC	3.82	0.342	18.784
10	CD	CD	7.90	0.000	17.936

نلاحظ أن الخوارزمية المعدلة تمكنت من تشخيص التداخل الحقيقي من الدرجة الثانية في 8 عينات وفشلت في حالتين فقط. كما نلاحظ أنه بالرغم من فشل الخوارزمية في تحديد التداخل الحقيقي في الحالتين السادسة والتاسعة، مع ذلك فإن الخوارزمية تمكنت من تحديد أحد العوامل الجينية الداخلة في تكوين التداخل المعتمد في توليد البينات. حيث تم تشخيص العامل E في العينة السادسة، وكذلك العامل A في العينة التاسع. ويلاحظ أيضاً أن التداخلين المذكورين غير معنويين، ما يدل على نجاعة الأسلوب المستخدم في التحقق من معنوية التداخلات المرشحة. ونلاحظ أيضا استقرار عمل الخوارزمية من ناحية الزمن المستهلك في التنفيذ، وهو وقت قصير جداً مقارنة فيما لو تم اعتماد اسلوب الاختبارات التباديلية التقليدي والذي يعتمد على توليد التوزيع التجريبي لـ $\overline{CVC}$ باستخدام 1000 عينة تباديلية.

عندما $n = 500$ والتداخل من الدرجة الثالثة

تم استخدام متجه المعلمات الآتي لتوليد مشاهدات المتغير المعتمد في جميع العينات العشرة:

$$[\alpha_1 \quad \alpha_2 \quad \beta_{ijk} \quad \beta_e] = [-2 \quad 1 \quad 5 \quad 1]$$

حيث أن α_1 و α_2 تمثلان معلمتي المقطع لفئتي المتغير المعتمد الأولى والثانية، وأن β_{ijk} تمثل معامل الانحدار للمتغير الذي يمثل التداخل بين العوامل الجينية i,j,k، وأن β_e تمثل معامل الحد العشوائي الذي تمت اضافته لضمان عشوائية عملية التوليد عملية التوليد.

وتم توليد الحد العشوائي من التوزيع الطبيعي بمتوسط صفري وانحراف معياري يساوي 3. ويوضح الجدول رقم 5 أدناه نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 500 والتداخل المسبب للمرض من الدرجة الثالثة.

الجدول رقم 5: نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 500 والتداخل المسبب للمرض من الدرجة الثالثة

Sample	True Model	Proposed Model	$\overline{CVC}$	$p-value$	Time (min)
1	BCE	ABE	3.22	0.545	24.22
2	AEF	AEF	7.14	0.000	23.82
3	ABD	ABF	2.89	0.613	23.43
4	CDE	CDE	4.15	0.036	24.38

5	DEF	DCF	2.97	0.598	23.76
6	BEF	BCD	1.84	0.748	24.68
7	ABC	ABC	6.98	0.000	24.01
8	CDF	CDF	7.67	0.000	23.68
9	ADF	BDE	1.65	0.765	23.21
10	BCE	CDF	2.03	0.730	24.95

واضح من النتائج أعلاه أن الخوارزمية لم توفق في معظم العينات في تشخيص التداخل الحقيقي من الدرجة الثالثة، وذلك على الرغم من تحديد عاملين جينيين بشكل صحيح في أغلب الحالات. ويعود السبب الرئيس في فشل الخوارزمية في تحديد التداخلات الحقيقية بشكل دقيق إلى قلة عدد المشاهدات المعتمدة في العينة. حيث أنه عند توزيع المشاهدات على الخلايا في الجدول التوافقي ثلاثي الأبعاد، يكون من الوارد جداً وقوع عدد صغير جدا من المشاهدات في بعض التداخلات الموقعية الموضحة في الجدول رقم 5، خصوصاً الخلايا الخاصة بالتداخلات الموقعية للألائل المتنحية. وكما هو معلوم إحصائياً، فإن عدم توفر معلومات كافية يؤدي إلى الحصول على نتائج غير دقيقة. مع ذلك، فإن جميع التداخلات المرشحة بشكل خاطئ تم تشخيصها على أنها غير معنوية إحصائياً. علماً أن الخوارزمية لم تتأثر بحجم العينة بشكل كبير عند تشخيص التداخلات من الدرجة الثانية وذلك لكون أن التداخلات الموقعية هي تسع تداخلات فقط. كما يلاحظ بأن زمن التنفيذ مستقر في معظم العينات المعتمدة مع زيادة طفيفة مقارنة بزمن التنفيذ في نتائج الدرجة الثانية. ويعود السبب في ذلك لكون أن النماذج الممكنة من الدرجة الثالثة أكثر منها بالنسبة للدرجة الثانية، والذي بدوره يضيف عبئاً حسابياً في عملية التحقيق التقاطعي.

ويلاحظ بشكل عام أن قيم $\overline{CVC}$ منخفضة نسبياً في الحالات التي تكون فيها حجم العينة 500 بغض النظر عن معنويتها، خصوصاً للنماذج من الدرجة الثالثة. وذلك يعود إلى صغر حجم العينة مما يقلل من فرصة تحديد التداخل الحقيقي. يذكر أن القيمة العظمى الممكنة لـ $\overline{CVC}$ هي 10 وذلك لأننا نستخدم عشر طويات في عملية التحقق المتقاطع.

يذكر أنه تم ظهور بعض الحالات المتباعدة divergence أثناء تنفيذ المحاكاة وذلك بسبب صغر حجم العينة وعدم توفر مشاهدات تكفي لظهور النمط الظاهري الترتيبي بشكل متزن ضمن مشاهدات المتغير المعتمد. ونقصد بالحالات المتباعدة هنا هي العينات التي فشلنا في بناء انموذج انحدار لوجستي ترتيبي فيها. علماً أن هذه المشكلة ظهرت في العينات ذات الحجم $n = 500$ فقط، خصوصاً في نماذج الدرجة الثالثة.

عندما $n = 1000$ والتداخل من الدرجة الثانية

تم استخدام متجه المعلمات الآتي لتوليد مشاهدات المتغير المعتمد في جميع العينات العشرة:

$$[\alpha_1 \quad \alpha_2 \quad \beta_{ij} \quad \beta_e] = [-1 \quad 2 \quad 4 \quad 1]$$

حيث تم توليد الحد العشوائي من التوزيع الطبيعي بمتوسط صفري وانحراف معياري يساوي 3. ويوضح الجدول رقم 6 أدناه نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 1000 والتداخل المسبب للمرض من الدرجة الثانية.

الجدول رقم 6: نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 1000 والتداخل المسبب للمرض من الدرجة الثانية

Sample	True Model	Proposed Model	$\overline{CVC}$	$p-value$	Time (min)
1	AB	AB	9.28	0.000	20.65
2	CF	CF	10.00	0.000	21.16
3	BE	BE	9.41	0.000	21.40
4	AD	AD	8.78	0.000	21.73
5	DF	DF	10.00	0.000	20.45
6	EF	EF	8.93	0.000	20.43
7	CE	CE	9.34	0.000	21.76
8	BC	BC	8.57	0.000	20.82
9	AE	AE	9.09	0.000	21.10
10	CD	CD	8.79	0.000	20.39

عندما $n = 1000$ والتداخل من الدرجة الثالثة

تم استخدام متجه المعلمات الآتي لتوليد مشاهدات المتغير المعتمد في جميع العينات العشرة:

$$[\alpha_1 \quad \alpha_2 \quad \beta_{ijk} \quad \beta_e] = [-1 \quad 2 \quad 4 \quad 1]$$

حيث تم توليد الحد العشوائي من التوزيع الطبيعي بمتوسط صفري وانحراف معياري يساوي 3. ويوضح الجدول رقم 7 أدناه نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 1000 والتداخل المسبب للمرض من الدرجة الثالثة.

الجدول رقم 7: نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 1000 والتداخل المسبب للمرض من الدرجة الثالثة

Sample	True Model	Proposed Model	$\overline{CVC}$	$p-value$	Time (min)
1	BCE	BCE	8.70	0.000	26.34
2	AEF	AEF	8.26	0.000	26.68
3	ABD	ABD	7.83	0.000	26.96
4	CDE	CDA	3.36	0.510	27.54
5	DEF	DEF	9.12	0.000	27.98
6	BEF	BEF	7.65	0.000	27.68
7	ABC	ABC	8.96	0.000	26.35
8	CDF	CDF	7.97	0.000	27.58
9	ADF	DEF	4.05	0.186	27.51
10	BCE	BCE	8.77	0.000	27.07

يلاحظ من النتائج في الجدول رقم 6 ورقم 7 عند مقارنتها بالنتائج الموضحة في الجدولين رقم 4 و5 تحسن أداء الخوارزمية عندما ازداد حجم العينة إلى 1000. حيث تمكنت الخوارزمية من تحديد التداخلات الحقيقية من الدرجة الثانية، وثمانية تداخلات حقيقية من الدرجة الثالثة. كما يلاحظ تحسن ملحوظ في قيم الـ $\overline{CVC}$ في كلتا الحالتين، حيث اقتربت القيم من 10 أكثر. وفيما يخص معنوية التداخلات المقترحة، فإن جميع التداخلات أظهرت معنوية إحصائية باستثناء التداخلين في العينتين الرابعة والتاسعة للتداخلات من الدرجة الثالثة، وذلك لكون أن الخوارزمية فشلت في تحديد التداخل الصحيح في هاتين العينتين. وأخيراً فإن زمن التنفيذ قد ارتفع بشكل طفيف وذلك لازدياد العبء الحسابي بسبب ازدياد حجم العينة.

عندما $n = 2000$ والتداخل من الدرجة الثانية

تم استخدام متجه المعلمات الآتي لتوليد مشاهدات المتغير المعتمد في جميع العينات العشرة:

$$[\alpha_1 \quad \alpha_2 \quad \beta_{ij} \quad \beta_e] = [-2 \quad 2 \quad 6 \quad 1]$$

حيث تم توليد الحد العشوائي من التوزيع الطبيعي بمتوسط صفري وانحراف معياري يساوي 3. ويوضح الجدول رقم 8 أدناه نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 2000 والتداخل المسبب للمرض من الدرجة الثانية.

الجدول رقم 8: نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 2000 والتداخل المسبب للمرض من الدرجة الثانية

Sample	True Model	Proposed Model	$\overline{CVC}$	$p-value$	Time (min)
1	AB	AB	10.00	0.000	24.86
2	CF	CF	10.00	0.000	24.62
3	BE	BE	9.86	0.000	24.87
4	AD	AD	10.00	0.000	24.99
5	DF	DF	10.00	0.000	25.48
6	EF	EF	10.00	0.000	25.84
7	CE	CE	9.92	0.000	24.95
8	BC	BC	10.00	0.000	25.03
9	AE	AE	10.00	0.000	25.79
10	CD	CD	10.00	0.000	24.75

عندما $n = 2000$ والتداخل من الدرجة الثالثة

تم استخدام متجه المعلمات الآتي لتوليد مشاهدات المتغير المعتمد في جميع العينات العشرة:

$$[\alpha_1 \quad \alpha_2 \quad \beta_{ijk} \quad \beta_e] = [-2 \quad 2 \quad 6 \quad 1]$$

حيث تم توليد الحد العشوائي من التوزيع الطبيعي بمتوسط صفري وانحراف معياري يساوي 3. ويوضح الجدول رقم 9 أدناه نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 2000 والتداخل المسبب للمرض من الدرجة الثالثة.

الجدول رقم 9: نتائج المحاكاة في الحالات التي يكون فيها حجم العينة 2000 والتداخل المسبب للمرض من الدرجة الثالثة

Sample	True Model	Proposed Model	$\overline{CVC}$	$p-value$	Time (min)
1	BCE	BCE	9.14	0.000	31.935
2	AEF	AEF	9.43	0.000	30.793
3	ABD	ABD	10.00	0.000	31.249
4	CDE	CDE	9.31	0.000	30.506
5	DEF	DEF	9.64	0.000	31.705
6	BEF	BEF	10.00	0.000	30.690
7	ABC	ABC	9.47	0.000	31.251
8	CDF	CDF	8.85	0.000	31.321
9	ADF	ADF	9.73	0.000	31.720
10	BCE	BCE	9.17	0.000	30.907

يلاحظ من النتائج في الجداول رقم 8 ورقم 9 عند مقارنتها بالنتائج الموضحة في الجداول رقم 4، 5، 6 و7 تحسن أداء الخوارزمية عندما ازداد حجم العينة الى 2000. حيث تمكنت الخوارزمية من تحديد التداخلات الحقيقية من الدرجة الثانية، والتداخلات الحقيقية من الدرجة الثالثة في جميع العينات المولدة. كما يلاحظ تحسن ملحوظ وكبير في قيم الـ $\overline{CVC}$ في كلتا الحالتين، حيث اقتربت القيم من 10 أكثر من السابق. وفيما يخص معنوية التداخلات المقترحة، فإن جميع التداخلات من الدرجتين الثانية والثالثة أظهرت معنوية إحصائية. وأخيراً فإن زمن التنفيذ قد ارتفع بشكل طفيف وذلك لازدياد العبء الحسابي بسبب ازدياد حجم العينة.

مما تقدم يمكن القول بأن الخوارزمية المعدلة أظهرت كفائتها في تشخيص التداخلات الحقيقة ذات العلاقة الوثيقة بظهور النمط الظاهري عندما يكون متغير الاستجابة من النوع الترتيبي. مع ذلك، فإن الخوارزمية تحتاج إلى عينات كبيرة نسبياً للعمل بكفاءة، خصوصاً عندما تكون التداخلات المؤثرة من الدرجات العليا. علماً أن الدراسات الجينية الحقيقية غالباً ما تعتمد على عينات كبيرة نوعاً ما وذلك لضمان انعكاس الأنماط الجينية الحقيقية المنتشرة في المجتمع في عينة الدراسة.

الاستنتاجات

رغم توافق النتائج أعلاه جزئياً مع لوائح المعهد الوطني الأمريكي للشيخوخة National Institute on Aging (NIA) (2023، Alzheimer's Disease Genetics Fact Sheet)، إلا أنه من غير المنطقي الاعتماد على هذه النتائج في اتخاذ قرارات تتعلق بتطور مشكلة المرونة الادراكية. إلا أن الجانب العملي أظهر فاعلية الخوارزمية المعدلة على التعامل مع البيانات الجينية عندما يكون متغير الاستجابة ترتيبياً.

حيث يمكن تلخيص أهم الاستنتاجات:

1. أظهرت نتائج المحاكاة نجاح الخوارزمية في التحقق من معنوية التداخلات الثنائية المرشحة.
2. الخوارزمية لم توفق في معظم العينات في تشخيص التداخل الحقيقي من الدرجة الثالثة، وذلك على الرغم من تحديد عاملين جينيين بشكل صحيح في أغلب الحالات. ويعود السبب الرئيس في فشل الخوارزمية في تحديد التداخلات الحقيقية بشكل دقيق إلى قلة عدد المشاهدات المعتمدة في العينة.
3. الخوارزمية تكون مستقرة من ناحية الزمن المستهلك في التنفيذ عند التداخلات الثنائية، وهو وقت قصير جداً فيما لو تم اعتماد اسلوب الاختبارات التقليدي.
4. الخوارزمية تكون مستقرة أيضاً من ناحية الزمن المستهلك في التنفيذ عند التداخلات الثلاثية، مع زيادة طفيفة مقارنة بزمن التنفيذ في نتائج الدرجة الثانية. ويعود السبب في ذلك لكون أن النماذج الممكنة من الدرجة الثالثة أكثر منها بالنسبة للدرجة الثانية، والذي بدوره يضيف عبئاً حسابياً في عملية التحقق التقاطعي.
5. الخوارزمية تكون فعالة جداً كلما ارتفع حجم العينة، حيث تتحسن قيم الـ $\overline{CVC}$ وتقترب كثيراً من الـ 10، وهذا يثبت فاعلية الطريقة المقترحة في تحديد التداخلات الأكثر تأثيراً.

Reference

1. Al-Khaledi, Z. T. (2019). "Serial Testing for Detection of Multilocus Genetic Interactions."
2. Alzheimer's Disease Genetics Fact Sheet, accessed 18 August 2023, https://www.nia.nih.gov/health/alzheimers-disease-genetics-fact-sheet.
3. Chauhan, W., Fatma, R., Wahab, A., & Afzal, M. (2022). Cataloging the potential SNPs (single nucleotide polymorphisms) associated with quantitative traits, viz. BMI (body mass index), IQ (intelligence quotient) and BP (blood pressure): an updated review. Egyptian Journal of Medical Human Genetics, 23(1), 57.
4. Gola, D., et al. (2016). "A roadmap to multifactor dimensionality reduction methods." Briefings in bioinformatics 17(2): 293-308.
5. Hua, X., et al. (2010). "Testing multiple gene interactions by the ordered combinatorial partitioning method in case–control studies." Bioinformatics 26(15): 1871-1878.
6. Millard SP (2013). _EnvStats: An R Package for Environmental Statistics_. Springer, New York. ISBN 978-1-4614-8455-4.
7. Nelson, M., et al. (2001). "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation." Genome research 11(3): 458-470.
8. Pattin, K. A., et al. (2009). "A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction." Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 33(1): 87-94.
9. Ritchie, M. D., et al. (2001). "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer." The American Journal of Human Genetics 69(1): 138-147.
10. Sofroniou, N. and G. D. Hutcheson (1999). "The multivariate social scientist: Introductory statistics using generalized linear models." The Multivariate Social Scientist: 1-288.
11. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
12. Yao, T., Sweeney, E., Nagorski, J., Shulman, J. M., & Allen, G. I. (2020). Quantifying cognitive resilience in alzheimer's disease: the alzheimer's disease cognitive resilience score. PLoS One, 15(11), e0241707.

Theory-based Model Validation in the Generalized Multifactor Dimensionality Reduction Algorithm for Ordinal Phenotypes

Mohammed Ibraheem Othman Zaid Tariq Saleh Al-Khaledi

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

Abstract: Clinical studies indicate a close relationship between some diseases and the presence of specific interactions between genetic factors. As is the case in many studies, revealing genetic interactions that have a significant impact on the emergence of genetic diseases requires extensive statistical analyses. Because of the enormous volume of genetic data in the human race, it was necessary to develop statistical methods adapted to deal with high-dimensional data. Multifactor Dimensionality Reduction (MDR) is one of the leading nonparametric algorithms in this field. The algorithm reduces the dimensions of genetic data to obtain the most important interaction that has a direct impact on increasing the likelihood of genetic diseases appearing. In its composition, the algorithm relies on a set of nonparametric procedures to diagnose genetic interference with the highest impact exclusively on binary response variables. Like any statistical method, this algorithm is not devoid of weaknesses and application limitations, so the algorithm had to be developed to overcome the obstacles. One of the weaknesses of this algorithm is that the algorithm cannot handle data sets with ordinal response variable. Some researchers have developed a generalization of the multifactor dimensionality reduction algorithm to enable it to work with ordinal data. However, the generalized algorithm is more complex than the original algorithm. Therefore, we proposed developing the original algorithm in a simple way by employing ordinal logistic regression to classify individuals in the sample, while keeping all steps of the original algorithm unchanged. On the other hand, the MDR algorithm adopts a non-parametric method to verify the significance of the interferences nominated in the algorithm. This nonparametric procedure is based on the idea of permutational tests, and it consumes a very long time compared to parametric procedures that relies on theoretical approaches. Some researchers have suggested using the generalized extreme value distribution to verify the statistical significance of candidate interactions, but this method has only been used with continuous and binary dependent variables. In this research, the theoretical method based on the generalized extreme value distribution was employed instead of the permutational tests adopted in the algorithm when the response variable is of the ordinal type.

Keywords: dimensionality reduction algorithm, ordinal logistic regression, genetic interactions, phenotypes.