University of Kentucky
UKnowledge

2019

Serial Testing for Detection of Multilocus Genetic Interactions

Zaid T. Al-Khaledi
University of Kentucky, zaidz80@gmail.com

Digital Object Identifier: https://doi.org/10.13023/etd.2019.168

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Al-Khaledi, Zaid T., "Serial Testing for Detection of Multilocus Genetic Interactions" (2019). *Theses and Dissertations--Statistics*. 37.
https://uknowledge.uky.edu/statistics_etds/37

This Doctoral Dissertation is brought to you for free and open access by the Statistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Statistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Zaid T. Al-Khaledi, Student

Dr. Richard Charnigo, Major Professor

Dr. Constance Wood, Director of Graduate Studies

Serial Testing for Detection of Multilocus Genetic Interactions

---

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Zaid T. Al-Khaledi
Lexington, Kentucky

Director: Dr. Richard Charnigo, Professor of Statistics
Lexington, Kentucky

2019

Copyright© Zaid T. Al-Khaledi 2019

ABSTRACT OF DISSERTATION

Serial Testing for Detection of Multilocus Genetic Interactions

A method to detect relationships between disease susceptibility and multilocus genetic interactions is the Multifactor-Dimensionality Reduction (MDR) technique pioneered by Ritchie et al. (2001). Since its introduction, many extensions have been pursued to deal with non-binary outcomes and/or account for multiple interactions simultaneously. Studying the effects of multilocus genetic interactions on continuous traits (blood pressure, weight, etc.) is one case that MDR does not handle. Culverhouse et al. (2004) and Gui et al. (2013) proposed two different methods to analyze such a case. In their research, Gui et al. (2013) introduced the Quantitative Multifactor-Dimensionality Reduction (QMDR) that uses the overall average of response variable to classify individuals into risk groups. The classification mechanism may not be efficient under some circumstances, especially when the overall mean is close to some multilocus means. To address such difficulties, we propose a new algorithm, the Ordered Combinatorial Quantitative Multifactor-Dimensionality Reduction (OQMDR), that uses a series of testings, based on ascending order of multilocus means, to identify best interactions of different orders with risk patterns that minimize the prediction error. Ten-fold cross-validation is used to choose from among the resulting models. Regular permutations testings are used to assess the significance of the selected model. The assessment procedure is also modified by utilizing the Generalized Extreme-Value distribution to enhance the efficiency of the evaluation process. We presented results from a simulation study to illustrate the performance of the algorithm. The proposed algorithm is also applied to a genetic data set associated with Alzheimer's Disease.

KEYWORDS: Multifactor dimensionality reduction; Cross Validation; Model selection; Continuous Trait; Continuous Phenotype; Ordered Combinatorial Partitioning

Author's signature: Zaid T. Al-Khaledi

Date: May 3, 2019

Serial Testing for Detection of Multilocus Genetic Interactions

By
Zaid T. Al-Khaledi

Director of Dissertation: Richard Charnigo

Director of Graduate Studies: Constance Wood

Date: May 3, 2019

I would like to dedicate my dissertation to my beloved parents and family

ACKNOWLEDGMENTS

I would like to express my most profound appreciation to my advisor, Dr. Richard Charnigo, for all the endless encouragement, continuous guidance, unlimited support, and infinite patience he offered throughout my dissertation coursework. Without his invaluable insights, the completion of this dissertation would have never been possible.

I'm extremely grateful to committee member Dr. David Fardo for his tremendous efforts to helping me understand genetic concepts and obtaining the needed dataset to complete my research.

I would also like to extend my deepest gratitude to the members of my committee, Dr. Mia Zhou, Dr. William Griffith, Dr. Arnold Stromberg, Dr. Katherine Thompson, and Dr. Chi Wang. The knowledge I gained from Dr. Mia Zhou and Dr. William Griffith has a significant role in writing my dissertation. Besides him accepting to substitute Dr. Mia Zhou, Dr. Arnold Stromberg's generosity and excitement to help all students made me more eager to accomplish my Ph.D. work. Special thanks to Dr. Katherine Thompson for teaching me solid simulation and coding background, which was much needed for this dissertation. Thanks should also go to Dr. Chi Wang for accepting to serve in my committee.

I must also thank Dr. Constance Wood for the care, feelings, and unwavering support she showed during the hard time I ran into while working on my Ph.D.

I would like to extend my sincere thanks to all professors at the Department of Statistics for the quality of education they provided, which excessively improved my background in Statistics.

Finally, I cannot leave the University of Kentucky without mentioning my friends and classmates Alejandro Villasante Tezanos and Lee Xu for the encouragement, support, and the good time we spent together during my studying.

TABLE OF CONTENTS

Acknowledgments iii

List of Tables vi

List of Figures viii

Chapter 1 A Review of Multifactor-Dimensionality Reduction 1
- 1.1 Introduction 1
- 1.2 Quantitative Multifactor-Dimensionality Reduction (QMDR) 7
- 1.3 This framework 9

Chapter 2 Ordered Combinatorial Partitioning and Quantitative Phenotypes 12
- 2.1 Introduction 12
- 2.2 Adaptation of OCP to Handle Continuous Phenotypes 13
- 2.3 Simulation Study 18
  - 2.3.1 Case 1: True Model = AB 20
  - 2.3.2 Case 2: True Model = ABD 26
  - 2.3.3 Case 3: True Model = BD 29
  - 2.3.4 Case 4: True Model = ABC 34
  - 2.3.5 Case 5: True Models = AB and AD 37
  - 2.3.6 Case 6: True Models = AB and CD 42

Chapter 3 Modification of The OQMDR Algorithm 47
- 3.1 Preliminary 47
- 3.2 The Generalized Extreme Value Distribution 48
- 3.3 Parameter Estimation 50
- 3.4 Utilizing The GEVD in OQMDR Algorithm 57
- 3.5 Numerical and Graphical Assessments 60
  - 3.5.1 Case 1: True model = AB 63
  - 3.5.2 Case 2: True model = ABD 81
  - 3.5.3 Case 3: True model = BD 94
  - 3.5.4 Case 4: True model = ABC 107
- 3.6 Summary 119

Chapter 4 Theoretical Findings 122
- 4.1 Derivation of MLE's required formulas 122
- 4.2 Validating the Law of Total Probability on Hau et al. paper [30] . . . 131
- 4.3 Theorem: Ordered Combinatorial Partitioning in OQMDR 134

Chapter 5 Real data analysis 139
- 5.1 Alzheimer's Disease (AD) overview 139

5.2 Data presentation . . . 140
5.3 Data analysis . . . 145
5.4 Conclusion . . . 154
5.5 Further work . . . 155

Appendix . . . 157
Presentation of other fitted distributions . . . 157

References . . . 175

Vita . . . 180

LIST OF TABLES

1.1 A 2×2^2 table represents the case:control data set with two interacting factors 6
1.2 The three 2×2 tables formed from the original 2×2^2 table 7

2.1 Data presentation of the interaction between A and B 15
2.2 Case 1: True model= AB, and $n = 500$ 22
2.3 Case 1: True model= AB, and $n = 1000$ 25
2.4 Case 1: True model= AB, and $n = 2000$ 25
2.5 Case 2: True model= ABD, and $n = 500$ 27
2.6 Case 2: True model= ABD, and $n = 1000$ 28
2.7 Case 2: True model= ABD, and $n = 2000$ 29
2.8 Case 3: True model= BD, and $n = 500$ 31
2.9 Case 3: True model= BD, and $n = 1000$ 33
2.10 Case 3: True model= BD, and $n = 2000$ 33
2.11 Case 4: True model= ABC, and $n = 500$ 35
2.12 Case 4: True model= ABC, and $n = 1000$ 36
2.13 Case 4: True model= ABC, and $n = 2000$ 37
2.14 Case 5: True models= AB and AD, and $n = 500$ 40
2.15 Case 5: True model= AB and AD, and $n = 1000$ 41
2.16 Case 5: True model= AB and AD, and $n = 2000$ 41
2.17 Case 6: True model= AB and CD, and $n = 500$ 44
2.18 Case 6: True model= AB and CD, and $n = 1000$ 45
2.19 Case 6: True model= AB and CD, and $n = 2000$ 45

3.1 Case 1: True model = AB, $n = 500$, and $m = m_1 = 30$ 64
3.2 Case 1: True model= AB, and $n = 1000$, and $m = m_1 = 30$ 71
3.3 Case 1: True model= AB, and $n = 2000$, and $m = m_1 = 30$ 77
3.4 Case 2: True model= ABD, and $n = 500$, and $m = m_1 = 30$ 82
3.5 Case 2: True model= ABD, and $n = 1000$, and $m = m_1 = 30$ 86
3.6 Case 2: True model= ABD, and $n = 2000$, and $m = m_1 = 30$ 90
3.7 Case 3: True model= BD, $n = 500$, and $m = m_1 = 30$ 95
3.8 Case 3: True model= BD, and $n = 1000$, and $m = m_1 = 30$ 99
3.9 Case 3: True model= BD, and $n = 2000$, and $m = m_1 = 30$ 103
3.10 Case 4: True model= ABC, and $n = 500$, and $m = m_1 = 30$ 108
3.11 Case 4: True model= ABC, and $n = 1000$, and $m = m_1 = 30$ 111
3.12 Case 4: True model= ABC, and $n = 2000$, and $m = m_1 = 30$ 115

4.1 The original penetrance of the two factors suggested by Hua et al. [30] . 131
4.2 The joint probabilities of disease and multilocus combinations of the two factors per the definition of the authors [30] 132
4.3 The suggested penetrance of the two factors 132

4.4 The joint probabilities of disease and multilocus combinations of the two factors per the definition of the authors [30] and our suggested penetrance 132

5.1 Genetic-variable list 142
5.2 Statistical summary 143
5.3 Model selection of the three cognitive scores 146
5.4 Proposed model evaluation 146
5.5 Principal Components analysis 148

LIST OF FIGURES

1.1 Interaction representation between two SNPs, and its two anticipated risk patterns 9

2.1 Representation of 4-factor interaction, each with three levels 14
2.2 Case 1: Risk pattern for the proposed 2-way models 23
2.3 Case 1: Risk pattern for the proposed 3-way models 24
2.4 Case 2: Risk patterns for the proposed 2-way models 27
2.5 Case 2: Risk pattern for the proposed 3-way models 27
2.6 Case 3: Risk patterns for the proposed 2-way models 31
2.7 Case 3: Risk patterns for the proposed 3-way models 32
2.8 Case 4: Risk patterns for the proposed 2-way models 35
2.9 Case 4: Risk patterns for the proposed 3-way models 36
2.10 Case 5: Risk pattern for the proposed 2-way models 40
2.11 Case 5: Risk patterns for the proposed 3-way models 40
2.12 Case 6: Risk pattern for the proposed 2-way models 44
2.13 Case 6: Risk pattern for the proposed 3-way models 44

3.1 Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 66
3.2 Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 67
3.3 Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 69
3.4 Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . . . 70
3.5 Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 72
3.6 Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 73
3.7 Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 74
3.8 Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 400 permuted p-values . . . 75
3.9 Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 78
3.10 Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 79
3.11 Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 80
3.12 Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . . . 81

3.13 Case 2: True model = ABD, n = 500; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 83
3.14 Case 2: True model = ABD, n = 500; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 84
3.15 Case 2: True model = ABD, n = 500; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values . . 85
3.16 Case 2: True model = ABD, n = 500; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . 86
3.17 Case 2: True model = ABD, n = 1000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 87
3.18 Case 2: True model = ABD, n = 1000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 88
3.19 Case 2: True model = ABD, n = 1000; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values . . 89
3.20 Case 2: True model = ABD, n = 1000; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . 90
3.21 Case 2: True model = ABD, n = 2000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 91
3.22 Case 2: True model = ABD, n = 2000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 92
3.23 Case 2: True model = ABD, n = 2000; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values . . 93
3.24 Case 2: True model = ABD, n = 2000; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 400 permuted p-values . 94
3.25 Case 3: True model = BD, n = 500; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 96
3.26 Case 3: True model = BD, n = 500; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 97
3.27 Case 3: True model = BD, n = 500; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 98
3.28 Case 3: True model = BD, n = 500; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 200 permuted p-values . . . 99
3.29 Case 3: True model = BD, n = 1000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 100
3.30 Case 3: True model = BD, n = 1000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 101
3.31 Case 3: True model = BD, n = 1000; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 102
3.32 Case 3: True model = BD, n = 1000; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . . . 103
3.33 Case 3: True model = BD, n = 2000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 104
3.34 Case 3: True model = BD, n = 2000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 105

3.35 Case 3: True model = BD, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 106
3.36 Case 3: True model = BD, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . . . 107
3.37 Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 108
3.38 Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 109
3.39 Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values . . 110
3.40 Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 200 permuted p-values . 111
3.41 Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 112
3.42 Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 113
3.43 Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values . . 114
3.44 Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 400 permuted p-values . 115
3.45 Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 116
3.46 Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores 117
3.47 Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values . . 118
3.48 Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values . 119

5.1 Empirical distribution of Cognitive Resilience compared to Normal distribution 143
5.2 Empirical distribution of Cognitive Reserve compared to Normal distribution 144
5.3 Empirical distribution of Global Resilience compared to Normal distribution 145
5.4 Risk patterns of the proposed 2-way interactions for each response 147
5.5 Risk patterns of the proposed 3-way interactions for each response 147
5.6 Cognitive Resilience; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 149
5.7 Cognitive Resilience; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 150
5.8 Cognitive Reserve; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 29 permuted t-scores 151
5.9 Cognitive Reserve; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 29 permuted p-values 152
5.10 Global Resilience; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores 153

5.11 Global Resilience; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values 154

Chapter 1 A Review of Multifactor-Dimensionality Reduction

1.1 Introduction

Disease susceptibility is considered to be substantially linked to multilocus genetics on the level of main effects and/or interaction effects [48]. Many parametric statistical methods have been used to model the relationship between disease susceptibility and genetic factors. The majority of these methods were derived from the concept of linear and generalized linear modeling [24]. Yet, due to the high dimensionality of genetic data and/or the relatively small sample size, these methods may not be efficient to work with under such circumstances. To see this, recall that the ordinary least squares (OLS) estimator of the vector of the linear regression coefficient (β) can be obtained according to equation 1.1:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{1.1}$$

where Y is an n-vector of the response variable, X is an $n \times p$ matrix of predictor variables, $\hat{\beta}$ is a p-vector of the OLS estimators of the regression coefficients, n is the number of observations, and p is the number of the regression coefficients in the fitted model.

When we run in a large p small n situation, i.e., the number of regression coefficients is larger than the number of observations $p > n$, then the rank of the matrix $X^T X$ is at most n. Which means there is a multicollinearity problem in the data. In such situation, the regular inverse for the matrix $X^T X$ does not exist. This implies the OLS method is no longer applicable. When a generalized inverse is used to calculate $\hat{\beta}$ according to equation 1.1, a unique estimator would not exist. Even if an approximated matrix inverse is used, the interpretation of the regression coefficients of the correlated predictors won't be accurate [31]. In genetic studies, p could get substantially large when an interaction of any order between genetic factors is considered in the analysis. For example, a second degree polynomial of a data set with ten genetic factors may contains $p = 56$ coefficients, which is the number of intercept, all main effects, and all 2-way effects coefficients in the model. This number gets larger exponentially when the number of factors increases or a higher degree interaction is considered in the study. Consequently, non-parametric alternatives have been developed to overcome the difficulties of using parametric methods.

Multifactor-Dimensionality Reduction (MDR) algorithm, originally introduced by Ritchie et al.[51], is one of the non-parametric methods that has been widely used and extended to describe the relationship between disease susceptibility and multilocus genetics interaction for case-control and discordant-sib-pair studies. The combinatorial partitioning method, described by Nelson et al. [46], motivated Ritchie and her colleagues to develop the MDR method. The main goal of the MDR method is to capture the single most significant multilocus genetic interaction by reducing the dimensionality of the genetics data to one single predictor via labeling each possible multilocus combination at high risk or low risk according to a certain criterion. Cross-validation is used to assess the validity of the proposed k-way interaction for $k = 2, 3, ..., N - 1$, where N is the number of factors in the data. Further, the significance of a final proposed interaction is verified using permutation testing.

The MDR method can be summarized in the following steps [51]:

1. First, identify N genes and/or the discrete environmental factors in the data.

2. Next, the frequency distribution of the data is displayed in a k-dimensional space for each considered k-way interaction. That is, the data of any 2-way interaction are visualized using a 2-way contingency table. Similarly, a 3-way contingency cube (or three 2-way contingency tables) is used to visualize the

data of any 3-way interaction, and so forth. The dimensions of these cross tabulations are determined by the number of levels in each factor. For instance, the frequency distribution for the interaction between two factors each with three levels is represented using a 3×3 contingency table. Each cell in the representation contains the frequencies of the cases and controls that correspond to a specific multilocus combination. A comparison between the case:control ratio in each cell and a previously specified threshold is used to determine whether the corresponding combination is considered high risk or low risk. The individuals in each cell (combination) are considered at high risk if the case:control ratio exceeds or equals to the specified threshold. Conversely, the individuals are labeled as low risk if the case:control ratio is inferior to the threshold. In their research, Ritchie and her colleagues suggested the threshold to be 1.0. The goal of the classification process is to reduce the dimensionality of the data space to a one-dimensional binary predictor variable.

3. Then, a proposed model (interaction) of order k is chosen as the one that has the smallest classification error (CE) for each possible k-way model. To obtain the CE for each model, the total number of misclassified individuals (patients labeled as low risk, and controls labeled as high risk) is recorded for each model. The misclassification of the patients is usually called false negative error (FN), whereas the number of incorrectly allocated controls is termed as false positive error (FP).

4. After that, in order to assess the validity of the proposed model, a 10-fold cross-validation (CV) is used for each k-way interaction. To perform the CV procedure, the data is randomly divided into ten approximately equally sized groups, such that each group has the exact same number of cases and controls to retain the case:control ratio equal to 1. In each fold, one group is excluded as a

testing data set, while the remaining 9 groups are deemed as a training data set. Later, the data classification and model selection procedures described in steps 2 and 3 are performed on the training data set. Next, individuals belong to the testing data set (the excluded group) are classified into high risk and low risk according to the binary predictor obtained from performing steps 2 and 3 on the training data set. The exclusion procedure is performed on each of the ten groups and the CEs are reported for each possible k-way model constructed using the training data sets. In a similar way to calculate CE, the prediction error (PE) is calculated for each excluded group in the ten folds. In particular, PE is the number of falsely classified individuals in the testing data set. To eliminate the possible effects of the random subsetting of the data, the entire CV procedure is repeated several times (e.g. five times). New random sub-grouping of the data into ten equally sized groups is carried out in each repetition. From all acquired CEs, the average CE ($\overline{\text{CE}}$) is calculated for each k-way model. Then, the models that minimize the $\overline{\text{CE}}$ for each degree of interaction are reported. Finally, the model that better represents the relationship between multilocus genetic interaction and disease susceptibility among all selected models is the one with the minimum average PE ($\overline{\text{PE}}$), where $\overline{\text{PE}}$ is calculated in a similar way to $\overline{\text{CE}}$ [24]. Cross-validation consistency (CVC) is used to evaluate the validity of the selected model. That is, MDR calculates how many times each specific model is selected from all ten folds. A final average cross-validation consistency ($\overline{\text{CVC}}$) is calculated for each proposed model based on the outcomes of all repetitions. The CVC is used to evaluate the validity of the final model because a true underlying effect should be recognizable regardless of the randomized subsetting of the data.

5. Finally, to verify the significance of the selected model, permutation testing is used with 1000 permuted data sets. Each time the labels of cases and controls

> are randomly shuffled while the remaining variables are kept untouched. To examine the statistical significance of the winner model, the $\overline{\text{CVC}}$ derived from the original data set is compared to the empirical distribution of the $\overline{\text{CVC}}$ generated by 1000 permutation testings. The proposed model is considered statistically significant if the permuted p-value is ≤ 0.05.

MDR method has been widely studied and extended to improve the overall algorithm and/or to address some of its drawbacks. As described by Gola et al. [24], these extensions generally focused on handling different phenotypic data [26, 36], different data structure [5, 25], risk labels allocation [39, 30], classification result evaluation [41, 7, 44], and p-value calculation procedures [43, 48, 17].

One common shortcoming of the MDR is that it only applies for evenly distributed samples, i.e., the controls and cases are equally observed in the data set. Velez et al. [53] proposed a few simple solutions to overcome the imbalanced data issue. The proposed remedies mainly depend on over-sampling, under-sampling, or using the cases:control ratio for the whole sample as a threshold. Another considerable weakness of the MDR algorithm is utilizing a constant threshold to classify individuals into high-risk and low-risk groups. Regardless of the benefits of using a fixed threshold, as it cuts down the computational burden, it may lead to a huge power loss [30].

Hua et al. [30] modified the MDR algorithm by using a threshold that maximizes the χ^2 test statistic among all possible ordered 2×2 contingency tables that are formed from a single $2 \times \Pi_1^k l_i$ table, where k represents the number of factors which interact, and l_i is the number of levels for the i^{th} factor. In each possible k-way interaction, there are $\Pi_{i=1}^k l_i - 1$ contingency tables of 2×2 dimensions, each table produces a single χ^2 test statistic. These $\Pi_{i=1}^k l_i - 1$ contingency tables represent different patterns of classifications of the data into risk groups. The partitioning and ordering procedures are mainly based on the idea of Ordered Combinatorial Partitioning (OCP) method [46]. Even though the OCP method considers only $\Pi_{i=1}^k l_i - 1$ partitions, it provides the same benefits of scanning all possible partitionings of the data [30].

In practice, consider the following illustration inspired by an example from Hua et al. [30]. Assume we have a case:control data set with two interacting factors, A and B, such that each factor has two levels. Let a_1 and a_2 be the levels for factor A, and b_1 and b_2 be the levels for factor B. We can represent the data of this interaction by a 2×2^2 table as shown in table 1.1 below.

Table 1.1: A 2×2^2 table represents the case:control data set with two interacting factors

	a_1b_1	a_2b_1	a_1b_2	a_2b_2	Total
Case	1	12	19	28	60
Control	11	13	20	16	60
Total	12	25	39	44	120

In this example, there are $2^2 - 1$ different 2×2 tables that can be formed from the original table, where three is the number of columns in the original table minus one. Before we construct the new tables, we need to reorder the columns of the original table in ascending order according to the case:control ratio in each column. Since the columns of the table in our example are already sorted, we can proceed to the next step. To form the first 2×2 table, we keep the first column as it is, while we merge the last three columns into one column. Then, combine the first two columns and the last two columns into two separate columns to create the second 2×2 table. Finally, the third table is formed by collapsing the data of the first three columns into one column and leaving the last column alone. Table 1.2 shows the three 2×2 tables formed from table 1.1.

Now, in each one of the three tables, we label the first column as low risk and the second column as high risk. This suggests that there are three different thresholds floating around, one threshold for each table. The threshold for table 1.2a falls between the case:control ratios of the first two columns of the original table, i.e., between 1/11 and 12/13. While for table 1.2b, its threshold is in between 12/13 and

Table 1.2: The three 2×2 tables formed from the original 2×2^2 table

(a) 1 vs. 2, 3, 4

	a_1b_1	a_2b_1, a_1b_2, a_2b_2	Total
Case	1	59	60
Control	11	49	60
Total	12	108	120

(b) 1, 2 vs. 3, 4

	a_1b_1, a_2b_1	a_1b_2, a_2b_2	Total
Case	13	47	60
Control	24	36	60
Total	37	83	120

(c) 1, 2, 3 vs. 4

	a_1b_1, a_2b_1, a_1b_2	a_2b_2	Total
Case	32	28	60
Control	44	16	60
Total	76	44	120

19/20, which are the ratios of the second and third columns in the original table. Finally, the third threshold, which is for table 1.2c, is larger than 19/20 and smaller than 28/16. The permuted p-values of the χ^2 tests of these three tables are 0.0049, 0.0447, and 0.039 respectively. All p-values are calculated from 10000 permutation testings using `R` software [50]. Obviously, the first table, and thus the first range of thresholds, maximizes the χ^2 test statistic among all three 2×2 tables. Accordingly, choosing any value between 1/11 and 12/13 as a cutoff point leads to maximizing the test statistic and therefore a more powerful test [30]. If the fixed threshold suggested by Ritchie et al., which is 1.0, were chosen to classify this data set, then the data would be classified in accordance to table 1.2c with a permuted p-value of 0.039. Thus, sticking with a constant threshold might lead one to propose a weaker model to capture the genetic predisposition.

1.2 Quantitative Multifactor-Dimensionality Reduction (QMDR)

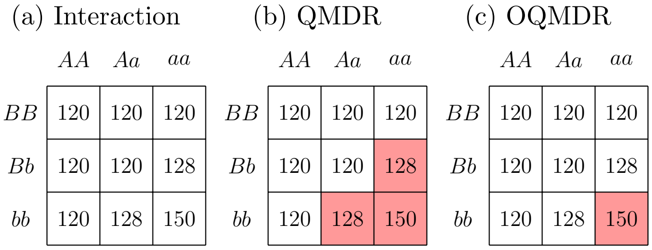
Another essential extension for the MDR method is to make it adequate for analyzing data sets with continuous phenotypes such as plasma triglyceride levels [46], blood pressure [14], and Body Mass Index [18]. In fact, the original MDR method can be utilized to analyze data sets with continuous phenotypes, but only after converting the continuous trait variable to a binary response variable according to a certain criterion or researcher prior experience. However, analyzing the data set with the original quantitative response would probably be more precise and informative. Generalized Multifactor-Dimensionality Reduction (GMDR) [39], Model-Based Multifactor-Dimensionality Reduction [8], and Quantitative Multifactor-Dimensionality Reduction (QMDR) [26] are some expansion algorithms of the original MDR approach.

The QMDR developed by Gui et al. [26] modified the original MDR by using the overall mean as the criterion of classifying the genotype combinations into high-risk and low-risk groups for each k-way model. In particular, each multilocus genotype combination in every possible k-way interaction is labeled high risk if its mean is higher than the overall mean of the response. Otherwise, the genotype combination is regarded as low risk. Similar to original MDR, all individuals will be placed in a high-risk group or a low-risk group according to the preceding classification to form a dichotomous predictor variable. A single Two-Sample t-Test for Equal Means is employed to compare the high-risk group vs. the low-risk group in each possible k-way model. The k-way interaction that maximizes the t-test statistic is selected as a proposed model for that specific order of interaction. To choose the model that better explains the variation in the continuous response among all suggested k-way models, 10-fold cross-validation with repetitions is performed to compute the cross-validation consistencies, t-scores, and Mean Squared Prediction Errors on testing data for each model. The model that maximizes the testing t-score is chosen as the best final model. The significance of the winner model is justified using permutation testings. Under the null hypothesis (i.e., no factors effects involved), the mean of the t-scores will approach zero. Thus, in their paper, Gui et al. [26] anticipated the empirical distribution of testing t-scores to be approximately normal and centered at zero. Hence, a normal distribution with a mean of zero was employed to estimate the empirical p-value of the final model as a replacement of the permuted p-value.

1.3 This framework

Despite its computational efficiency concerning fast evaluation, QMDR algorithm may lead one to select a weaker model to explain the variation in the response variable. For instance, let's consider the genetic data with two biallelic single-nucleotide polymorphisms (SNPs), with A and B being the major alleles for each SNP, and a and b are the minor alleles (see figure 1.1a). The numbers in figure 1.1a represent the means for multilocus interactions between the two SNPs, in the absence of statistical noise. According to the QMDR algorithm, every cell with a mean greater than the overall mean, which is 125.11 in this example, will be regarded as high risk. This suggests that all individuals with a mean of 128 are assumed at high risk of manifesting the disease as shown in figure 1.1b. However, we may think that a mean of 128 is not sufficiently large to classify the corresponding individuals at high risk; whereas, only cells with means of 150 would probably be considered at high risk (see figure 1.1c). Hence, we proposed a new algorithm to handle such cases. We named our algorithm as the Ordered Combinatorial Quantitative Multifactor-Dimensionality Reduction (OQMDR)

Figure 1.1: Interaction representation between two SNPs, and its two anticipated risk patterns



(a) Interaction

	AA	Aa	aa
BB	120	120	120
Bb	120	120	128
bb	120	128	150

(b) QMDR

	AA	Aa	aa
BB	120	120	120
Bb	120	120	128
bb	120	128	150

(c) OQMDR

	AA	Aa	aa
BB	120	120	120
Bb	120	120	128
bb	120	128	150

* Individuals in highlighted cells are at high risk.

In chapter 2, we will extend the idea of Ordered Combinatorial Partitioning method introduced by [30] to data sets with quantitative traits to perform a series of t-tests to capture the genetic predisposition. For each possible k-way model, there will be $\Pi_{i=1}^{k} l_i - 1$ different t-tests, where k is the degree of the interaction, and l_i is the number of levels of the i^{th} factor for $i = 1, 2, ..., k$. Each t-test corresponds to a specific pattern to classify the data into high-risk and low-risk groups. From each possible k-way model, we propose the pattern that corresponds to the largest t-statistic among all $\Pi_{i=1}^{k} l_i - 1$ computed t-statistics. From the pool of all maximum t-statistics derived for all possible k-way models, a single maximum of the maximums t-statistic will be selected, and the corresponding model along with its risk pattern is considered our proposed model for that specific degree of interaction.

A 10-fold cross-validation procedure with five repetitions is carried out to calculate the average cross-validation consistency ($\overline{\text{CVC}}$), average testing t-score, and average Mean Squared Prediction Errors ($\overline{\text{MSPE}}$) for the proposed k-way models. A final single model that maximizes the average testing t-score is selected as a winner model. Average cross-validation consistency is used as a tiebreaker in case if the proposed models of various orders end up with the same average testing t-score. A most parsimonious model is selected when all criteria are tied between the selected models of different degrees. Permutation testings with 1000 permuted data sets are used to justify the significance of the final model. The p-value is calculated by comparing the average permuted testing t-scores to the average testing t-score of the proposed model. A comparison between the output from our method and QMDR is performed for six different cases. Each case is repeated ten times at three different sample sizes with a different simulated data set. The method that captures the actual model, where applicable, and has smaller $\overline{\text{MSPE}}$ is considered better in each case.

In chapter 3, we modified the OQMDR algorithm to overcome the time consumption issue and to increase the accuracy of model evaluation. The adjustment involves utilizing the Generalized Extreme-Value Distribution (GEVD) to justify model significance, which was used by Pattin et al. [48] and Hua et al. [30] for the same purpose. The approach is initially suggested to reduce computation burdens of using regular permutation testings. We adapted the GEVD approach to fit with OQMDR to assess both the test statistic and its p-value for further justification. The GEVD is considered because the final model is selected upon maximizing the test statistic. Permuting a small set of test statistics could be used to approximate the null distribution of the test statistic. Analogous to the test statistic, the significance of the p-value of each suggested model is justified using the same approach. The idea follows from the fact that a p-value is a smooth decreasing transformation of a test statistic. Therefore, we tested multiple different distributions as well as three different transformations of the p-value to find the best fit according to the graphical representation. Among all considered distributions and/or transformations, the GEVD of $-\log(-\log(p\text{-value}))$ is chosen to verify the validity of the p-value. The uniform(0,1) distribution of the identity transformation shows a huge enhancement when a large number of permuted samples is considered. However, the behavior of the GEVD of the $-\log(-\log(p\text{-value}))$ is globally better than all other choices. A simulation assessment is carried out on 120 different data sets to evaluate the new procedure. The output is compared to the findings from chapter 2 regarding efficiency and significance of proposed interactions.

In chapter 4, we presented some simple theoretical findings. Finally, in chapter 5, we applied the OQMDR algorithm to Alzheimer Disease data set with three continuous responses.

Copyright© Zaid T. Al-Khaledi, 2019.

Chapter 2 Ordered Combinatorial Partitioning and Quantitative Phenotypes

2.1 Introduction

Quantitative Multifactor-Dimensionality Reduction (QMDR) is a modified version of the MDR algorithm. The QMDR is suggested to handle genetic data sets with continuous phenotypes [26]. The method uses the overall mean of the continuous trait as a threshold to classify individuals into high-risk and low-risk groups in each multilocus combination. A single Two-Sample t-Test for Equal Means is used to evaluate the difference between the means of the two groups for each possible interaction of a specific order. The best model of order $k = 2, 3, ..., N - 1$, where N is the total number of factors in the data, with a particular risk pattern is the one that maximizes the t-test statistic among all calculated t-statistics of all possible interactions. Under certain conditions, utilizing the overall mean of the continuous variable might lead to choosing a weaker model to explain the genetic predisposition. Even when the QMDR method picks the most important interaction, it might miss the risk pattern that better represents the relationship between the phenotype and the genetic factors.

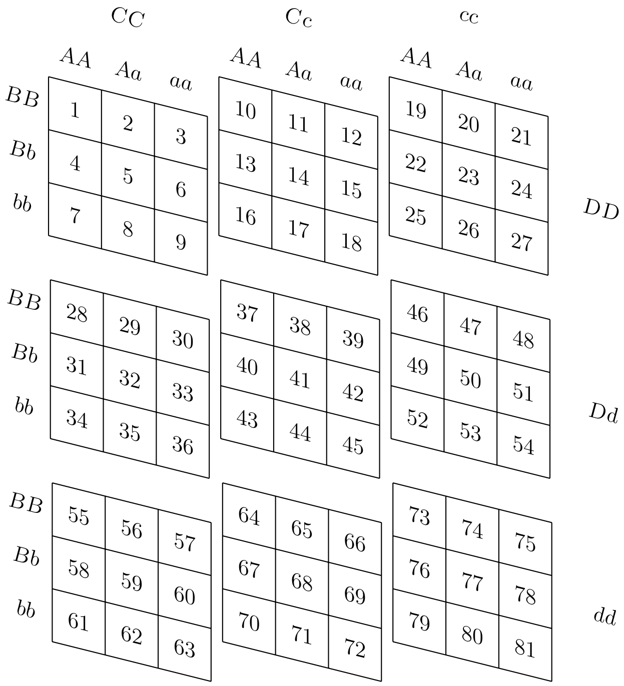
To overcome the weaknesses in such cases, we developed a new algorithm, the Ordered Combinatorial Quantitative Multifactor-Dimensionality Reduction (OQMDR), that considers all logical risk patterns for each interaction based on the idea of Combinatorial Partitioning (CP) [46]. To reduce the computational burden, we adapted the Ordered Combinatorial Partitioning (OCP) strategy introduced by Hua et al. [30] to work with continuous variables. The use of the OCP is anticipated to give the same maximum t statistic obtained when the exhaustive testing over the set of all possible Combinatorial Partitionings is performed.

2.2 Adaptation of OCP to Handle Continuous Phenotypes

The new algorithm can be described as follows. First, determine the total number of factors, N, in the data set. Then, reorganize the data into an N-dimensional array, such that each element in the array contains data that belong to a certain combination between N factors levels. Figure 2.1 shows the representation of the data when the total number of the selected factors is four. Next, all possible k-way interactions are considered for $k = 2, 3, ..., N - 1$, and a k-dimensional array is used to represent the data. Then, the means of all possible multilocus combinations (cell means) are calculated. There are $\Pi_{i=1}^{k} l_i$ possible multilocus combinations for each interaction of order k, where l_i is the number of levels of the i^{th} factor, and k is the number of interacting factors. Then, we use the OCP procedure to capture the single most important k-way interaction that better explains the variation in the continuous phenotype. That is, the multilocus combinations are sorted in an ascending order according to their means. then a set of size $\Pi_{i=1}^{k} l_i - 1$ tables are formed by collapsing the sorted cells into two groups (high-risk and low-risk groups). After that, a series of $\Pi_{i=1}^{k} l_i - 1$ t-testings are performed between the high-risk group versus the low-risk group from each partitioning. Each calculated t-statistic corresponds to a certain risk pattern of the multilocus interaction. The OCP procedure is applied to each possible interaction of order $k = 2, 3, ..., N - 1$. This produces $N - 2$ sets of size $\binom{N}{k}$ models along with their risk patterns, in which each model is maximizing the t-statistic. Afterward, a single model is selected from each of the $N - 2$ sets, such that the selected model is maximizing the maximized t-statistics.

For better illustration of risk pattern selection, consider the following example of three interacting factors (i.e., $k = 3$) with quantitative phenotype. Let X be the continuous variable of interest, and assume there are three interacting factors A, B, and C, where each factor has three levels, i.e., $l_i = 3, i = 1, 2, 3$. Let $(a_1, a_2, a_3), (b_1, b_2, b_3)$, and (c_1, c_2, c_3) be the levels of factor A, B, and C respectively.

Figure 2.1: Representation of 4-factor interaction, each with three levels



In this example, we have three possible 2-way interactions, which are (AB, AC, and BC). Thus, if we consider the 2-way interaction between A and B, then we would have $\Pi_{i=1}^{k} l_i = 9$ different multilocus combinations between these two factors. Therefore, the data can be represented in a 3×3 table (table 2.1). Next, calculate the mean for each possible multilocus combinations, and let $\bar{X}_j$ and $\bar{X}_{(j)}$ be the mean and the ordered mean of the j^{th} combination for $j = 1, 2, ..., 9$. After that, we reorder the cells of the 3×3 table based on their means in an ascending order. Now, the data in each cell will be treated as a single subset from the original data set, which means we divide the data into nine different groups $(g_{(1)}, g_{(2)}, ..., g_{(9)})$ in this example, where $g_{(j)}$ is the group of individuals that belong to the j^{th} ordered cell. Next, we aggregate the groups to perform a series of eight ($\Pi_{i=1}^{k} l_i - 1 = 8$) different t-tests. Each aggregation gives one distinct risk pattern. The first t-test will be between the data from $g_{(1)}$ as a first sample and the data from ($g_{(2)}, g_{(3)}$, and $g_{(9)}$) combined together as a second sample. Whereas, the second t-test will be between the data from

$g_{(1)}$, and $g_{(2)}$ grouped together as a first sample and the data from ($g_{(3)}, g_{(4)}$, and $g_{(9)}$) combined together as a second sample, and so forth. Finally, the first eight groups are treated as one sample and tested against the data from the ninth group. Among the eight risk patterns we have in this example, the one that maximizes the t-test statistic will be chosen as the proposed risk pattern for the interaction between the factors A and B. The procedure is repeated for each possible 2-way interaction, and a single risk pattern is selected. The interaction that maximizes the maximized t-statistics is selected as our proposed 2-way model in this example.

Table 2.1: Data presentation of the interaction between A and B

	a_1	a_2	a_3
b_1	Data with a_1, b_1 $\bar{X}_{11}$	Data with a_2, b_1 $\bar{X}_{21}$	Data with a_3, b_1 $\bar{X}_{31}$
b_2	Data with a_1, b_2 $\bar{X}_{12}$	Data with a_2, b_2 $\bar{X}_{22}$	Data with a_3, b_2 $\bar{X}_{32}$
b_3	Data with a_1, b_3 $\bar{X}_{13}$	Data with a_2, b_3 $\bar{X}_{23}$	Data with a_3, b_3 $\bar{X}_{33}$

A 10- fold cross-validation procedure with five repetitions is performed to justify the validity of the selected model. That is, we divide the data into ten approximately equal sized subsets, then we exclude one subset as a testing data set while treat the remaining nine subsets as a training data set. The cross-validation consistency (CVC) is calculated for the selected model from each order k for each repetition, and the average CVC over repetitions is reported. The CVC is calculated as the number of times out of ten the proposed model with its risk pattern from the original data is completely reproduced from the cross-validation. Testing t-score, as well as Mean Squared Prediction Errors, are calculated for each repetition. The testing t-scores are the calculated t-test statistics from the testing data set that are classified in accordance with the risk patterns obtained from performing the cross-validation procedure on the training data set. Similarly, the Mean Squared Prediction Errors are calculated according to equation 2.1. A final $\overline{\text{CVC}}$, average testing t-score, and $\overline{\text{MSPE}}$ from all five repetitions are reported for each of the $N-2$ selected models. Where $\overline{\text{CVC}}$, and $\overline{\text{MSPE}}$ are the average of the five CVC's, and MSPE's obtained from each repetition, respectively.

$$\text{MSPE} \;=\; n^{-1}\sum_{j=1}^{10}\left[\sum_{i=1}^{n_{j.low}}\left(Y_{ij.low}-\hat{Y}_{j.low}\right)^2+\sum_{i=1}^{n_{j.high}}\left(Y_{ij.high}-\hat{Y}_{j.high}\right)^2\right] \quad (2.1)$$

Above, n is the number of individuals in the whole data set, $n_{j.low}$ and $n_{j.high}$ are the numbers of individuals classified as low risk and high risk respectively when they are treated as a testing group in the j^{th} fold, $Y_{ij.low}$ and $Y_{ij.high}$ are the observed values of the continuous trait variable correspond to individuals classified at low and high risk respectively in the j^{th} fold, and $\hat{Y}_{j.low}$ and $\hat{Y}_{j.high}$ are the means response of individuals classified at low and high risk respectively when they are treated as a training group in the j^{th} fold.

According to the results from cross-validation, the model with the corresponding risk pattern that maximizes the average testing t-score is chosen as the best model among all $N-2$ proposed models of any degree. The $\overline{\text{CVC}}$ is used as a tiebreaker when the average testing t-scores are tied between the selected models. Finally, a most parsimonious model is chosen when both the average testing t-scores and the $\overline{\text{CVC}}$s are tied for two models.

The significance of the winner model is justified by permuting the original data 1000 times, and the average permuted CVC, the average permuted testing t-score, and average permuted MSPE are calculated from each permuted data set. The p-value, which is denoted by p_t, is calculated according to equation 2.2, and it represents how many times the calculated average testing t-score is smaller than the average permuted testing t-score divided by 1000. The examined model is considered significant if its

p-value is less than 0.05.

$$p_t \quad = \tfrac{1}{1000} \quad \sum_{perm=1}^{1000} \mathbb{I}_{\{\bar{t^*} < \bar{t^*}_{perm}\}} \tag{2.2}$$

where $\bar{t^*}$ is the average testing t-score calculated from the original data set, $\bar{t^*}_{perm}$ is the average permuted testing t-score, and p_t is the empirical p-value.

To summarize the model selection process in the OQMDR, let $t_{1,j,k}, t_{2,j,k}, \ldots, t_{r,j,k}$ be the t statistics of the ordered risk patterns of the j^{th} k-way interaction, where $r = \Pi_i^k l_i - 1$ is the total number of the considered risk patterns of the j^{th} k-way interaction, where l_i is the number of levels of the i^{th} factor. Then, for each possible interaction of any order k, we choose the risk pattern that maximizes the test statistic:

$$t_{max,j,k} \quad := \quad \max(t_{1,j,k}, t_{2,j,k}, \ldots, t_{r,j,k})$$

where $t_{max,j,k}$ is the largest test statistic produced from all examined risk patterns of the j^{th} k-way interaction.

Later, we choose the best k-way interaction by optimizing over all maximized test statistics:

$$t_{max,max,k} \quad := \quad \max(t_{max,1,k}, t_{max,2,k}, \ldots, t_{max,m,k})$$

where $t_{max,max,k}$ is the largest t-statistic produced from all possible k-way interactions, and $m = \binom{N}{k}$ is the number of all possible k-way interactions.

Finally, we choose the final best model by optimizing over the testing t-scores of the proposed models of order $k = 2, 3, ..., N-1$. That is, if t^*_k is the testing t-score of the proposed k-way interaction, then the final best interaction with its selected risk pattern is the one with a testing score $t^*_{k_{max}}$, such that:

$$t^*_{k_{max}} \quad := \quad \max(t^*_2, t^*_3, \ldots, t^*_{N-1})$$

Comparing to QMDR method, that method will only consider one risk pattern for each possible interaction of any degree, which is based on the overall mean of the continuous trait variable. The QMDR requires a single t-test for each interaction, and it selects the interaction with the largest t statistic among other examined interactions. The two algorithms will likely end up proposing the same model and risk pattern when the cause of the variation in the response is tremendously distinguishable. Yet, in many cases, some combinations have means that are very close to the overall mean, which might lead to increase prediction error if QMDR is employed.

In the next section, we applied the OQMDR algorithm on several simulated data sets. The QMDR algorithm also applied to the same simulated data sets to assess the ability of the OQMDR to capture the correct model and/or to select a model with a smaller MSPE comparing to QMDR.

2.3 Simulation Study

We tested our proposed method using multiple simulated data sets. The main goal of the simulation study is to examine the ability of the OQMDR method to spot the most important interaction and whether that captured model coincides with the actual model that used to generate the data. In addition to that, we compared the performance of both the QMDR and OQMDR methods in all simulated data sets. Each simulated data set consists of five variables in which four of them contain individuals information about four genetic factors, each with three levels. In this work, we are using upper case letters to represent genetic factors, i.e. $A, B, C..$, etc. Whereas, the levels (allele combinations) of factor A are presented as (AA, Aa, aa), and (BB, Bb, bb) for factor B, and so on. Finally, the fifth variable contains the continuous phenotypic information of individuals.

The simulation procedure is accomplished as follows. First, the genetic information is generated in accordance with the Hardy-Weinberg principle [16]. That is, assuming each gene has two alleles (for example A and a) with a single locus frequencies of $p(A) = p$ and $p(a) = q$. Hence $p(AA) = p^2$, $p(Aa) = 2pq$, and $p(aa) = q^2$. In all simulated data sets, genetic information are generated using $p = q = 0.5$. After that, all possible combinations of all factor levels are represented in a four dimensional space as shown in figure 2.1.

Next, the continuous trait variable is generated based on six different scenarios in which each scenario links the high phenotype status of individuals to a certain combination of the genetic factors. Ten different data sets for each sample size of 500, 1000, and 2000 are randomly generated according to each scenario. In our simulation study, all data sets are generated based on either one or two 2-way interaction(s) (equation 2.3), or a single 3-way interaction (equation 2.4) as the actual disease predisposition interaction(s). Both QMDR and OQMDR algorithms are applied to each of the generated data sets.

$$Y_i \;=\; \mu + \sum_{1 \le a < b}^{4} \sum_{l_a, l_b \ge 1}^{3} \alpha_{ab.l_a l_b} \mathbb{I}_{\{X_{ai}=l_a, X_{bi}=l_b\}} + \epsilon_i \tag{2.3}$$

Above, Y_i is the simulated value of the trait variable of the i^{th} individual, μ is a baseline mean, and it's considered known for the purpose of simulation, α_{ab} is a 3×3 matrix of coefficients of the 2-way interaction between the a^{th} and the b^{th}

factors (there are a total of four factors in each simulated data set) and it's considered known for the purpose of simulation, $\mathbb{I}_{\{\cdot\}}$ is the indicator function, X_{ai} and X_{bi} are the generated allele combinations for the a^{th} and b^{th} factors respectively of the i^{th} individual, l_a and l_b are the allele combinations (the levels) of the a^{th} and b^{th} factors respectively, and ϵ_i is the random error term of the i^{th} individual, and $\epsilon_i \overset{iid}{\sim} N(0, 400)$.

$$Y_i \;=\; \mu + \sum_{1\leq a<b<c}^{4} \sum_{l_a,l_b,l_c,\geq 1}^{3} \beta_{abc.l_a l_b l_c} \mathbb{I}_{\{X_{ai}=l_a, X_{bi}=l_b, X_{ci}=l_c\}} + \epsilon_i \qquad (2.4)$$

Above,Y_i is the simulated value of the trait variable of the i^{th} individual, μ is a baseline mean, and it's considered known for the purpose of simulation, β_{abc} is a $3 \times 3 \times 3$ array of coefficients of the 3-way interaction between the a^{th}, b^{th}, and c^{th} factors and it's considered known for the purpose of simulation, $X_{ai}, X_{bi},$ and X_{ci} are the generated allele combinations of the a^{th}, b^{th}, and c^{th} factors respectively of the i^{th} individual, l_a, l_b, and l_c are the allele combinations (the levels) of the a^{th}, b^{th}, and c^{th} factors respectively, and ϵ_i is the random error term of the i^{th} individual, and $\epsilon_i \overset{iid}{\sim} N(0, 400)$.

For the purpose of simulation, the allele combinations are defined as numbers instead of letters. For example in factor A, we coded its allele combinations (or levels) as $(AA, Aa, aa) = (0, 1, 2)$.

The output of the six different simulated scenarios is demonstrated in the following subsections.

2.3.1 Case 1: True Model = AB

In the first case, the continuous phenotype variable is generated according to equation 2.3 with α_{12} be the only non-zero matrix in this case and it's given below. This matrix of coefficients will make most of the variation in the response variable due to the 2-way interaction between factors A and B. Therefore, AB is the anticipated proposed 2-way interaction, and one of ABC and ABD is likely to be selected as the proposed 3-way interaction because both of these interactions contains the true 2-way interaction (AB) that causes the disease. The reason behind choosing a simply spotted 2-way interaction is to assess the ability of the OQMDR method to capture the actual model and to see whether it gives the same output given by QMDR or not. First, we run both OQMDR and QMDR algorithms on ten different data sets of size 500, the output are summarized in table 2.2. Then, using the same model defined in 2.3 with the matrix α_{12} defined in 2.5, two different data sets of sizes 1000 and 2000 respectively are randomly generated and analyzed using both algorithms. The summarized results are presented in tables 2.3 and 2.4 respectively.

$$\alpha_{12} = \begin{bmatrix} 20 & 20 & 20 \\ 20 & 0 & 0 \\ 20 & 0 & 0 \end{bmatrix} \tag{2.5}$$

Table 2.2: Case 1: True model= AB, and $n = 500$

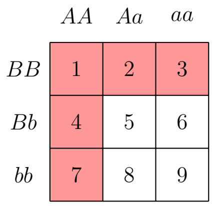
	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Final Model		
Set	New	QMDR	New	QMDR	New	QMDR	New	QMDR	New	QMDR	p_t
1	AB	AB	11.7287	11.7287	10	10	408.4470	408.4470	AB	AB	0.000
	ABD	ABD	10.3067	11.3290	4.2	8.6	427.7721	414.4013			
2	AB	AB	10.0317	10.0317	10	10	388.1858	388.1858	AB	AB	0.000
	ABD	ABD	8.5569	9.1548	4.6	6	405.6381	399.4966			
3	AB	AB	9.2992	9.5056	9.8	10	415.1356	413.2127	AB	AB	0.000
	ABC	ABC	7.5275	8.0303	2.4	3	441.5868	434.3282			
4	AB	AB	10.1738	10.1738	10	10	394.1987	394.1987	AB	AB	0.002
	ABD	ABD	8.5796	9.0350	4.4	3.2	414.5262	410.0920			
5	AB	AB	8.8108	8.8108	10	10	384.8257	384.8257	AB	AB	0.001
	ABC	ABC	7.6971	8.0768	2.8	4	398.9528	394.6174			
6	AB	AB	9.4734	9.5716	9.8	10	412.9209	411.4114	AB	AB	0.000
	ABD	ABD	9.4018	9.4370	6.4	7	413.7644	413.8329			
7	AB	AB	9.3303	9.8125	9	10	434.5222	426.9884	AB	AB	0.000
	ABC	ABC	8.0843	8.3846	4.6	1.8	448.9833	448.7589			
8	AB	AB	11.8964	11.8964	10	10	385.3691	385.3691	AB	AB	0.001
	ABD	ABD	10.2532	11.2503	4.6	7.6	409.5222	394.6645			
9	AB	AB	11.9762	11.9762	10	10	357.8918	357.8918	AB	AB	0.000
	ABD	ABD	11.1572	11.5281	6.4	6.2	369.0396	363.5977			
10	AB	AB	10.0830	10.083	10	10	395.8701	395.8701	AB	AB	0.002
	ABD	ABD	8.4653	8.3005	3.4	1.8	417.2304	421.1289			

Both OQMDR and QMDR proposed the interaction AB with the pattern shown in figure 2.2, which minimizes the prediction error and coincides with the real risk pattern, as the best 2-way model in most of the ten data sets. In fact, QMDR captured the risk pattern shown in figure 2.2 from all 10 samples. On the other hand, our method failed to spot the risk pattern that minimizes the prediction error in two cases, and it selected another risk pattern. This is showing that the QMDR performs better (and faster) when the risk pattern is recognizable, especially when the sample size is relatively small comparing to the number of multilocus combinations in the data set.

For the proposed 3-way model, ABC and ABD are chosen as the best 3-way models with the given risk pattern in figure 2.3, which minimizes the prediction error and matches the actual risk pattern used to generate the data set. The proposed risk pattern of the 3-way models shown in figure 2.3 is reproduced two out of ten and three out of ten times from OQMDR and QMDR respectively. Notice that both

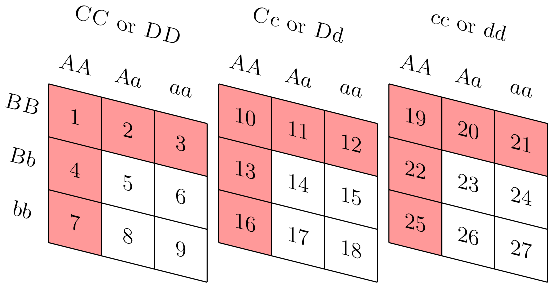
ABC and ABD models proposed the same risk patterns, this is mainly because factors C and D do not have a considerable effect on the continuous trait, and much of the variation is originally from the 2-way interaction between A and B. This can be seen clearly by looking at the testing $\bar{t}$-scores, as well as the cross-validation consistencies, of the proposed 2-way and 3-way models, where the 2-way models are favored from all ten generated data sets. Both algorithms perform similarly when capturing the best 2-way model, with QMDR performing better in samples 6 and 7, where the $\overline{\text{MSPE}}$s are smaller for QMDR in these two cases. The reason why the $\overline{\text{MSPE}}$s are smaller is because the cross-validation consistencies are larger, which means various risk patterns are proposed in some folds of the cross-validation procedure for OQMDR. This will make the predicted values in equation 2.1 calculated by OQMDR different from the ones predicted by QMDR, which in turn make the two $\overline{\text{MSPE}}$s different. Even though QMDR has lower $\overline{\text{MSPE}}$ in two data sets, the two algorithms selected the same risk pattern of the model AB from the remaining eight data sets. The p-values for the winner models from the OQMDR method are calculated using equation 2.2 and reported in table 2.2. All proposed models show a statistical significance at $\alpha = 0.05$.

Figure 2.2: Case 1: Risk pattern for the proposed 2-way models



	AA	Aa	aa
BB	1	2	3
Bb	4	5	6
bb	7	8	9

Figure 2.3: Case 1: Risk pattern for the proposed 3-way models



Tables 2.3 and 2.4 show the results of analyzing data sets of size 1000 and 2000 generated in accordance with equation 2.3 with the matrix α_{12} defined in 2.5. Again, both algorithms selected AB with the pattern shown in figures 2.2 as the best 2-way interaction, and as the best final model from all ten data sets. All final ten best models are minimizing the $\overline{\text{MSPE}}$s, and are statistically significant at $\alpha = 0.05$. Similarly, ABC and ABD with risk pattern shown in figure 2.3 are proposed from all data sets except sample 6 when $n = 1000$, where both algorithms failed to capture the true risk pattern. As the sample size gets bigger, both algorithms propose 3-way interactions with $\overline{\text{CVC}}$s, $\bar{t}$-scores, and $\overline{\text{MSPE}}$ssimilar to the ones of the selected 2-way models in most cases. This is mainly because, as sample size increases, enough data for multilocus combination is generated to capture the true risk pattern. However, none of the selected 3-way models beat the chosen 2-way models in all cases. This justifies the ability of both algorithms to detect the most important interaction among all possible interactions of any degree.

Table 2.3: Case 1: True model= AB, and $n = 1000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	New	QMDR	New	QMDR	New	QMDR	New	QMDR	New	QMDR	p_t
1	AB	AB	16.2091	16.2091	10	10	391.3797	391.3797	AB	AB	0.000
	ABD	ABD	16.0485	16.2091	9.6	10	393.1263	391.3797			
2	AB	AB	14.9832	14.9832	10	10	394.2057	394.2057	AB	AB	0.000
	ABD	ABC	14.3735	14.4725	4.4	5.6	399.5710	399.077			
3	AB	AB	16.8702	16.8702	10	10	415.5575	415.5575	AB	AB	0.000
	ABC	ABC	16.0420	16.8702	6.6	10	425.2365	415.5575			
4	AB	AB	16.1570	16.1570	10	10	385.8023	385.8023	AB	AB	0.000
	ABD	ABD	16.1570	16.0306	10	9.6	385.8023	386.9053			
5	AB	AB	15.7406	15.7406	10	10	418.5548	418.5548	AB	AB	0.000
	ABD	ABD	15.1607	15.1955	6	6.8	424.3849	424.0121			
6	AB	AB	14.3067	14.3067	10	10	416.5734	416.5734	AB	AB	0.000
	ABC	ABC	13.4176	13.8459	4.4	5	425.8612	421.7158			
7	AB	AB	15.7926	15.7926	10	10	396.3913	396.3913	AB	AB	0.000
	ABD	ABD	15.7926	15.6330	10	9.6	396.3913	397.7595			
8	AB	AB	14.7499	14.7499	10	10	387.1103	387.1103	AB	AB	0.000
	ABC	ABC	13.9113	14.4371	4.4	8.2	395.2781	390.1936			
9	AB	AB	15.1977	15.1977	10	10	386.0439	386.0439	AB	AB	0.000
	ABD	ABD	14.5515	14.4102	5.8	6	391.7452	392.7678			
10	AB	AB	13.6602	13.6602	10	10	400.5872	400.5872	AB	AB	0.000
	ABD	ABD	12.9908	13.2186	5.6	8.4	407.2909	405.0866			

Table 2.4: Case 1: True model= AB, and $n = 2000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	New	QMDR	New	QMDR	New	QMDR	New	QMDR	New	QMDR	p_t
1	AB	AB	20.8692	20.8692	10	10	400.5952	400.5952	AB	AB	0.000
	ABD	ABD	20.8311	20.8692	9.8	10	400.8953	400.5952			
2	AB	AB	22.1992	22.1992	10	10	419.1182	419.1182	AB	AB	0.000
	ABD	ABD	22.1992	22.1992	10	10	419.1182	419.1182			
3	AB	AB	23.0960	23.0960	10	10	397.5907	397.5907	AB	AB	0.000
	ABC	ABC	23.0960	23.0960	10	10	397.5907	397.5907			
4	AB	AB	22.2190	22.2190	10	10	403.6428	403.6428	AB	AB	0.000
	ABC	ABC	22.2190	22.2190	10	10	403.6428	403.6428			
5	AB	AB	23.7884	23.7884	10	10	395.2871	395.2871	AB	AB	0.000
	ABC	ABC	23.7884	23.7884	10	10	395.2871	395.2871			
6	AB	AB	23.3298	23.3298	10	10	409.8670	409.8670	AB	AB	0.000
	ABD	ABD	23.7884	23.3298	9.8	10	410.2410	409.8670			
7	AB	AB	22.3034	22.3034	10	10	417.6942	417.6942	AB	AB	0.000
	ABC	ABC	23.2794	22.3034	9.8	10	417.7373	417.6942			
8	AB	AB	22.9130	22.9130	10	10	380.0712	380.0712	AB	AB	0.000
	ABC	ABC	22.2509	22.7141	9.6	9.2	380.8257	381.4304			
9	AB	AB	22.7459	22.7459	10	10	384.0000	384.0000	AB	AB	0.000
	ABC	ABC	22.7077	22.7459	8.2	10	385.4606	384.0000			
10	AB	AB	22.0691	22.0691	10	10	401.2103	401.2103	AB	AB	0.000
	ABD	ABD	22.0691	22.0691	10	10	401.2103	401.2103			

2.3.2 Case 2: True Model = ABD

In this case, the data is generated according to equation 2.4 to produce a 3rd-degree interaction between A, B and D. The $3 \times 3 \times 3$ array of coefficients of the 3-way interaction between factors A, B and D is defined in array β_{124} below (2.6), which is the only non-zero array in equation 2.4. The proposed 2-way model is anticipated to be either AB, AD, or BD because all of them are related to the true 3-way model. The two algorithms are applied to ten different data sets of sizes 500, 1000, and 2000 (i.e., a total of 30 different samples). The results are summarized in tables 2.5, 2.6, and 2.7 for each distinct sample size, respectively.

$$\beta_{124} \quad = \quad \begin{bmatrix} 20 & 20 & 20 \\ 20 & 0 & 0 \\ 20 & 0 & 0 \end{bmatrix} \text{¦} \begin{bmatrix} 20 & 20 & 20 \\ 20 & 0 & 0 \\ 20 & 0 & 0 \end{bmatrix} \text{¦} \begin{bmatrix} 20 & 20 & 20 \\ 20 & 20 & 20 \\ 20 & 20 & 20 \end{bmatrix} \tag{2.6}$$

Figure 2.4 shows the risk pattern of the proposed 2nd-degree interactions. Both algorithms choose either AB, AD, or BD with the same pattern from nine different data sets. The selected risk pattern, in fact, coincides with the true risk pattern for the proposed models. That is, the data of the highlighted combinations in figure 2.4 are originally generated from normal distribution with $\mu = 140$, so they are anticipated to be at high risk. For the 3-way interaction, ABD with the risk pattern shown in figure 2.5 are chosen as the best 3rd-degree interaction in both methods from all simulated data sets. When $n = 500$, the OQMDR was able to catch the true 3rd-degree interaction with the true risk pattern as a best final model three times out of ten generated data sets. Similarly the QMDR did, however, OQMDR minimized the $\overline{\text{MSPE}}$ six times comparing to three times for QMDR. This suggests that, compared to Case 1 results, the OQMDR detects higher order interactions better than QMDR when a considerable portion of the variation is linked to higher degree models. Permutation testing validates the significance of all final models at $\alpha = 0.05$.

Table 2.5: Case 2: True model= ABD, and $n = 500$

	Model		t^*_k-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	New	QMDR	New	QMDR	New	QMDR	New	QMDR	New	QMDR	p_t
1	AB	AB	6.5730	6.1691	7	6.2	455.9759	458.6748	ABD	ABD	0.000
	ABD	ABD	10.7520	10.6366	9.6	9.2	398.7606	400.4112			
2	BD	BD	10.4849	9.9858	8	8	448.4179	453.5067	ABD	ABD	0.001
	ABD	ABD	14.0163	13.357	9.8	8.2	390.9129	401.4832			
3	AB	AB	8.7770	8.9778	7.4	8.4	452.556	449.0203	ABD	ABD	0.001
	ABD	ABD	12.9996	12.9996	9.8	10	386.422	386.4220			
4	AB	AB	6.8381	7.3701	7	8.6	456.0006	446.8594	ABD	ABD	0.000
	ABD	ABD	11.3547	11.2875	9.6	9.8	392.2859	393.2927			
5	BD	BD	7.1715	7.1715	5.8	5.8	432.2053	432.2053	ABD	ABD	0.002
	ABD	ABD	9.8944	9.8585	8.4	8	396.0920	396.8831			
6	AD	AD	8.2396	8.2396	10	10	447.0274	447.0274	ABD	ABD	0.001
	ABD	ABD	10.1203	9.8981	9.6	9.4	422.0019	424.1579			
7	AB	AB	8.3741	8.7376	8	8.6	460.101	453.6067	ABD	ABD	0.001
	ABD	ABD	12.3105	11.9923	8.4	7.8	403.2968	407.4371			
8	BD	BD	6.4064	6.5146	4	4.8	483.439	479.9112	ABD	ABD	0.001
	ABD	ABD	8.1019	10.9400	7	8	443.2720	416.6436			
9	AD	AD	6.3084	6.6935	3	5.4	423.0105	418.6122	ABD	ABD	0.005
	ABD	ABD	8.2252	8.5616	3.4	2	401.3948	396.7326			
10	AD	AD	10.8318	10.8318	10	10	448.6341	448.6341	ABD	ABD	0.000
	ABD	ABD	12.2139	12.0012	9.8	9	422.3265	425.3305			

Figure 2.4: Case 2: Risk patterns for the proposed 2-way models

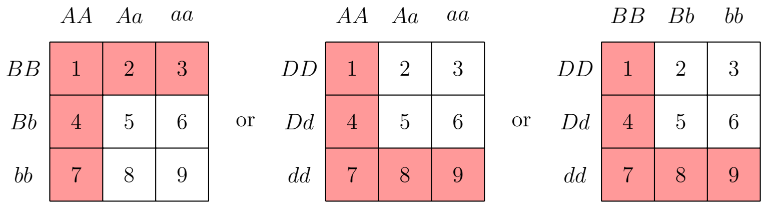
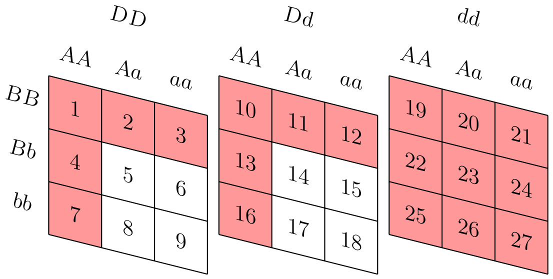


Figure 2.5: Case 2: Risk pattern for the proposed 3-way models



As the sample size increases, both algorithms were able to catch the true 3rd order interaction with true risk pattern in most cases. OQMDR did slightly better than QMDR in most cases, which can be seen by looking at the $\overline{\text{MSPE}}$. Similarly, for 2-way interaction, both algorithms show better performance when $n = 1000$ and $n = 2000$.

Table 2.6: Case 2: True model= ABD, and $n = 1000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	New	QMDR	New	QMDR	New	QMDR	New	QMDR	New	QMDR	p_t
1	AD	AD	10.5574	10.2271	9	9	441.5641	442.5816	ABD	ABD	0.000
	ABD	ABD	15.1906	14.8834	9.8	8.8	396.6424	399.3344			
2	AB	AB	11.7607	12.1116	8.6	9	448.8092	445.3460	ABD	ABD	0.000
	ABD	ABD	17.4943	17.1770	9.6	8	391.4856	394.3848			
3	AB	AB	10.1154	10.1833	8.4	8.6	449.8666	448.0307	ABD	ABD	0.000
	ABD	ABD	15.2729	15.0277	10	9.4	402.2826	404.4142			
4	AB	AB	11.8645	11.8339	9.6	9.6	446.2381	445.5723	ABD	ABD	0.000
	ABD	ABD	16.5129	16.4661	9.6	9.4	399.4253	399.8401			
5	AD	AD	11.0368	11.2362	7.6	8.2	419.1949	416.9439	ABD	ABD	0.000
	ABD	ABD	16.4543	16.3527	9.4	8.8	367.1102	368.0519			
6	AB	AB	9.7263	10.1839	8	8.8	445.8096	441.5744	ABD	ABD	0.000
	ABD	ABD	14.8798	14.5676	9.6	8.2	400.7890	403.5115			
7	AD	AD	12.7311	12.7311	9.8	9.8	461.7784	461.7784	ABD	ABD	0.000
	ABD	ABD	16.3577	16.2499	9.8	9.6	419.8441	420.9200			
8	AD	AD	12.2878	12.2878	9.4	9.4	444.8018	444.8018	ABD	ABD	0.000
	ABD	ABD	16.8883	16.8883	10	10	397.9913	397.9913			
9	AB	AB	11.3598	11.2036	8.8	9.2	430.9790	430.9132	ABD	ABD	0.000
	ABD	ABD	15.0487	14.9694	10	9.8	394.3994	395.1003			
10	AB	AB	13.3458	13.3458	10	10	419.7528	419.7528	ABD	ABD	0.000
	ABD	ABD	16.6502	16.6893	9.6	9.8	386.2950	385.9145			

Table 2.7: Case 2: True model= ABD, and $n = 2000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	New	QMDR	New	QMDR	New	QMDR	New	QMDR	New	QMDR	p_t
1	AD	AD	14.8645	15.0224	8.8	9	444.0846	442.5293	ABD	ABD	0.000
	ABD	ABD	21.731	21.6413	10	9.8	397.6133	398.1759			
2	AD	AD	17.0628	16.9337	9.4	9.6	440.5959	440.621	ABD	ABD	0.000
	ABD	ABD	22.2545	22.2545	10	10	403.5848	403.5848			
3	AD	AD	15.1775	15.3682	8.2	8.6	434.3665	432.7851	ABD	ABD	0.000
	ABD	ABD	22.5347	22.5347	10	10	385.8353	385.8353			
4	BD	BD	14.7748	14.2952	3.8	3.8	426.5887	428.0045	ABD	ABD	0.000
	ABD	ABD	23.6509	23.6509	10	10	366.8095	366.8095			
5	AD	AD	15.7685	15.9937	4	4.8	493.5584	491.1328	ABD	ABD	0.000
	ABD	ABD	25.3687	25.3687	10	10	418.3394	418.3394			
6	AB	AB	17.2735	17.5525	9.2	9.6	459.3593	457.415	ABD	ABD	0.000
	ABD	ABD	24.0497	24.0497	10	10	409.8031	409.8031			
7	BD	BD	14.4874	13.9843	8.4	8.4	440.5774	442.247	ABD	ABD	0.000
	ABD	ABD	21.6227	21.6227	10	10	394.0755	394.0755			
8	BD	BD	15.8059	16.081	9	9	439.1698	437.3144	ABD	ABD	0.000
	ABD	ABD	21.9942	21.9942	10	10	394.0679	394.0679			
9	BD	BD	15.511	15.511	6.6	6.6	478.2585	478.2585	ABD	ABD	0.000
	ABD	ABD	22.2766	22.187	8.6	5.4	427.2937	428.0481			
10	BD	BD	15.4857	15.4857	5.4	5.4	429.6563	429.6563	ABD	ABD	0.000
	ABD	ABD	21.9574	21.9574	10	10	384.8686	384.8686			

The first two cases, in which the variation is generated to be spotted easily, prove the ability of OQMDR method to spot the true source of variation precisely. However, besides its fast performance, QMDR performs slightly better in a few cases in terms of precision, especially when a low order of interaction is considered and the sample size is relatively small. In the next two cases, we will test the ability of OQMDR of detecting the true models when it is somewhat ambiguous.

2.3.3 Case 3: True Model = BD

In this case, we are considering an interaction with a risk pattern that is not easy to identify. The data is generated using equation 2.3 such that the 2-way interaction between factors B and D with the multilocus coefficient matrix α_{24} defined in equation 2.7 is causing the variation. However, in this scenario, some multilocus combinations are hard to tell whether they are at high risk or not because their averages are very close to the overall mean of the continuous variable. Since QMDR is using the overall mean as a threshold to classify individuals, we expect these combinations to be identified at high risk when QMDR is employed. This is because the fixed overall mean of the simulated data is 125.112, and is slightly lower than 128, the fixed average of the multilocus combinations *bbDd* and *Bbdd*. On the other hand, since the OQMDR is optimizing over the t statistic as a criterion for classification, these cells likely would not be classified as high risk when OQMDR is employed because 125.112 is not too far from 128.

$$\alpha_{24} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 8 \\ 0 & 8 & 30 \end{bmatrix} \tag{2.7}$$

The simulation study shows that with samples of size $n = 500$, OQMDR algorithm selected the 2nd-degree model BD with a risk pattern that labels the combination *bbdd* at high risk, and leaves the remaining eight combinations at low risk (figure 2.6a). Conversely, QMDR detected a risk pattern that assumes individuals with *bbDd*, *Bbdd*, or *bbdd* combinations are at high risk, while the remaining individuals get a low-risk label (figure 2.6b). Results in table 2.8 show that the OQMDR algorithm proposed models with lower $\overline{\text{MSPE}}$ than the one suggested by QMDR in eight different data sets of size $n = 500$. OQMDR mistakenly proposed a three-way interaction from one simulated data set. We believe this happened mainly due to over-fitting of the three-way interaction. All proposed interactions show a statistical significance under $\alpha = 0.05$ level of significance.

Table 2.8: Case 3: True model= BD, and $n = 500$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	BD	BD	12.5683	4.1713	10	2.8	434.0705	504.8569	BCD	BD	0.000
	BCD	BCD	12.9882	3.3897	9.2	1	485.8083	518.4198			
2	BD	BD	8.7738	5.4023	10	7.8	408.2276	422.9147	BD	BD	0.005
	ABD	BCD	7.6926	4.5184	9.6	3.4	410.2450	435.5922			
3	BD	BD	9.1315	4.4981	10	2.8	416.5782	457.2411	BD	BD	0.002
	BCD	BCD	6.6742	3.4628	6.2	1	435.0735	475.9284			
4	BD	BD	7.9738	6.7334	10	9.2	449.8849	464.1022	BD	BD	0.004
	ABD	ABD	4.3174	5.1466	5.6	2.2	479.8738	485.3993			
5	BD	BD	8.6924	5.9639	10	9.4	449.0318	470.4264	BD	BD	0.003
	BCD	BCD	7.0404	5.3489	7.4	2.6	460.5493	479.8260			
6	BD	BD	9.8547	6.5190	10	10	399.2730	437.8976	BD	BD	0.001
	ABD	BCD	6.1132	3.8583	4.2	0.6	444.9145	470.8817			
7	BD	BD	5.5904	7.3218	5	9.6	410.0513	394.3539	BD	BD	0.006
	BCD	BCD	5.0521	6.5264	2.6	3.4	420.7200	405.2157			
8	BD	BD	7.0933	3.9189	10	8.2	426.6266	458.4983	BD	BD	0.002
	ABD	ACD	2.0475	3.8924	4.4	2.8	465.5235	464.1423			
9	BD	BD	5.3883	5.7548	8.4	7.6	390.8981	397.7204	BD	BD	0.008
	BCD	BCD	3.4575	4.5889	1	1.2	421.3202	411.4107			
10	BD	BD	7.5865	7.9369	9.2	9.2	407.7313	407.0326	BD	BD	0.002
	ABD	ABD	6.7918	7.3485	4.2	5.2	418.4133	415.3837			

Figure 2.6: Case 3: Risk patterns for the proposed 2-way models

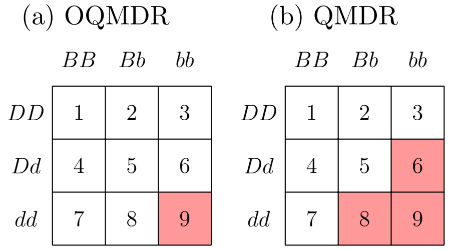
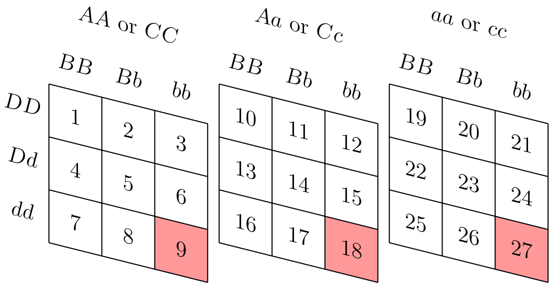

(a) OQMDR (b) QMDR

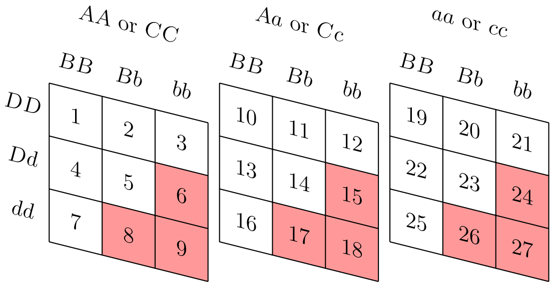
Figure 2.7 shows the suggested risk patterns for the proposed 3-way interaction from both methods (for QMDR, only when $n = 2000$). These risk patterns coincide with the one proposed for the 2-way interaction. However, our method selected the one that minimizes the prediction error six times out of ten. That is, OQMDR considers individuals from six different data sets with *AAbbdd*, *Aabbdd*, and *aabbdd* only at high risk when ABD is selected (similarly, when BCD is selected) as the best 3-way interaction (figure 2.7a). On the other hand, QMDR could not select a certain risk pattern more than once, i.e., there was a distinct risk pattern for each simulated data set.

Figure 2.7: Case 3: Risk patterns for the proposed 3-way models

(a) OQMDR



(b) QMDR



Eventually, as sample size increases, both methods steadily proposed the risk pattern of the chosen 2-way model described in figure 2.6. QMDR shows a higher $\overline{\text{MSPE}}$ in eight out of ten different data sets. Likewise, OQMDR was able to capture the pattern shown in 2.7a more frequently for the chosen 3-way model. While QMDR failed to propose a frequent pattern when $n = 1000$, it suggested the one shown in figure 2.7b from two different data sets of size $n = 2000$.

Table 2.9: Case 3: True model= BD, and $n = 1000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	BD	BD	11.0682	7.8971	10	10	380.8382	402.2249	BD	BD	0.000
	BCD	BCD	10.4367	4.1576	9.4	0.6	383.8686	425.6393			
2	BD	BD	12.9821	9.8492	10	10	422.2082	435.4096	BD	BD	0.000
	BCD	BCD	11.4207	7.6621	7.4	2.2	433.6829	451.3270			
3	BD	BD	6.3368	8.2012	7	10	442.1930	436.1841	BD	BD	0.000
	ABD	BCD	2.9712	6.5175	5	3.8	461.4515	448.9053			
4	BD	BD	9.3511	6.7128	10	6.2	425.7843	439.5918	ABD	BD	0.000
	ABD	BD	11.4646	5.7312	10	2.6	421.2707	446.2048			
5	BD	BD	12.9657	6.7591	10	7.2	432.7799	474.0929	BD	BD	0.000
	BCD	BCD	12.8884	6.2155	9.8	1.6	433.1348	479.1019			
6	BD	BD	11.3902	8.8044	10	9.4	370.0498	383.7267	BD	BD	0.000
	ABD	BCD	8.7654	8.1114	8.8	4.2	375.9737	387.5567			
7	BD	BD	10.2836	6.8308	10	8.2	408.8597	434.9666	BD	BD	0.000
	ABD	BD	10.2836	6.7667	10	1.4	408.8597	436.3363			
8	BD	BD	10.0232	8.0489	10	8.6	414.0131	434.9144	BD	BD	0.000
	ABD	BD	8.6775	7.1249	4.2	1.6	429.7187	442.5955			
9	BD	BD	8.4548	8.4579	9.6	10	425.9161	425.8094	BD	BD	0.000
	ABD	BD	5.7513	8.3393	5.2	5	438.4999	427.1634			
10	BD	BD	8.5341	8.4838	9.4	9.4	419.9538	419.0005	BD	BD	0.000
	BCD	BCD	6.7702	5.7943	3	1.4	431.3286	437.6684			

Table 2.10: Case 3: True model= BD, and $n = 2000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	BD	BD	11.6430	11.8266	9.6	10	396.1322	394.8831	BD	BD	0.000
	ABD	ABD	8.6305	10.7238	5.6	5.8	405.4561	399.2102			
2	BD	BD	12.2938	14.0354	9.2	10	416.3105	415.0608	BD	BD	0.000
	BCD	ABD	11.8194	12.5054	9	4.6	417.2390	421.9241			
3	BD	BD	14.3724	11.5847	10	10	392.2242	404.8032	BD	BD	0.000
	ABD	ABD	14.3724	11.3747	10	6	392.2242	405.3948			
4	BD	BD	13.5646	11.3259	10	10	387.7859	396.5111	BD	BD	0.000
	BCD	BCD	13.5646	11.0208	10	5.6	387.7859	397.9806			
5	BD	BD	19.5815	12.4247	10	10	406.9310	426.2756	BD	BD	0.000
	ABD	ABD	19.5815	12.1834	10	5.4	406.9310	427.9580			
6	BD	BD	13.5869	11.5905	10	10	412.4245	420.8890	BD	BD	0.000
	ABD	BCD	13.5869	10.2778	10	6.8	412.4245	426.7064			
7	BD	BD	15.2050	12.1491	10	10	418.1521	429.5760	BD	BD	0.000
	BCD	ABD	15.2050	10.8154	10	3.4	418.1521	435.5389			
8	BD	BD	15.3370	10.2874	10	10	401.5109	419.0839	BD	BD	0.000
	BCD	BCD	15.3370	8.7304	10	2.8	401.5109	426.8023			
9	BD	BD	16.1559	12.1680	10	10	414.2017	433.6866	BD	BD	0.000
	BCD	ABD	16.1559	10.5314	10	2.6	414.2017	441.7923			
10	BD	BD	13.7691	8.4757	10	5.2	426.6336	456.4637	BD	BD	0.000
	BCD	BCD	13.3682	7.6388	8.6	1.4	429.3549	460.3287			

This case clearly shows the ability of the OQMDR method to capture the true

2-way model with the risk pattern that minimizes the prediction error comparing to the one selected by QMDR.

2.3.4 Case 4: True Model = ABC

In this case, we will inspect the behavior of both methods when the data is generated using a true 3-way interaction with a vague risk pattern. That is, the data is generated using equation 2.4 to make much of the variation comes from a 3-way interaction between A, B, and C with the non-zero array β_{123} shown below (equation 2.8). Similar to case 3, some individuals with a certain multilocus combination seem to be affected by the interaction but not to the point where they can be diagnosed at high risk. These individuals are the ones with a multilocus coefficient of 8 in β_{123}. Once again, we expect these individuals to be recognized at high risk of developing the disease when the data is analyzed using QMDR. On the other hand, we think that classifying these individuals at low risk could benefit the prediction error of the proposed model.

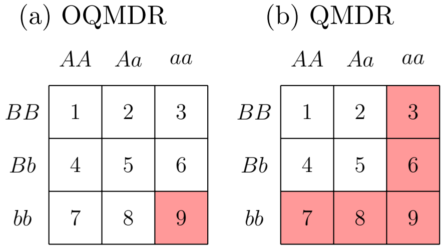
$$\beta_{123} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 30 \end{bmatrix} \vdots \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 8 \\ 0 & 8 & 30 \end{bmatrix} \vdots \begin{bmatrix} 0 & 0 & 30 \\ 0 & 8 & 30 \\ 30 & 30 & 30 \end{bmatrix} \tag{2.8}$$

Tables 2.11, 2.12, and 2.13 show the summarized result of all simulated data sets for this case. We can see from these tables that the OQMDR method is able to capture the true model with high $\overline{\text{CVC}}$ from all generated data sets regardless of sample size. On the other hand, QMDR method couldn't spot the right interaction from two samples of size $n = 500$. In addition, $\overline{\text{CVC}}$ is too low for 3-way models comparing to 2-way models, especially for small data sets. QMDR performance enhanced when $n = 2000$ comparing to its performance with smaller samples. Even when QMDR catches the true 3-way interaction as the best final model, the proposed risk pattern (figure 2.9b) still not similar to the one proposed by OQMDR (figure 2.9a). Therefore, the calculated $\overline{\text{MSPE}}$ by QMDR is bigger in most cases comparing to the calculated $\overline{\text{MSPE}}$ when our method is employed. The selected risk patterns for the 2-way models (figure 2.8) coincide with the 3-way risk patterns suggested by both algorithms. All final models are statistically significant at $\alpha = 0.05$.

Table 2.11: Case 4: True model= ABC, and $n = 500$

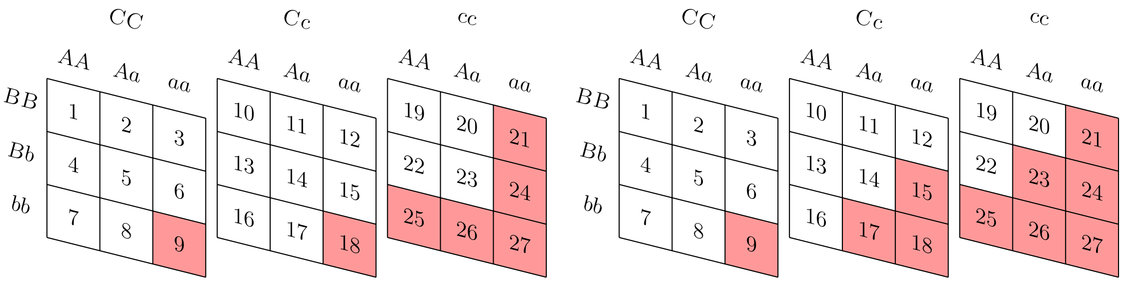
	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	AB	AB	8.1592	4.3038	9.6	5.2	449.3350	463.3199	ABC	ABC	0.004
	ABC	ABC	9.4425	6.1780	10	1.6	425.6279	448.3547			
2	AC	AB	6.7711	5.1319	7.8	3.0	491.2154	496.8703	ABC	ABC	0.000
	ABC	ABC	13.0731	5.5268	10	1.6	422.3193	493.7586			
3	AB	AB	7.5528	6.7622	7.6	8.4	469.8370	474.9827	ABC	ABC	0.000
	ABC	ABC	12.1367	10.7735	8.2	7.6	390.6237	408.4659			
4	BC	BC	7.2838	8.0747	8.2	9.6	443.4005	433.8597	ABC	BC	0.002
	ABC	ABC	10.0161	7.6555	8.8	2	404.5911	443.1291			
5	AB	AB	7.0792	7.2978	7.8	8.0	419.0948	414.6803	ABC	ABC	0.000
	ABC	ABC	9.4804	5.5522	6.2	1.8	388.8457	440.0572			
6	AC	BC	7.9512	5.2632	10	8.4	414.9305	431.8647	ABC	ABC	0.000
	ABC	ABC	10.0555	5.4055	9.4	2.6	375.7701	433.6660			
7	AC	BC	3.4077	3.8853	2.6	2.8	496.2005	488.5971	ABC	ABC	0.003
	ABC	ABC	9.4207	6.0516	9.4	3.6	421.7769	465.6457			
8	AC	AC	4.5705	6.1200	5.4	8.2	492.0345	481.0760	ABC	ABC	0.000
	ABC	ABC	7.3506	8.5925	3.8	4.6	464.3512	448.5742			
9	AB	AC	7.5584	4.9344	9.2	6	413.3126	429.5232	ABC	ABC	0.000
	ABC	ABC	11.3497	9.1189	9.6	7	356.4875	378.7932			
10	AB	BC	3.6775	7.3431	3	8.6	481.1166	445.0819	ABC	BC	0.000
	ABC	ABC	12.0756	7.0335	10	1.6	392.1839	454.1246			

Figure 2.8: Case 4: Risk patterns for the proposed 2-way models


(a) OQMDR

(b) QMDR

Figure 2.9: Case 4: Risk patterns for the proposed 3-way models


(a) OQMDR

(b) QMDR, $n = 2000$

Table 2.12: Case 4: True model= ABC, and $n = 1000$

	Model		t^*_k-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	AB	AC	5.3255	9.2554	5.6	10	437.9143	415.8413	ABC	ABC	0.000
	ABC	ABC	14.6746	10.3055	9.2	1.6	375.0088	406.6042			
2	AC	AB	8.147	8.0186	7.2	6.4	477.4761	475.508	ABC	ABC	0.000
	ABC	ABC	16.5818	9.4974	10	1.4	405.113	465.0731			
3	AB	AC	6.2362	8.6327	5.2	7	468.3818	457.2100	ABC	ABC	0.000
	ABC	ABC	15.2867	10.331	10	3.2	402.3473	442.2342			
4	AC	BC	11.0101	8.1911	9.8	9	491.1019	500.1286	ABC	ABC	0.000
	ABC	ABC	16.6098	9.5573	10	1.4	428.3733	489.5901			
5	AB	AB	6.2034	8.5717	3.4	7.4	464.496	447.0618	ABC	ABC	0.000
	ABC	ABC	14.3892	12.0804	10	10	399.3435	417.3327			
6	AB	BC	7.7195	7.4988	7.6	4.4	520.3213	517.9386	ABC	ABC	0.000
	ABC	ABC	15.5735	10.7579	8	2.8	447.6252	484.1171			
7	AB	AC	11.1854	9.2312	9.8	7.8	460.1744	466.6482	ABC	ABC	0.000
	ABC	ABC	16.8849	14.0872	10	6.6	398.714	419.2578			
8	AC	AC	8.0931	7.9741	5.2	9	476.6604	469.4303	ABC	ABC	0.000
	ABC	ABC	15.2774	9.7153	10	1.6	410.097	454.5966			
9	AC	BC	8.4073	6.0474	4.8	5.2	507.5589	522.2604	ABC	ABC	0.000
	ABC	ABC	16.8113	11.807	10	5	419.8534	460.1074			
10	BC	BC	8.9735	7.2974	9	6.6	459.5296	465.8615	ABC	ABC	0.000
	ABC	ABC	14.3775	10.557	10	4.8	409.0999	439.5665			

Table 2.13: Case 4: True model= ABC, and $n = 2000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	AC	BC	12.7407	13.3746	4.6	7	493.361	480.8834	ABC	ABC	0.000
	ABC	ABC	26.5165	18.2417	10	7.6	406.868	445.3642			
2	AC	BC	13.9946	10.7861	8.8	5.4	483.6113	493.7696	ABC	ABC	0.000
	ABC	ABC	23.1537	17.0146	10	8.2	410.5839	451.5221			
3	BC	BC	14.4859	9.3923	10	5.4	475.9016	491.8175	ABC	ABC	0.000
	ABC	ABC	21.8987	16.9541	10	8	413.9034	444.4705			
4	BC	AB	12.6492	11.7796	7.8	7.8	512.8168	507.7509	ABC	ABC	0.000
	ABC	ABC	22.9638	17.1027	10	9.6	435.1123	470.0695			
5	BC	AB	13.9567	11.5544	10	9.4	484.6121	488.8457	ABC	ABC	0.000
	ABC	ABC	20.6543	15.9273	10	4.2	427.4643	454.6603			
6	AB	BC	16.9326	14.4704	10	10	499.1951	495.6421	ABC	ABC	0.000
	ABC	ABC	24.7238	19.1987	10	5.4	425.9687	457.8688			
7	AB	AB	14.7683	12.1203	9.2	8.4	469.6782	471.5572	ABC	ABC	0.000
	ABC	ABC	23.5707	17.4366	10	9.4	404.9740	437.7395			
8	AC	AB	9.1476	13.0252	7.6	8.4	509.5984	494.44	ABC	ABC	0.000
	ABC	ABC	22.8113	17.1016	10	4.2	428.3051	464.533			
9	BC	BC	14.6272	10.3086	10	4.4	473.948	489.892	ABC	ABC	0.000
	ABC	ABC	22.4449	17.8233	10	8.4	418.9029	441.5572			
10	BC	BC	11.6949	12.5488	4.6	6	455.6321	452.4679	ABC	ABC	0.000
	ABC	ABC	21.8562	17.5682	10	8.6	394.3332	418.256			

The last two cases show that OQMDR method is superior to QMDR method in terms of selecting the true model with a more realistic risk pattern that minimizes the prediction error. However, OQMDR algorithm, similar to QMDR, attributes the variation in the continuous phenotype to a single interaction, which is usually the most significant interaction. In the following two cases, we will investigate the drawback of capturing the true model when the true model comprises multiple gene-gene interactions.

2.3.5 Case 5: True Models = AB and AD

We generated the data sets in accordance with equation 2.3 with the non-zero matrices α_{12} and α_{14} given in 2.9 and 2.10, receptively. The way we generated the data makes the variation mainly due to the 2-way interaction between factors A and D. It also attributes a considerable portion of the variation to the 2-way interaction between factors A and B, but not as potent as AD; however, it should be easily recognizable. Hence, we have two different 2-way interactions that can be deemed as the primary sources of the variation in the continuous variable. Since OQMDR and QMDR algorithms can only propose a single most significant interaction, we expect the 2-way interaction between A and D, and the 3-way interaction between A, B, and D to be detected as the best 2-way and best 3-way models, respectively. Notice that factor A is a common factor in both true 2-way interactions, therefore, if we combine the effects of the two 2-way interactions, we could end up with a true 3-way interaction of ABD with the coefficient array given in 2.11. In fact, the means of the cells with coefficients of 15 in β_{124} defined in 2.11 are slightly lower than 139.44, the overall fixed mean of the response variable. In this case, we expect OQMDR, opposite to QMDR, to propose a 3-way interaction with a risk pattern that deems these cells at high risk rather than low risk, and this will likely benefit the prediction error afterward.

$$\alpha_{12} = \begin{bmatrix} 15 & 15 & 15 \\ 15 & 0 & 0 \\ 15 & 0 & 0 \end{bmatrix} \tag{2.9}$$

$$\alpha_{14} = \begin{bmatrix} 0 & 0 & 20 \\ 0 & 0 & 20 \\ 20 & 20 & 20 \end{bmatrix} \tag{2.10}$$

$$\beta_{124} = \begin{bmatrix} 15 & 15 & 35 \\ 15 & 0 & 20 \\ 15 & 0 & 20 \end{bmatrix} \vdots \begin{bmatrix} 15 & 15 & 35 \\ 15 & 0 & 20 \\ 15 & 0 & 20 \end{bmatrix} \vdots \begin{bmatrix} 35 & 35 & 35 \\ 35 & 20 & 20 \\ 35 & 20 & 20 \end{bmatrix} \tag{2.11}$$

Simulation results summarized in tables 2.14, 2.15, and 2.16 show the struggle of both algorithms to spot a consistent model, particularly for small data sets. However, both algorithms can recognize AD more frequently than AB as the single most significant 2-way interaction, which agrees with the original model used to generate the data. Similarly for the 3-way models, the interaction ABD is almost always selected as the best 3-way model. In many cases, OQMDR favors the 3-way model ABD over the 2-way model AD, which could be considered as an evidence of the ability of the OQMDR method to detect most of the significant variations in the response variable, which in turn, reduces the prediction error. Besides the struggle of choosing the same model repeatedly, the two methods also struggled to select a typical risk pattern for both studied orders of interaction for small samples. With the sample size gets larger, the outcomes of selecting the 2-way models become more stable, and both algorithms propose AD with the same risk pattern shown in figure 2.10. While for 3-way models, a more frequent risk pattern (figure 2.11a), that coincides with the array in 2.11, is steadily selected when OQMDR algorithm is used. Figure 2.11b shows the most frequent risk pattern suggested by QMDR, which does not recognize cells with coefficients of 15 in β_{124} given in 2.11 at high risk. Clearly, the risk pattern shown in figure 2.11a is enhancing the $\overline{\mathrm{MSPE}}$ for models suggested by OQMDR comparing to QMDR. The final assessment shows a statistical significance for all selected models under $\alpha = 0.05$ level of significance.

Table 2.14: Case 5: True models= AB and AD, and $n = 500$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	AD	AD	11.7314	9.7983	10	4.4	440.3874	468.9433	AD	ABD	0.000
	ABD	ABD	9.4894	10.6699	3.2	3.4	468.8971	456.4106			
2	AD	AD	9.2082	8.9597	9.8	9.4	398.6205	400.5740	AD	AD	0.003
	ACD	ACD	5.4320	7.7104	1.4	2.4	431.8312	418.2683			
3	AD	AD	6.5039	7.2341	7.2	7	473.7473	465.9969	AD	AD	0.008
	ABD	ABD	5.0696	6.1065	0.2	1.6	495.1787	484.8954			
4	AD	AD	10.2775	10.1762	10	9.8	479.6339	482.1524	AD	AD	0.000
	ABD	ABD	8.7677	7.7517	5.0	1.8	505.1238	522.5698			
5	AD	AD	6.1591	5.8608	2.6	4.6	476.7516	478.0475	ABD	ABD	0.004
	ABD	ABD	7.3457	8.0023	2.4	3	461.2534	452.2448			
6	AD	AD	9.4226	9.3614	8	8.2	467.8078	468.3114	AD	AD	0.000
	ABD	ABD	7.9864	7.3039	5.2	0.6	473.1573	503.1038			
7	AD	AD	9.6197	9.9817	7.2	9	431.1400	426.8027	ABD	ABD	0.000
	ABD	ABD	10.5822	11.1982	2.2	6.8	419.0597	409.5560			
8	AD	AD	10.5357	9.2378	10	5.8	419.6213	435.2594	ABD	AD	0.000
	ABD	ABD	11.5489	8.3165	9.6	1.2	407.8584	452.2779			
9	AB	AB	10.4754	10.7267	9.2	9.6	427.7591	423.9420	ABD	AD	0.000
	ABD	ABD	12.0899	9.6891	8.2	2.2	409.9213	442.8068			
10	AD	AD	8.3815	7.6558	5.8	4	478.9403	495.3767	ABD	AD	0.000
	ABD	ACD	11.4136	7.5647	9	2.6	439.0593	498.6284			

Figure 2.10: Case 5: Risk pattern for the proposed 2-way models

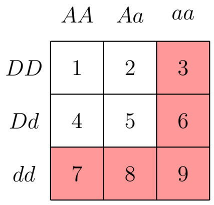
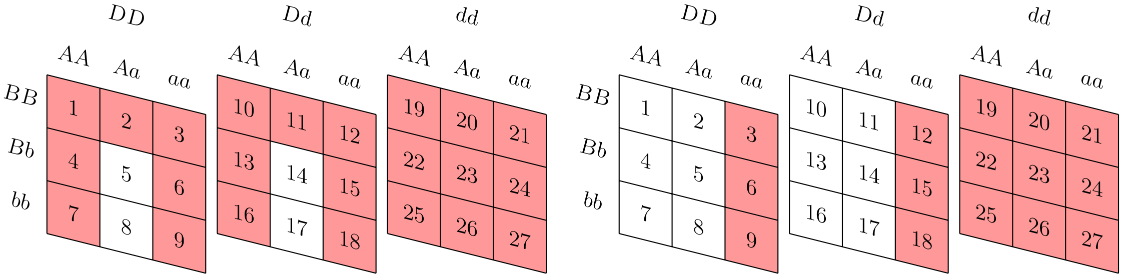


Figure 2.11: Case 5: Risk patterns for the proposed 3-way models


(a) OQMDR

(b) QMDR, $n = 2000$

Table 2.15: Case 5: True model= AB and AD, and $n = 1000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	AD	AD	13.2596	13.1028	6.8	5	425.0486	429.4854	ABD	ABD	0.000
	ABD	ABD	13.4813	14.4820	2.4	5.4	424.7562	417.4259			
2	AD	AD	12.9223	13.6241	7.2	9.4	440.5278	434.3656	AD	AD	0.000
	ABD	ABC	12.1719	12.8069	3	5.6	447.2645	443.1458			
3	AD	AD	13.4657	12.6078	10	8.8	425.9823	432.2383	ABD	AD	0.000
	ABD	ABD	14.9638	10.6318	10	0.8	417.5412	454.2127			
4	AD	AD	13.8768	12.6944	9.4	5.6	438.6083	450.0987	AD	ABD	0.000
	ABD	ABD	12.7438	13.3979	3.4	2	447.9546	443.7380			
5	AD	AD	13.7833	13.1832	9.8	8.2	501.0047	507.0657	ABD	ABD	0.000
	ABD	ABD	14.8234	13.2919	6	2	492.5713	508.7934			
6	AD	AD	13.7852	11.8361	10	4.8	471.9331	489.2642	ABD	ABD	0.000
	ABD	ABD	14.1047	11.9281	5	3.4	471.0434	489.4899			
7	AB	AB	9.6469	9.7269	5.2	5.6	439.6363	437.3603	ABD	AD	0.000
	ABD	ABD	13.3888	9.6629	9.6	1	405.5949	443.5348			
8	AD	AD	13.2426	11.9883	9.6	8	415.0823	422.7717	ABD	AD	0.000
	ABD	ABD	13.6737	11.7685	8.2	3.4	416.6501	425.1808			
9	AD	AD	14.8311	14.8311	10	10	407.6052	407.4881	ABD	AD	0.000
	ABD	ABC	15.3357	13.1694	5.6	4	402.7284	422.4762			
10	AD	AD	11.7728	11.8629	6.8	6.2	459.4256	459.6078	AD	AD	0.000
	ABD	ACD	11.5622	11.7843	8.8	1.4	451.3953	460.4617			

Table 2.16: Case 5: True model= AB and AD, and $n = 2000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	AD	AD	19.5487	18.2254	10	6.8	436.7792	445.8609	ABD	AD	0.000
	ABD	ABD	21.0317	17.6228	7.4	1.8	426.0936	451.9332			
2	AD	AD	20.7218	20.7218	10	10	430.2648	430.3555	ABD	AD	0.000
	ABD	ABD	22.2465	19.4373	10	4.6	425.5934	439.8688			
3	AD	AD	18.4839	16.9137	9.6	5.2	457.5057	468.3477	AD	ACD	0.000
	ABD	ACD	18.4156	17.3341	8.2	3.4	459.2096	465.4065			
4	AD	AD	15.6074	15.4906	8.4	6.8	472.9608	474.4019	ABD	AD	0.000
	ABD	ABD	17.3775	15.0600	5.8	1.8	456.6554	477.7161			
5	AD	AD	18.2416	19.0254	7.4	10	442.8265	436.8549	AD	AD	0.000
	ABD	ACD	17.1150	17.4120	7.2	5.2	441.0806	447.5459			
6	AD	AD	16.6251	17.6843	4.6	7.6	443.2278	438.9623	ABD	ACD	0.000
	ABD	ACD	19.9968	18.1588	4.6	6	425.5457	436.0290			
7	AD	AD	19.8670	18.7648	10	6.6	452.6704	459.0022	ABD	ABD	0.000
	ABD	ABD	22.0097	19.1501	10	2.8	442.1378	456.0802			
8	AD	AD	19.5662	18.8089	9.2	6.4	483.2015	488.6749	ABD	ACD	0.000
	ABD	ACD	21.0144	19.1270	9.6	5.2	476.8636	487.4600			
9	AD	AD	18.8469	19.3615	7.4	10	465.6927	461.9549	ABD	AD	0.000
	ABD	ABD	20.9963	17.1455	7.4	2.2	453.7394	478.1857			
10	AD	AD	19.2096	19.2096	10	10	431.6748	431.8430	AD	AD	0.000
	ABD	ABD	19.1753	18.5053	6	3.4	435.6491	437.3304			

2.3.6 Case 6: True Models = AB and CD

In this last simulated scenario, we intend to examine the OQMDR behavior when there are two distinct 2-way interactions affecting the response, i.e., no common factor between the two interactions. Hence, the model given in equation 2.3 is utilized along with the non-zero matrices α_{12} and α_{34} listed in 2.12 and 2.13, respectively, to generate the response variable such that certain combinations of AB and CD are causing the variation. Once again, none of the two algorithms can report a set of the most significant interactions; therefore, CD is expected to be selected as the most significant 2-way interaction because it has more weight on the Y. However, due to the drawback of both algorithms to capture more than one interaction, it's feasible to end up with an interaction that does not agree with any of the components of the actual model. Notice that the effects of both 2-way interactions can be combined to form a single 4-way interaction with the array γ_{1234} shown in 2.14. Accordingly, the all-way interaction has a higher chance to be proposed over lower order interactions. However, we only consider all possible 2-way and 3-way interactions; thus, all-way interactions are not an area of interest in this study.

$$\alpha_{12} = \begin{bmatrix} 0 & 0 & 15 \\ 0 & 0 & 15 \\ 15 & 15 & 15 \end{bmatrix} \tag{2.12}$$

$$\alpha_{34} = \begin{bmatrix} 0 & 0 & 20 \\ 0 & 0 & 20 \\ 20 & 20 & 20 \end{bmatrix} \tag{2.13}$$

$$\gamma_{1234} = \begin{bmatrix} \begin{bmatrix} 0 & 0 & 15 \\ 0 & 0 & 15 \\ 15 & 15 & 15 \end{bmatrix} \text{¦} \begin{bmatrix} 0 & 0 & 15 \\ 0 & 0 & 15 \\ 15 & 15 & 15 \end{bmatrix} \text{¦} \begin{bmatrix} 20 & 20 & 35 \\ 20 & 20 & 35 \\ 35 & 35 & 35 \end{bmatrix} \\ \\ \begin{bmatrix} 0 & 0 & 15 \\ 0 & 0 & 15 \\ 15 & 15 & 15 \end{bmatrix} \text{¦} \begin{bmatrix} 0 & 0 & 15 \\ 0 & 0 & 15 \\ 15 & 15 & 15 \end{bmatrix} \text{¦} \begin{bmatrix} 20 & 20 & 35 \\ 20 & 20 & 35 \\ 35 & 35 & 35 \end{bmatrix} \\ \\ \begin{bmatrix} 20 & 20 & 35 \\ 20 & 20 & 35 \\ 35 & 35 & 35 \end{bmatrix} \text{¦} \begin{bmatrix} 20 & 20 & 35 \\ 20 & 20 & 35 \\ 35 & 35 & 35 \end{bmatrix} \text{¦} \begin{bmatrix} 20 & 20 & 35 \\ 20 & 20 & 35 \\ 35 & 35 & 35 \end{bmatrix} \end{bmatrix} \tag{2.14}$$

Simulation results in tables 2.17, 2.17, and 2.17 show that the 2-way interaction CD is almost always selected from both algorithms, regardless of sample size. Notice that the $\overline{\text{MSPE}}$ of the 2-way model is smaller, in most samples, comparing to the $\overline{\text{MSPE}}$ of the proposed 3-way interactions; yet, it is not as small as the $\overline{\text{MSPE}}$ for previous cases. The reason why the $\overline{\text{MSPE}}$ is higher in this scenario is that the suggested interaction does not explain all the distinction in the response variable. On the other hand, both algorithms selected ACD or BCD as the best 3-way interaction, which agrees with the actual model to some extent, because CD is stronger in the real model than the other 2-way. Figure 2.12 shows that the risk pattern of the proposed 2-way interaction coincides with the coefficients in α_{34} shown in 2.10. Finally, permutation testing shows a statistical significance of all proposed model under $\alpha = 0.05$.

Table 2.17: Case 6: True model= AB and CD, and $n = 500$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	BD	AB	9.0069	7.0102	9.8	7.2	473.8596	473.471	BD	ABD	0.001
	BCD	ABD	4.0438	10.7973	8.8	5.8	503.8610	422.9145			
2	CD	CD	11.8300	11.8300	10	10	489.5151	489.5151	CD	CD	0.000
	ACD	ACD	9.2065	10.3003	2	4.8	539.7714	517.0491			
3	CD	CD	10.6991	10.6991	10	10	440.9293	440.9293	CD	CD	0.002
	BCD	BCD	8.5849	9.7989	2	5.8	473.0339	453.8657			
4	CD	CD	11.3164	11.3164	10	10	475.2049	475.2049	CD	CD	0.001
	BCD	ACD	6.6323	9.1011	2.4	3	552.0917	512.3414			
5	CD	CD	10.0542	10.313	9.6	10	452.8280	447.6891	CD	CD	0.001
	BCD	BCD	9.1363	8.9036	3.6	2.6	466.6728	470.9822			
6	CD	CD	8.9645	9.0815	9.8	10	457.8634	456.4878	CD	CD	0.003
	ABC	ABC	5.6587	7.9537	3.2	6.6	496.4742	474.6744			
7	CD	CD	11.4007	11.4007	10	10	462.4258	462.4258	CD	CD	0.000
	ACD	ACD	9.7958	10.1877	3.4	4.8	490.1020	484.1013			
8	CD	CD	10.4819	10.4819	10	10	476.3193	476.3193	ACD	ACD	0.001
	ACD	ACD	10.9939	10.7249	6.8	6.6	466.1766	470.4285			
9	CD	CD	10.9124	10.9124	10	10	424.9754	424.9754	CD	ACD	0.002
	ACD	ACD	10.6844	11.0481	6.6	7.4	429.7659	423.9798			
10	CD	CD	7.7286	7.5909	5.6	5.4	489.9941	490.2981	BCD	BCD	0.000
	BCD	BCD	10.0815	10.4887	5.8	7.4	449.8219	444.0032			

Figure 2.12: Case 6: Risk pattern for the proposed 2-way models

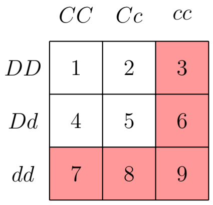


Figure 2.13: Case 6: Risk pattern for the proposed 3-way models

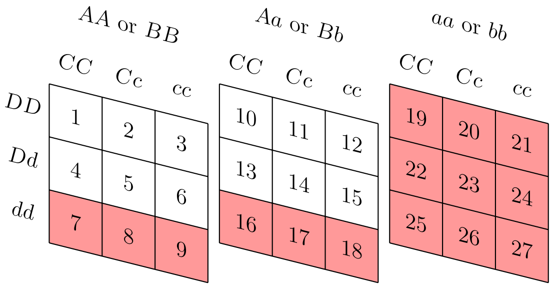


Table 2.18: Case 6: True model= AB and CD, and $n = 1000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	CD	CD	15.8692	15.8692	10	10	446.3464	446.3464	ACD	ACD	0.000
	ACD	ACD	15.8810	16.3463	4.6	8.6	447.6558	441.7190			
2	CD	CD	15.4883	15.4883	10	10	439.9703	439.9703	CD	CD	0.000
	BCD	BCD	13.4539	13.4441	3.4	2.2	463.9966	463.3633			
3	CD	CD	14.9038	14.9038	10	10	405.4913	405.4913	CD	CD	0.000
	BCD	BCD	14.0032	13.4048	7.2	5.2	414.6212	419.7204			
4	CD	CD	14.8358	14.8358	10	10	460.2263	460.2263	CD	CD	0.000
	BCD	BCD	14.0720	14.3682	2.2	1	469.7958	465.2314			
5	CD	CD	14.7896	14.7896	10	10	429.2789	429.2789	CD	ACD	0.000
	ACD	ACD	14.1741	14.8106	7.6	9	436.9781	429.8223			
6	CD	CD	13.3759	13.3759	10	10	450.7268	450.7268	CD	CD	0.000
	ABD	ABD	12.2997	13.0095	4.6	5.2	461.1326	452.5986			
7	CD	CD	11.1334	10.8941	9.2	9	494.0292	496.4412	BCD	BCD	0.000
	BCD	BCD	11.0531	11.1747	5.8	3.8	492.6775	492.4458			
8	CD	CD	15.0628	15.0628	10	10	463.5435	463.5435	CD	CD	0.000
	BCD	BCD	14.7724	13.8812	6.4	5.2	467.5203	477.2938			
9	CD	CD	14.5080	14.5080	10	10	478.4491	478.4491	CD	CD	0.000
	BCD	BCD	13.4828	13.8369	5.8	5.6	489.0802	486.0552			
10	CD	CD	15.5132	15.5132	10	10	440.6888	440.6888	CD	CD	0.000
	BCD	BCD	14.8498	14.1504	6	4.2	448.3133	456.5098			

Table 2.19: Case 6: True model= AB and CD, and $n = 2000$

	Model		t_k^*-score		$\overline{\text{CVC}}$		$\overline{\text{MSPE}}$		Best Model		
Set	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	NEW	QMDR	p_t
1	CD	CD	21.1057	21.1057	10	10	430.2597	430.2597	CD	CD	0.000
	ACD	ACD	20.0069	20.1394	3	6.6	438.8670	437.5555			
2	CD	CD	20.1353	20.1353	10	10	473.6575	473.6575	CD	CD	0.000
	BCD	BCD	18.3900	18.7932	2.2	3.4	486.3417	483.3645			
3	CD	CD	22.0497	22.0497	10	10	455.8369	455.8369	CD	CD	0.000
	BCD	BCD	20.7965	18.8983	5.6	2.4	466.3449	481.0303			
4	CD	CD	21.8425	21.8425	10	10	467.6960	467.6960	CD	CD	0.000
	ACD	ACD	20.8895	21.7709	4.2	7.0	475.5016	467.9811			
5	CD	CD	21.8637	21.8637	10	10	431.9311	431.9311	CD	CD	0.000
	ACD	ACD	20.4798	20.1306	3.6	5.8	442.8755	444.8904			
6	CD	CD	20.8501	20.8501	10	10	435.0721	435.0721	CD	CD	0.000
	BCD	BCD	19.9172	19.2205	4.2	2.8	442.4099	448.0044			
7	CD	CD	20.1937	20.1937	10	10	460.0336	460.0336	ACD	ACD	0.000
	ACD	ACD	20.7405	20.9611	8	9.6	455.9089	454.3679			
8	CD	CD	20.9419	20.9419	10	10	469.1425	469.1425	CD	CD	0.000
	BCD	BCD	19.1011	19.0341	2.2	3.8	483.0796	483.8075			
9	CD	CD	19.8520	19.8520	10	10	442.7585	442.7585	ACD	CD	0.000
	ACD	ACD	20.4027	18.5823	8.6	3.2	440.0263	452.3520			
10	CD	CD	19.9674	19.9674	10	10	456.0138	456.0138	BCD	BCD	0.000
	BCD	BCD	20.7447	20.2638	7	5.4	450.4488	453.7931			

Finally, it's important to mention that all proposed models from applying the

QMDR algorithm showed a statistical significance at $\alpha = 0.05$, regardless of whether they minimize the prediction error or not.

Copyright© Zaid T. Al-Khaledi, 2019.

Chapter 3 Modification of The OQMDR Algorithm

3.1 Preliminary

In chapter 2, we presented our new suggested algorithm to analyze genetic data sets with a continuous phenotypic response. We showed that the risk patterns of the proposed models by the OQMDR algorithm minimize the prediction error (smaller MSPE) compared to the risk patterns of the models suggested by the QMDR algorithm when both methods are applied to same data sets. However, taking into account the new algorithm digs deeper into the data to detect the final risk pattern for each interaction, the computation time can be substantial. Recalling that the OQMDR algorithm, similar to MDR and some other MDR-based algorithms, uses 1000 permutation testings to justify the significance of the final model. Therefore, this part of the algorithm has the lion's share when talking about time consumption. Besides, the computational burden gets heavier with bigger data sets. It is also affected by the complexity of the examined models. Coding experience shows that evaluating a 3-way model requires almost twice the time as long as a 2-way model does with the same data set analyzed on the same machine. Accordingly, finding a time-efficient replacement procedure to the permutation testing may benefit the proposed approach.

In 2009, Pattin et al. [48], and later in 2010, Hua et al. [30] both introduced a time-effective procedure that uses a theory-based technique to evaluate the proposed model instead of using the regular machine learning procedure. They suggested using the Generalized Extreme Value Distribution (GEVD), described by Jenkinson in 1955 [32], as an approximated theoretical distribution of the test statistic of the proposed model. The suggested approach does not eliminate the permutation testing procedure completely; instead, it reduces the number of permuted data sets required to assess the final model to 20 permutations [48], or at most 50 permutations [30]. The idea behind employing the GEVD to evaluate the final model is merely based on the fact that the statistic of the final model is chosen as the maximum of all computed statistics of all examined models. Hence, we can generate a set of permuted maximized statistics to estimate the parameters of the approximated theoretical distribution of the original maximized statistic. Since we are choosing between models by optimizing over the testing t-score in our work, we think that utilizing the GEVD is applicable, and it would likely improve the computation speed of our algorithm.

3.2 The Generalized Extreme Value Distribution

The Generalized Extreme Value Distribution (GEVD) initially described by Jenkinson in 1955 [32] is used to model the maximum (or minimum) of a sequence of independent and identically distributed random variables. That is, let $X_1, X_2, X_3, ...$ be a sequence of independent and identically distributed random variables. And define Y_n to be the largest order statistic:

$$Y_n \quad := \quad \max(X_1, X_2, ..., X_n)\, \forall n \in \mathbb{Z}^+$$

Then for some constants $a_n > 0$ and $b_n \in \mathbb{R}$, we have $(Y_n - b_n)/a_n$ has a limiting distribution called the Generalized Extreme Value distribution with the cumulative distribution function (CDF) given in equation 3.1[15]. That is:

$$P\left(\frac{Y_n - b_n}{a_n} \leq y\right) \quad \longrightarrow \quad F_Y(y), \text{ as } n \to \infty$$

where F is the CDF of the GEVD and it is defined as follows:

$$F_Y(y) \quad = \quad \begin{cases} \exp\left[-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right] & \text{for } \xi \neq 0 \\ \\ \exp\left[-\exp\left(-\frac{y-\mu}{\sigma}\right)\right] & \text{for } \xi = 0 \end{cases} \tag{3.1}$$

defined on $1+\xi\left(\frac{y-\mu}{\sigma}\right) > 0$ for $\xi \neq 0$, and $y \in (-\infty, \infty)$ for $\xi = 0$, where $\mu \in (-\infty, \infty)$ is the location parameter, $\sigma > 0$ is the scale parameter, and $\xi \in (-\infty, \infty)$ is the shape parameter of the distribution. In fact, three different distributions can be derived from the GEVD. These distributions are Weibull distribution when $\xi < 0$, Fréchet distribution when $\xi > 0$, and Gumbel distribution as $\xi \to 0$ [15]. Some references [32, 11] use a different parametrization to the one shown in equation 3.1 by defining the shape parameter as $k = -\xi$. This reparametrization does not affect the maximum likelihood estimates of the parameters except for the sign of the estimated value of the shape parameter, $\hat{\xi}$. In this work, we will consider the parametrization given in equation (3.1) when deriving the maximum likelihood estimators of the GEVD. Adequate changes are considered when we used `R` functions that rely on the alternative parametrization.

The mean, the variance, and the skewness of a random variable following the GEVD can be obtained as follows [22]:

$$\text{Mean} \quad = \quad \begin{cases} \mu + \sigma\frac{g_1 - 1}{\xi} & \text{if } \xi \neq 0, \xi < 1 \\ \mu + \sigma\gamma & \text{if } \xi = 0 \\ \infty & \text{if } \xi \geq 1 \end{cases} , \tag{3.2}$$

$$\text{Variance} = \begin{cases} \sigma^2 + \frac{g_2 - g_1^2}{\xi^2} & \text{if } \xi \neq 0, \xi < 0.5 \\ \sigma^2 \frac{\pi}{6} & \text{if } \xi = 0 \\ \infty & \text{if } \xi \geq 0.5 \end{cases}, \tag{3.3}$$

$$\text{Skewness} = \begin{cases} \operatorname{sgn}(\xi) \frac{g_3 - 3g_1 g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}} & \text{if } \xi \neq 0, \xi < \frac{1}{3} \\ \frac{12\sqrt{6}\zeta(3)}{\pi^3} & \text{if } \xi = 0 \\ \infty & \text{if } \xi \geq \frac{1}{3} \end{cases} \tag{3.4}$$

where $g_i = \Gamma(1 - i\xi)$ for $i = 1, 2, 3, ...$, γ is the Euler's constant, $\operatorname{sgn}(\cdot)$ is the sign function, and $\zeta(x) = \sum_{n=1}^{\infty} n^{-x}$ is the Euler-Riemann zeta function.

We will make use of these three measures to initiate the estimation process of the parameters of the GEVD.

3.3 Parameter Estimation

The Generalized Extreme Value distribution with the CDF given in equation 3.1 has three parameters, the location μ, the shape σ, and the scale ξ. These parameters can be estimated by the Probability-Weighted Moments method [29], or the maximum likelihood estimator (MLE) method [47, 49, 34]. The estimation procedure has to be done numerically, for example by using the multivariate version of the Newton-Raphson algorithm [38] because the derivatives of the log-likelihood cannot be solved for the three parameters. Otherwise, we may use the profile likelihood function with a fixed range of values assigned to ξ, then calculate the regular MLE's of the other two functions [15]. In this work, the analytical approach is utilized to obtain the MLE's of the three parameters. To proceed with the calculation of the MLE's, we need to derive the formulas of the gradient vector, $g(\theta)$, and the inverse of the Hessian matrix of the log-likelihood, H^{-1}. Next, we calculate the MLE's iteratively according to the formula defined in 3.5 below:

$$\hat{\theta}_t = \hat{\theta}_{t-1} - H^{-1}(\hat{\theta}_{t-1})g(\hat{\theta}_{t-1}) \tag{3.5}$$

where:

$$\theta = \begin{bmatrix} \mu & \sigma & \xi \end{bmatrix}',$$

$$g(\theta) = \begin{bmatrix} \frac{\partial l(\theta)}{\partial \mu} & \frac{\partial l(\theta)}{\partial \sigma} & \frac{\partial l(\theta)}{\partial \xi} \end{bmatrix}',$$

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 l(\theta)}{\partial \mu^2} & \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma} & \frac{\partial^2 l(\theta)}{\partial \mu \partial \xi} \\ \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma} & \frac{\partial^2 l(\theta)}{\partial \sigma^2} & \frac{\partial^2 l(\theta)}{\partial \sigma \partial \xi} \\ \frac{\partial^2 l(\theta)}{\partial \mu \partial \xi} & \frac{\partial^2 l(\theta)}{\partial \sigma \partial \xi} & \frac{\partial^2 l(\theta)}{\partial \xi^2} \end{bmatrix},$$

and the index t denotes iterations for $t = 1, 2, \ldots$.

Notice that the Hessian matrix is symmetric (i.e., $H = H^T$). The final forms of the elements of $g(\theta)$ and $H(\theta)$ are given by Joe in an unpublished technical report [33]. We decided to verify the derivation of all derivatives needed to calculate the MLE's, where the case of $\xi \neq 0$ is considered in the derivation.

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of independent and identically distributed random variables that follow the GEVD with the CDF defined in equation 3.1 for $\xi \neq 0$. Therefore, the common probability density function (PDF) can be written as:

$$f_Y(y;\mu,\sigma,\xi) \quad = \quad \frac{1}{\sigma}\left[1+\xi\frac{y-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})} e^{-\left[1+\xi\frac{y-\mu}{\sigma}\right]^{-\frac{1}{\xi}}}$$

defined on $1+\xi\left(\frac{y-\mu}{\sigma}\right) > 0$ for $\xi \neq 0$, $\mu \in (-\infty,\infty)$ is the location parameter , $\sigma > 0$ is the scale parameter, and $|\xi| > 0$ is the shape parameter.

Thus, the likelihood function for $Y_1, Y_2, ..., Y_n$ is:

$$\begin{aligned}
L(\mu,\sigma,\xi) &= \Pi_{i=1}^n f_{Y_i}(y_i;\mu,\sigma,\xi) \\
&= \Pi_{i=1}^n \frac{1}{\sigma}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})} e^{-\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}}} \\
&= \frac{1}{\sigma^n}\left[\Pi_{i=1}^n\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\right] e^{-\sum_{i=1}^n\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}}} \\
&= \frac{1}{\sigma^n}\left[\Pi_{i=1}^n(1+\xi z_i)^{-(1+\frac{1}{\xi})}\right] e^{-\sum_{i=1}^n(1+\xi z_i)^{-\frac{1}{\xi}}}
\end{aligned}$$

where $z_i = \frac{y_i-\mu}{\sigma}$.

Then, the log-likelihood is:

$$l(\mu,\sigma,\xi) \quad = \quad -n\log\sigma - (1+\frac{1}{\xi})\sum_{i=1}^n \log(1+\xi z_i) - \sum_{i=1}^n (1+\xi z_i)^{-\frac{1}{\xi}} \qquad (3.6)$$

and the elements of the gradient are:

$$\frac{\partial l}{\partial \mu} \quad = \quad \frac{\xi+1}{\sigma}\sum_{i=1}^n (1+\xi z_i)^{-1} - \frac{1}{\sigma}\sum_{i=1}^n (1+\xi z_i)^{-(1+\frac{1}{\xi})}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\xi+1}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-1} z_i - \frac{1}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})} z_i$$

$$\begin{aligned}\frac{\partial l}{\partial \mu} &= \frac{1}{\xi^2}\sum_{i=1}^{n}\log(1+\xi z_i) - (1+\frac{1}{\xi})\sum_{i=1}^{n}(1+\xi z_i)^{-1} z_i \\ &\quad +\frac{1}{\xi}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})} z_i - \frac{1}{\xi^2}\sum_{i=1}^{n}(1+\xi z_i)^{-\frac{1}{\xi}}\log(1+\xi z_i)\end{aligned}$$

and the Hessian matrix elements are:

$$\frac{\partial^2 l}{\partial \mu^2} = \frac{\xi(\xi+1)}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-2} - \frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-(2+\frac{1}{\xi})}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \mu \partial \sigma} &= -\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-1} + \frac{\xi(\xi+1)}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-2} z_i \\ &\quad +\frac{1}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})} - \frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-(2+\frac{1}{\xi})} z_i\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \mu \partial \xi} &= \frac{1}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-1} - \frac{\xi+1}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-2} z_i \\ &\quad +\frac{1+\xi^{-1}}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-(2+\frac{1}{\xi})} z_i - \frac{1}{\xi^2\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})}\log(1+\xi z_i)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2} - 2\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-1}z_i + \frac{\xi(\xi+1)}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-2}z_i^2 \\ &\quad + \frac{2}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})}z_i - \frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}(1+\xi z_i)^{-(2+\frac{1}{\xi})}z_i^2\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \sigma \partial \xi} &= -\frac{\xi+1}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-2}z_i^2 + \frac{1}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-1}z_i \\ &\quad + \frac{1+\xi^{-1}}{\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-(2+\frac{1}{\xi})}z_i^2 - \frac{1}{\xi^2\sigma}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})}\log(1+\xi z_i)z_i\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \xi^2} &= (1+\frac{1}{\xi})\sum_{i=1}^{n}(1+\xi z_i)^{-2}z_i^2 - (1+\frac{1}{\xi})\frac{1}{\xi}\sum_{i=1}^{n}(1+\xi z_i)^{-(2+\frac{1}{\xi})}z_i^2 \\ &\quad + \frac{2}{\xi^2}\sum_{i=1}^{n}(1+\xi z_i)^{-1}z_i - \frac{2}{\xi^2}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})}z_i \\ &\quad + \frac{2}{\xi^3}\sum_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})}\log(1+\xi z_i)z_i - \frac{2}{\xi^3}\sum_{i=1}^{n}\log(1+\xi z_i) \\ &\quad + \frac{2}{\xi^3}\sum_{i=1}^{n}(1+\xi z_i)^{-\frac{1}{\xi}}\log(1+\xi z_i) - \frac{1}{\xi^4}\sum_{i=1}^{n}(1+\xi z_i)^{-\frac{1}{\xi}}\left(\log(1+\xi z_i)\right)^2\end{aligned}$$

The MLE's are calculated numerically according to equation 3.5 using `R`. Due to the poor behavior of the likelihood function of the GEVD, the procedure requires the initial values of the three parameters to be chosen deliberately close to the final estimated values. Otherwise, the estimation process may diverge in some cases [33]. Since Gumbel distribution is a special case of the GEVD and can be obtained by letting $\xi \longrightarrow 0$, Castillo et al. [11] suggested using the MLE formulas of the location and scale parameters of Gumbel distribution to estimate μ_0 and σ_0, respectively, and set $\xi_0 := 0$ to initiate the iterative estimation process. This could be an easy way to determine the initial values of θ. However, coding experience shows that it is very unusual to end up with an empirical distribution with $\hat{\xi}_{MLE} \approx 0$. Hence, setting $\xi_0 := 0$ may not lead to convergence always. Accordingly, we think that solving the mean, the variance, and the skewness, which are given in equations 3.2, 3.3, and 3.4, receptively, would provide a more logical selection of the starting points. Solving these equations requires calculating the mean, $\bar{Y}$, the variance, S_Y^2, and the skewness, $\hat{\eta}_3$ from an observed sample. Here, the coefficient of skewness is the third standardized central moment and is defined as follows:

$$\eta_3 = \frac{E(Y - E(Y))^3}{[E(Y - E(Y))^2]^{3/2}}$$

and it can be estimated as follows [42]:

$$\hat{\eta}_3 = \frac{n \sum_{i=1}^{n} (y_i - \bar{y})^3}{(n-1)(n-2) S_Y^3}$$

where

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Next, we use the `uniroot` function in `R` to obtain the root of equation 3.7 below:

$$\operatorname{sgn}(\xi) \frac{g_3 - 3g_1 g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}} - \hat{\eta}_3 = 0 \tag{3.7}$$

The root of equation 3.7 only exists when we assume $\xi < \frac{1}{3}$. This root will be used in the Newton-Raphson algorithm as ξ_0. To obtain μ_0 and σ_0, we solve equation

3.2 for μ, and equation 3.3 for σ, respectively, to get the following formulas:

$$\begin{aligned}\sigma_0 &= \sqrt{\frac{S^2}{g_2 - g_1^2}\xi_0^2}\\ \mu_0 &= \bar{Y} - (g_1 - 1)\frac{\sigma_0}{\xi_0}\end{aligned}$$

where g_1 and g_2 are calculated using ξ_0.

Finally, the approach of selecting the initial values described above does not assure convergence of the Newton-Raphson algorithm always. Therefore, we might need to adjust the initial values ($\hat{\theta}_0$), or the value of $\hat{\theta}_s$ for $s < t$, manually until we achieve the convergence [33]. A further modification to the estimation algorithm is recommended by Prescott and Walden [49] and by Otten and Van Montfort [47]. The adjustment, which involves adding an optional correction step to the analytical estimation process, was mainly proposed to reduce the number of iterations required to achieve the maximum of the likelihood and to increase the chance of convergence.

We write our own `R` code to compute the MLE's of the GEVD. The code involves using the library `EnvStats` [42] to call the function `Skewness` required to calculate the skewness of the sample. The results of our code are compared to the output of the function `fitdist` from the library `fitdistrplus` [19], which can be used to obtain the MLE of the GEVD. Both codes are supplemented with the same set of initial values produced from the mechanism described earlier. Despite returning the same MLE's, both codes fail to converge in some cases, especially when the number of permuted statistics is less than 20. In our code, we impose some constraints on the value of $\hat{\theta}_s$ for $s < t$ within the algorithm to reduce the chance of divergence by keeping the value of $\hat{\theta}_s$ under control.

3.4 Utilizing The GEVD in OQMDR Algorithm

As mentioned earlier, the GEVD has been used to assess the significance of the proposed interactions as a time-efficient replacement to the regular permutation testings procedure in some MDR-based algorithms [48, 30]. Since we are optimizing over many t-test statistics, we think that employing the GEVD in the OQMDR algorithm would likely benefit the efficiency of model assessment. That is, the GEVD can be used to approximate the behavior of the maximized testing t-score. To do this, we generate a relatively small number of permuted samples and report the permuted maximized testing t-score from each permutation. The number of permuted data sets needed to estimate the approximated distribution of the optimized test statistic could be as low as 20 permuted data sets [48], or 50 permuted data sets [30] instead of the 1000 permuted data sets we used in chapter 2.

To proceed with the calculation, assume we have a data set of size n. And let Y be the continuous response variable of interest in the data set, and let N be the total number of genetic factors in the data set. The selection of the final model process is going to be similar to the approach described in chapter 2; therefore, we will skip directly to the model assessment component of the algorithm. Let $t_k^{*(0)}$ be the testing t-score of the proposed k-way interaction when computed from the original data set, for $k = 2, 3, \ldots, N-1$. Now, define $T_{k_{max}}^{*(0)}$ to be the largest order statistic of the random variable $T_k^{*(0)}$, i.e.:

$$T_{k_{max}}^{*(0)} := \max(T_2^{*(0)}, T_3^{*(0)}, \ldots, T_{N-1}^{*(0)})$$

Since $T_{k_{max}}^{*(0)}$ is a maximum of a sequence of random variables, we assume that the GEVD would be a plausible approximation to model the behavior of $T_{k_{max}}^{*(0)}$. To estimate the approximated null distribution of $T_{k_{max}}^{*(0)}$, we permute the original data set m

times and re-perform the OQMDR algorithm on each of the permuted data sets to get $t^{*(1)}_{k_{max}}, t^{*(2)}_{k_{max}}, \ldots, t^{*(m)}_{k_{max}}$. Next, we apply the numerical estimation algorithm described earlier in equation 3.5 on the permuted sample of t-scores $(t^{*(1)}_{k_{max}}, t^{*(2)}_{k_{max}}, \ldots, t^{*(m)}_{k_{max}})$ to obtain the MLE's of the parameters characterizing the null distribution of $T^{*(0)}_{k_{max}}$. Once we are done estimating the GEVD parameters, we can calculate the approximated p-value of $t^{*(0)}_{k_{max}}$ as follows (cf. Hua et al., 2010 [30]):

$$p^{(0)}_{k_{max}} = 1 - F_{T^{*(0)}_{k_{max}}}(t^{*(0)}_{k_{max}}; \hat{\mu}_{t^{*(0)}_{k_{max}}}, \hat{\sigma}_{t^{*(0)}_{k_{max}}}, \hat{\xi}_{t^{*(0)}_{k_{max}}})$$

where $F_{T^{*(0)}_{k_{max}}}$ is the GEVD distribution function of the random variable $T^{*(0)}_{k_{max}}$ evaluated at $T^{*(0)}_{k_{max}} = t^{*(0)}_{k_{max}}$, and $\hat{\mu}_{t^{*(0)}_{k_{max}}}$, $\hat{\sigma}_{t^{*(0)}_{k_{max}}}$, and $\hat{\xi}_{t^{*(0)}_{k_{max}}}$ are the MLE's of the parameters of the GEVD of $T^{*(0)}_{k_{max}}$.

Next, we need to justify the validity of $p^{(0)}_{k_{max}}$, which can be done by approximating the null distribution of $P^{(0)}_{k_{max}}$. Since $p^{(0)}_{k_{max}} \in (0,1)$ and is a monotone decreasing function of $T^{*(0)}_{k_{max}}$, we thought we could consider following Hua et al. [30] and utilizing the GEVD again to approximate the distribution of $-\log(P^{(0)}_{k_{max}})$ in order to verify the validity of the computed p-value. However, based on numerical investigation, the GEVD doesn't seem to be an appropriate choice to approximate the distribution of $-\log(P^{(0)}_{k_{max}})$ in our case. Therefore, we tested a few other distributions to find the best fit for the null distribution of $P^{(0)}_{k_{max}}$, $-\log(P^{(0)}_{k_{max}})$, and $-\log(-\log(P^{(0)}_{k_{max}}))$. We tested uniform and beta distributions for $P^{(0)}_{k_{max}}$, Weibull and GEV distributions for $-\log(P^{(0)}_{k_{max}})$, and GEVD for $-\log(-\log(P^{(0)}_{k_{max}}))$. Among all considered distributions and transformations, GEVD for $-\log(-\log(P^{(0)}_{k_{max}}))$ appears to be the best choice per the graphical representation of the simulated data. Notice that because $p^{(0)}_{k_{max}}$ is monotone decreasing in $t^{(0)}_{k_{max}}$, the transformation $-\log(-\log(p^{(0)}_{k_{max}}))$ is monotone decreasing in $t^{(0)}_{k_{max}}$. The reason why we consider the $-\log(-\log(P^{(0)}_{k_{max}}))$ transformation is because typically the null distribution of the p-value is uniform(0,1). Therefore, since we know that $p^{(0)}_{k_{max}} \in (0,1)$, we can assume that the null distribution of $P^{(0)}_{k_{max}}$ is approximately uniform(0,1). Hence, under this assumption, the random variable $-\log(P^{(0)}_{k_{max}})$ would follow the exponential distribution with a scale parameter $\sigma = 1$. Now, since the exponential distribution is a special case of Weibull distribution with a shape parameter $\mu = 1$ and a scale parameter $\sigma = 1$, we may assume that the random variable $-\log(P^{(0)}_{k_{max}})$ is distributed as Weibull$(1,1)$ [10]. Subsequently, the log transformation (so as the $-\log$ transformation) of a Weibull$(1,1)$ random variable follows Gumbel distribution, which is a special case of the GEVD when $\xi \to 0$ [10]. To see this, let $X \sim$ Weibull(μ,σ) with the CDF defined as follows:

$$F_X(x;\mu,\sigma) \quad = \quad 1 - \exp(-\frac{x}{\sigma})^{\mu} \text{ for } x > 0; \mu,\sigma > 0$$

Now, let $Y = g(X) = \mu(1 - \sigma\log\frac{X}{\sigma})$. Since Y is a monotonic decreasing transformation on X, the CDF of Y can be obtained using the monotone transformation formula [10]. That is:

$$\begin{aligned} F_Y(y) &= 1 - F_X(g^{-1}(y)) \\ &= 1 - F_X(\sigma\exp(-\frac{y-\mu}{\sigma})^{\frac{1}{\mu}}) \\ &= \exp\left[-\frac{\sigma\exp(-\frac{y-\mu}{\sigma})^{\frac{1}{\mu}}}{\sigma}\right]^{\mu} \\ &= \exp\left[-\exp\left(-\frac{y-\mu}{\sigma}\right)\right] \text{ for } y \in \mathbb{R}; \mu,\sigma > 0 \end{aligned}$$

The latter form of $F_Y(y)$ is the CDF of the Gumbel distribution as defined in equation 3.1. Therefore, the GEVD, which includes Gumbel distribution as a particular case, would be a plausible candidate to describe the behavior of $-\log(-\log(P^{(0)}_{k_{max}}))$.

The preceding described assessment can be practically done by applying the OQMDR algorithm on m_1 permuted data sets to get $p^{(1)}_{k_{max}}, p^{(2)}_{k_{max}}, \ldots, p^{(m_1)}_{k_{max}}$. Then, this permuted sample of minimized p-values is used to approximate the null GEV distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ using the multivariate Newton-Raphson method for parameter estimation defined in 3.5. The final assessment of $p^{(0)}_{k_{max}}$ is given in the form:

$$p_v = F_V(v; \hat{\mu}_v, \hat{\sigma}_v, \hat{\xi}_v)$$

where $V = -\log(-\log(P^{(0)}_{k_{max}}))$, $\hat{\mu}_v, \hat{\sigma}_v$, and $\hat{\xi}_v$ are the MLE's of the location, the scale, and the shape parameters of the distribution of V, respectively, and p_v is the CDF of V calculated at $v = -\log(-\log(p^{(0)}_{k_{max}}))$.

Finally, the p-value $p^{(0)}_{k_{max}}$ of the model with $t^{*(0)}_{k_{max}}$ is considered statistically significant if $p_v \leq 0.05$.

A simulation study will be discussed in the next section to demonstrate the described assessment approach and compare the result to our finding in chapter 2.

3.5 Numerical and Graphical Assessments

In this section, we regenerated all data sets of the first four cases from section 2.3 to carefully examine the modified component of the OQMDR algorithm. The simulation process is performed in R using the same mechanism that we described in chapter 2. That is, the two alleles' frequencies that are used to generate the factors' information are $p = q = 0.5$. Similarly, we generated the continuous response using one of the models described in equations 2.3 and 2.4, depending on the desired order of interaction. Then, we applied the modified approach on each of the 120 simulated data sets of cases $1-4$ described in sections 2.3.1, 2.3.2, 2.3.3, and 2.3.4 to evaluate the effectiveness of the suggested GEVD procedure for assessing the significance of the proposed models. Furthermore, a comparison between the permutation testings and the GEVD procedures, in terms of the statistical significance and calculation time, is carried out for all simulated data sets.

For each generated data set, the described approach in equation 3.5 is employed to estimate the null GEVD of the testing t-score of the final model, $T^{*(0)}_{k_{max}}$, using m permuted testing t-scores, where m is a relatively small number of permuted samples. Coding experience shows that a permuted sample of any size less than 30 might cause the analytical estimation process of the MLE's to diverge more frequently. In details, the m permuted t-scores $(t^{*(1)}_{k_{max}}, t^{*(2)}_{k_{max}}, \ldots, t^{*(m)}_{k_{max}})$ are utilized to estimate the location, the scale, and the shape parameters of the GEVD using our own written `R` program. The outputs are verified with the results obtained by applying the function `fitdist` from the library `fitdistrplus` [19] on the same permuted samples.

In addition to obtaining the MLE's, we established a graphical representation of the empirical and theoretical null distributions of $T^{*(0)}_{k_{max}}$. The graphical representation comprises four different plots: a histogram with empirical and theoretical densities' curves overlaid, empirical and theoretical cumulative probabilities against quantiles plot (CDF plot), a quantile-quantile (Q-Q) plot, and a probability-probability (P-P) plot. The plotted empirical and theoretical densities are obtained using the functions `density` from `R` based library, and `dgevd` from the library `EnvStats` [42], respectively. Similarly, the CDF curves are produced using the function `cdfcomp` from the library `fitdistrplus` [19], with the function `pgevd` from the library `EnvStats` [42]. Furthermore, the Q-Q plot, which plots the quantiles of the empirical distribution against the quantiles produced from the theoretical distribution [13], is schemed using the function `qqplot` from `R` based library, with the function `qgevd` from the library `EnvStats` [42]. Finally, the P-P plot, which compares the empirical CDF versus the theoretical CDF [13], is created by sketching the empirical cumulative probabilities against the probabilities from the theoretical CDF. Notice that the empirical cumulative probabilities in the CDF plots and the P-P plots are calculated using $(1 : m - 0.5)/m$, where m is the number of permuted statistics [19, 42]. Due to space limitations, the graphs are plotted only for one data set for each case and a distinct sample size (AB & $n = 500$, AB & $n = 1000$, ..., ABC & $n = 2000$). The graphical representations are produced for the same selected data sets with 1000 permuted t-scores to justify the selection of the GEVD to model the behavior of $T^{*(0)}_{k_{max}}$ for large numbers of permutations.

Analogous to $T^{*(0)}_{k_{max}}$, the null GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$ is approximated using m_1 permuted p-values $(p^{(1)}_{k_{max}}, p^{(2)}_{k_{max}}, \ldots, p^{(m_1)}_{k_{max}})$. The estimated parameters of the null distributions are reported for each of the transformed $P^{(0)}_{k_{max}}$ that corresponds to a certain $T^{*(0)}_{k_{max}}$ in various data sets corresponding to a given case. In addition, the four-plot schemes (see previous paragraph) are carried out for selected set of samples. The probability plots are initially produced from m_1 permuted p-values, whit m_1 being a relatively small number, such that each permuted p-value is originated from a distinct permuted sample of size m. Later, a larger number of permuted p-values is considered to generate the plots. Moreover, we fit a set of different distributions of $P^{(0)}_{k_{max}}$ or a transformation of $P^{(0)}_{k_{max}}$ to compare to the GEVD of the $-\log(-\log(P^{(0)}_{k_{max}})$, and the result is partially presented in the appendices. The approximated null distributions of $T^{*(0)}_{k_{max}}$ and $-\log(-\log(P^{(0)}_{k_{max}}))$ are utilized to determine the significance of the calculated t-score, $t^{*(0)}_{k_{max}}$. Finally, the results from this simulation study are compared to the findings in chapter 2 regarding calculation time and significance of suggested models. All simulations are done using `R` software [50] installed in a machine powered by an Intel Core i7-4500u CPU.

3.5.1 Case 1: True model = AB

Refer to section 2.3.1, the same ten generated data sets of each sample size (500, 1000, and 2000) are used again to evaluate the modification of the OQMDR algorithm. The summarized outputs are listed in tables 3.1, 3.2, and 3.3. Similar to the regular permutation testing, all proposed models show a statistical significance at $\alpha = 0.05$ regardless of the sample size. It can be seen by looking at the $p_{k_{max}}^{(0)}$ values in each table. In addition, the GEVD assessment of these $p_{k_{max}}^{(0)}$'s is carried out and the final theoretical p-values are listed under the p_v column. Once again, all $p_{k_{max}}^{(0)}$'s are considered significant at $\alpha = 0.05$ level of significance except for one case when $n = 2000$ (case 8, table 3.3), where the MLE approach fails to converge even after trying many different initial values.

Table 3.1: Case 1: True model = AB, $n = 500$, and $m = m_1 = 30$

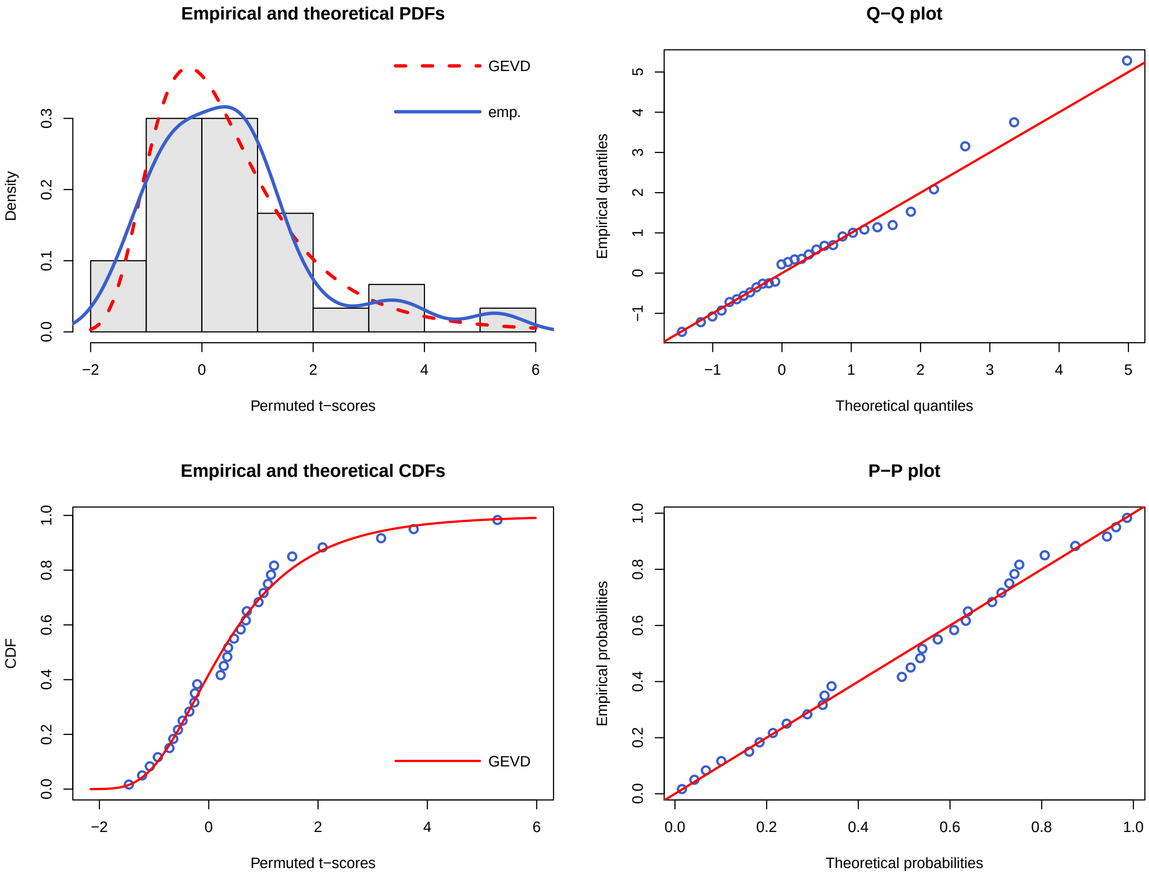
	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	11.7287	0.1945	1.0630	0.2844	0.0071	-0.1564	1.2699	0.2295	0.0239	29.6343	0.000	29.6763
2	10.0317	-0.1092	1.1286	-0.1420	0.0000	0.5375	0.9291	0.1437	0.0000	30.6164	0.000	30.3414
3	9.2992	-0.2410	0.9569	-0.3926	0.0000	-0.1275	1.2581	-0.0058	0.0000	28.6911	0.000	31.4487
4	10.1738	-0.0730	1.2086	-0.1575	0.0000	0.1610	1.1165	-0.0390	0.0000	30.0305	0.000	32.0555
5	8.8108	-0.0071	1.1363	-0.2591	0.0000	-0.1611	0.6310	0.209	0.0000	29.8299	0.002	30.6594
6	9.4734	0.1556	1.3150	-0.3104	0.0000	-0.1160	0.9250	0.2105	0.0000	28.8816	0.001	32.1799
7	9.3303	0.0202	1.3270	-0.0583	0.0001	0.5502	1.1931	0.1403	0.0000	28.2546	0.000	31.0413
8	11.8964	-0.0357	0.9912	-0.0356	0.0000	0.0684	1.0479	0.1989	0.0000	29.7952	0.001	29.8212
9	11.9762	-0.1403	0.9967	0.1082	0.0004	0.2656	1.1461	0.0801	0.0001	30.5986	0.000	29.6083
10	10.0830	-0.1037	1.2199	-0.0105	0.0002	-0.0284	1.0644	0.0063	0.0005	30.7604	0.002	31.1963

* Column headers are defined as follows:

- $t^{*(0)}_{k_{max}}$: The maximized testing t-score for the proposed model of the k^{th} order, calculated from the original data set.
- $\hat{\mu}_{t^{*(0)}_{k_{max}}}$: The MLE of the location parameter of the null GEVD of $T^{*(0)}_{k_{max}}$ observed at $t^{*(0)}_{k_{max}}$.
- $\hat{\sigma}_{t^{*(0)}_{k_{max}}}$: The MLE of the scale parameter of the null GEVD of $T^{*(0)}_{k_{max}}$ observed at $t^{*(0)}_{k_{max}}$.
- $\hat{\xi}_{t^{*(0)}_{k_{max}}}$: The MLE of the shape parameter of the null GEVD of $T^{*(0)}_{k_{max}}$ observed at $t^{*(0)}_{k_{max}}$.
- $p^{(0)}_{k_{max}}$: The theoretical p-value of $t^{*(0)}_{k_{max}}$ obtained from the null GEVD of $T^{*(0)}_{k_{max}}$.
- $\hat{\mu}_v$: The MLE of the location parameter of the null GEVD of V observed at v.
- $\hat{\sigma}_v$: The MLE of the scale parameter of the null GEVD of V observed at v.
- $\hat{\xi}_v$: The MLE of the shape parameter of the null GEVD of V observed at v.
- p_v: The theoretical p-value of $p^{(0)}_{k_{max}}$ obtained from the null GEVD of $-\log(-\log(P^{(0)}_{k_{max}}))$.
- Time: The required time to apply the algorithm on each data set in minutes.
- p-value: The simulated p-value from the regular permutation testing.

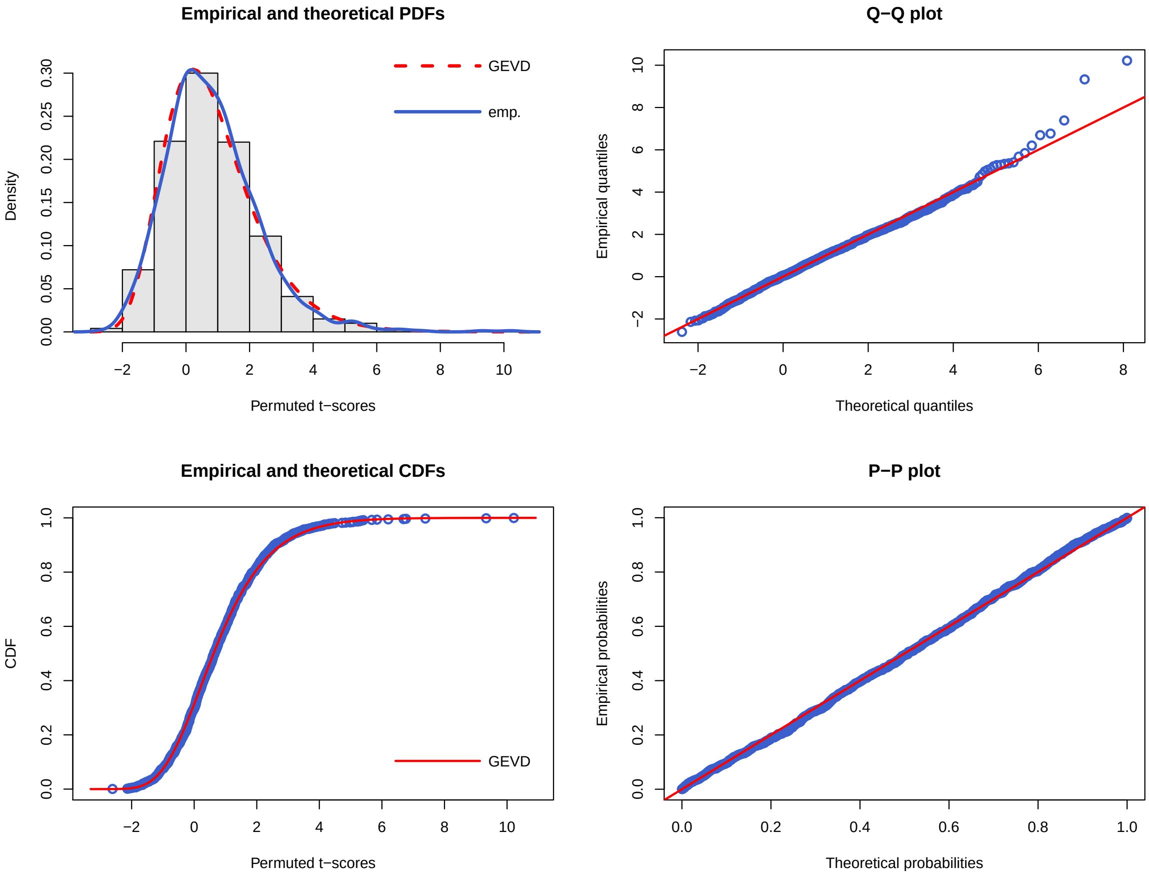
The graphical representation (figures 3.1 and 3.2) compares the empirical behavior of $T^{*(0)}_{k_{max}}$ with the theoretical GEVD when $n = 500$. Looking at figure 3.1, where only 30 permuted t-scores are used to approximate the distribution, all four plots show a decent fit between the empirical distribution and the theoretical GEVD of $T^{*(0)}_{k_{max}}$. In fact, the fit between the two PDFs may not look ideal when we look at the histogram with PDF curves; However, this lack of fit is likely due to the small number of permuted statistics used to establish the fit. Besides the slight deficiency between the two PDF curves, there is not a considerable migration from the fit that can be spotted from the other three plots. Likewise, figure 3.2 provides a better evidence to deem the GEVD as a plausible choice to explain the variation in $T^{*(0)}_{k_{max}}$. Further, the Q-Q plot shows a minor migration from the 45-degree reference line on the right tail of the distribution, which also agrees with the long right tail shown in the histogram. This slight departure from the fit on the right tail is due to observing a few large quantiles, as we can see from the CDF plot. Other than that, it seems there is no doubt that $T^{*(0)}_{k_{max}}$ behaves approximately per the GEVD, which can be inferred by looking at the P-P plot.

Figure 3.1: Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 for details. The histogram with the empirical PDFs shows how the empirical distribution of $T^{*(0)}_{k_{max}}$ looks like compared to the theoretical GEVD. The Q-Q plot reveals whether there is any shifting in location or scale between the two distributions, and detects outliers. The empirical and theoretical CDFs plot displays the nature of the empirical CDF compared to the theoretical CDF. The P-P plot shows whether there is a departure from the fitted GEVD or not [13].

Figure 3.2: Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores


Empirical and theoretical PDFs

Q–Q plot

Empirical and theoretical CDFs

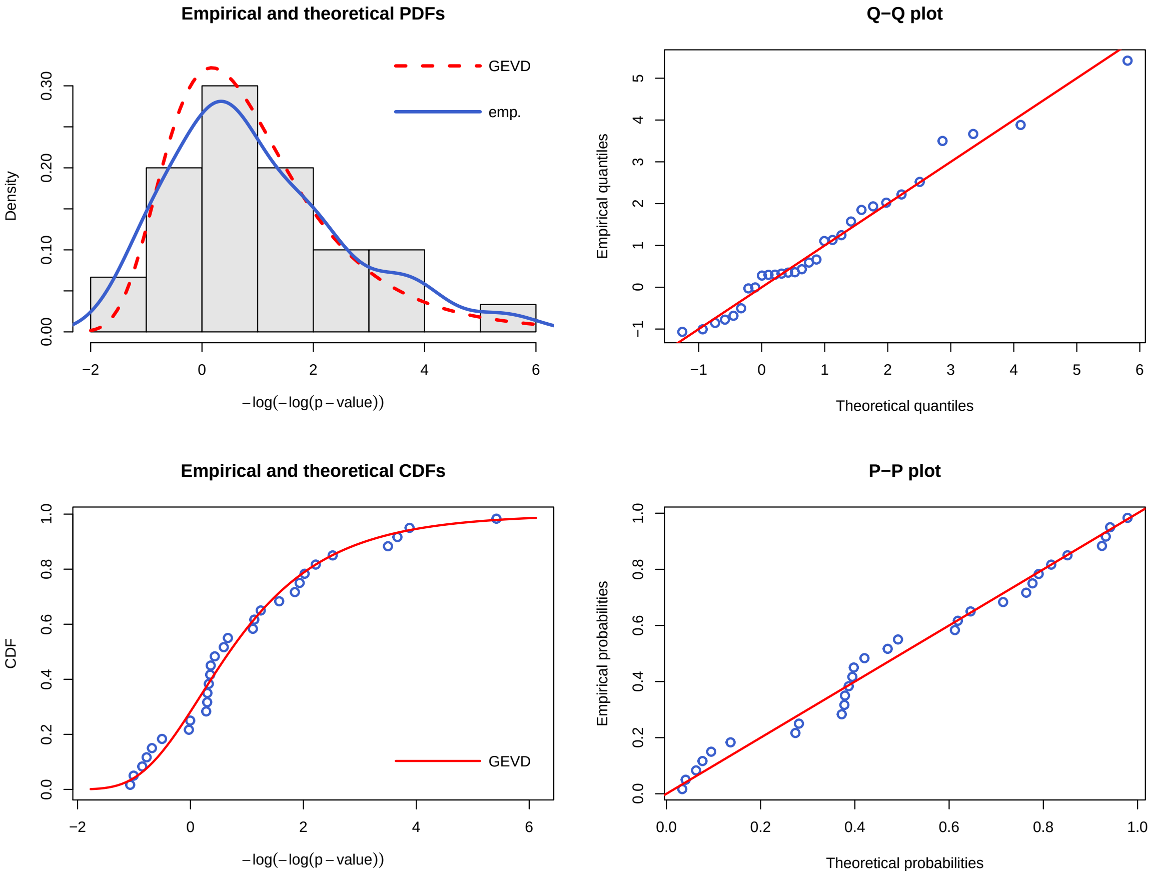
P–P plot

* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

On the other hand, the graphical comparison between the empirical distribution and the theoretical GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$ is presented on figure 3.3 with $m_1 = 30$ permuted p-values, and on figure 3.4 with $m_1 = 500$ permuted p-values, respectively. Originally, we considered multiple different transformations and/or distributions besides the GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$. It turns out that the GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$ does a decent job explaining the variation in the response compared to other considerations, especially for small number of permuted p-values ($m_1 = 30$). Keep in mind that the generated distribution of the p-values is somehow affected by the generated distribution of the permuted t-scores because each permuted p-value is originated from a particular permuted sample of t-scores; Hence, we might not have a smooth fit unless we simulate a large enough number of permuted t-scores in the first place ($m \geq 30$). That explains why we observe a distinguishing smooth fit in figure 3.4 compared to the fit in figure 3.3. For the same reason, we considered different numbers of permuted t-scores and p-values with many different scenarios to detect the permutation size that provides enough information to maintain a remarkable fit.

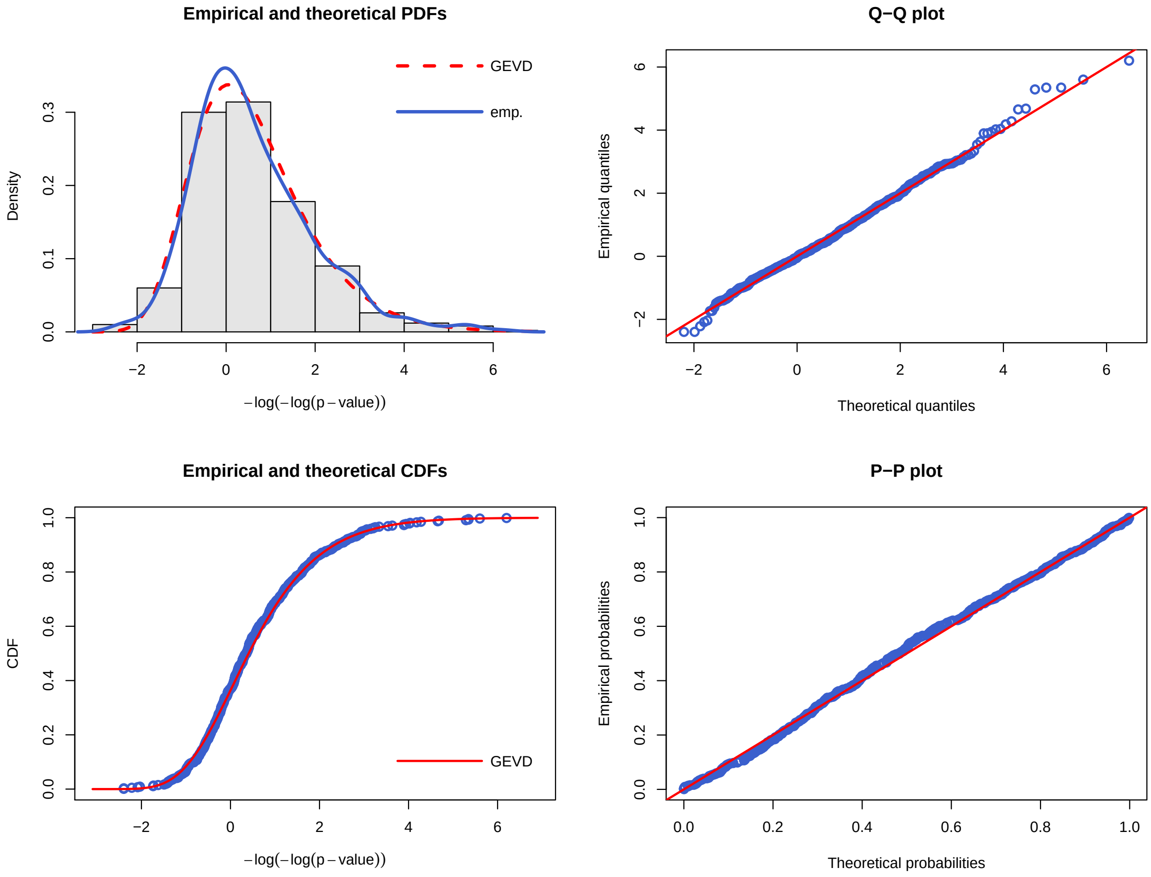
For a fact, we still need to choose a relatively small number of permutations to retain a reasonable calculation time. Therefore, the time of calculation and the precision of fit are the main two elements that we kept in mind when we decide which size is ideal. Subsequently, the approximation in figure 3.3 is done using $m_1 = 30$ permuted p-values, such that each p-value is originated from a set of $m = 30$ permuted t-scores, which is what we ended up choosing after many simulation attempts. To examine the behavior of the $-\log(-\log(P^{(0)}_{k_{max}}))$ for large number of permutations, we also tested a multiple different large numbers of permuted p-values that are generated from a fairly small number of t-scores. Among the ones we considered, which are (40, 200), (50, 200), (50, 500), (60, 400), and (200, 200) for the number of permuted t-scores (m) and the number of p-values (m_1), respectively. Simulation experience shows that a set of size $m \geq 50$ permuted t-scores is enough to produce a well behaved p-value (smoothly follow the fit). Accordingly, we considered using $m = 50$ with $m_1 = 500$, $m = 60$ with $m_1 = 400$, or $m = 200$ with $m_1 = 200$ to closely examine the nature of the $-\log(-\log(P^{(0)}_{k_{max}}))$.

Figure 3.3: Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.4: Case 1: True model = AB, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Afterward, the figures 3.3 and 3.4 clearly show that the GEVD is a legitimate choice to approximate the distribution of the transformed p-value. In fact, the selected transformation of the p-value better facilitates the GEVD compared to other transformations.

On the contrary, increasing the original sample size (the number of individuals in the data set) from 500 to 1000 or 2000 did not help improving the precision of the estimation process nor the quality of the fitting. Indeed, as we increase the sample size, the divergence problem of the MLE process becomes more frequent than when $n = 500$, which might seem counterintuitive. However, the ambiguity will be revealed if we recall that the testing t-score tends to be proportional to the square root of the sample size; thus, as n increases, outliers become more influential on the fitting process. Therefore, we encountered a divergence problem in about 10% of the cases when $n = 1000$, and about 30% of the cases for $n = 2000$. Anyhow, we overcome the divergences in the Newton-Raphson algorithm by suppressing the value of the estimated shape parameter, $\hat{\xi}_t$, from getting larger than 1/3 within each iteration until we reach a complete convergence. Notice that the permuted p-values are inversely related to the t-scores; therefore, we expected to encounter a more frequent divergence while approximating the distribution of the transformed p-values. After all, the suppression adjustment does help achieving the convergence in almost all problematic cases (16 out of 17 different data sets of sizes 1000 and 2000) except for one data set of size 2000 (see table 3.3).

Table 3.2: Case 1: True model= AB, and $n = 1000$, and $m = m_1 = 30$

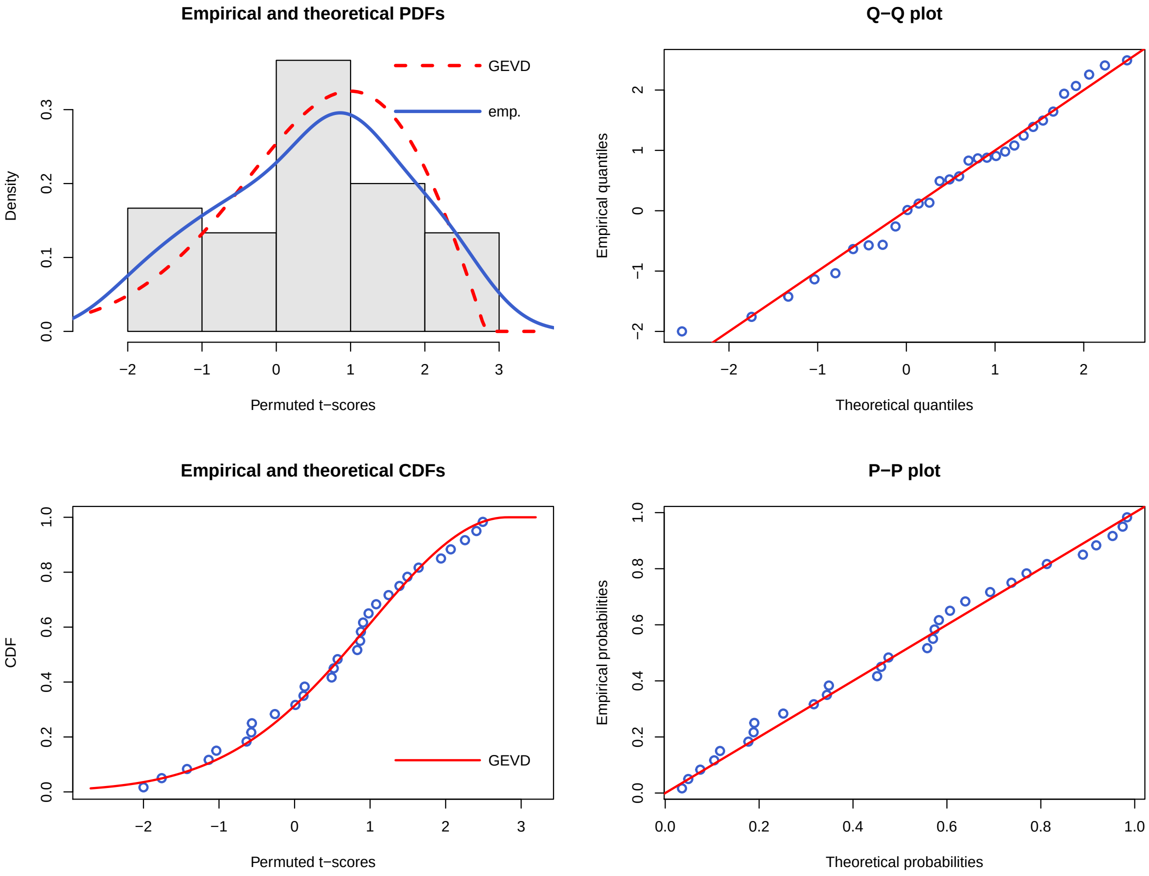
	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	16.2091	-0.0445	1.1873	-0.0985	0.0000	0.2032	1.3173	-0.0360	0.0000	40.6546	0.000	44.0068
2	14.9832	0.2131	1.1328	0.2649	0.0035	-0.1580	1.1380	-0.0532	0.0225	40.3236	0.000	43.4187
3	16.8702	0.2045	1.3272	-0.5083	0.0000	0.0750	1.1692	-0.1453	0.0000	40.3071	0.000	43.2199
4	16.1570	0.2231	1.5331	-0.3267	0.0000	0.1361	0.8681	0.0778	0.0000	43.5178	0.000	42.8729
5	15.7406	0.0244	1.2803	-0.2479	0.0000	-1.6937	1.8093	-0.1436	0.0000	38.8186	0.000	42.4825
6	14.3067	0.0222	0.8995	0.1408	0.0002	0.2123	0.9702	-0.1482	0.0004	38.7839	0.000	42.3103
7	15.7926	-0.1886	1.0923	-0.1232	0.0000	0.7177	1.2127	-0.3567	0.0000	38.8247	0.000	42.3003
8	14.7499	0.2934	1.3505	-0.2286	0.0000	-0.0187	1.2324	-0.0194	0.0000	38.7957	0.000	42.1112
9	15.1977	-0.0190	1.3157	-0.2970	0.0000	-0.2423	0.9008	0.0811	0.0000	40.2208	0.000	42.2298
10	13.6602	-0.2700	1.1455	-0.1818	0.0000	-0.3651	1.9043	-0.2396	0.0000	38.9530	0.000	42.0930

* Column headers are defined as in table 3.1.

Either way, even though increasing the sample size adds a little bit of complication to the estimation process, it does not worsen the the quality of fitting besides to the increased chance of having more influential extreme values. This can be seen by looking at figures 3.5 and 3.6 for the approximated distribution of the t-score, and 3.7 and 3.8 for the $-\log(-\log(P^{(0)}_{k_{max}}))$. Similarly, we have a reasonable fit when $n = 2000$ for both the t-score and the transformed p-value (see figures 3.9, 3.10, 3.11, and 3.12).

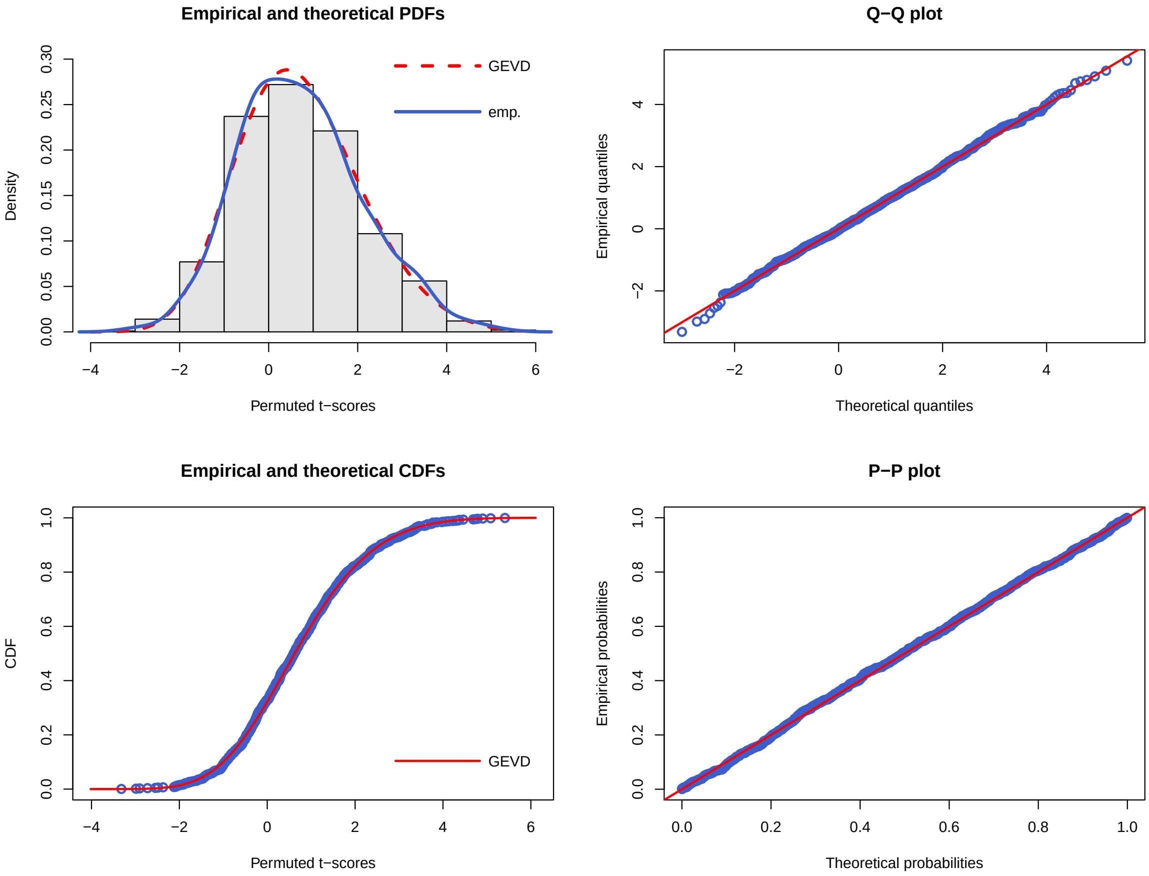
It's important to point out that, in some cases, we have some permuted p-values are practically equal to zero, which would make the log transformation undefined for some p-values. Accordingly, we add an infinitesimal quantity to the zero p-values before applying the transformation. The Q-Q plot in figure 3.12 shows three points at the very bottom end of the 45° reference line, where these points are originally zero p-values. These values do not influence or change the approximated distribution substantially because usually there is a tiny number of them, plus they are not too far in distance from other permuted p-values.

Figure 3.5: Case 1: True model = AB, n = 1000; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



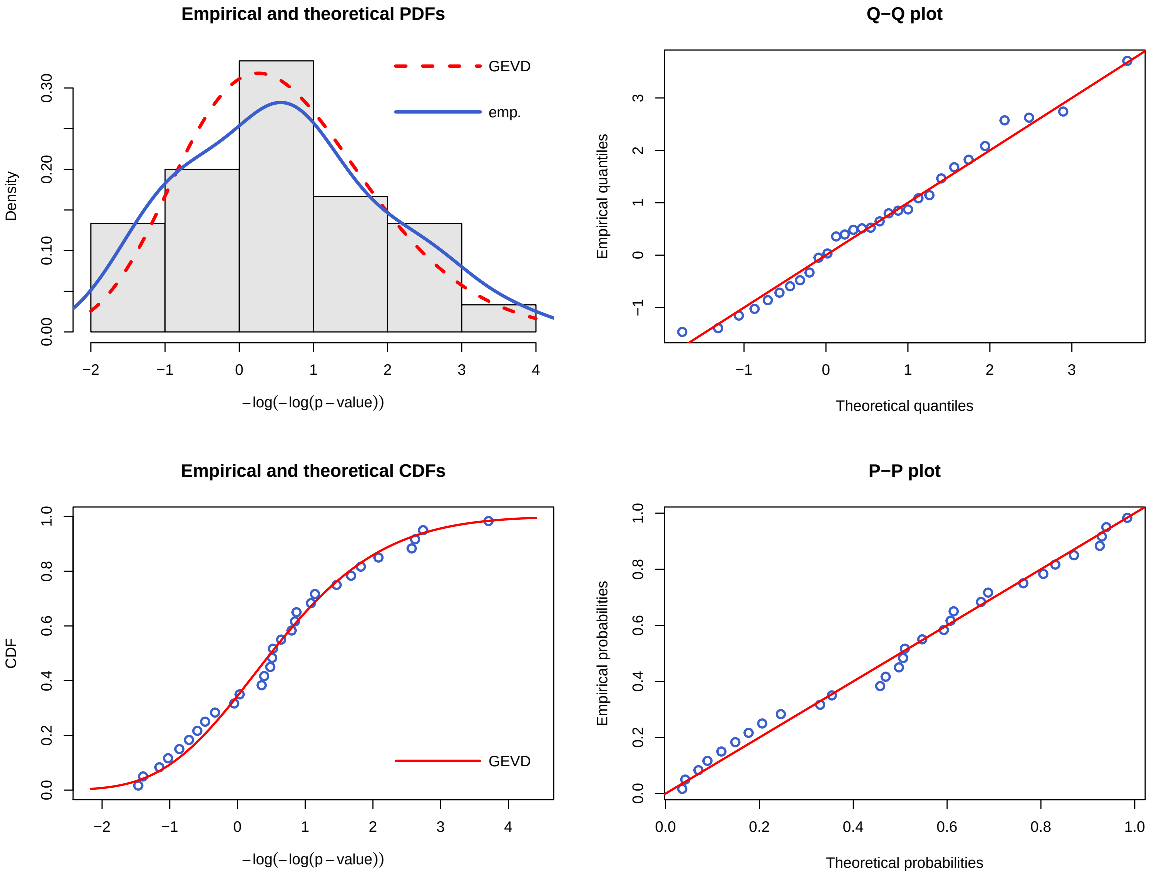
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.6: Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



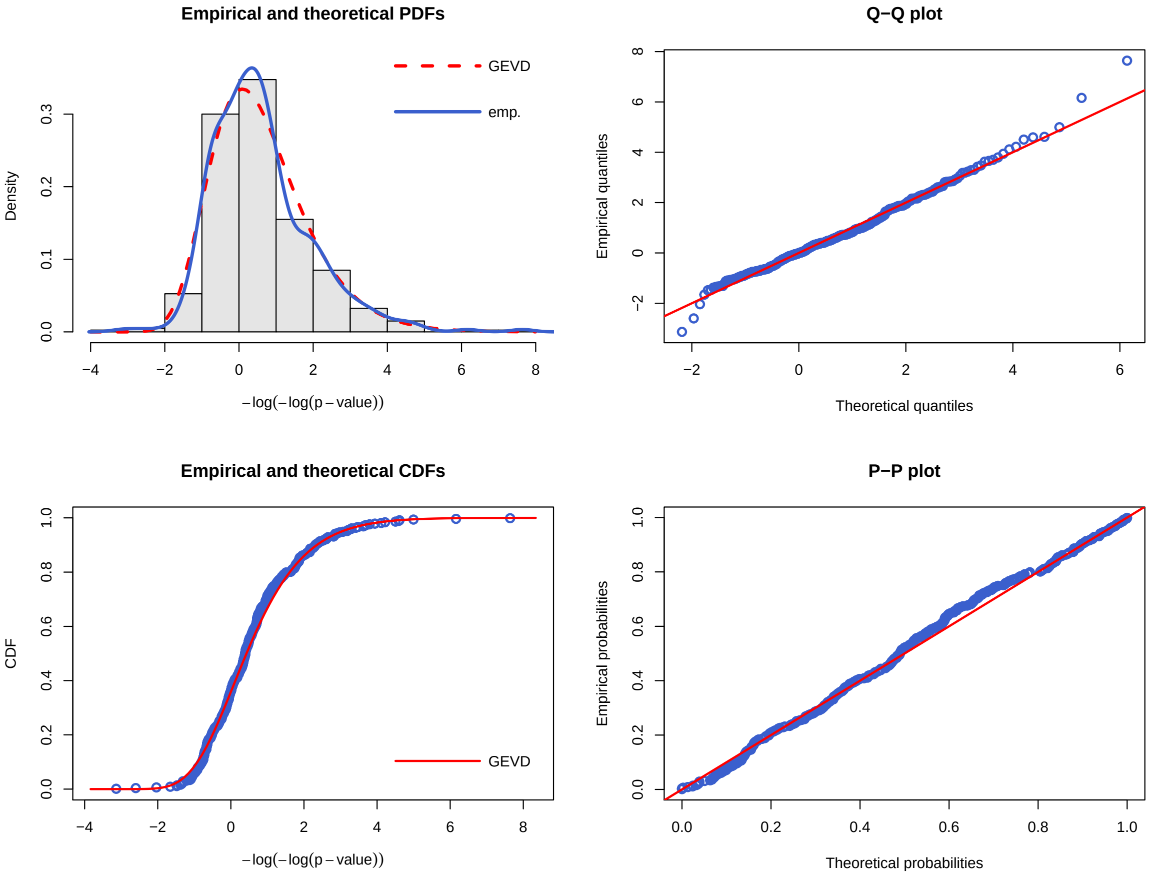
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.7: Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.8: Case 1: True model = AB, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 400 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Finally, and probably most importantly, even though the suggested GEVD approach successfully helps to evaluate the proposed models that agree with the underlying interactions used to generate the data, the new proposal does not benefit, compared to the ordinary permutations method, the calculation time aspect of the OQMRD, which is opposed to what we anticipated. This can be inferred by comparing the calculation times between the GEVD and the regular permutations from tables 3.1, 3.2, and 3.3. Yet, after digging deeper into what caused the new procedure to fail dominating the original assessment technique, we come out with a few elements that could influence the calculation time aspect of the GEVD procedure.

Besides coding complication of the new approach compared to the permutation procedure, we think that the leading cause of raising the calculation time is the evaluation of the p-value of the examined model. This component has been added to the algorithm, after introducing the GEVD approach, to make a rigorous decision about the reliability of the chosen model. In fact, the assessment of the p-value portion of the GEVD proposal absorbs an enormous amount of time compared to the test score evaluation, i.e., obtaining the p-value itself. Although it's possible, the regular permutation procedure does not validate the p-value of the suggested model because it would consume a tremendous amount of time (re-permute the 1000 permutations many times to obtain the null distribution of the p-value).

On the other hand, the GEVD approach can provide a more accurate p-value than the permuted p-value. That is, the permuted p-value can be reported up to three decimal places only; whereas, the new approach can provide a p-value as small as $2.225074E - 308$, which is the machine epsilon in `R`, yet no additional time is needed. However, in our simulation, we rounded all outputs to four decimal places for the sake of space limitation. Once again, to obtain a more exact permuted p-value, we need to permute the original data set beyond 1000 times, which in turn would exceedingly increase the computation burden.

Another aspect that influences the computation time is the selection of the GEVD over uniform(0,1) distribution to evaluate the p-value, which is inspired by Hua et al., 2010 [30]. That is, if we assume that $P^{(0)}_{k_{max}}$ follows a continuous uniform(0,1) distribution, then we wouldn't need to estimate the null GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$, which ingests about 95% of the calculation time. This assumption seems to be reasonable to some extent, particularly when the number of the permuted p-values is large enough ($m_1 > 400$). With this intention, we graphically examined the behavior of $P^{(0)}_{k_{max}}$ with respect to the uniform(0,1) distribution, and the results is briefly presented in the appendix. From the output presented in this chapter and in the appendix, we certainly can presume that the GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$ surpasses other considered distributions when the number of the permuted p-values is relatively small ($m_1 = 30$). However, the performance of the uniform(0,1) distribution, contrary to the other fitted distributions, enhanced substantially with larger number of permuted samples. In addition, if we utilize the uniform(0,1) distribution to evaluate the observed value of $P^{(0)}_{k_{max}}$, then the observed value of $P^{(0)}_{k_{max}}$ numerically matches its p-value in up to more than ten decimal places, which agrees with the Probability Integral Transformation principle of a standard uniform random variable [10]. In short, employing a uniform(0,1) distribution seems feasible as it helps reducing the computation time; however, it has to be done with caution, specifically for small number of permuted samples.

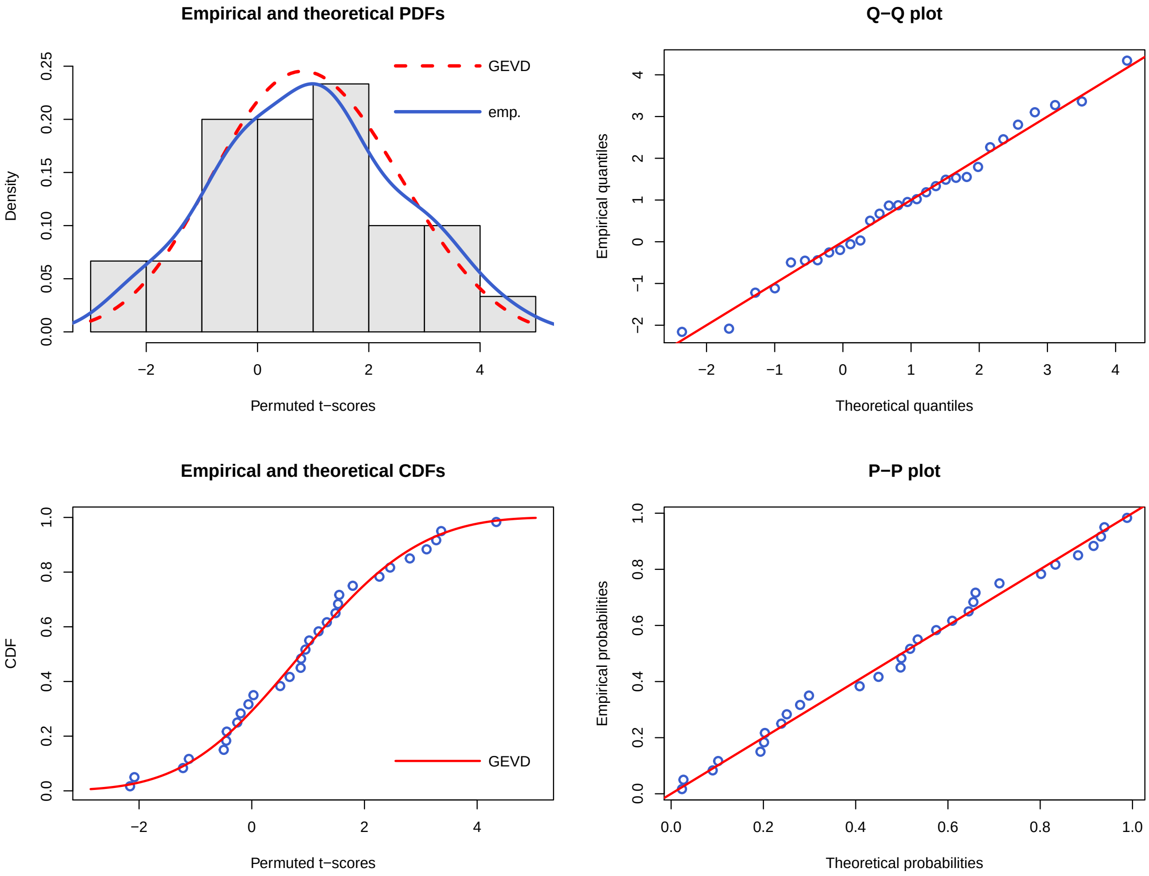
Under those aforementioned circumstances, we think that the suggested GEVD assessment is still predominating the regular permutation testing approach in terms of time and precision. However, further investigation might lead to a more efficient approach to evaluating the selected interactions.

Table 3.3: Case 1: True model= AB, and $n = 2000$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	20.8692	-0.4323	1.2426	-0.1523	0.0000	0.1250	0.9952	-0.0963	0.0000	61.8658	0.000	59.3812
2	22.1992	0.4960	1.4690	-0.5333	0.0000	-0.1322	1.2277	0.2275	0.0000	61.3719	0.000	59.7195
3	23.0960	0.1785	1.0632	-0.1225	0.0000	0.4089	1.2740	-0.2172	0.0000	60.0602	0.000	59.8337
4	22.2190	0.5902	1.6001	-0.4643	0.0000	-0.4802	1.2874	-0.0896	0.0000	59.4864	0.000	60.8675
5	23.7884	0.3335	1.5638	-0.2755	0.0000	-0.3527	1.0300	0.1059	0.0000	59.5922	0.000	59.7674
6	23.3298	-0.0323	1.0901	-0.0872	0.0000	-0.0283	1.3411	-0.2265	0.0000	59.7866	0.000	59.8626
7	22.3034	0.1401	1.1893	-0.2912	0.0000	0.0336	1.2075	0.0470	0.0000	59.6857	0.000	59.6379
8	22.9130	0.5555	1.3248	-0.3482	0.0000	NA	NA	NA	NA	59.2123	0.000	59.7593
9	22.7459	-0.0507	1.3052	-0.1459	0.0000	-0.0840	1.2365	-0.3177	0.0000	59.4446	0.000	59.8364
10	22.0691	-0.0590	1.2525	-0.1865	0.0000	-0.1803	0.9519	0.1870	0.0000	59.3913	0.000	59.9434

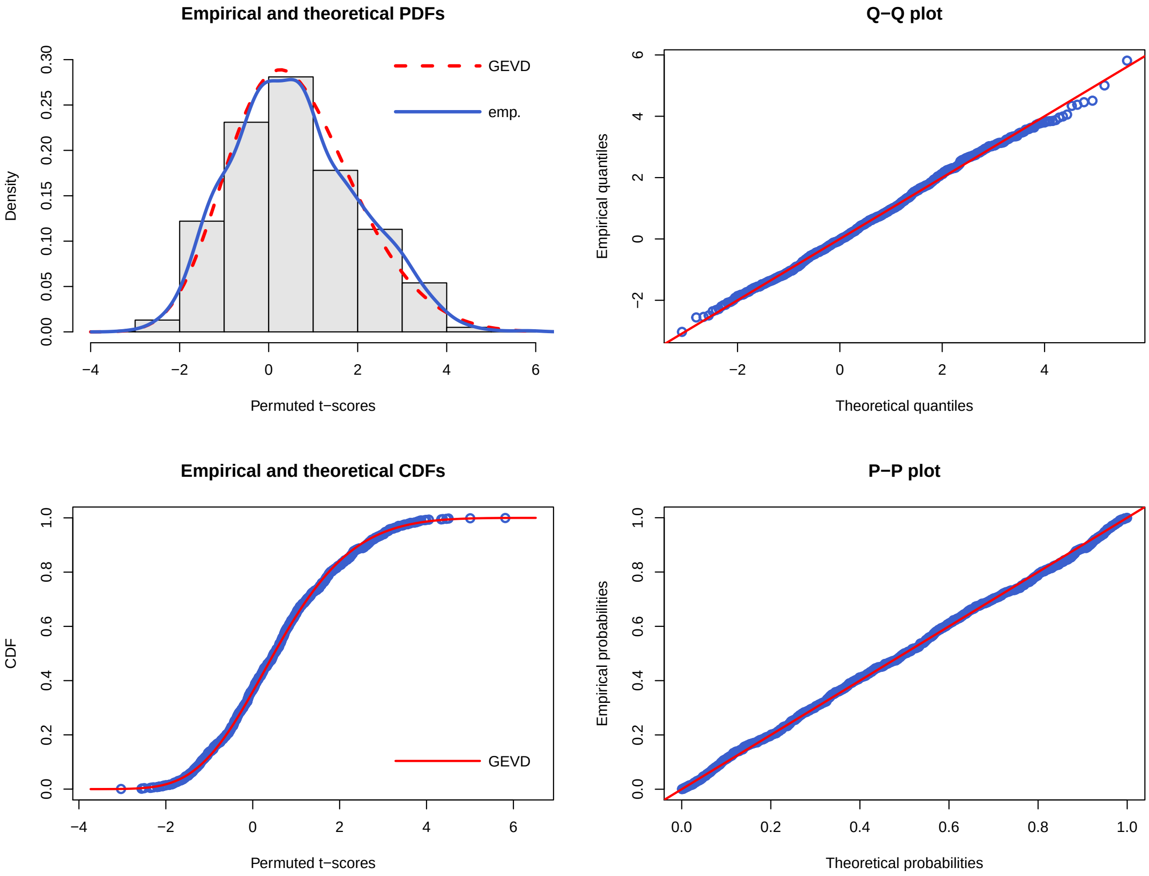
* Column headers are defined as in table 3.1.

Figure 3.9: Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



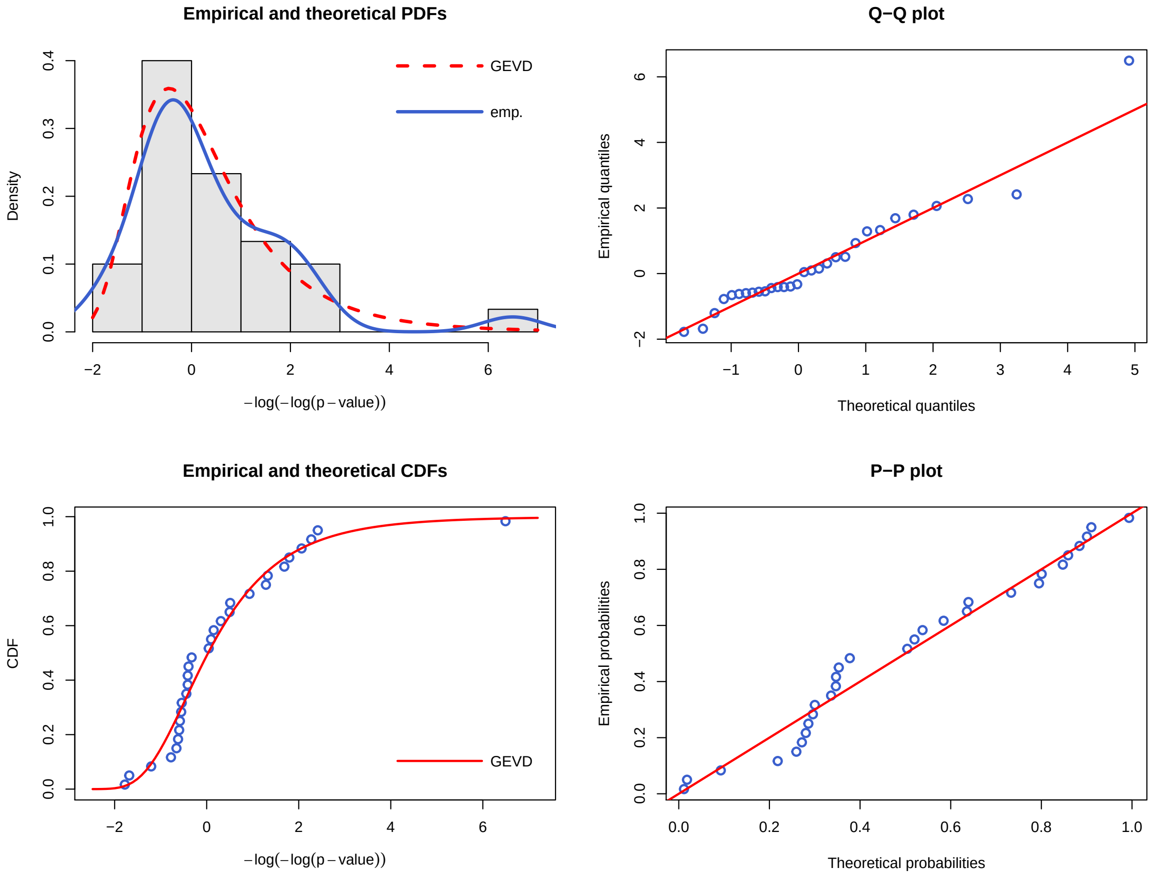
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.10: Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.11: Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values


Empirical and theoretical PDFs

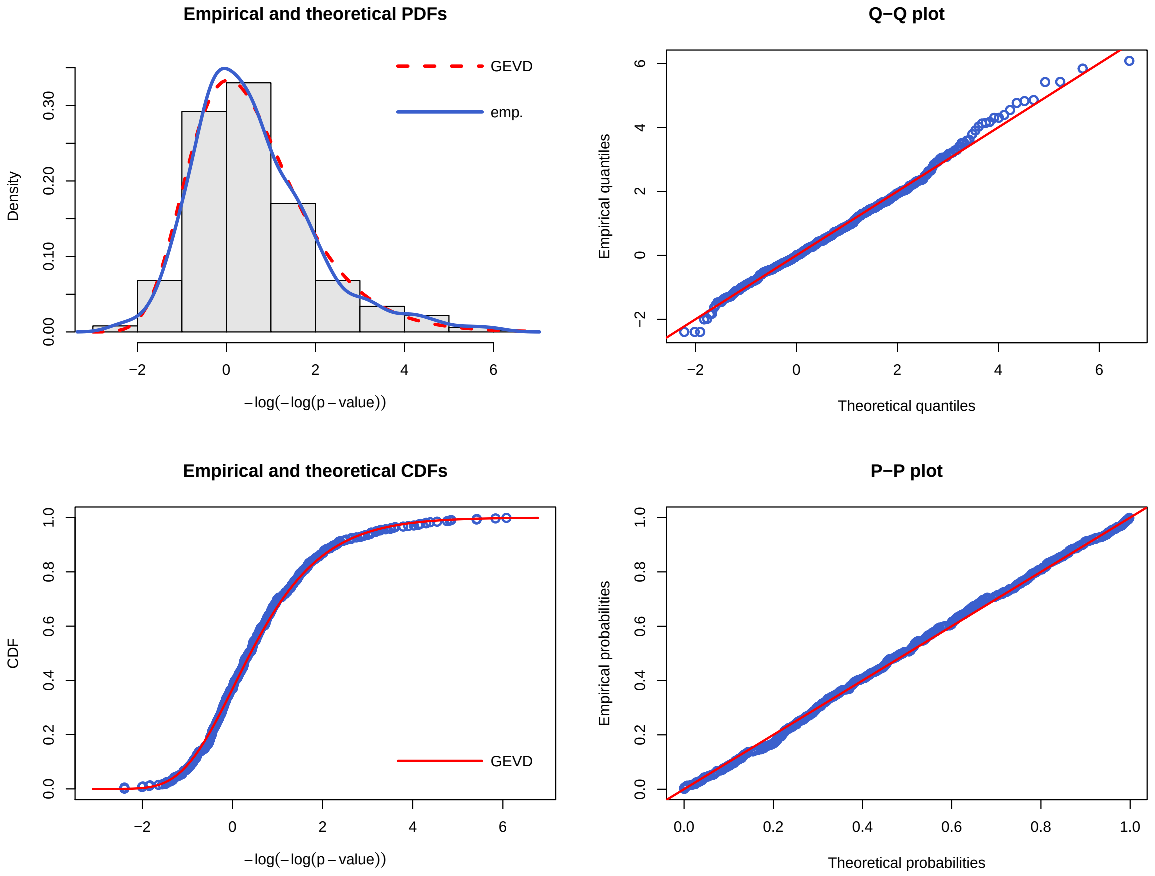
Q–Q plot

Empirical and theoretical CDFs

P–P plot

* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.12: Case 1: True model = AB, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

3.5.2 Case 2: True model = ABD

Once again, we regenerated the data sets that we used in section 2.3.2, where the underlying interaction is ABD, to evaluate the modification of the OQMDR algorithm. We listed all results in tables 3.4, 3.5, and 3.6.

The results in table 3.4 show that $p^{(0)}_{k_{max}}$ of the suggested model from the ninth data set is not significant ($p_v = 0.0736 > 0.05$). Whereas, the same interaction is considered significant when evaluated using the regular permutation testing. While this issue could occur more often with small samples, yet the real issue is not the model evaluation procedure itself. It's, in fact, the selected risk pattern from this data set does not coincide with the true risk pattern used to generate the data (figure 2.5). Knowing that the proposed risk pattern is chosen from both algorithms, QMDR and OQMDR, which means that the issue is from the data generation in the first place. Recall that a 3rd-degree interaction has 27 different allele combinations; hence, we might end up with very few observations in some combinations, which would affect the risk status of individuals in these combinations. Accordingly, the proposed risk pattern could be misleading under such circumstances. As a matter of fact, the GEVD approach does not recognize this risk pattern as a valid risk pattern, while the conventional approach does. This could be considered as a strength favoring the GEVD approach; however, we would need to do more investigation for confirmation.

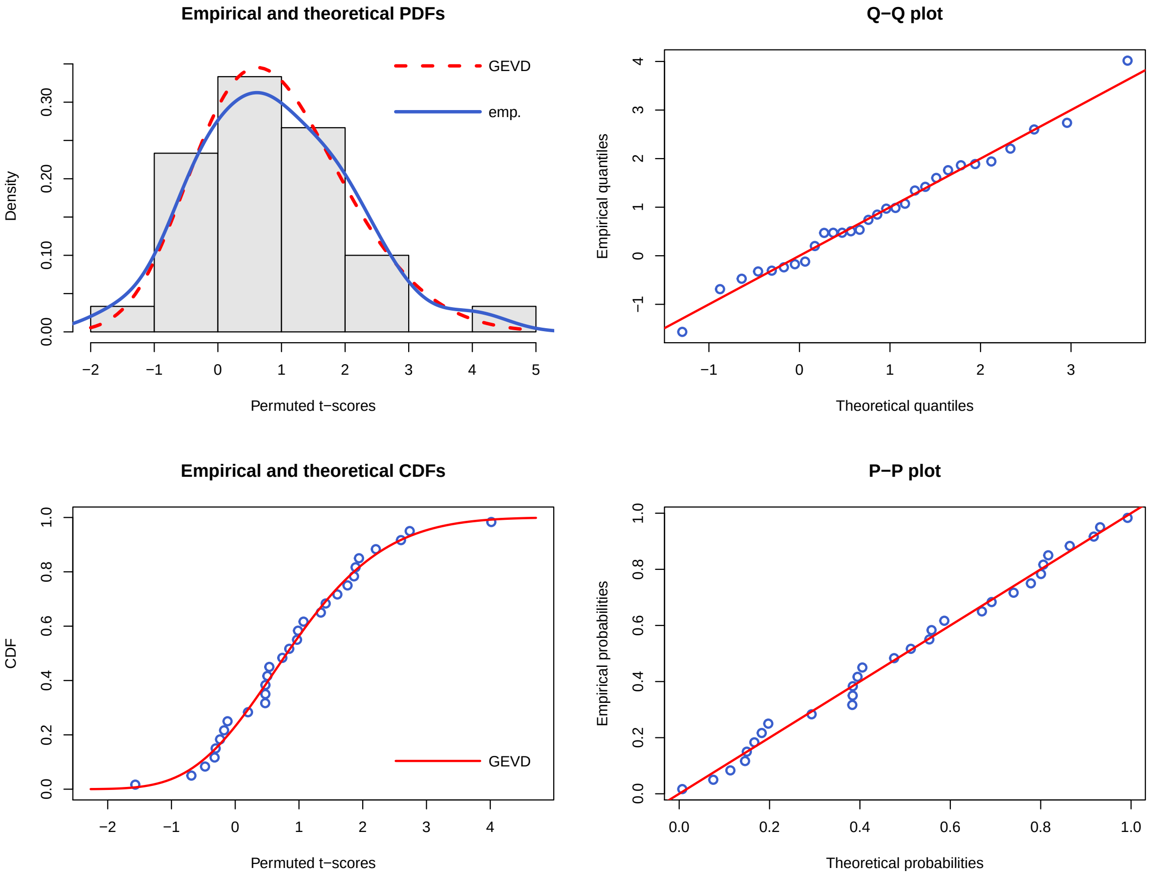
Table 3.4: Case 2: True model= ABD, and $n = 500$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	10.5026	0.4516	1.1913	-0.1902	0.0000	-0.1173	0.9337	-0.0489	0.0000	29.7165	0.000	29.4977
2	14.0163	0.4222	1.3083	0.0014	0.0000	-0.1447	0.9801	0.0328	0.0000	28.8601	0.001	29.4172
3	12.8563	0.2675	1.1150	-0.1480	0.0000	0.0555	1.0197	0.0295	0.0000	28.7999	0.001	29.4198
4	11.3547	0.2179	1.0836	0.0603	0.0003	-0.1761	0.8652	0.0385	0.0000	28.8446	0.000	29.3971
5	9.8944	-0.3290	1.0816	0.1321	0.0022	-0.3280	0.9294	0.1096	0.0030	28.7975	0.002	29.3120
6	10.1203	0.2842	0.8315	-0.0255	0.0000	-0.2759	1.4626	0.0160	0.0057	28.8144	0.001	29.4565
7	12.3105	0.7376	1.3469	-0.5025	0.0000	0.1243	0.9594	0.0521	0.0000	28.7748	0.001	30.4116
8	8.1019	0.5767	1.6743	-0.6075	0.0000	0.1750	1.6290	-0.0849	0.0000	28.8483	0.001	29.8514
9	8.2252	-0.4021	1.1873	0.2409	0.0149	0.1759	1.6697	-0.0146	0.0736	29.6636	0.005	29.4499
10	12.2139	0.4257	1.0817	-0.1678	0.0000	0.4180	1.2153	-0.3737	0.0000	30.1630	0.000	29.5340

* Column headers are defined as in table 3.1.

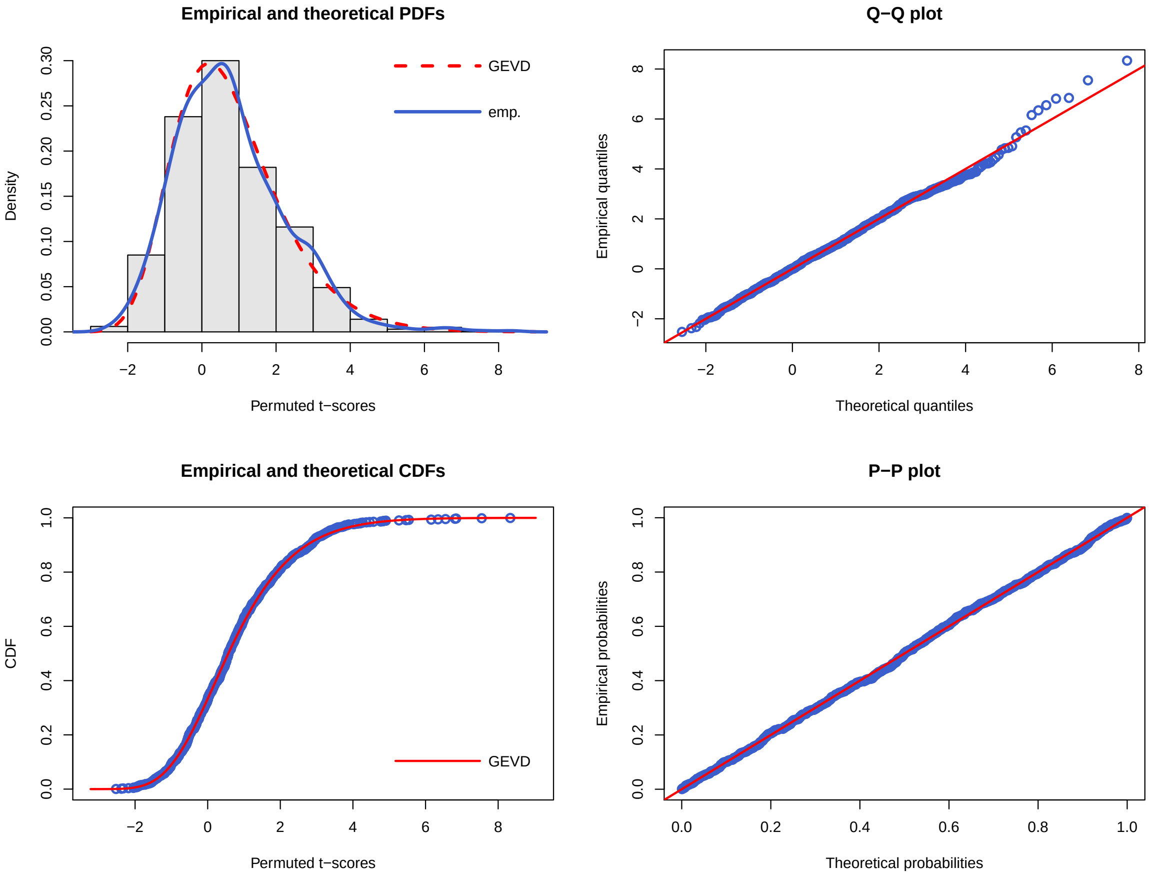
On the other hand, similar to case 1, all graphs show that the GEVD is nicely approximating the distributions of both $T^{*(0)}_{k_{max}}$ and $-\log(-\log(P^{(0)}_{k_{max}}))$. Once again, changing the sample size doesn't have any noticeable influence on the quality of fitting. Similarly, the order of the examined interaction does not affect the approximation process, which can be inferred by comparing the output of this case with case 1 results.

Figure 3.13: Case 2: True model = ABD, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



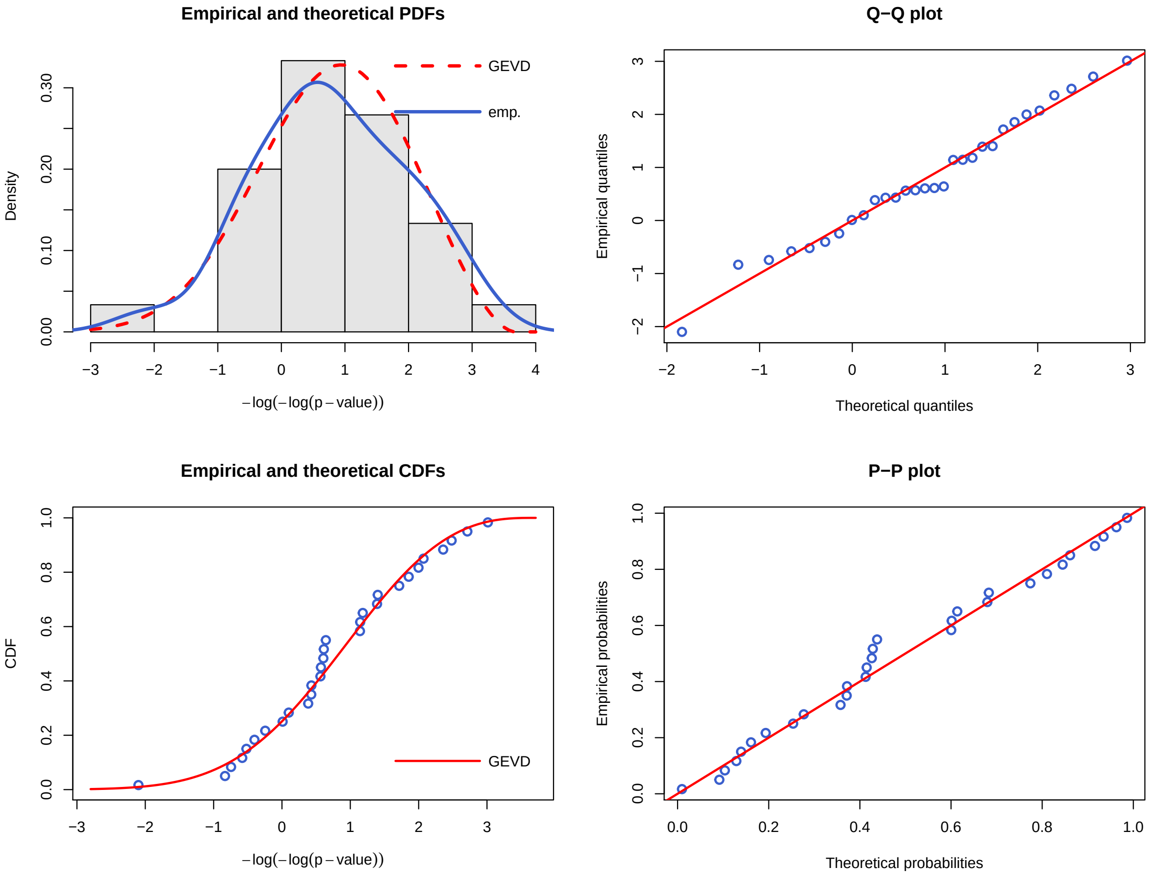
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.14: Case 2: True model = ABD, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

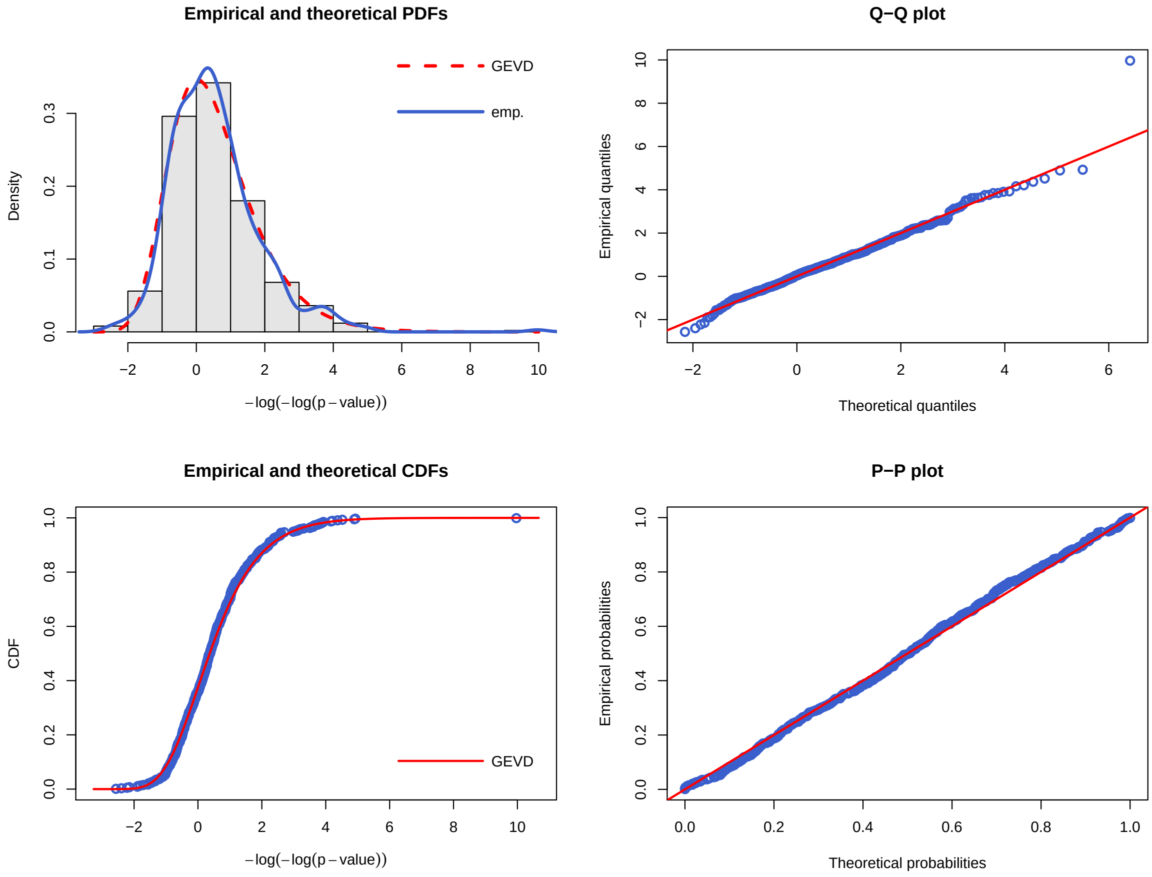
Figure 3.15: Case 2: True model = ABD, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



---

[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.16: Case 2: True model $= ABD$, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values



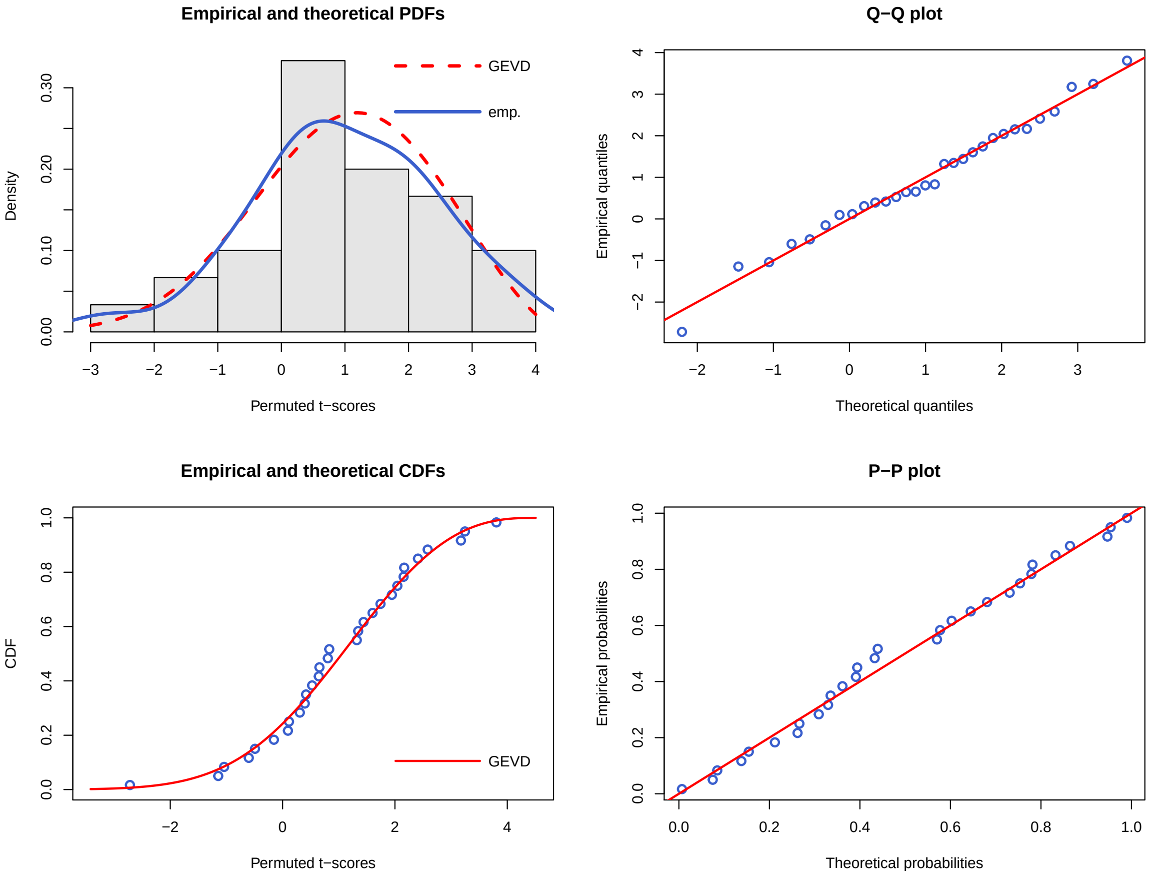
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Table 3.5: Case 2: True model$= ABD$, and $n = 1000$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	15.1322	-0.0821	1.2545	-0.2901	0.0000	-0.0012	0.9089	0.1936	0.0000	39.6350	0.000	39.3132
2	17.3736	-0.1201	0.9489	-0.0843	0.0000	0.2674	0.9975	0.1897	0.0000	39.3425	0.000	39.1848
3	15.2270	-0.0243	1.3561	-0.3010	0.0000	-0.1651	1.3763	-0.0733	0.0000	39.9399	0.000	40.3263
4	16.6369	0.1515	1.3768	-0.4173	0.0000	0.1778	1.1726	-0.1029	0.0000	42.3773	0.000	39.1652
5	16.5188	0.0180	1.3193	-0.1383	0.0000	-0.2254	0.8843	0.2302	0.0000	41.7511	0.000	39.1250
6	14.8783	0.6393	1.2252	-0.3503	0.0000	-0.2497	0.8293	0.2476	0.0000	40.7742	0.000	39.1751
7	16.3628	0.5519	1.4818	-0.3753	0.0000	0.2564	1.0828	0.0534	0.0000	42.1615	0.000	39.0741
8	16.8883	0.0523	1.2285	0.0091	0.0000	0.1017	1.2990	-0.1449	0.0025	41.1971	0.000	39.1690
9	15.0487	0.0027	1.1268	-0.2555	0.0000	-0.2092	0.8168	0.1044	0.0000	40.5250	0.000	39.1192
10	16.6091	0.1178	1.2885	-0.3730	0.0000	-0.0452	1.2771	0.0162	0.0000	41.9034	0.000	39.2857

* Column headers are defined as in table 3.1.

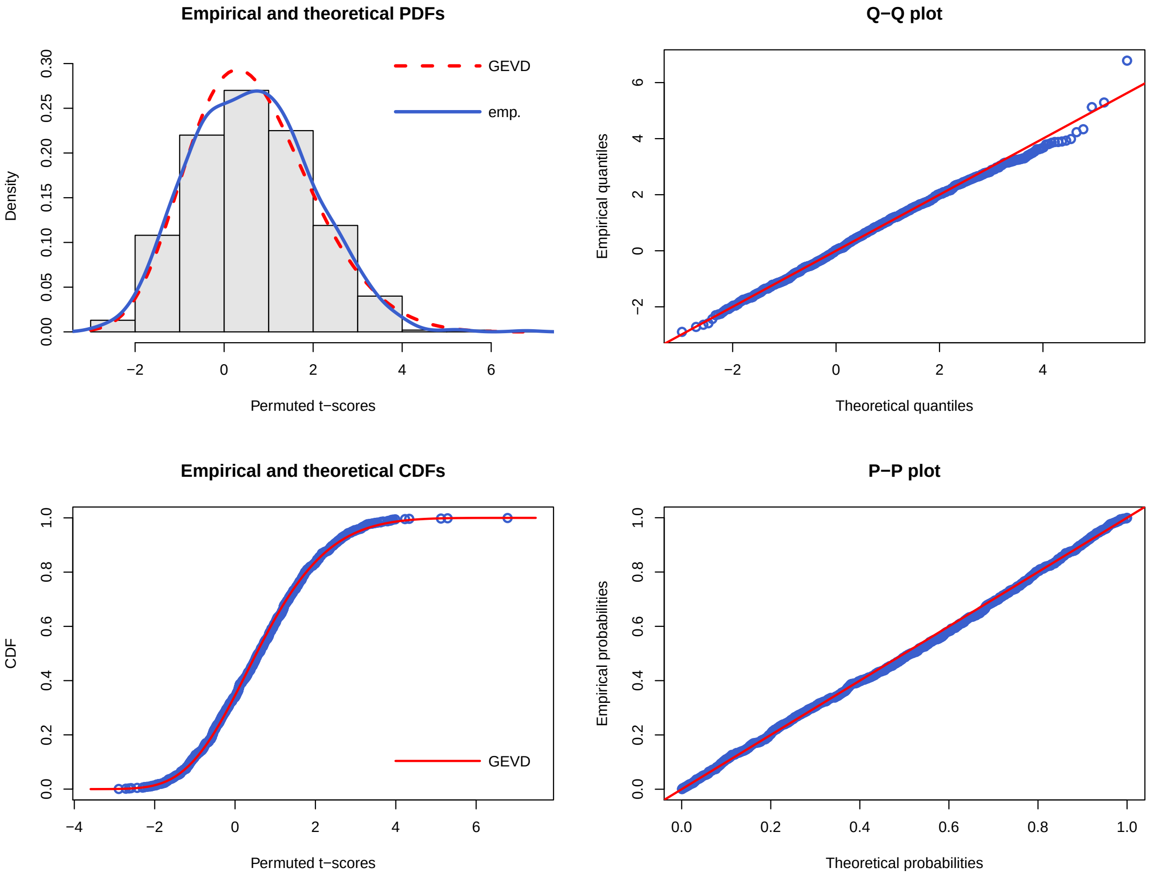
Figure 3.17: Case 2: True model = ABD, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



---

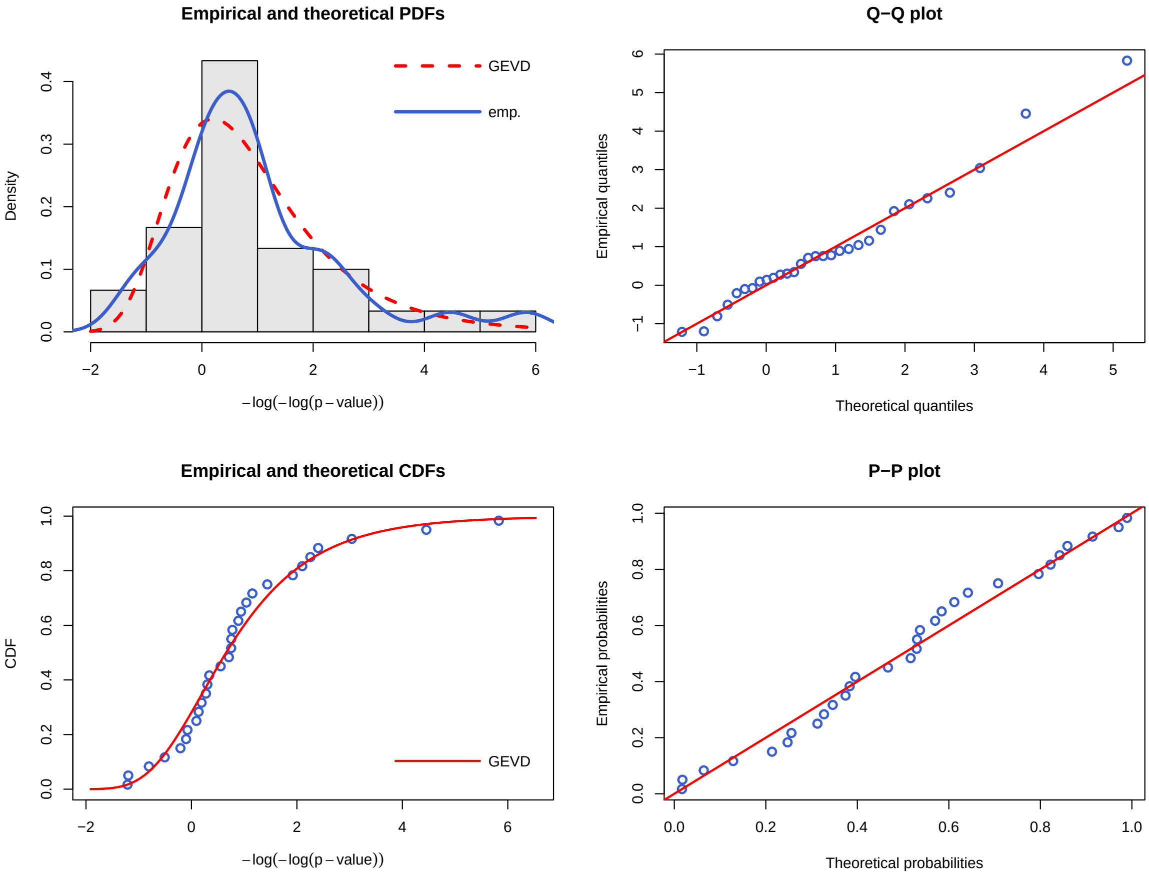
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.18: Case 2: True model = ABD, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



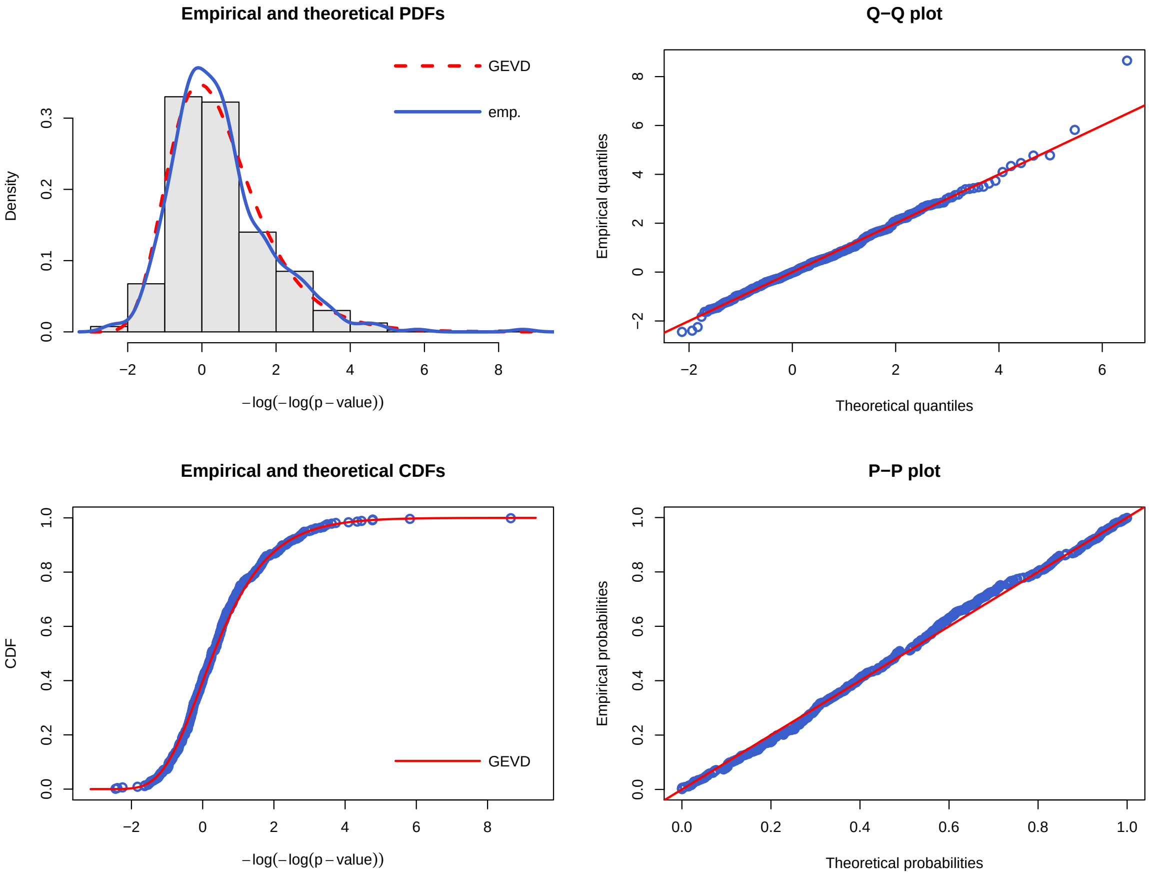
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.19: Case 2: True model = ABD, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.20: Case 2: True model = ABD, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values



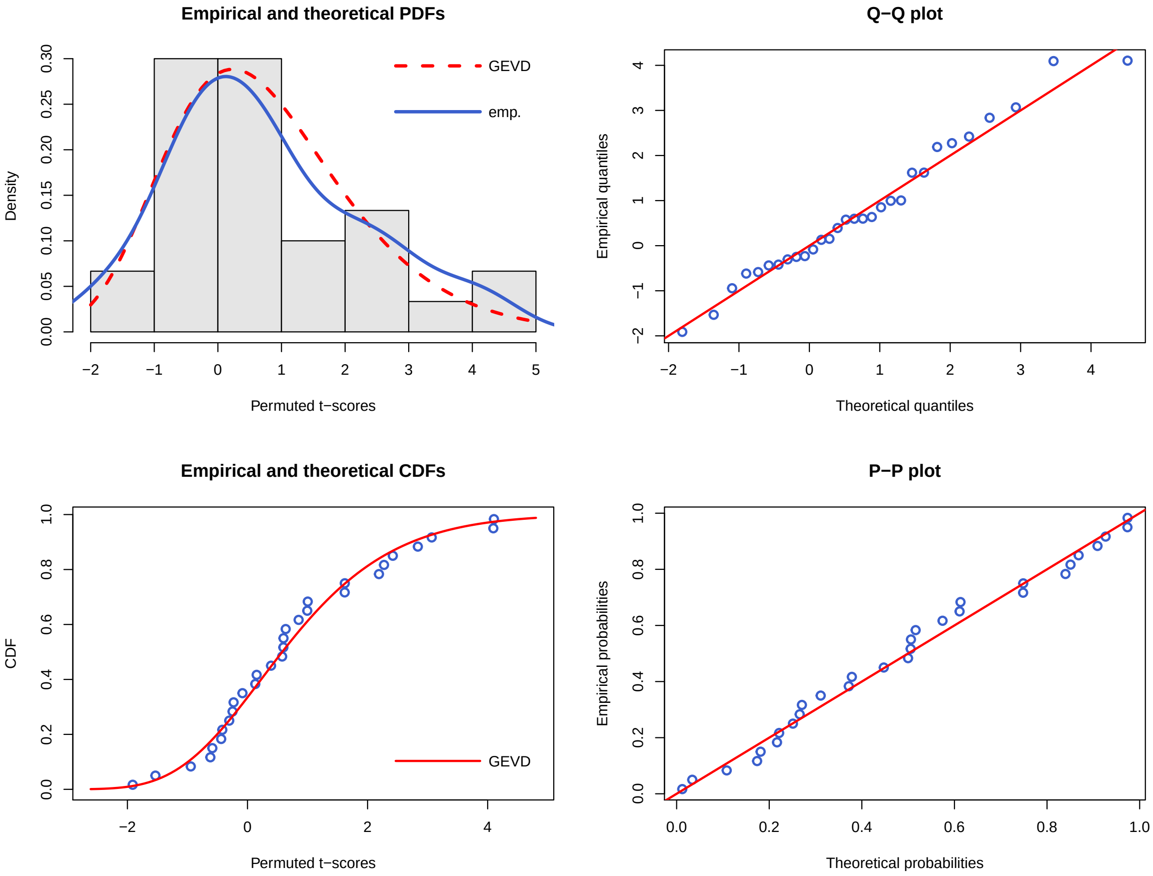
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Table 3.6: Case 2: True model= ABD, and $n = 2000$, and $m = m_1 = 30$

	GEVD procedure									Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	21.6356	0.3843	1.2571	-0.2417	0.0000	0.1747	1.5326	-0.1341	0.0000	60.6626	0.000	61.0031
2	22.2545	0.3189	1.3060	-0.2431	0.0000	-0.0167	0.9431	0.3275	0.0000	60.6943	0.000	59.9010
3	22.5347	0.0560	1.2461	-0.2748	0.0000	0.3476	0.9832	0.3668	0.0000	59.9240	0.000	60.6521
4	23.6509	0.1157	1.2808	-0.0871	0.0000	-0.0634	1.2076	-0.0828	0.0000	60.5431	0.000	59.7909
5	25.3687	1.0186	1.4955	-0.4187	0.0000	-0.2790	0.9754	0.0860	0.0000	60.7385	0.000	59.6732
6	24.0497	0.9217	1.2136	-0.4777	0.0000	-0.0663	0.9990	0.0078	0.0000	60.5031	0.000	59.8435
7	21.6227	0.2589	1.0186	-0.0087	0.0000	-0.3698	0.8484	-0.0006	0.0000	60.3570	0.000	60.0786
8	21.9942	-0.1910	1.1896	0.0245	0.0000	-0.0330	1.0626	0.1146	0.0000	61.4807	0.000	60.3793
9	22.2720	0.2202	1.1159	-0.3454	0.0000	0.6094	1.2382	-0.0615	0.0000	60.3046	0.000	59.7811
10	21.9574	0.2130	1.4370	-0.1831	0.0000	-0.0967	0.9100	0.1297	0.0000	60.7451	0.000	59.7573

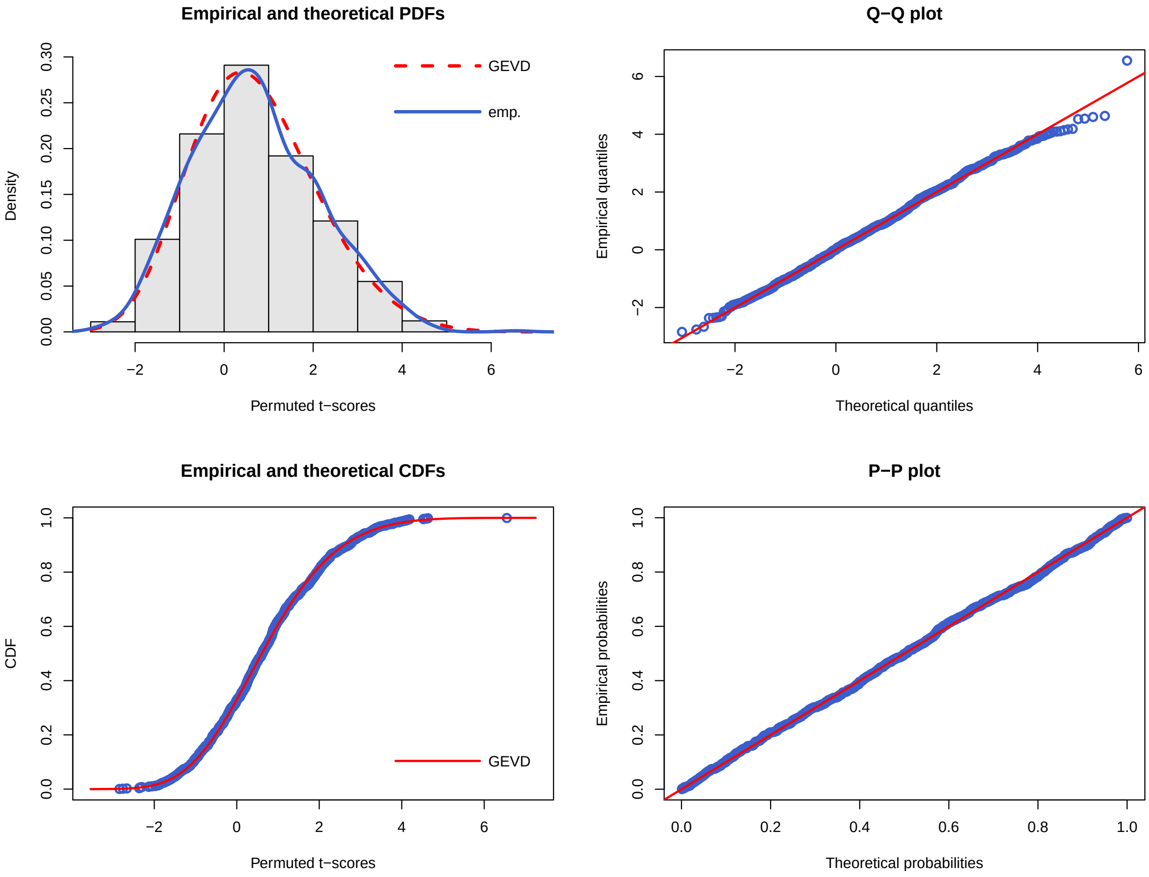
* Column headers are defined as in table 3.1.

Figure 3.21: Case 2: True model = ABD, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



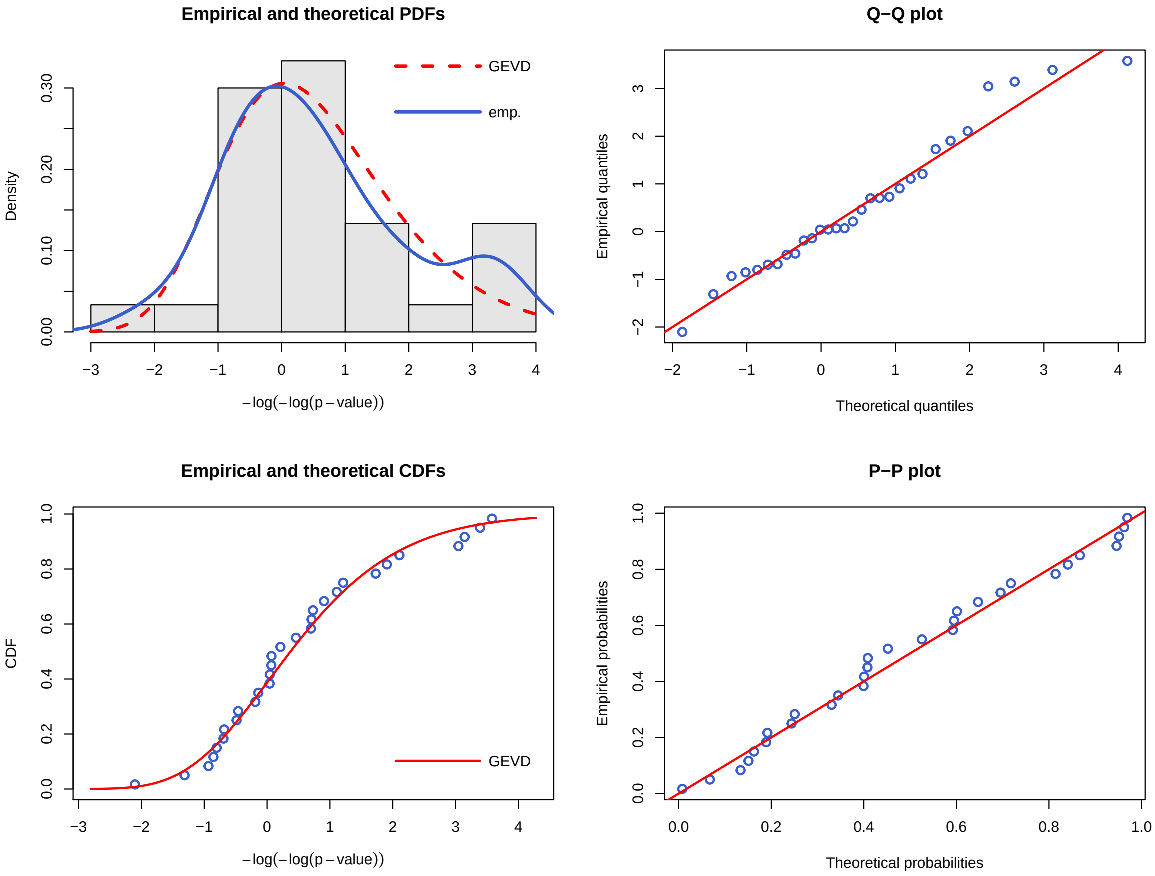
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.22: Case 2: True model = ABD, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

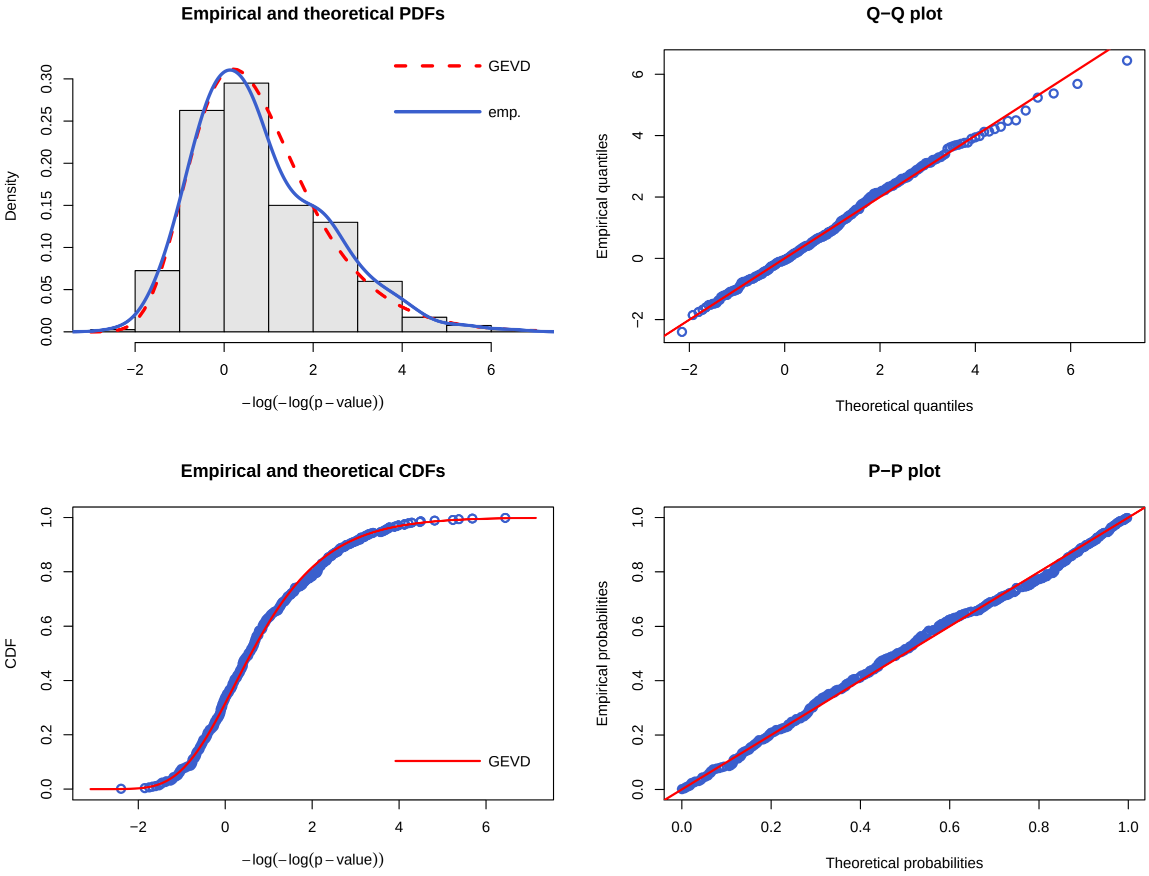
Figure 3.23: Case 2: True model = ABD, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



---

[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.24: Case 2: True model = ABD, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 400 permuted p-values



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

3.5.3 Case 3: True model = BD

Another second degree interaction is considered in this case for additional confirmation of the validity of the GEVD approach. Basically, we applied the GEVD procedure on the data sets simulated in section 2.3.3. Tables 3.7, 3.8, and 3.9 summarize the output from these data sets.

Similar to the previous case, we experience two insignificant proposed interactions when we employ the GEVD to evaluate the suggested models, and it happens only when $n = 500$ (see table 3.7, sets 7 and 10). After investigation, we discovered that these two examined models are the only two models that suggested the interaction

BD with the risk pattern shown in figure 2.6b; whereas the rest suggested the risk pattern presented in figure 2.6a. Recall that we learned from chapter 2 that the risk pattern shown in figure 2.6a minimizes the MSPE compared to the other risk pattern from 2.6b (see table 2.8), which means, there are better models than the examined ones embedded in these data sets. Therefore, the GEVD procedure might deem these two models as insignificant for that same reason. On the other hand, the regular testing procedure recognizes all interactions as significant, regardless of the proposed risk patterns. Once again, this point could suggest that the GEVD procedure does a better job evaluating the significance of the suggested models compared to the permutation testings.

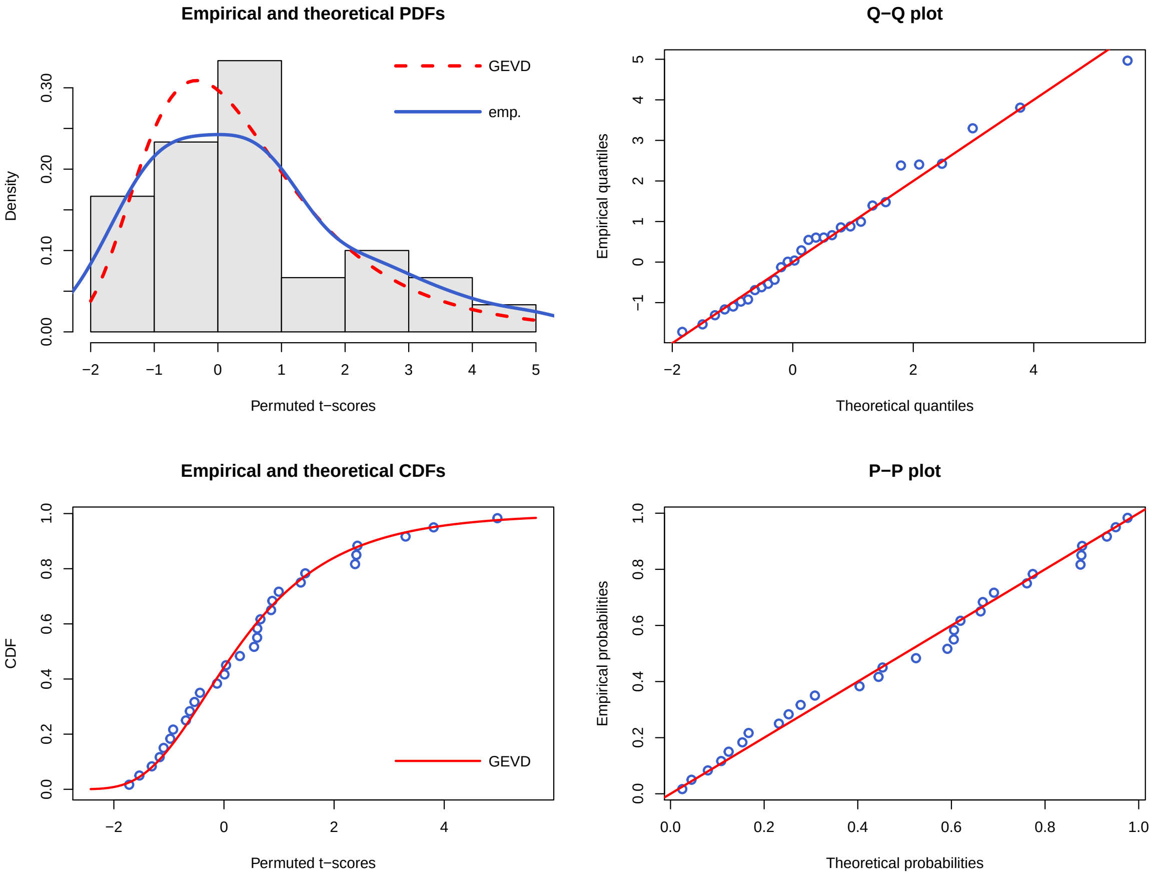
Table 3.7: Case 3: True model= BD, $n = 500$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	12.9882	0.7660	1.1297	-0.2731	0.0000	0.2540	0.9493	0.0597	0.0000	29.0405	0.000	29.5849
2	8.7738	-0.0150	1.1439	0.1630	0.0068	0.6481	1.6289	0.0551	0.0147	28.9430	0.005	29.3935
3	9.1315	0.2059	1.2138	-0.2635	0.0000	-0.1481	0.8663	0.1473	0.0000	29.0621	0.002	29.5588
4	7.9738	0.2739	1.3288	-0.1470	0.0000	0.3673	0.9211	0.0279	0.0000	30.7999	0.004	29.6322
5	8.6924	0.1631	1.0311	0.1730	0.0059	-0.1094	1.0436	-0.0665	0.0176	30.2448	0.003	33.6082
6	9.8547	-0.2433	1.1950	0.0820	0.0016	0.0563	0.9907	-0.1222	0.0034	29.6046	0.001	34.6892
7	5.5904	0.1358	1.0554	-0.0452	0.0028	0.0562	1.3315	-0.5744	0.0637	29.2609	0.006	31.6696
8	7.0933	0.1428	1.1205	-0.0640	0.0004	-0.0118	1.0452	0.1023	0.0001	29.1880	0.002	30.9975
9	5.3883	0.2930	0.9601	-0.1205	0.0002	-0.1909	0.6152	0.5063	0.0000	29.2257	0.008	30.8168
10	7.5865	0.1655	1.1768	0.1325	0.0102	-0.1917	1.2223	-0.2984	0.0766	29.2464	0.002	30.7754

* Column headers are defined as in table 3.1.

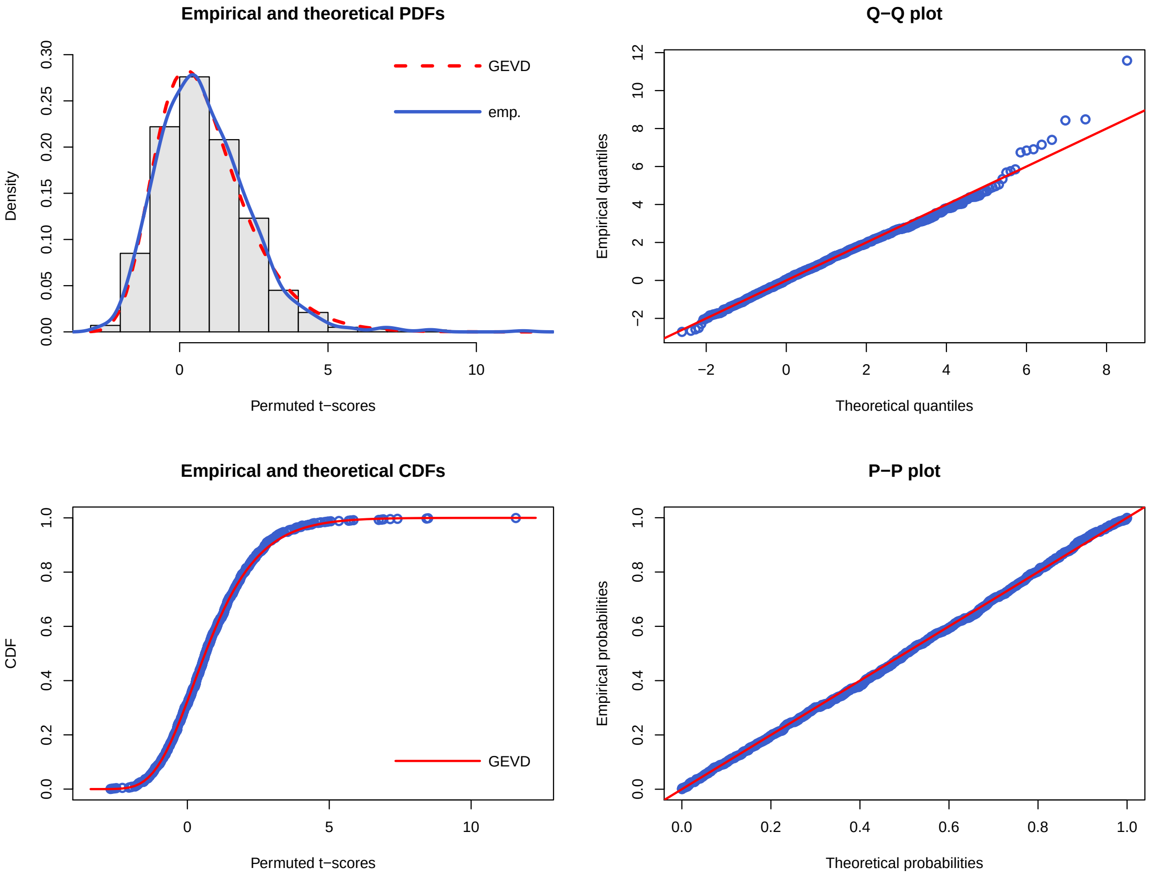
Finally, the graphical representation did not reveal any noticeable migration from the fit neither for $T^{*(0)}_{k_{max}}$ nor for $-\log(-\log(P^{(0)}_{k_{max}}))$ except for a slight lack of fit that could be spotted from a small number of permutations (see the histograms on figures 3.25, 3.27, 3.29, 3.31, 3.33, and 3.35). Besides, the GEVD seems a very reasonable choice to evaluate both $T^{*(0)}_{k_{max}}$ and $P^{(0)}_{k_{max}}$.

Figure 3.25: Case 3: True model = BD, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.26: Case 3: True model = BD, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores


Empirical and theoretical PDFs

Q–Q plot

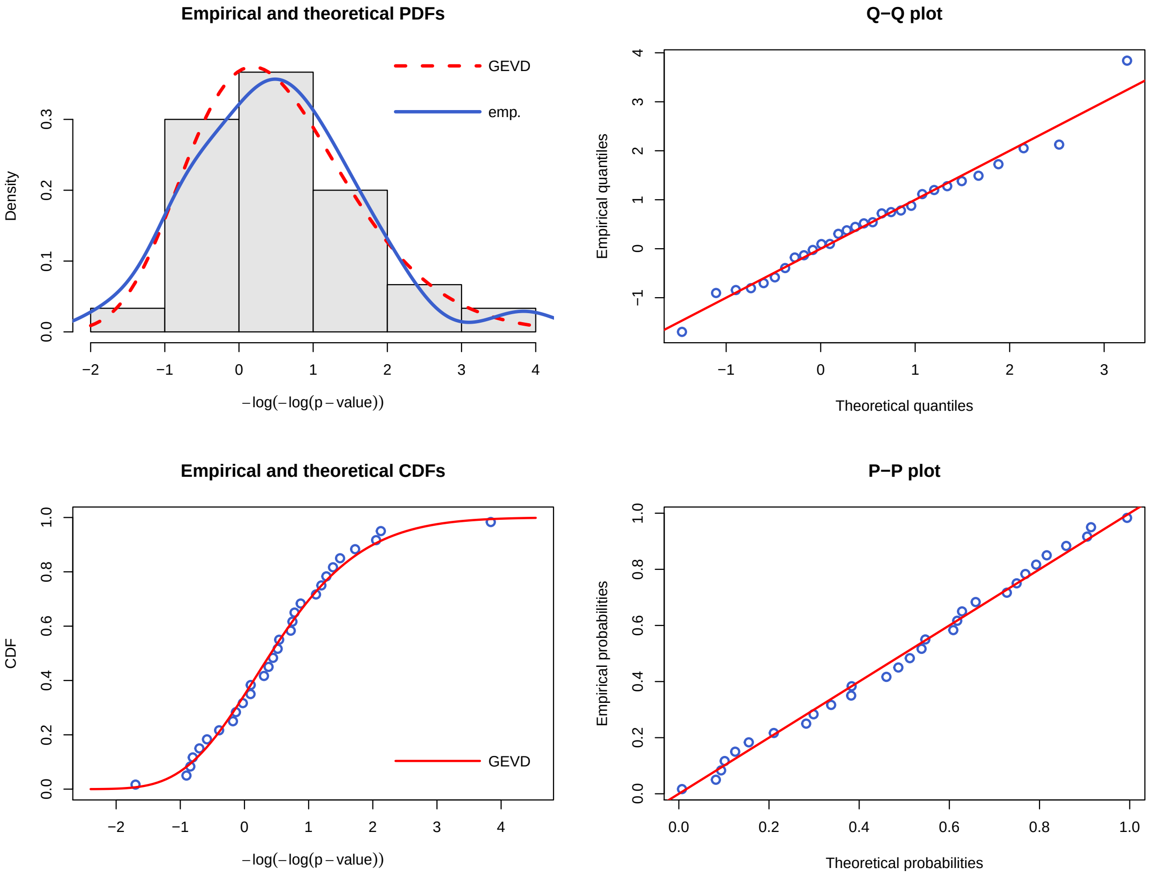
Empirical and theoretical CDFs

P–P plot

---

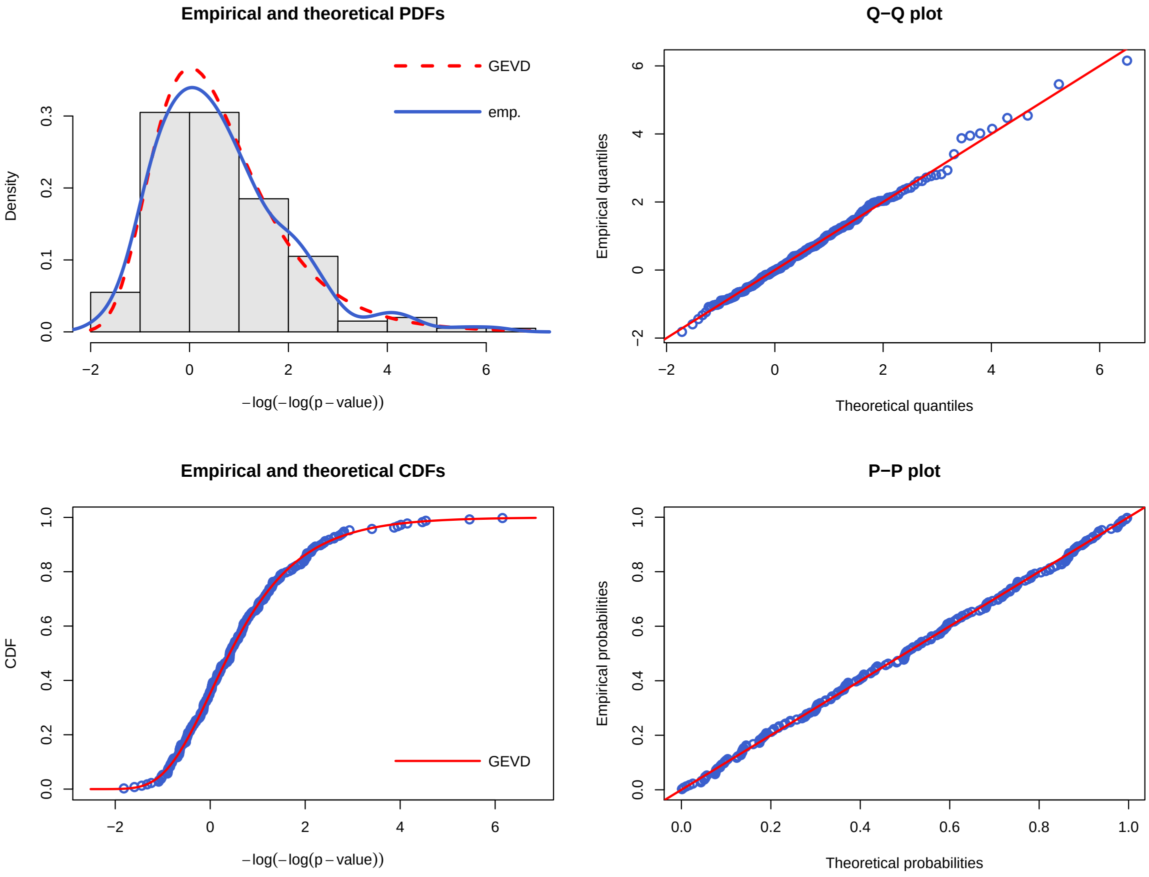
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.27: Case 3: True model $= BD$, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.28: Case 3: True model = BD, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 200 permuted p-values



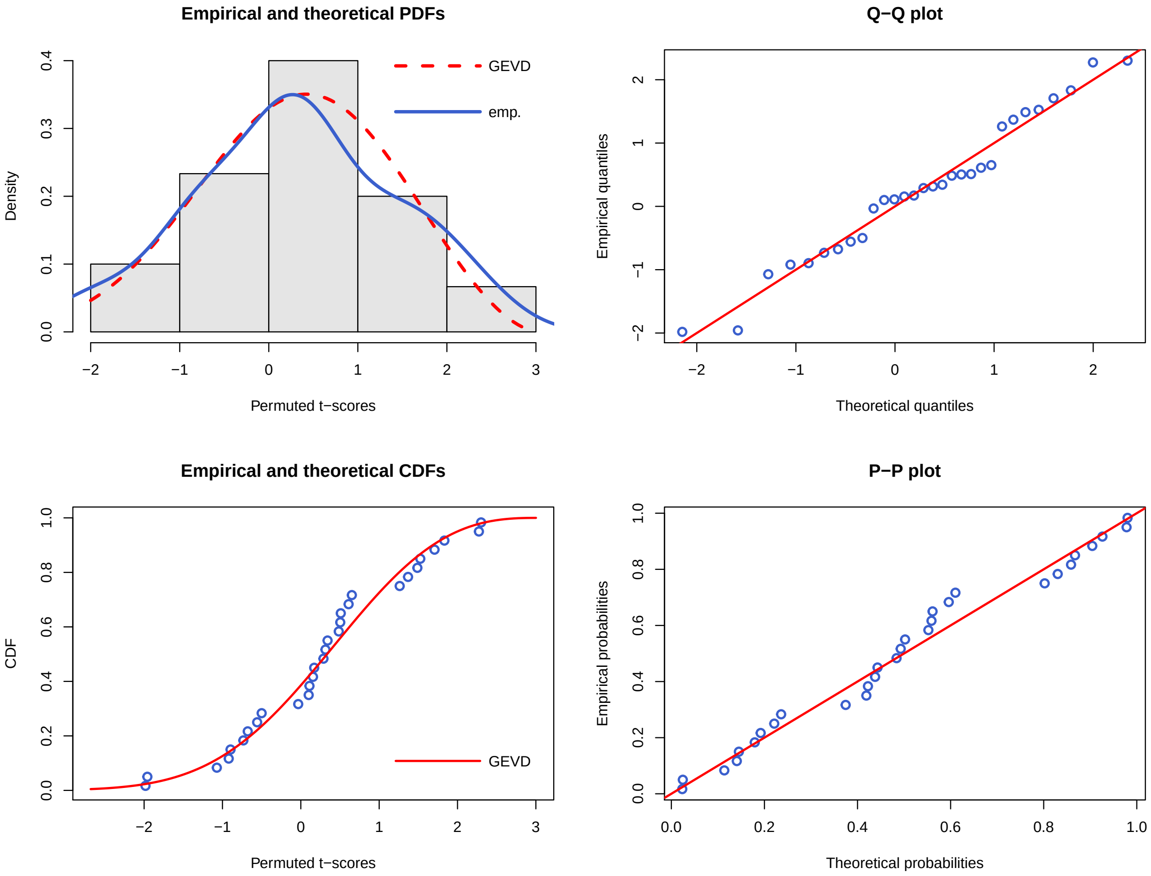
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Table 3.8: Case 3: True model= BD, and $n = 1000$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	11.0682	0.7401	1.4035	-0.2022	0.0000	0.1948	1.0660	-0.0195	0.0000	43.0924	0.000	41.9711
2	12.9821	0.4320	1.4931	-0.4267	0.0000	0.1701	0.9123	0.0147	0.0000	41.2146	0.000	40.7905
3	6.3368	-0.1106	1.2203	-0.0767	0.0011	-0.0537	1.2973	-0.1499	0.0257	40.9804	0.000	41.7732
4	11.4646	0.3605	1.1072	-0.1243	0.0000	0.0711	1.6141	-0.2356	0.0000	41.4339	0.000	42.7985
5	12.9657	0.2149	1.2809	-0.2058	0.0000	0.0209	1.0337	-0.0775	0.0000	41.6909	0.000	45.7260
6	11.3902	-0.0538	1.1338	-0.3668	0.0000	0.1543	1.0346	0.0867	0.0000	44.1730	0.000	42.0594
7	10.2836	0.3404	1.0226	-0.3115	0.0000	-0.0272	1.2894	-0.3138	0.0000	41.3094	0.000	42.1066
8	10.0232	0.0516	1.2395	-0.0617	0.0000	0.0135	1.3031	-0.2907	0.0120	41.6469	0.000	43.0061
9	8.4548	0.1327	1.3908	-0.2435	0.0000	-0.1412	0.7047	0.2337	0.0000	41.7274	0.000	44.0856
10	8.5341	0.2593	1.4394	-0.3475	0.0000	-0.5682	0.8772	0.0597	0.0000	39.9479	0.000	42.2784

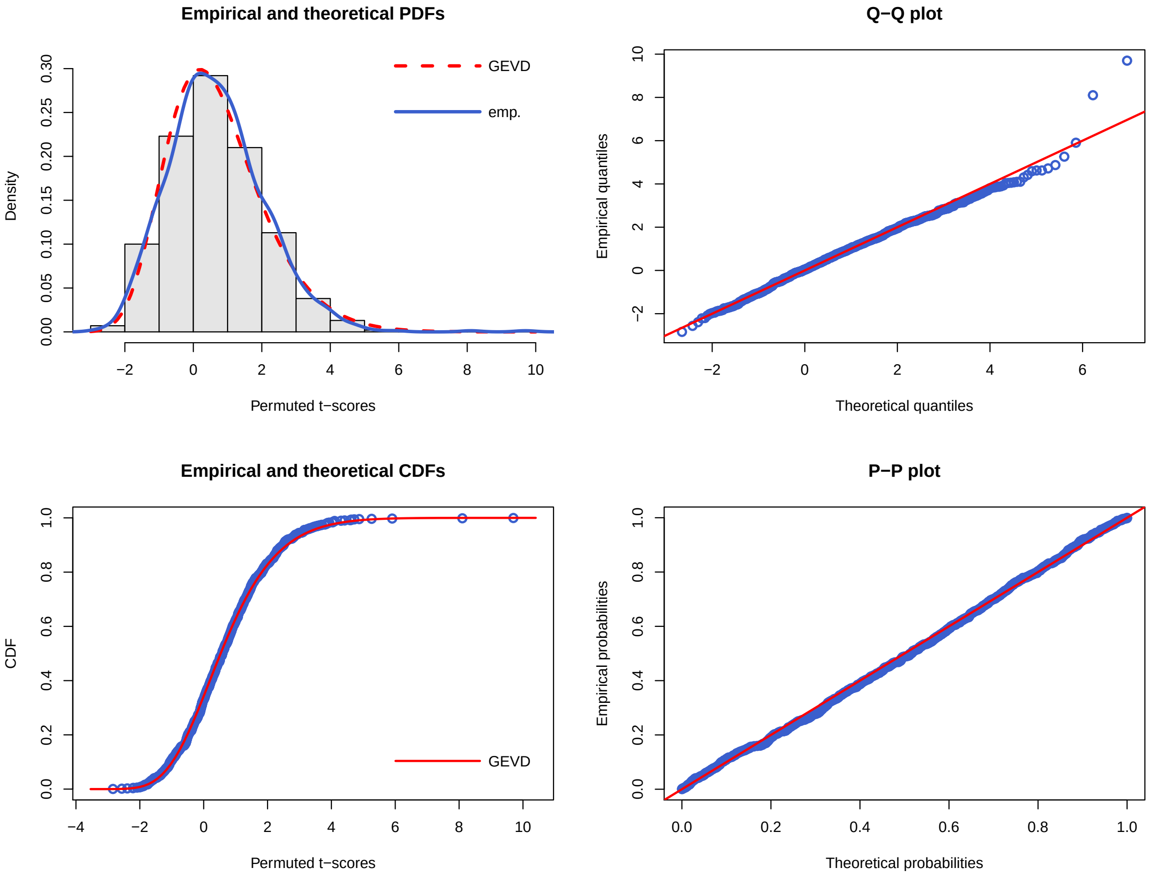
* Column headers are defined as in table 3.1.

Figure 3.29: Case 3: True model = BD, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



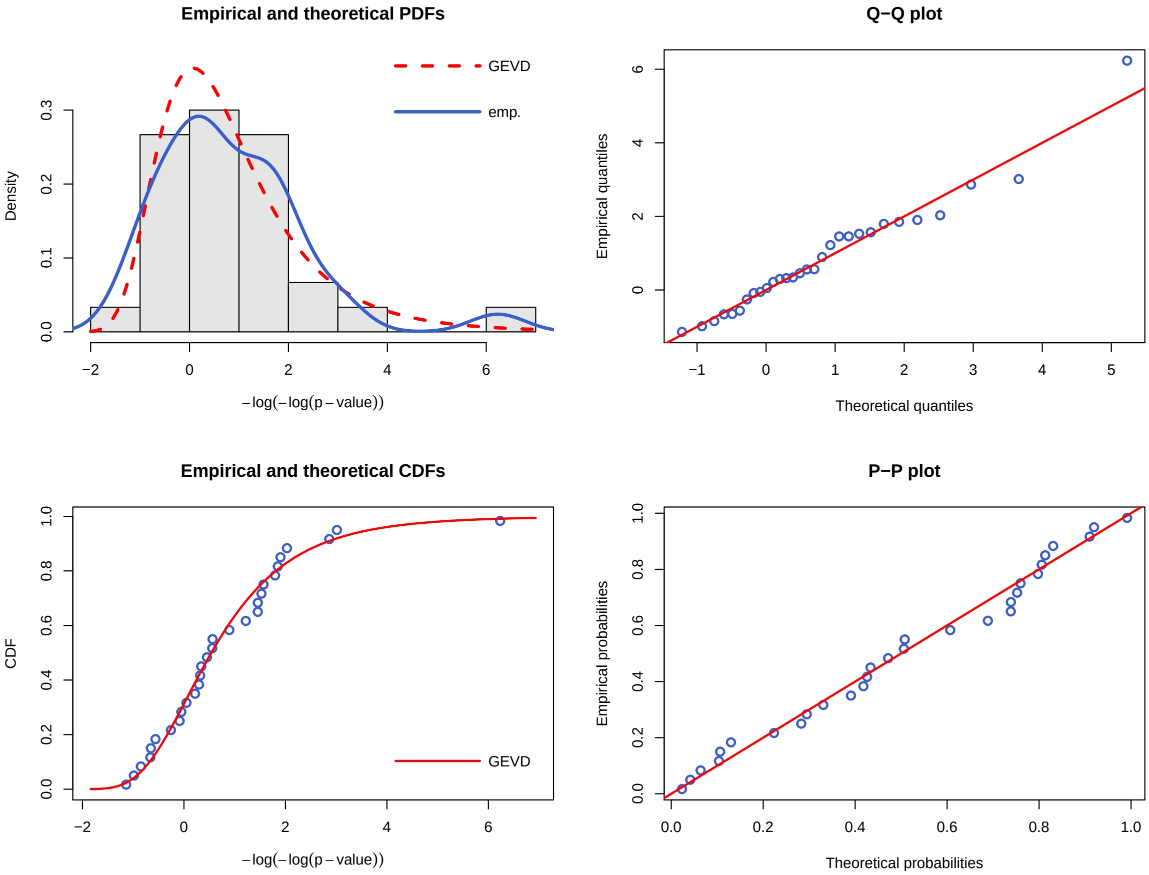
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.30: Case 3: True model = BD, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.31: Case 3: True model = BD, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values


Empirical and theoretical PDFs

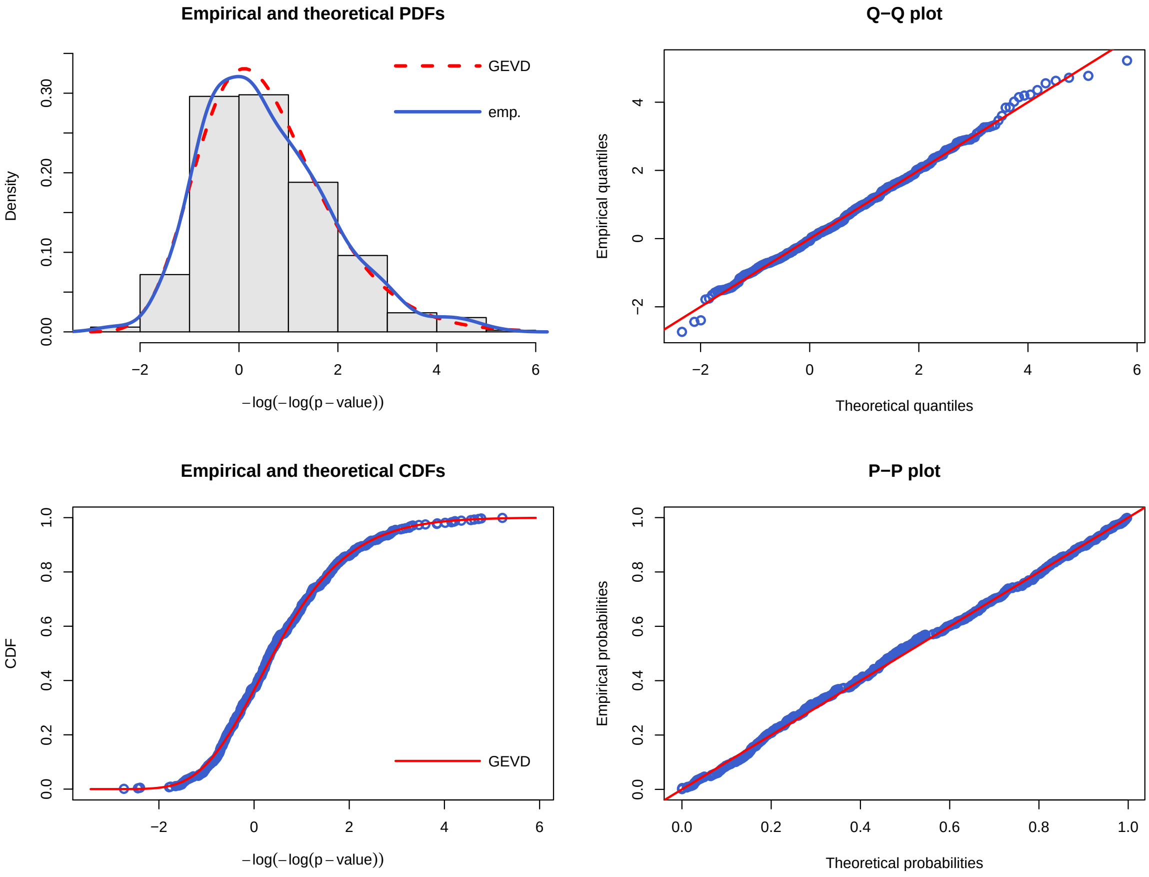
Q–Q plot

Empirical and theoretical CDFs

P–P plot

* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.32: Case 3: True model $= BD$, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values



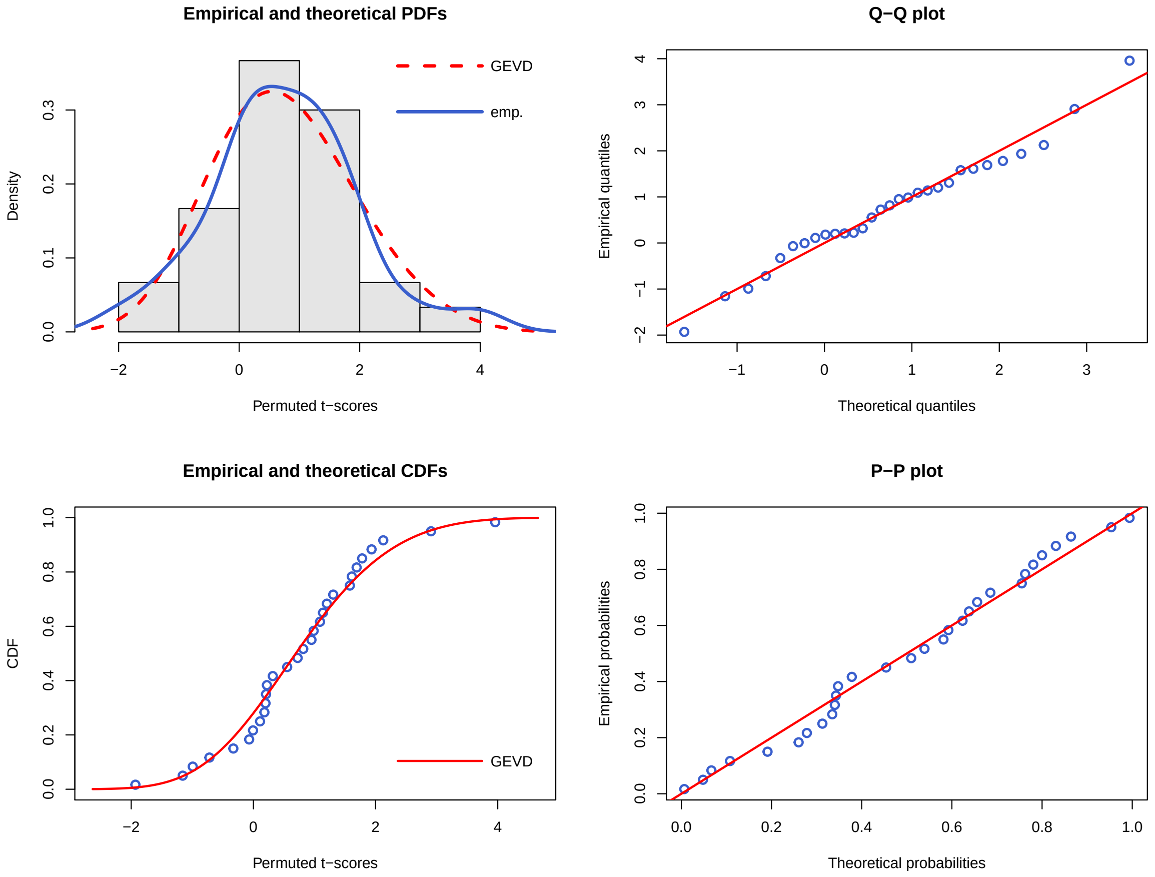
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Table 3.9: Case 3: True model$= BD$, and $n = 2000$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	11.6430	0.2859	1.1569	-0.2044	0.0000	-0.1837	1.5392	-0.2311	0.0000	64.9627	0.000	64.1316
2	12.2938	0.2190	1.1905	-0.2233	0.0000	0.0688	0.9159	-0.1721	0.0000	65.1735	0.000	63.9616
3	14.3724	0.1580	1.1395	-0.2001	0.0000	0.1110	0.7740	-0.1198	0.0000	64.8242	0.000	63.7335
4	13.5646	-0.1723	1.1201	-0.3950	0.0000	0.5113	1.4053	-0.8522	0.0008	64.2101	0.000	63.5998
5	19.5815	-0.0342	1.0346	-0.1840	0.0000	0.0819	0.9170	-0.0923	0.0000	62.0592	0.000	62.5083
6	13.5869	-0.0984	1.3070	-0.4082	0.0000	-0.3689	0.8679	0.1498	0.0000	60.8961	0.000	62.1566
7	15.2050	0.4950	1.4679	-0.6029	0.0000	-0.2058	1.3508	-0.1262	0.0000	60.8463	0.000	62.2635
8	15.3370	0.3098	1.3324	-0.2556	0.0000	0.2848	1.2500	-0.2846	0.0000	60.8577	0.000	62.1972
9	16.1559	0.8818	1.3687	-0.5594	0.0000	0.2727	1.2604	-0.1606	0.0000	60.4628	0.000	62.9126
10	13.7691	-0.0942	1.4842	-0.2872	0.0000	0.1223	1.2820	-0.2542	0.0000	60.4286	0.000	62.0273

* Column headers are defined as in table 3.1.

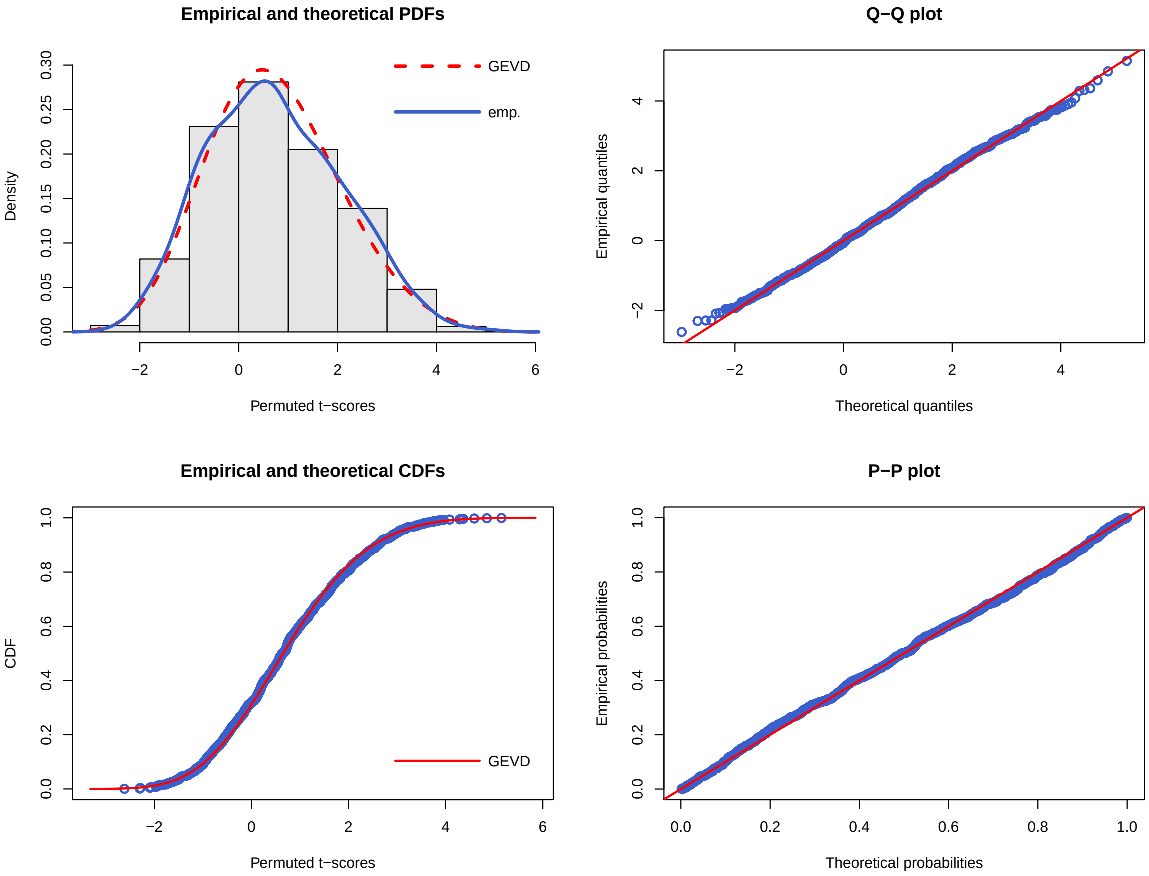
Figure 3.33: Case 3: True model = BD, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



---

* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.34: Case 3: True model = BD, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores


Empirical and theoretical PDFs

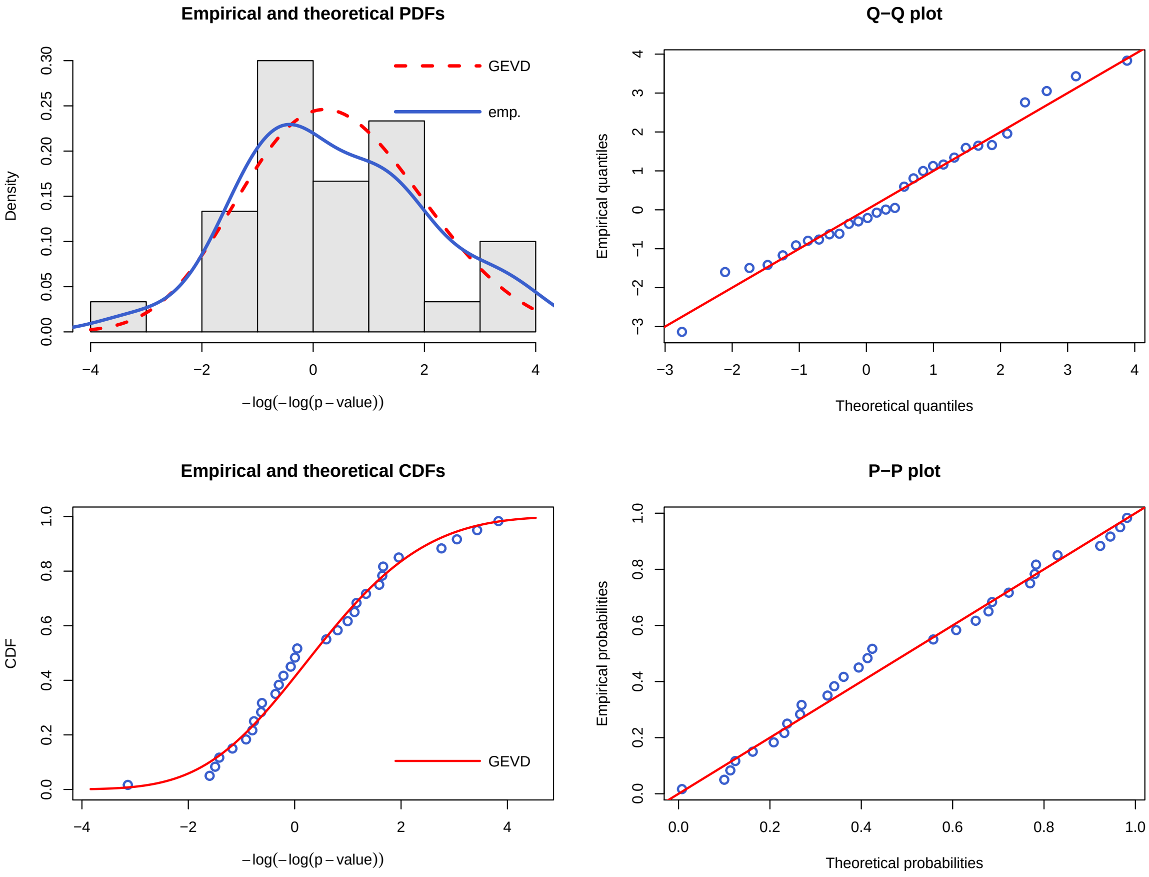
Q–Q plot

Empirical and theoretical CDFs

P–P plot

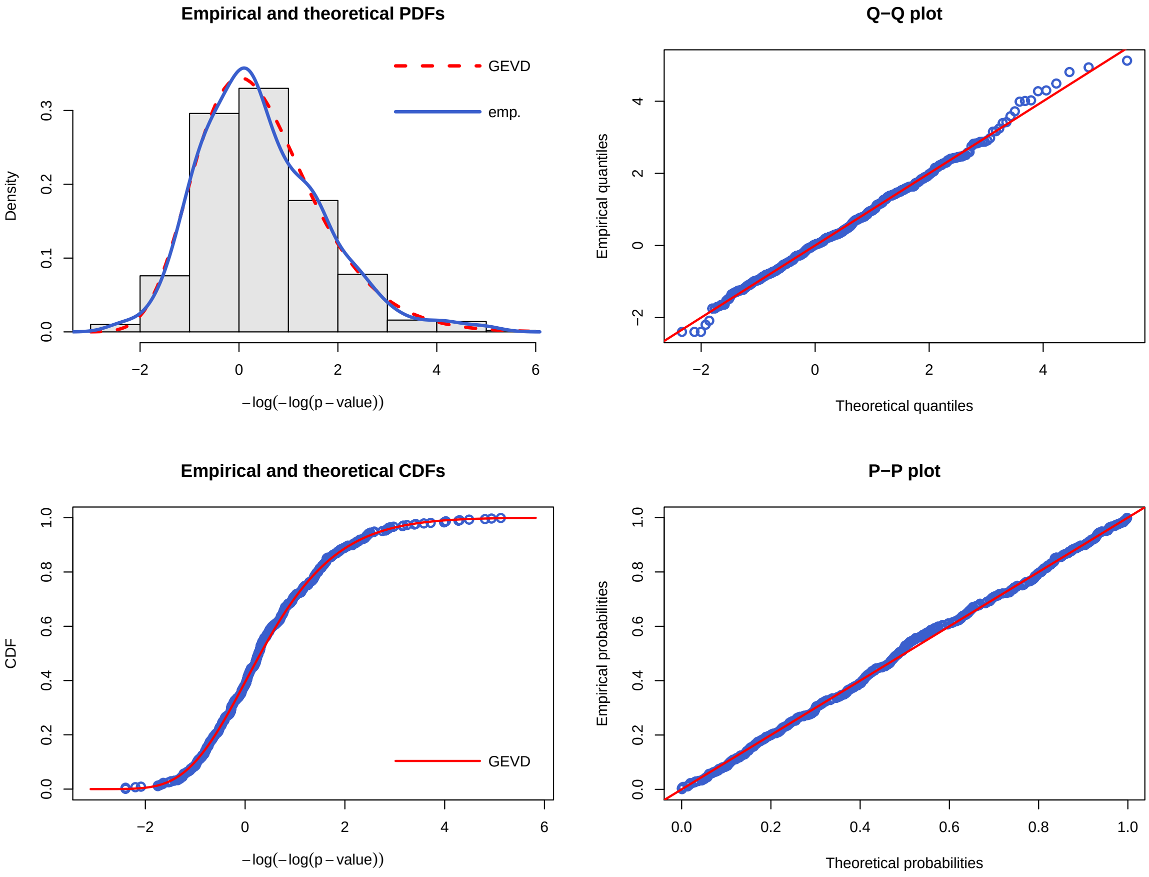
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.35: Case 3: True model = BD, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.36: Case 3: True model = BD, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values


Empirical and theoretical PDFs

Q−Q plot

Empirical and theoretical CDFs

P−P plot

* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

3.5.4 Case 4: True model = ABC

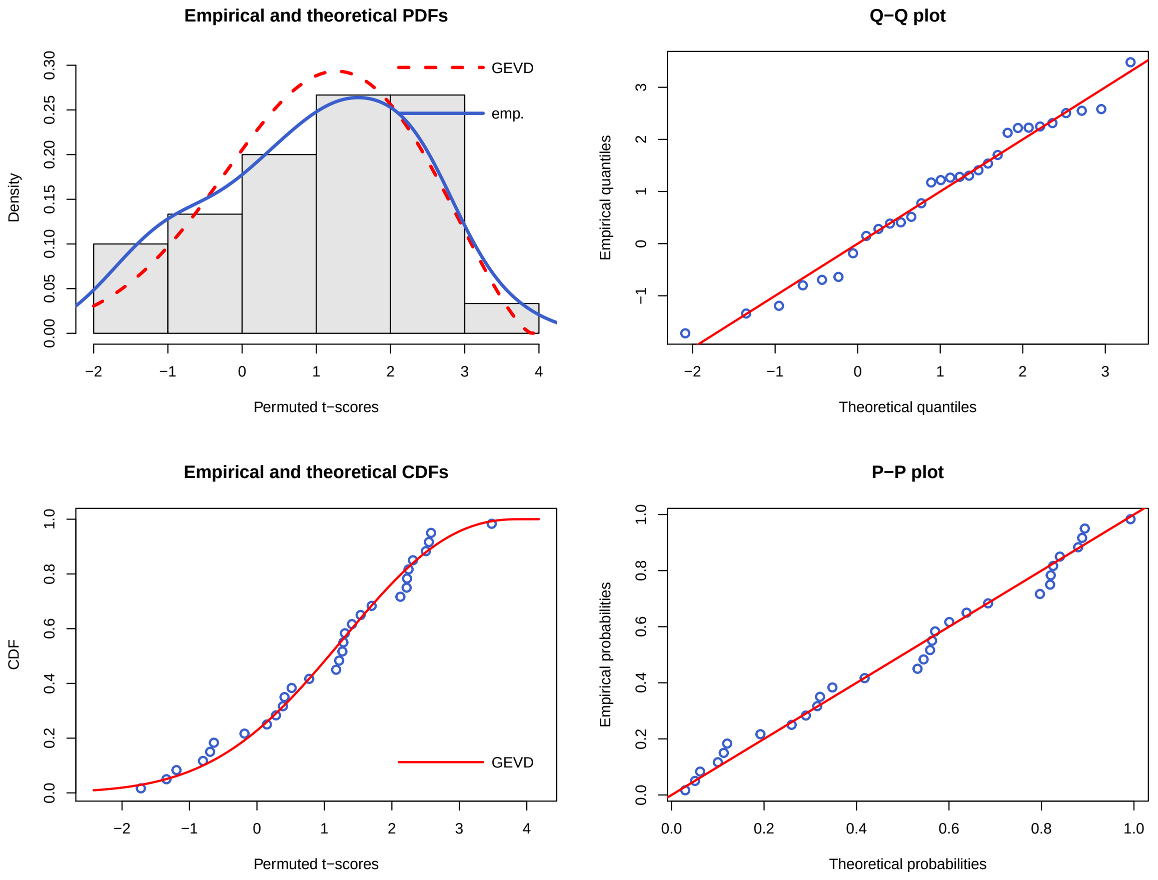
In this last case, we evaluate the GEVD approach using the generated data sets from section 2.3.4, which uses ABC as the true disease predisposition. The output coincides with the findings from previous cases, which suggest that all proposed models are significant, and the GEVD is a reliable replacement of the permutation testings to study the behavior of the test statistic and its p-value.

Table 3.10: Case 4: True model= ABC, and $n = 500$, and $m = m_1 = 30$

	GEVD procedure									Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	9.4425	0.5531	1.5466	-0.3279	0.0000	0.2812	0.8262	-0.1939	0.0000	30.4723	0.004	30.0196
2	13.0731	0.5741	1.1391	-0.1103	0.0000	0.0087	1.3556	-0.4597	0.0000	31.9494	0.000	31.1732
3	12.1367	0.2399	1.1930	0.1283	0.0016	0.4865	1.3148	-0.2867	0.0146	30.3629	0.000	29.6430
4	10.0161	0.5916	1.3926	-0.4213	0.0000	-0.2422	1.0116	0.1057	0.0000	30.7973	0.002	29.6669
5	9.4804	0.2371	1.2477	-0.2209	0.0000	0.1889	1.2612	0.0570	0.0000	29.0575	0.000	29.6187
6	10.0555	0.5593	1.5019	-0.5503	0.0000	-0.0829	0.8782	0.0146	0.0000	30.8309	0.000	29.6438
7	9.4207	-0.1154	1.1543	0.1353	0.0039	0.4073	1.4548	-0.4648	0.0477	29.5905	0.003	29.6449
8	7.3506	-0.0983	1.2103	-0.1924	0.0000	0.0590	1.4408	-0.0593	0.0000	30.8074	0.000	31.0272
9	11.3497	0.1141	1.3360	-0.3719	0.0000	0.0325	1.3855	-0.2364	0.0000	29.0184	0.000	29.7044
10	12.0756	0.2474	1.1116	-0.0975	0.0000	-0.1997	0.8700	-0.0961	0.0000	28.2528	0.000	29.5465

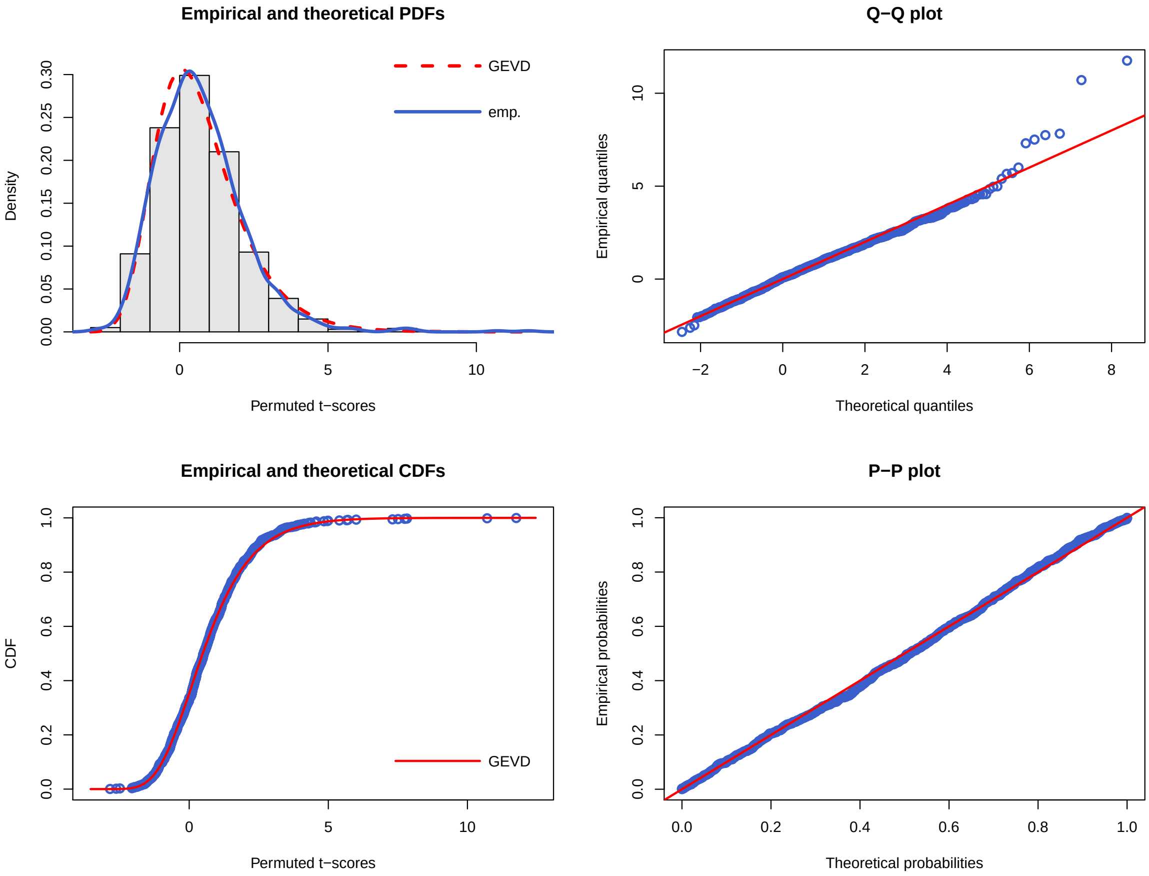
* Column headers are defined as in table 3.1.

Figure 3.37: Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



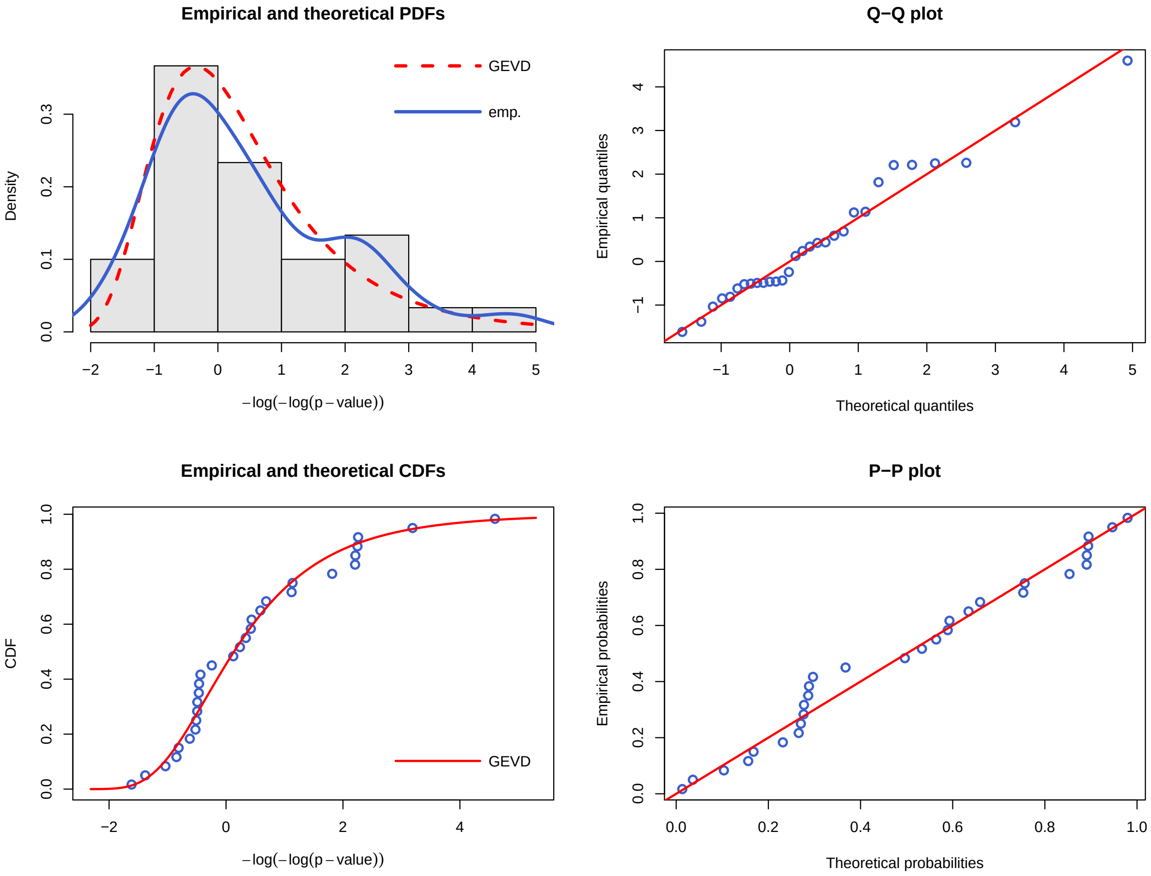
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.38: Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

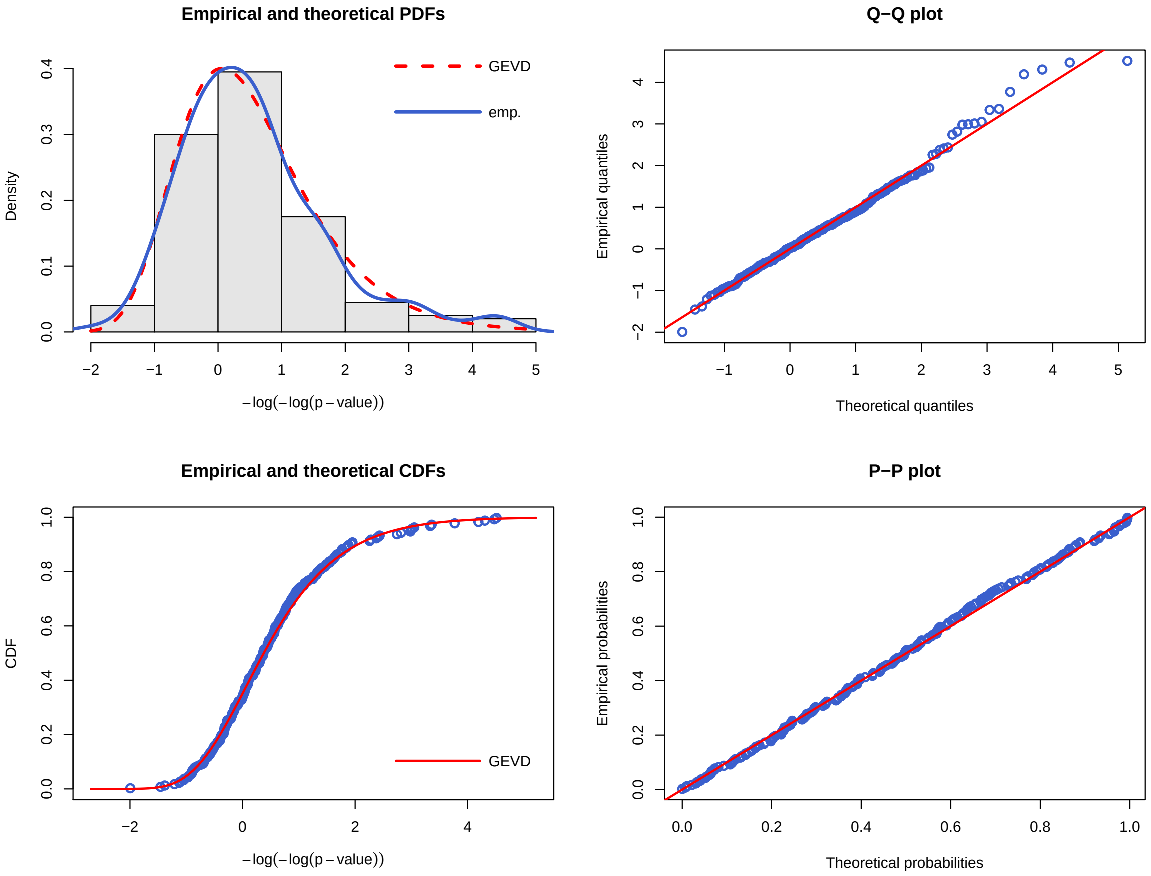
Figure 3.39: Case 4: True model = ABC, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



---

[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.40: Case 4: True model $= ABC$, $n = 500$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 200 permuted p-values



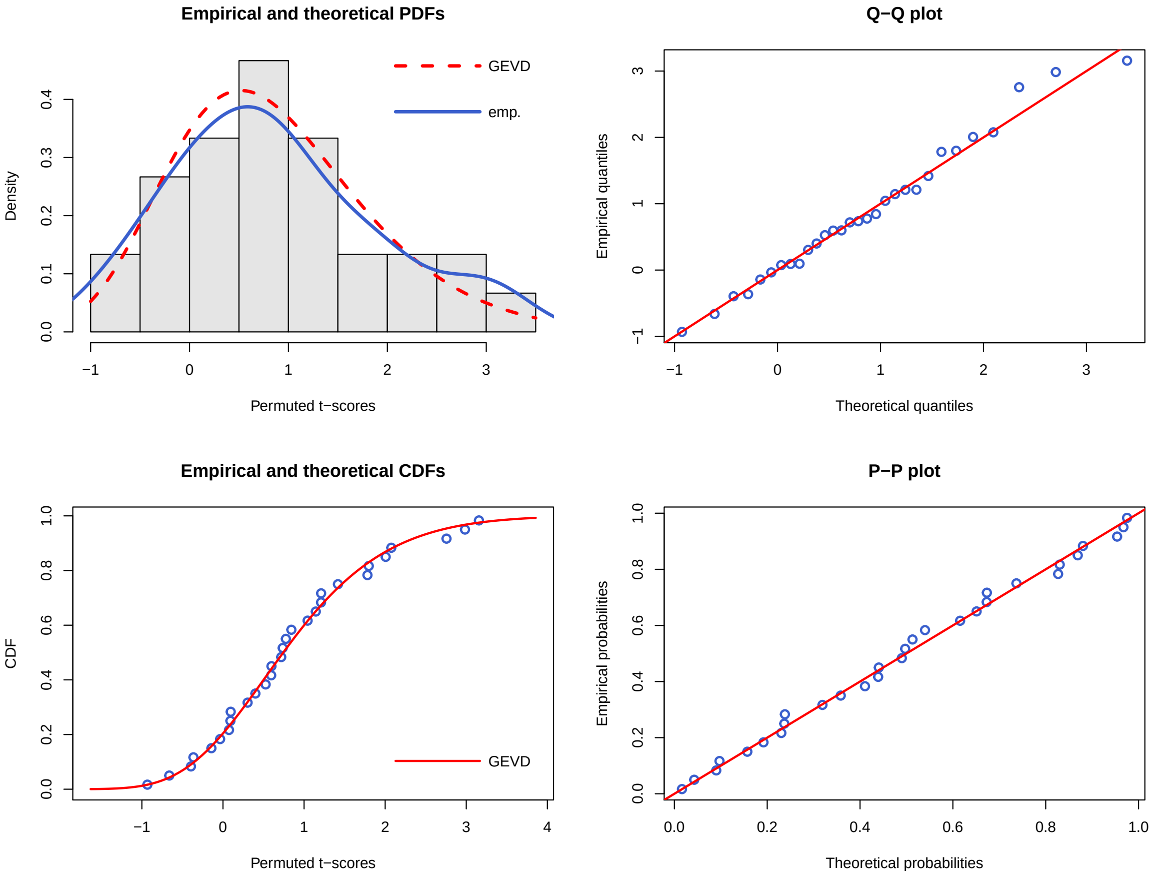
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Table 3.11: Case 4: True model$= ABC$, and $n = 1000$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	14.8271	-0.1240	1.2508	-0.2212	0.0000	-0.3596	0.9174	0.0056	0.0000	39.0442	0.000	39.6043
2	16.5818	0.1330	0.8206	-0.1821	0.0000	0.4874	1.1700	0.1070	0.0000	39.0856	0.000	39.5326
3	15.2867	0.0679	1.1469	-0.0586	0.0000	0.0807	1.3305	-0.3509	0.0024	39.1061	0.000	39.4339
4	16.6098	0.0140	1.4397	-0.0597	0.0000	-0.0767	1.0196	0.2809	0.0000	39.0468	0.000	39.5011
5	14.3892	0.1779	1.0973	-0.1218	0.0000	-0.3756	1.4787	-0.3100	0.0000	39.2719	0.000	39.4693
6	15.5734	0.4229	0.8905	-0.1027	0.0000	0.0195	0.9763	0.1767	0.0000	39.1094	0.000	39.4865
7	16.8849	0.1002	1.0957	-0.1849	0.0000	0.4833	0.9222	0.2140	0.0000	40.3319	0.000	39.5207
8	15.2774	-0.6211	1.1544	-0.0600	0.0000	-0.1937	1.3755	-0.1884	0.0011	39.0103	0.000	39.5045
9	16.8113	-0.0017	1.1742	-0.1877	0.0000	0.0840	0.7758	0.2378	0.0000	39.0729	0.000	39.5146
10	14.0856	-0.0377	1.3131	-0.6788	0.0000	-0.2527	0.9694	0.1891	0.0000	39.0179	0.000	39.5048

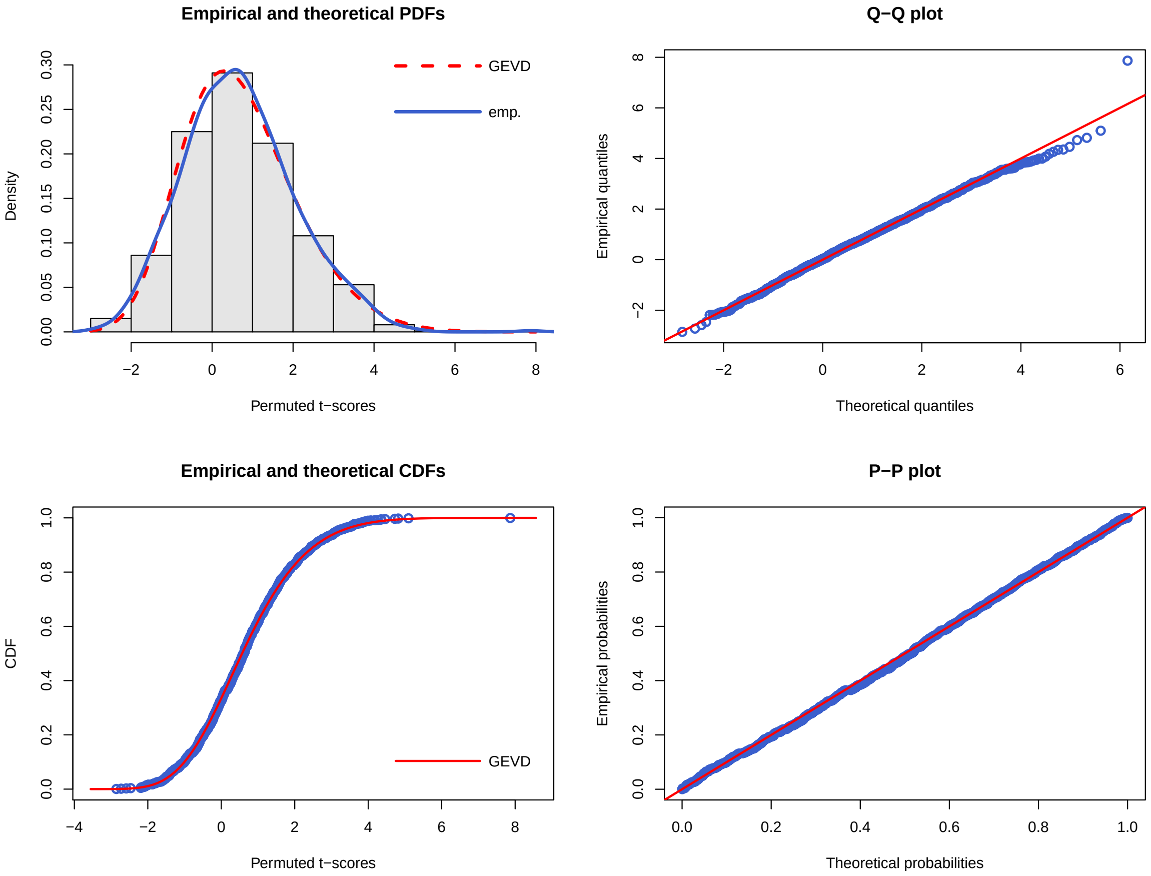
* Column headers are defined as in table 3.1.

Figure 3.41: Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



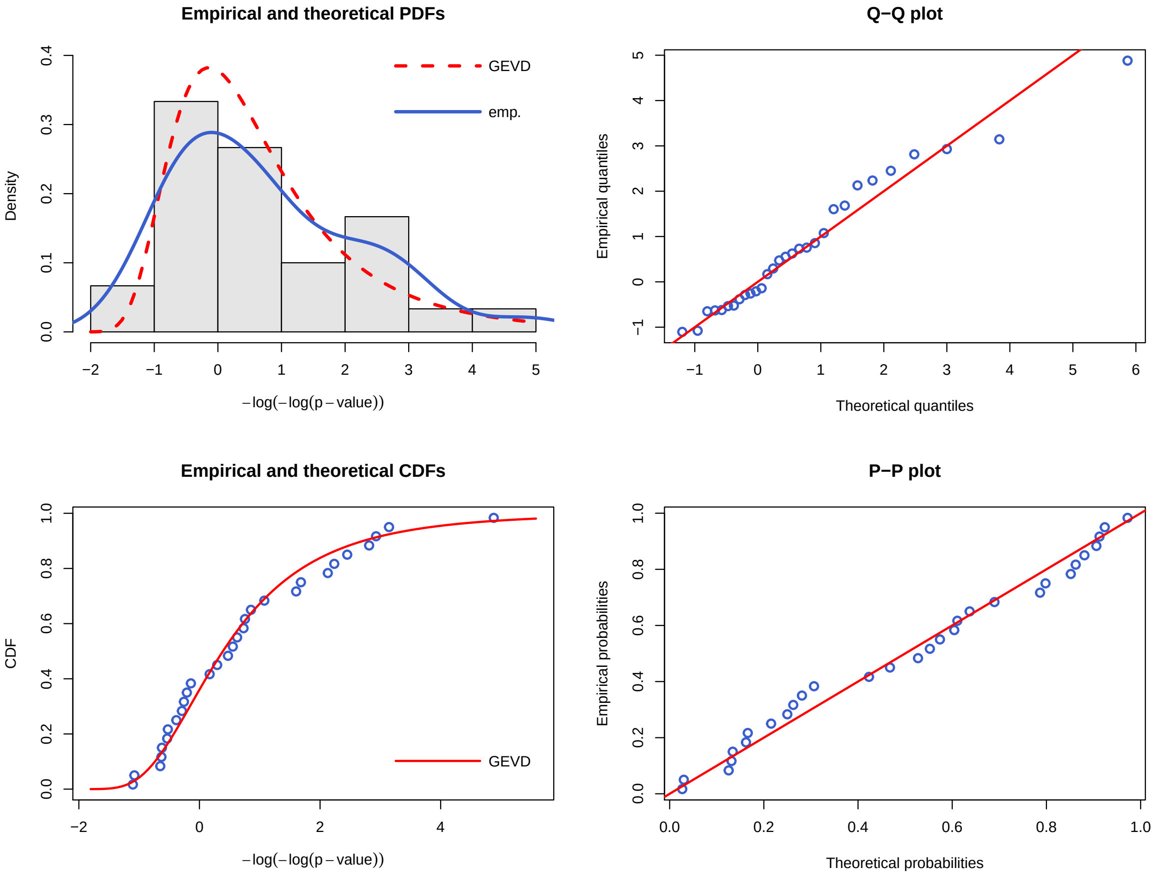
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.42: Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



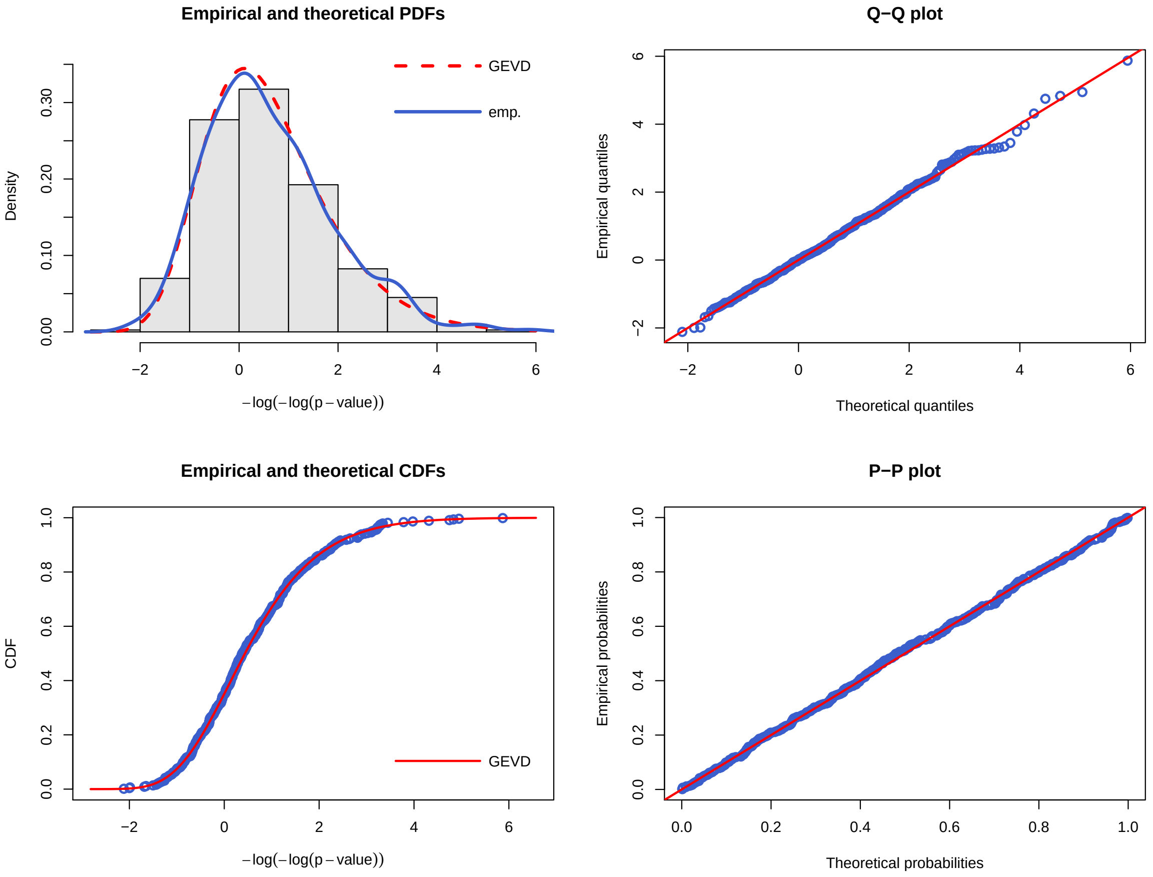
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.43: Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.44: Case 4: True model = ABC, $n = 1000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 400 permuted p-values



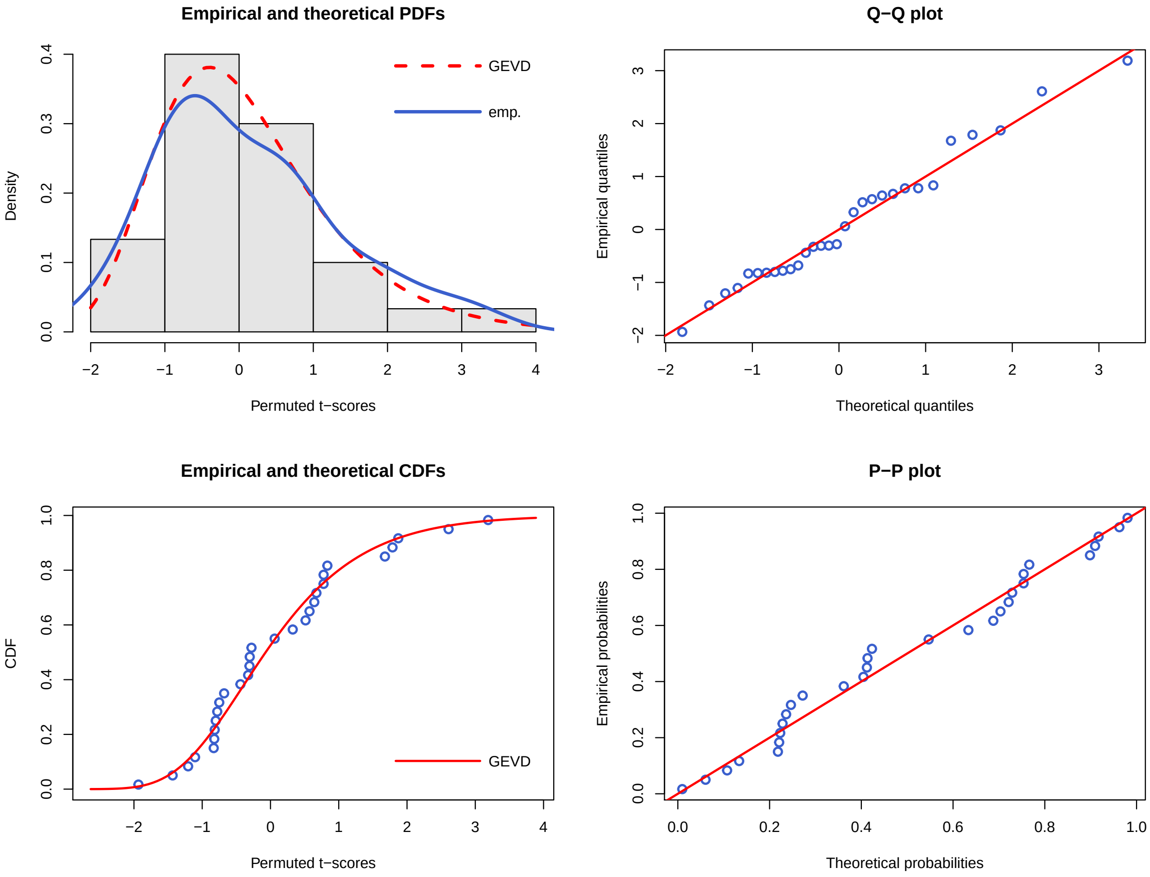
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Table 3.12: Case 4: True model= ABC, and $n = 2000$, and $m = m_1 = 30$

	GEVD procedure										Permutation	
Set	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	Time (min)	p_t	Time (min)
1	26.5165	-0.1392	1.6560	0.3515	0.0000	0.3468	1.0736	-0.3888	0.0000	60.9639	0.000	60.4848
2	23.1537	-0.4237	0.9656	0.0246	0.0000	0.1836	1.5028	0.4012	0.0033	60.6772	0.000	60.4309
3	21.8987	-0.2714	1.5365	0.1351	0.0000	0.5274	1.5966	0.0936	0.0000	60.7499	0.000	60.4446
4	22.9638	0.2467	1.3306	0.4938	0.0000	-0.2975	1.5185	0.3015	0.0000	60.9498	0.000	60.4343
5	20.6543	0.4167	1.6659	0.2720	0.0000	-0.0020	1.4295	0.1710	0.0000	61.5221	0.000	60.5223
6	24.7238	0.0681	1.1493	0.2970	0.0000	0.1034	0.9738	-0.1946	0.0000	60.6067	0.000	60.3343
7	23.5707	0.1032	1.4913	0.2549	0.0000	0.2271	1.0676	0.0752	0.0000	60.4514	0.000	60.4536
8	22.8113	0.3707	1.1092	0.1829	0.0000	0.1750	1.4578	0.3079	0.0000	60.5447	0.000	60.4828
9	22.4449	-0.0021	1.1052	0.2827	0.0000	0.0977	0.8729	0.1447	0.0000	60.5211	0.000	60.5020
10	21.8562	0.5149	1.2811	0.4050	0.0000	0.1427	0.9142	0.0719	0.0000	60.7040	0.000	60.3734

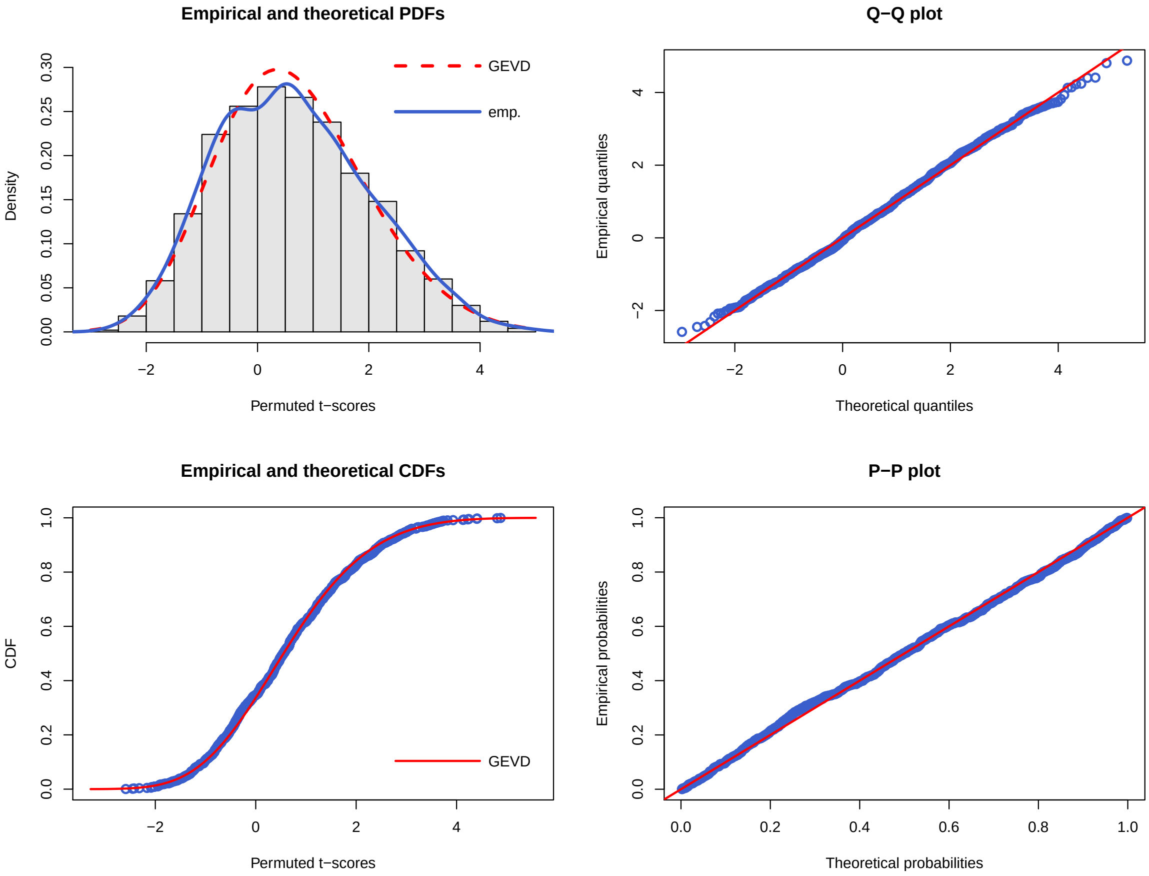
* Column headers are defined as in table 3.1.

Figure 3.45: Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



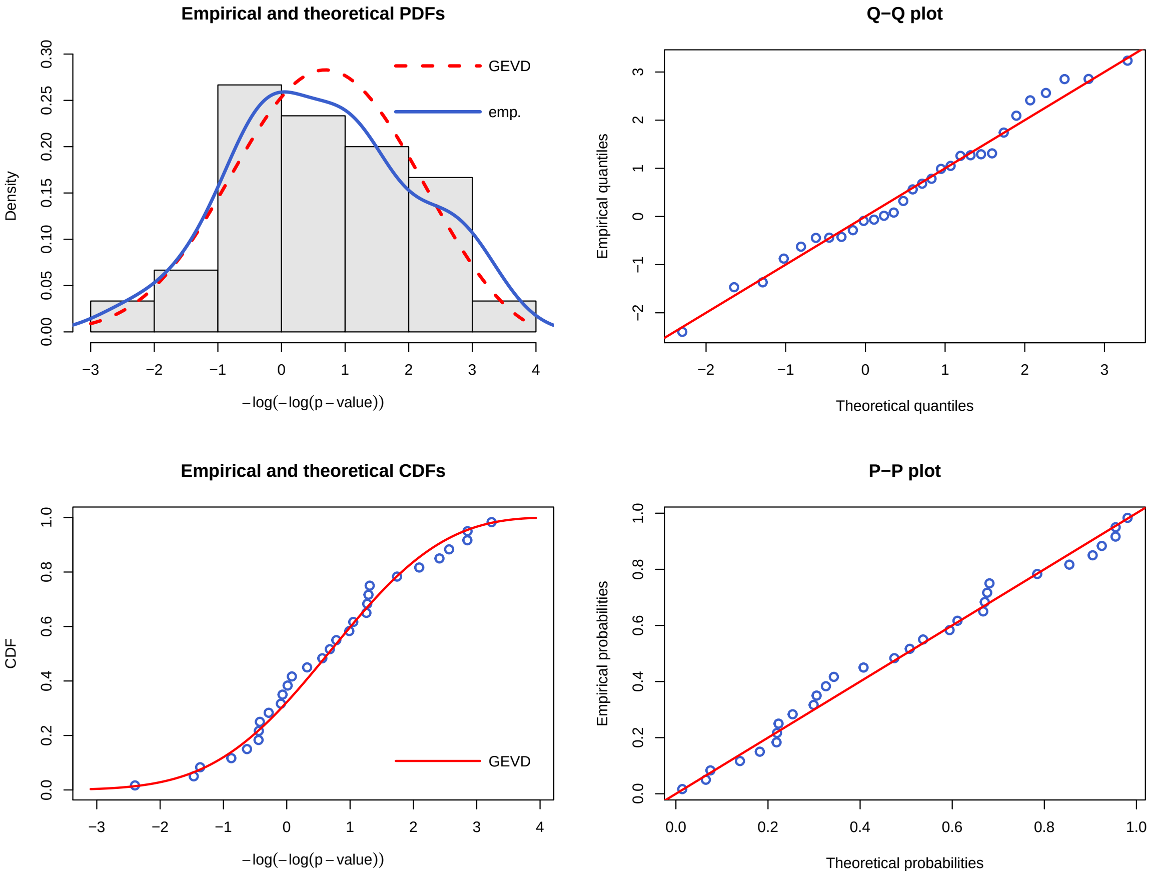
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.46: Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 1000 permuted t-scores



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

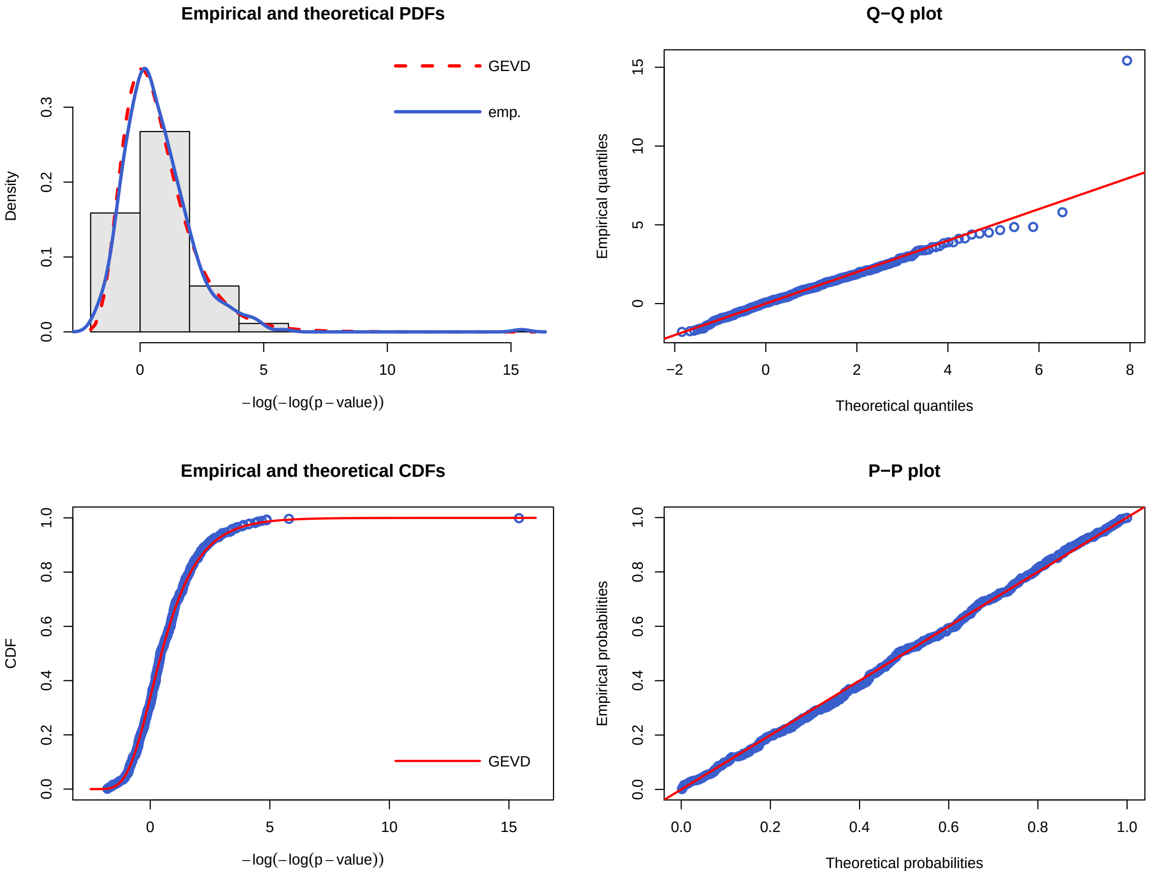
Figure 3.47: Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



---

[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3.48: Case 4: True model = ABC, $n = 2000$; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 500 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

3.6 Summary

A GEVD approach of model evaluation on MDR-based approaches was initially proposed by Pattin et al. [48], and then by Hua et al. [30]. An approach was suggested to evaluate the t-test statistic in Pattin et al. [48] using 20 permuted data sets; they assumed that the p-value is distributed as uniform(0,1). On the other hand, Hua et al. [30] used 50 permuted samples to utilize the GEVD to explain the variation of the χ^2 statistic of the 2×2 contingency table. Further, the GEVD parameters of the $-\log(p)$ were estimated using a set of 50 permuted p-values to validate the observed

p-value.

In our research, we adapted the GEVD approach to the OQMDR algorithm to assess the significance of the proposed models. After considering several permutation sizes, we deduced that a set of 20 permuted test statistics provides insufficient information to maintain a high chance of convergence when obtaining the MLE's. In contrast, even though using 50 permuted samples could improve the quality of the approximation, it also leads to a substantial inflation in the calculation time. Therefore, we think that a set of 30 permuted t-statistics is sufficient to ensure convergence and to produce satisfactory approximation at an acceptable pace. Similarly, a set of 30 permuted p-values is used to obtain the MLE's of the approximated distribution of the p-value.

From the simulation results, the GEVD is demonstrated to be a plausible choice, compared to other examined distributions in this study, to evaluate the observed t-score and its p-value. Our study shows that a double logarithmic transformation of the p-value fits better than a single logarithm (see the appendix), which was suggested by Hua et al. [30].

On the other hand, the GEVD approach is primarily proposed to reduce the computation burden and enhance the efficiency of the OQMDR algorithm. However, the simulation study did not reveal a significant improvement in this aspect. Regardless, the GEVD procedure increased the precision of the calculated p-values, which requires a huge amount of time if the regular permutations are employed. Another consideration is that the evaluation of the p-value portion, which consumes about 95% of the time, seems to be necessary, especially for small number of permutations because our study shows that the assumption of the p-value following a uniform(0,1) is invalid for small number of permutations. In addition, the GEVD approach showed a more realistic evaluation than the regular permutations do, specifically, for the cases where wrong risk patterns are selected. However, a further investigation is required to confirm this feature because it wasn't our primary intention in this study.

Copyright© Zaid T. Al-Khaledi, 2019.

Chapter 4 Theoretical Findings

4.1 Derivation of MLE's required formulas

In this section we will give a full step-by-step derivation of the first and second derivatives of the log-likelihood function of the GEVD (equation 3.6) with respect to each of its three parameters. The case where $\xi \neq 0$ in equation 3.1 is considered in the derivation. The MLE's of Gumbel distribution (when $\xi \to 0$) are easy to obtain with no further analytical approach needed.

Let $Y_1, Y_2, ..., Y_n$ be a sequence of independent and identically distributed random variables that follow the GEVD with the CDF defined in equation 3.1 for $\xi \neq 0$. Therefore, the common probability density function (PDF) can be written as:

$$f_Y(y; \mu, \sigma, \xi) \quad = \quad \frac{1}{\sigma}\left[1 + \xi\frac{y-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})} e^{-\left[1+\xi\frac{y-\mu}{\sigma}\right]^{-\frac{1}{\xi}}}$$

defined on $1 + \xi\left(\frac{y-\mu}{\sigma}\right) > 0$ for $\xi \neq 0$, $\mu \in (-\infty, \infty)$ is the location parameter , $\sigma > 0$ is the scale parameter, and $|\xi| > 0$ is the shape parameter.

Thus, the likelihood function for $Y_1, Y_2, ..., Y_n$ is:

$$\begin{aligned}
L(\mu, \sigma, \xi) &= \Pi_{i=1}^{n} f_{Y_i}(y_i; \mu, \sigma, \xi) \\
&= \Pi_{i=1}^{n} \frac{1}{\sigma}\left[1 + \xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})} e^{-\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}}} \\
&= \frac{1}{\sigma^n}\left[\Pi_{i=1}^{n}\left[1 + \xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\right] e^{-\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}}} \\
&= \frac{1}{\sigma^n}\left[\Pi_{i=1}^{n}(1+\xi z_i)^{-(1+\frac{1}{\xi})}\right] e^{-\sum_{i=1}^{n}(1+\xi z_i)^{-\frac{1}{\xi}}}
\end{aligned}$$

where $z_i = \frac{y_i - \mu}{\sigma}$.

Then, the log-likelihood is:

$$\begin{aligned} l(\mu,\sigma,\xi) &= -n\log\sigma - (1+\frac{1}{\xi})\sum_{i=1}^{n}\log\left[1+\xi\frac{y_i-\mu}{\sigma}\right] - \sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}} \\ &= -n\log\sigma - (1+\frac{1}{\xi})\sum_{i=1}^{n}\log(1+\xi z_i) - \sum_{i=1}^{n}(1+\xi z_i)^{-\frac{1}{\xi}} \end{aligned}$$

Therefore, the first derivative of the log-likelihood function with respect to (w.r.t.) μ is:

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{\partial}{\partial \mu}\left[-n\log\sigma - (1+\frac{1}{\xi})\sum_{i=1}^{n}\log\left[1+\xi\frac{y_i-\mu}{\sigma}\right] - \sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right] \\ &= 0 + (1+\frac{1}{\xi})\frac{\xi}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1} - \frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\left(\frac{\xi}{\sigma}\right) \\ &= \frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})} \\ &= \frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-(1+\frac{1}{\xi})} \end{aligned}$$

Then, w.r.t. σ is:

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left[-n\log\sigma-(1+\frac{1}{\xi})\sum_{i=1}^{n}\log\left[1+\xi\frac{y_i-\mu}{\sigma}\right]-\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right]\\
&= -\frac{n}{\sigma}+(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}\xi\frac{y_i-\mu}{\sigma^2}\\
&\quad -\frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\xi\frac{y_i-\mu}{\sigma^2}\\
&= -\frac{n}{\sigma}+\frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}\frac{y_i-\mu}{\sigma}-\frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\\
&= -\frac{n}{\sigma}+\frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1}z_i-\frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i
\end{aligned}
$$

Finally, w.r.t. ξ is:

$$
\begin{aligned}
\frac{\partial l}{\partial \xi} &= \frac{\partial}{\partial \xi}\left[-n\log\sigma-(1+\frac{1}{\xi})\sum_{i=1}^{n}\log(1+\xi z_i)-\sum_{i=1}^{n}(1+\xi z_i)^{-\frac{1}{\xi}}\right]\\
&= \frac{\partial}{\partial \xi}\left[-n\log\sigma-(1+\frac{1}{\xi})\sum_{i=1}^{n}\log(1+\xi z_i)-\sum_{i=1}^{n}e^{-\frac{1}{\xi}\log(1+\xi z_i)}\right]\\
&= 0-(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1}z_i+\frac{1}{\xi^2}\sum_{i=1}^{n}\log(1+\xi z_i)\\
&\quad -\sum_{i=1}^{n}e^{-\frac{1}{\xi}\log(1+\xi z_i)}\left[-\frac{1}{\xi}\left[1+\xi z_i\right]^{-1}z_i+\frac{1}{\xi^2}\log(1+\xi z_i)\right]\\
&= \frac{1}{\xi^2}\sum_{i=1}^{n}\log(1+\xi z_i)-(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1}z_i\\
&\quad -\frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\log(1+\xi z_i)+\frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i
\end{aligned}
$$

Now, differentiating the log-likelihood function w.r.t. μ twice yields:

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \mu^2} &= \frac{\partial}{\partial \mu}\left[\frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\right] \\
&= \frac{\xi+1}{\sigma}\frac{\xi}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-2} - \frac{1}{\sigma}(1+\frac{1}{\xi})\frac{\xi}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(2+\frac{1}{\xi})} \\
&= \frac{\xi(\xi+1)}{\sigma^2}\sum_{i=1}^{n}[1+\xi z_i]^{-2} - \frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}[1+\xi z_i]^{-(2+\frac{1}{\xi})}
\end{aligned}
$$

And w.r.t. σ twice yields:

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma}\left[-\frac{n}{\sigma} + \frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}\frac{y_i-\mu}{\sigma}\right. \\
&\quad \left. -\frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\right]
\end{aligned}
$$

$$
\begin{aligned}
1^{st} \ \& \ 2^{nd} \text{ terms} &= \frac{\partial}{\partial\sigma}\left[-\frac{n}{\sigma}+\frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}\frac{y_i-\mu}{\sigma}\right]\\
&= \frac{\partial}{\partial\sigma}\left[-\frac{n}{\sigma}+\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}(y_i-\mu)\right]\\
&= \frac{n}{\sigma^2}+\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-2}\left[\xi\frac{y_i-\mu}{\sigma}\right]^{2}\\
&\quad -2\frac{\xi+1}{\sigma^3}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}(y_i-\mu)\\
&= \frac{n}{\sigma^2}+\frac{\xi^2(\xi+1)}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-2}z_i^2-2\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1}z_i\\
3^{rd} \text{ term} &= \frac{\partial}{\partial\sigma}\left[-\frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\right]\\
&= -\frac{1}{\sigma^2}(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(2+\frac{1}{\xi})}\left[\xi\frac{y_i-\mu}{\sigma}\right]^{2}\\
&\quad +\frac{2}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\\
&= -\frac{1}{\sigma^2}(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(2+\frac{1}{\xi})}\left[\xi z_i\right]^{2}+\frac{2}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\sigma^2} &= \frac{n}{\sigma^2}+\frac{\xi^2(\xi+1)}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-2}z_i^2-2\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1}z_i\\
&\quad -\frac{1}{\sigma^2}(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(2+\frac{1}{\xi})}\left[\xi z_i\right]^{2}+\frac{2}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i
\end{aligned}
$$

Then, w.r.t. ξ two times:

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \xi^2} &= \frac{\partial}{\partial \xi}\left[\frac{1}{\xi^2}\sum_{i=1}^{n}\log(1+\xi z_i) - (1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1} z_i\right. \\
&\qquad \left. -\frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\log(1+\xi z_i) + \frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})} z_i\right]
\end{aligned}
$$

$$
\begin{aligned}
1^{st}\&\ 2^{nd}\ \text{terms} &= \frac{\partial}{\partial \xi}\left[\frac{1}{\xi^2}\sum_{i=1}^{n}\log(1+\xi z_i) - (1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1} z_i\right] \\
&= \frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1} z_i - \frac{2}{\xi^3}\sum_{i=1}^{n}\log(1+\xi z_i) \\
&\quad +(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-2} z_i^2 + \frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1} z_i
\end{aligned}
$$

$$
\begin{aligned}
3^{rd}\ \text{term} &= \frac{\partial}{\partial \xi}\left[-\frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\log(1+\xi z_i)\right] \\
&= \frac{\partial}{\partial \xi}\left[-\frac{1}{\xi^2}\sum_{i=1}^{n} e^{-\frac{1}{\xi}\log(1+\xi z_i)}\log(1+\xi z_i)\right] \\
&= -\frac{1}{\xi^2}\sum_{i=1}^{n} e^{-\frac{1}{\xi}\log(1+\xi z_i)}\left[1+\xi z_i\right]^{-1} z_i \\
&\quad -\frac{1}{\xi^2}\sum_{i=1}^{n}\log(1+\xi z_i) e^{-\frac{1}{\xi}\log(1+\xi z_i)}\left[-\frac{1}{\xi}\left[1+\xi z_i\right]^{-1} z_i + \frac{1}{\xi^2}\log(1+\xi z_i)\right] \\
&\quad +\frac{2}{\xi^3}\sum_{i=1}^{n} e^{-\frac{1}{\xi}\log(1+\xi z_i)}\log(1+\xi z_i) \\
&= -\frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})} z_i + \frac{1}{\xi^3}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})} z_i \log(1+\xi z_i) \\
&\quad -\frac{1}{\xi^4}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\left[\log(1+\xi z_i)\right]^2 + \frac{2}{\xi^3}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\log(1+\xi z_i)
\end{aligned}
$$

$$
\begin{aligned}
4^{th}\text{ term} &= \frac{\partial}{\partial\xi}\left[\frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i\right]\\
&= \frac{\partial}{\partial\xi}\left[\frac{1}{\xi}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}z_i\right]\\
&= \frac{1}{\xi}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}z_i\left[-(1+\frac{1}{\xi})\left[1+\xi z_i\right]^{-1}z_i+\frac{1}{\xi^2}\log(1+\xi z_i)\right]\\
&\quad -\frac{1}{\xi^2}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}z_i\\
&= -(1+\frac{1}{\xi})\frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(2+\frac{1}{\xi})}z_i^2+\frac{1}{\xi^3}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i\log(1+\xi z_i)\\
&\quad -\frac{1}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\xi^2} &= \frac{2}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-1}z_i-\frac{2}{\xi^3}\sum_{i=1}^{n}\log(1+\xi z_i)+(1+\frac{1}{\xi})\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-2}z_i^2\\
&\quad -\frac{2}{\xi^2}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i+\frac{2}{\xi^3}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(1+\frac{1}{\xi})}z_i\log(1+\xi z_i)\\
&\quad -\frac{1}{\xi^4}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\left[\log(1+\xi z_i)\right]^2+\frac{2}{\xi^3}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-\frac{1}{\xi}}\log(1+\xi z_i)\\
&\quad -(1+\frac{1}{\xi})\frac{1}{\xi}\sum_{i=1}^{n}\left[1+\xi z_i\right]^{-(2+\frac{1}{\xi})}z_i^2
\end{aligned}
$$

Now, the second derivative of the log-likelihood function w.r.t. σ first, then w.r.t. μ is:

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\mu\partial\sigma} &= \frac{\partial}{\partial\mu}\left[-\frac{n}{\sigma}+\frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-1}\frac{y_i-\mu}{\sigma}\right.\\
&\quad\left.-\frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\right]\\
&= \frac{\partial}{\partial\mu}\left[-\frac{n}{\sigma}+\frac{\xi+1}{\sigma}\sum_{i=1}^{n}\left[\left[\frac{y_i-\mu}{\sigma}\right]^{-1}+\xi\right]^{-1}\right.\\
&\quad\left.-\frac{1}{\sigma}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\right]\\
&= -\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[\left[\frac{y_i-\mu}{\sigma}\right]^{-1}+\xi\right]^{-2}\left[\frac{y_i-\mu}{\sigma}\right]^{-2}+\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\\
&\quad-\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(2+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}\\
&= -\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-2}+\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(1+\frac{1}{\xi})}\\
&\quad-\frac{\xi+1}{\sigma^2}\sum_{i=1}^{n}\left[1+\xi\frac{y_i-\mu}{\sigma}\right]^{-(2+\frac{1}{\xi})}\frac{y_i-\mu}{\sigma}
\end{aligned}
$$

And w.r.t. ξ first, then w.r.t. μ is:

$$\begin{aligned}
\frac{\partial^2 l}{\partial\xi\partial\mu} &= \frac{\partial}{\partial\xi}\left[\frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-(1+\frac{1}{\xi})}\right] \\
&= \frac{\partial}{\partial\xi}\left[\frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}\right] \\
&= -\frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-2}\,z_i + \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1} \\
&\quad -\frac{1}{\sigma}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}\left[-(1+\frac{1}{\xi})\,[1+\xi z_i]^{-1}\,z_i + \frac{1}{\xi^2}\log(1+\xi z_i)\right] \\
&= -\frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-2}\,z_i + \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1} \\
&\quad +\frac{1}{\sigma}(1+\frac{1}{\xi})\sum_{i=1}^{n}[1+\xi z_i]^{-(2+\frac{1}{\xi})}\,z_i - \frac{1}{\sigma\xi^2}\sum_{i=1}^{n}[1+\xi z_i]^{-(1+\frac{1}{\xi})}\log(1+\xi z_i)
\end{aligned}$$

Finally, w.r.t. to σ then for ξ yields:

$$\begin{aligned}
\frac{\partial^2 l}{\partial\xi\partial\sigma} &= \frac{\partial}{\partial\xi}\left[-\frac{n}{\sigma} + \frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1}\,z_i - \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-(1+\frac{1}{\xi})}\,z_i\right] \\
&= \frac{\partial}{\partial\xi}\left[-\frac{n}{\sigma} + \frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1}\,z_i - \frac{1}{\sigma}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}z_i\right] \\
&= 0 - \frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-2}\,z_i^2 + \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1}\,z_i \\
&\quad -\frac{1}{\sigma}\sum_{i=1}^{n}e^{-(1+\frac{1}{\xi})\log(1+\xi z_i)}z_i\left[-(1+\frac{1}{\xi})\,[1+\xi z_i]^{-1}\,z_i + \frac{1}{\xi^2}\log(1+\xi z_i)\right] \\
&= -\frac{\xi+1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-2}\,z_i^2 + \frac{1}{\sigma}\sum_{i=1}^{n}[1+\xi z_i]^{-1}\,z_i \\
&\quad +\frac{1}{\sigma}(1+\frac{1}{\xi})\sum_{i=1}^{n}[1+\xi z_i]^{-(2+\frac{1}{\xi})}\,z_i^2 - \frac{1}{\sigma\xi^2}\sum_{i=1}^{n}[1+\xi z_i]^{-(1+\frac{1}{\xi})}\,z_i\log(1+\xi z_i)
\end{aligned}$$

Since the Hessian matrix is symmetric (i.e., $H = H^T$), therefore, we don't need to derive the remaining three elements. All derived formulas are used in `R` code to obtain the MLE's of the GEVD according to the Newton's method (equation 3.5).

4.2 Validating the Law of Total Probability on Hau et al. paper [30]

In their simulation study, Hua et al. [30] generated an interaction effect between two of the ten simulated biallelic genetic factors (table 4.1). The association was generated so that no main effect is appreciated.

Table 4.1: The original penetrance of the two factors suggested by Hua et al. [30]

		Factor A		
		AA	Aa	aa
Factor B	BB	ϕK	$(1+\frac{p_1}{2q_1}(1-\phi))K$	K
	Bb	$(1+\frac{p_1}{2q_1}(1-\phi))K$	$(1-\frac{p_1p_2}{4q_1q_2}(1-\phi))K$	K
	bb	K	K	K

* p_1, p_2, q_1, and q_2 are the minor and the major allele frequencies of factors A and B, respectively; K is the proportion of individuals with the disease; and ϕ is a tuning parameter [30].

Since the entries of table 4.1 represent the disease penetrance on the population, the total probability of acquiring the disease has to add up to K, which was defined as the population prevalence. Given that the authors assigned p_1 and p_2 as the minor allele frequencies of factors A and B, respectively; then, per the Law of Total Probability, the probability of having the disease ($P(D)$) can be calculated as follows:

$$P(D) = \sum_{i\in\{AA,Aa,aa\}} \sum_{j\in\{BB,Bb,bb\}} P(D|ij)P(ij) \qquad (4.1)$$

where the conditional probabilities of disease given a specific multilocus combination ($P(D|ij)'s$) are given in table 4.1 listed earlier; whereas, the joint probabilities of multilocus combinations ($P(D|ij)P(ij)'s$) are defined in table 4.2 below.

Table 4.2: The joint probabilities of disease and multilocus combinations of the two factors per the definition of the authors [30]

		Factor A		
		AA	Aa	aa
Factor B	BB	$q_1^2q_2^2\phi K$	$2p_1q_1q_2^2(1+\frac{p_1}{2q_1}(1-\phi))K$	$p_1^2q_2^2K$
	Bb	$2q_1^2p_2q_2(1+\frac{p_1}{2q_1}(1-\phi))K$	$4p_1q_1p_2q_2(1-\frac{p_1p_2}{4q_1q_2}(1-\phi))K$	$2p_1^2p_2q_2K$
	bb	$q_1^2p_2^2K$	$2p_1q_1p_2^2K$	$p_1^2p_2^2K$

* Refer to table 4.1 for details.

However, the $P(D)$ defined in equation 4.1 cannot add up to K with the given penetrance in table 4.1. We can simply show the contradiction by substituting any set of values of p_1, p_2, and ϕ. For instance, $K^{-1}P(D) = 0.595$ when $p_1 = p_2 = 0.05$ and $\phi = 0.5$, while it should be sum up to 1. Therefore, we proposed an alteration (tables 4.3 and 4.4) to the penetrance provided in the paper that should fix the imbalances in table 4.1.

Table 4.3: The suggested penetrance of the two factors

		Factor A		
		AA	Aa	aa
Factor B	BB	$(1+\phi p_1p_2)K$	$(1-\phi\frac{q_1p_2}{2})K$	K
	Bb	$(1-\phi\frac{p_1q_2}{2})K$	$(1+\phi\frac{q_1q_2}{4})K$	K
	bb	K	K	K

* Refer to table 4.1 for details.

Table 4.4: The joint probabilities of disease and multilocus combinations of the two factors per the definition of the authors [30] and our suggested penetrance

		Factor A		
		AA	Aa	aa
Factor B	BB	$q_1^2q_2^2(1+\phi p_1p_2)K$	$2p_1q_1q_2^2(1-\phi\frac{q_1p_2}{2})K$	$p_1^2q_2^2K$
	Bb	$2q_1^2p_2q_2(1-\phi\frac{p_1q_2}{2})K$	$4p_1q_1p_2q_2(1+\phi\frac{q_1q_2}{4})K$	$2p_1^2p_2q_2K$
	bb	$q_1^2p_2^2K$	$2p_1q_1p_2^2K$	$p_1^2p_2^2K$

* Refer to table 4.1 for details.

Now, the joint probabilities of disease and multilocus combinations are balanced, and they sum up to K as we can see from the vitrification listed below:

$$\begin{aligned}
P(D) &= \sum_{i\in\{AA,Aa,aa\}} \sum_{j\in\{BB,Bb,bb\}} P(D|ij)P(ij) \\
&= K\Big[q_1^2q_2^2(1+\phi p_1p_2) + 2q_1^2p_2q_2(1-\phi\frac{p_1q_2}{2}) + q_1^2p_2^2 \\
&\quad +2p_1q_1q_2^2(1-\phi\frac{q_1p_2}{2}) + 4p_1q_1p_2q_2(1+\phi\frac{q_1q_2}{4}) \\
&\quad +2p_1q_1p_2^2 + p_1^2q_2^2 + 2p_1^2p_2q_2 + p_1^2p_2^2\Big] \\
\Longrightarrow\ K^{-1}P(D) &= q_1^2q_2^2 + \cancel{\phi p_1q_1^2p_2q_2^2} + 2q_1^2p_2q_2 - \cancel{\phi p_1q_1^2p_2q_2^2} + q_1^2p_2^2 \\
&\quad +2p_1q_1q_2^2 - \cancel{\phi p_1q_1^2p_2q_2^2} + 4p_1q_1p_2q_2 + \cancel{\phi p_1q_1^2p_2q_2^2} \\
&\quad +2p_1q_1p_2^2 + p_1^2q_2^2 + 2p_1^2p_2q_2 + p_1^2p_2^2
\end{aligned}$$

Now, recall that $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. So:

$$\begin{aligned}
K^{-1}P(D) &= (1-p_1)^2(1-p_2)^2 + 2(1-p_1)^2p_2(1-p_2) + (1-p_1)^2p_2^2 \\
&\quad +2p_1(1-p_1)(1-p_2)^2 + 4p_1(1-p_1)p_2(1-p_2) + 2p_1(1-p_1)p_2^2 \\
&\quad +p_1^2(1-p_2)^2 + 2p_1^2p_2(1-p_2) + p_1^2p_2^2 \\
&= \underset{P(AABB)}{\underline{1 - \overset{(1)}{2p_2} + \overset{(2)}{p_2^2} - \overset{(3)}{2p_1} + \overset{(4)}{4p_1p_2} - \overset{(5)}{2p_1p_2^2} + \overset{(6)}{p_1^2} - \overset{(7)}{2p_1^2p_2} + \overset{(8)}{p_1^2p_2^2}}} \\
&\quad + \underset{P(AABb)}{\underline{\overset{(1)}{2p_2} - \overset{(2)}{2p_2^2} - \overset{(4)}{4p_1p_2} + \overset{(5)}{4p_1p_2^2} + \overset{(7)}{2p_1^2p_2} - \overset{(8)}{2p_1^2p_2^2}}} \\
&\quad + \underset{P(AAbb)}{\underline{\overset{(2)}{p_2^2} - \overset{(5)}{2p_1p_2^2} + \overset{(8)}{p_1^2p_2^2}}} + \underset{P(AaBB)}{\underline{\overset{(3)}{2p_1} - \overset{(6)}{2p_1^2} - \overset{(4)}{4p_1p_2} + \overset{(7)}{4p_1^2p_2} + \overset{(5)}{2p_1p_2^2} - \overset{(8)}{2p_1^2p_2^2}}} \\
&\quad + \underset{P(AaBb)}{\underline{\overset{(4)}{4p_1p_2} - \overset{(5)}{4p_1p_2^2} - \overset{(7)}{4p_1^2p_2} + \overset{(8)}{4p_1^2p_2^2}}} + \underset{P(Aabb)}{\underline{\overset{(5)}{2p_1p_2^2} - \overset{(8)}{2p_1^2p_2^2}}} \\
&\quad + \underset{P(aaBB)}{\underline{\overset{(6)}{p_1^2} - \overset{(7)}{2p_1^2p_2} + \overset{(8)}{p_1^2p_2^2}}} + \underset{P(aaBb)}{\underline{\overset{(7)}{2p_1^2p_2} - \overset{(8)}{2p_1^2p_2^2}}} + \underset{P(aabb)}{\underline{\overset{(8)}{p_1^2p_2^2}}} \\
&= 1 \\
\Longrightarrow\ P(D) &= K
\end{aligned}$$

4.3 Theorem: Ordered Combinatorial Partitioning in OQMDR

Assume that we have a data set of size n with a continuous response variable Y and a single categorical covariate with three levels. Therefore, there are three possible Combinatorial Partitionings that can be applied to Y, $\{1\}$ versus $\{2,\ 3\}$, $\{1,\ 2\}$ versus $\{3\}$, and $\{1,\ 3\}$ versus $\{2\}$. Let n_i and $\bar{Y}_i$; for $i = 1, 2, 3$, be the sample size and the arithmetic mean of the data from the i^{th} level (or category) of the covariate, respectively. Now, without loss of generality, let $\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3$; then, per the Ordered Combinatorial Partitioning principle, $t_{2|13} < \max(t_{3|12}, t_{23|1})$, given that:

$$\begin{aligned}
t_{2|13} &= \frac{\bar{Y}_2 - \bar{Y}_{13}}{\sqrt{\frac{S_2^2}{n_2} + \frac{S_{13}^2}{n_{13}}}} \\
t_{3|12} &= \frac{\bar{Y}_3 - \bar{Y}_{12}}{\sqrt{\frac{S_3^2}{n_3} + \frac{S_{12}^2}{n_{12}}}} \\
t_{23|1} &= \frac{\bar{Y}_{23} - \bar{Y}_1}{\sqrt{\frac{S_{23}^2}{n_{23}} + \frac{S_1^2}{n_1}}}
\end{aligned}$$

where n_{ij}, $\bar{Y}_{ij}$, and S_{ij}^2 represent the sample size, the average, and the variance of the combined data from the i^{th} and the j^{th} level of the covariate, respectively.

Proof:

For a more tractable situation, let $n_1 = n_2 = n_3 = \bar{n}$. Also assume that all group variances are equal and known, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$. Hence:

$$\begin{aligned}
t_{2|13} &= \frac{\bar{Y}_2 - \bar{Y}_{13}}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{S_{13}^2}{n_{13}}}} \\
t_{3|12} &= \frac{\bar{Y}_3 - \bar{Y}_{12}}{\sqrt{\frac{\sigma_3^2}{n_3} + \frac{S_{12}^2}{n_{12}}}} \\
t_{23|1} &= \frac{\bar{Y}_{23} - \bar{Y}_1}{\sqrt{\frac{\sigma_{23}^2}{n_{23}} + \frac{S_1^2}{n_1}}}
\end{aligned}$$

where:

$$\begin{aligned}
\bar{Y}_{ij} &= \frac{n_i\bar{Y}_i + n_j\bar{Y}_j}{n_i + n_j} \\
&= \frac{\bar{n}(\bar{Y}_i + \bar{Y}_j)}{2\bar{n}} \\
&= \frac{\bar{Y}_i + \bar{Y}_j}{2}
\end{aligned}$$

and

$$\begin{aligned}
S_{ij}^2 &= \frac{(n_i - 1)\sigma_i^2 + (n_j - 1)\sigma_j^2 + n_i\bar{Y}_i^2 + n_j\bar{Y}_j^2 - (n_i + n_j)\bar{Y}_{ij}^2}{n_1 + n_2 - 1} \\
&= \frac{(\bar{n} - 1)\sigma^2 + (\bar{n} - 1)\sigma^2 + \bar{n}\bar{Y}_i^2 + \bar{n}\bar{Y}_j^2 - (\bar{n} + \bar{n})\left(\frac{\bar{Y}_i + \bar{Y}_j}{2}\right)^2}{\bar{n} + \bar{n} - 1} \\
&= \frac{2(\bar{n} - 1)\sigma^2 + \bar{n}\bar{Y}_i^2 + \bar{n}\bar{Y}_j^2 - 2\bar{n}\frac{\bar{Y}_i^2 + \bar{Y}_j^2 + 2\bar{Y}_i\bar{Y}_j}{4}}{2\bar{n} - 1} \\
&= \frac{2(\bar{n} - 1)\sigma^2 + \bar{n}\bar{Y}_i^2 + \bar{n}\bar{Y}_j^2 - \frac{\bar{n}}{2}\bar{Y}_i^2 - \frac{\bar{n}}{2}\bar{Y}_j^2 - \bar{n}\bar{Y}_i\bar{Y}_j}{2\bar{n} - 1} \\
&= \frac{2(\bar{n} - 1)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_i^2 + \bar{Y}_j^2 - 2\bar{Y}_i\bar{Y}_j)}{2\bar{n} - 1} \\
&= \frac{2(\bar{n} - 1)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_i - \bar{Y}_j)^2}{2\bar{n} - 1}
\end{aligned}$$

Now:

$$\begin{aligned}
t_{2|13} &= \frac{\bar{Y}_2 - \bar{Y}_{13}}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{S_{13}^2}{n_{13}}}} \\
&= \frac{\bar{Y}_2 - \frac{\bar{Y}_1 + \bar{Y}_3}{2}}{\sqrt{\frac{\sigma^2}{\bar{n}} + \frac{2(\bar{n}-1)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_1 - \bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\
&= \frac{\bar{Y}_2 - \frac{\bar{Y}_1 + \bar{Y}_3}{2}}{\sqrt{\frac{2(2\bar{n}-1)\sigma^2 + 2(\bar{n}-1)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_1 - \bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\
&= \frac{\bar{Y}_2 - \frac{\bar{Y}_1 + \bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_1 - \bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}}
\end{aligned}$$

Similarly:

$$t_{3|12} = \frac{\bar{Y}_3 - \frac{\bar{Y}_1 + \bar{Y}_2}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_1 - \bar{Y}_2)^2}{2\bar{n}(2\bar{n}-1)}}}$$

$$t_{23|1} = \frac{\frac{\bar{Y}_2 + \bar{Y}_3}{2} - \bar{Y}_1}{\sqrt{\frac{(6\bar{n}-4)\sigma^2 + \frac{\bar{n}}{2}(\bar{Y}_2 - \bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}}$$

Notice that:

$$\begin{aligned} t_{2|13} &= \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\ &< \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_2-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \end{aligned}$$

follows by $\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3$

Therefore:

$$t_{23|1} - t_{2|13} > t_{23|1} - \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_2-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}}$$

$$\begin{aligned} \Longrightarrow t_{23|1} - t_{2|13} &> \frac{\frac{\bar{Y}_2+\bar{Y}_3}{2} - \bar{Y}_1}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_2-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} - \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_2-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\ &= \frac{\frac{\bar{Y}_2+\bar{Y}_3}{2} - \bar{Y}_1 - \bar{Y}_2 + \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_2-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\ &= \frac{\bar{Y}_3 - \frac{\bar{Y}_1+\bar{Y}_2}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_2-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\ &> 0 \end{aligned}$$

again, follows by $\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3$

Similarly:

$$\begin{aligned} t_{2|13} &= \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_3)^2}{2\bar{n}(2\bar{n}-1)}}} \\ &< \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_2)^2}{2\bar{n}(2\bar{n}-1)}}} \end{aligned}$$

follows by $\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3$

Therefore:

$$t_{3|12} - t_{2|13} > t_{3|12} - \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_2)^2}{2\bar{n}(2\bar{n}-1)}}}$$

$$\begin{aligned}
\implies t_{3|12} - t_{2|13} &> \frac{\bar{Y}_3 - \frac{\bar{Y}_1+\bar{Y}_2}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_2)^2}{2\bar{n}(2\bar{n}-1)}}} - \frac{\bar{Y}_2 - \frac{\bar{Y}_1+\bar{Y}_3}{2}}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_2)^2}{2\bar{n}(2\bar{n}-1)}}} \\
&= \frac{\frac{3}{2}(\bar{Y}_3 - \bar{Y}_2)}{\sqrt{\frac{(6\bar{n}-4)\sigma^2+\frac{\bar{n}}{2}(\bar{Y}_1-\bar{Y}_2)^2}{2\bar{n}(2\bar{n}-1)}}} \\
&> 0
\end{aligned}$$

again, follows by $\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3$

Hence, $t_{2|13} < \max(t_{3|12}, t_{23|1})$.

In fact, under the aforementioned restriction about the equality of group sizes and variances we assumed at the beginning of the proof, we achieved a stronger conclusion than we claimed. That is, we just proved that $t_{2|13} < \min(t_{3|12}, t_{23|1})$.

Copyright© Zaid T. Al-Khaledi, 2019.

Chapter 5 Real data analysis

5.1 Alzheimer's Disease (AD) overview

Since the beginning of the past century, life expectancy has dramatically increased in the United States. The United States Life Tables of 2004 show a steady positive trend from 1900 to 2004 where life expectancy jumped from 49.24 years in 1900 to 68.07 years in 1950, then to 77.8 years in 2004 (a total of 28.56 years increase) [4]. As a result, age-related medical conditions became more frequent than before, especially neurodegenerative[1] issues [52]. Alzheimer's Disease (AD) is one of these neurodegenerative disorders that affects about 10% of people aged 65+ years [2]. The disease, which was first identified by the German psychiatrist Alois Alzheimer in 1906 [28], progresses over time causing many mental and physical health complications to patients. Soon after it occurs, AD causes brain cell loss, which leads to brain size shrinkage, which in turn reduces a patient's brain capability to function normally. Consequently, AD patients could face short term memory impairment, talking difficulties, struggling to remember well-known people and places, problems accomplishing daily living and self-care activities, and eventually mental disability and dementia [45]. Thus far, no medical treatment has been proven to help to reverse or suppress Alzheimer's Disease from advancing to late stage; however, some treatment might help reducing the symptoms of AD [3].

The majority of Alzheimer's Disease cases (about 95%) occur after age 65 (late-onset). The remaining 5% of the cases occur in younger people, often after age 30. On the other hand, about 75% of Alzheimer's Disease cases occur sporadically (only one patient in a family), which is known as Sporadic Alzheimer's Disease. Whereas,

---

[1] Neurodegenerative refers to a degeneration of human brain neurons, which results in a degradation in human cognitive functions [27].

25% of the cases are family related (multiple cases in one family), which is, therefore, known as Familial Alzheimer's Disease. Most of the early-onset cases are Familial AD. Both types, sporadic AD and familial AD, are believed to be linked to mutations in certain genes or occurrence of certain combinations of genes. Namely, the early-onset AD is linked to mutations in amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) genes. Whereas the late-onset AD is significantly linked with the apolipoprotein E (ApoE) gene, specifically the ApoE-ε4 allele [40, 23].

Besides symptoms, patients of Alzheimer's Disease typically develop some other biochemical characteristics, for instance, significant elevations in the level of the cerebrospinal fluid tau (CSF) and the urine neuronal thread protein (NTP) in patients with AD compared to controls [35]. Many studies have been conducted to investigate connections between AD and genetics [6]. The majority of the studies modeled the relationship between specific gene information and certain measured indicators of AD or comparing the gene expressions in case versus control groups.

In our research, we are interested in investigating an effect of combinations of genetic factors on some continuous response. Therefore, Alzheimer's Disease data set with a continuous biochemical marker would satisfy our need. In fact, we are inspecting the connection between three different continuous measures of cognition impairment and a set of bio-markers that are linked to AD [37]. The data set is explored in details in the next section.

5.2 Data presentation

As we pointed out earlier, we are trying to discover a relationship between human cognition and some genetic factors in AD patients. Accordingly, we obtained our data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI: http://adni.loni.ucla.edu/) with help from Dr. David Fardo. The size of the original data set is 612 individuals with 32 measured variables; three of them are continuous responses. These responses are cognitive resilience (CRs), cognitive reserve (CRv), and global resilience (GRs). In general cognitive resilience refers to the human ability to overcome or resist the negative impacts of specific life circumstances, like poverty [21], family crises [12], and aging-related issues [54]. However, in our data set, the cognitive variables are outcomes of cognitive performance evaluation compared to what is expected, given the underlying pathology and other AD risk factors. Essentially, they are different residuals for cognitive performance after adjusting in various ways [20]. Subjects underwent various neuropsychological assessments, and an overall memory score was derived [9]. The rest of the variables provide information about some environmental (or biological) markers and genetic information.

The effects of all variables but the genetics (except the ApoE-ε4) have been regressed out through multiple linear regression models, and hence the three responses in the given data set represent residuals from each fitted model [20]. Thus, we eliminated all non-gene markers and the ApoE-ε4 factor from the data set, which reduces the total number of variables to 25 instead of 32. Besides, since our proposed method does not handle data sets with missing or unreported gene information, we eliminated all variables that contain 40 or more NA entries from the data set. Finally, we perform the analysis on cases with no missing information only. The exclusion reduces the total number of observations to 480 with 15 variables, in which three of them are responses. The entries of all genetic variables are {0, 1, or 2}, which represent allele combinations of each factor. We also use Latin alphabet letters to label all factors for easy presentation. Refer to table 5.1 for details about included and excluded SNPs in the study.

Table 5.1: Genetic-variable list

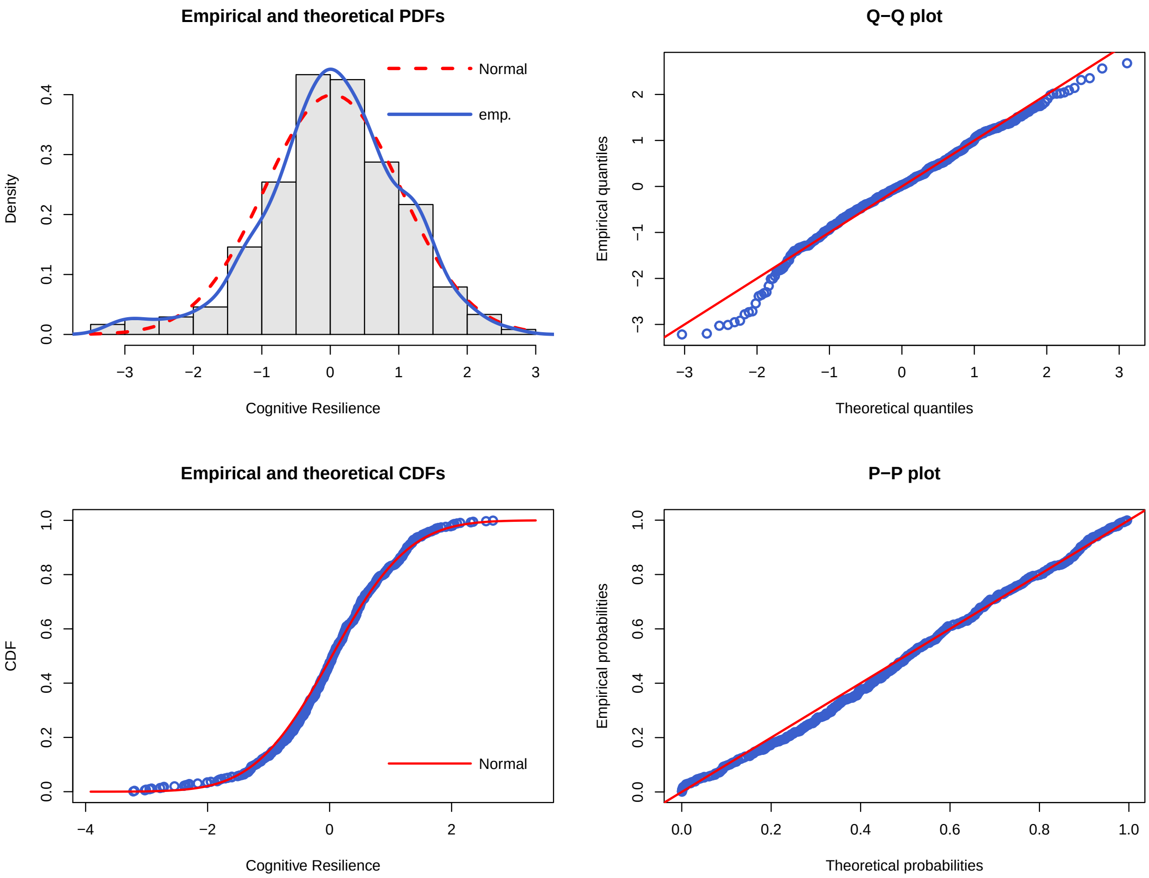
SNP	Gene	Missing	Included	Label
rs6656401	CR1	118	No	—
rs6733839	BIN1	323	No	—
rs35349669	INPP5D	246	No	—
rs190982	MEF2C	114	No	—
rs75932628	TREM2	33	Yes	A
rs10948363	CD2AP	0	Yes	B
rs2718058	NME8	39	Yes	C
rs1476679	ZCWPW1	65	No	—
rs11771145	EPHA1	0	Yes	D
rs28834970	PTK2B	70	No	—
rs9331896	CLU	323	No	—
rs10838725	CELF1	1	Yes	E
rs983392	MS4A6A	131	No	—
rs10792832	PICALM	4	Yes	F
rs11218343	SORL1	7	Yes	G
rs17125944	FERMT2	0	Yes	H
rs17125721	PSEN1	29	Yes	I
rs10498633	SLC24A4	0	Yes	J
rs8093731	DSG2	53	No	—
rs4147929	ABCA7	46	No	—
rs3865444	CD33	0	Yes	K
rs7274581	CASS4	27	Yes	L

Refer to table 5.2 and figures 5.1, 5.2, and 5.3 for data exploration of the three response variables in the data set. Recall that these responses are stored as residuals after regressing out the effects of all non-genetic factors and the ApoE-ε4 from the data set. Also, recall that errors from multiple linear regression models are assumed normally distributed with mean zero and positive standard deviation. Now, from looking at the summary statistics, it seems like all three responses have a mean close to zero and a standard deviation of one. Therefore, it seems like normally-distributed errors is a valid assumption about the responses. However, the medians and the graphical representations suggest that the marginal distributions of the three variables are slightly skewed to the left, which suggests that there might be some non-spotted variation left in the residuals, mainly for the CRv variable. Regardless, this unexplained variation might not be strong enough to be caught, especially for the CRs and GRs variables, in which their distributions are almost symmetric.

Table 5.2: Statistical summary

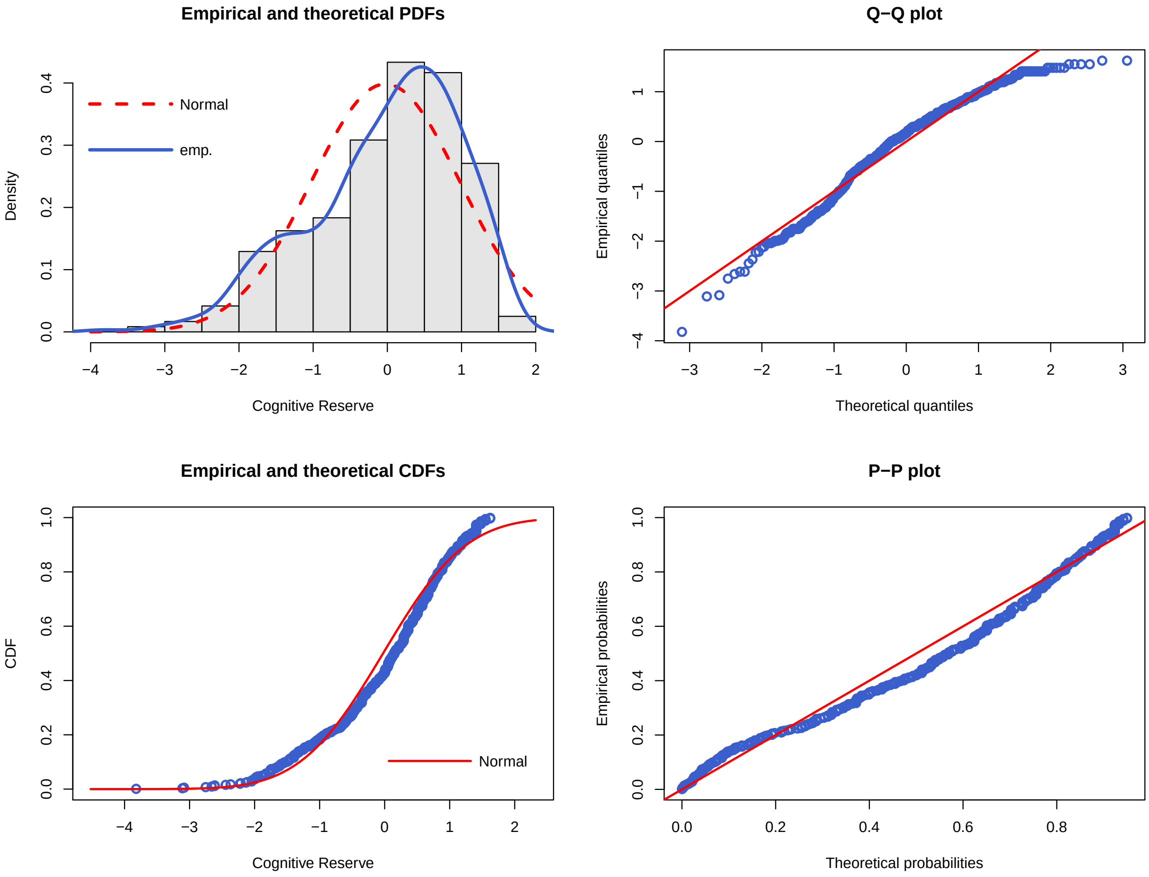
Statistics	Cognitive Resilience	Cognitive Reserve	Global Resilience
n	480	480	480
min	-3.2178	-3.8218	-3.1058
max	2.6790	1.6232	2.4172
range	5.8968	5.4450	5.5230
median	0.0487	0.1493	0.1160
mean	0.0315	-0.0318	0.0012
SD	0.9998	1.0032	0.9885
middle 95%	(-2.3193, 1.8280)	(-2.1106, 1.4083)	(-2.1610, 1.6691)

Figure 5.1: Empirical distribution of Cognitive Resilience compared to Normal distribution



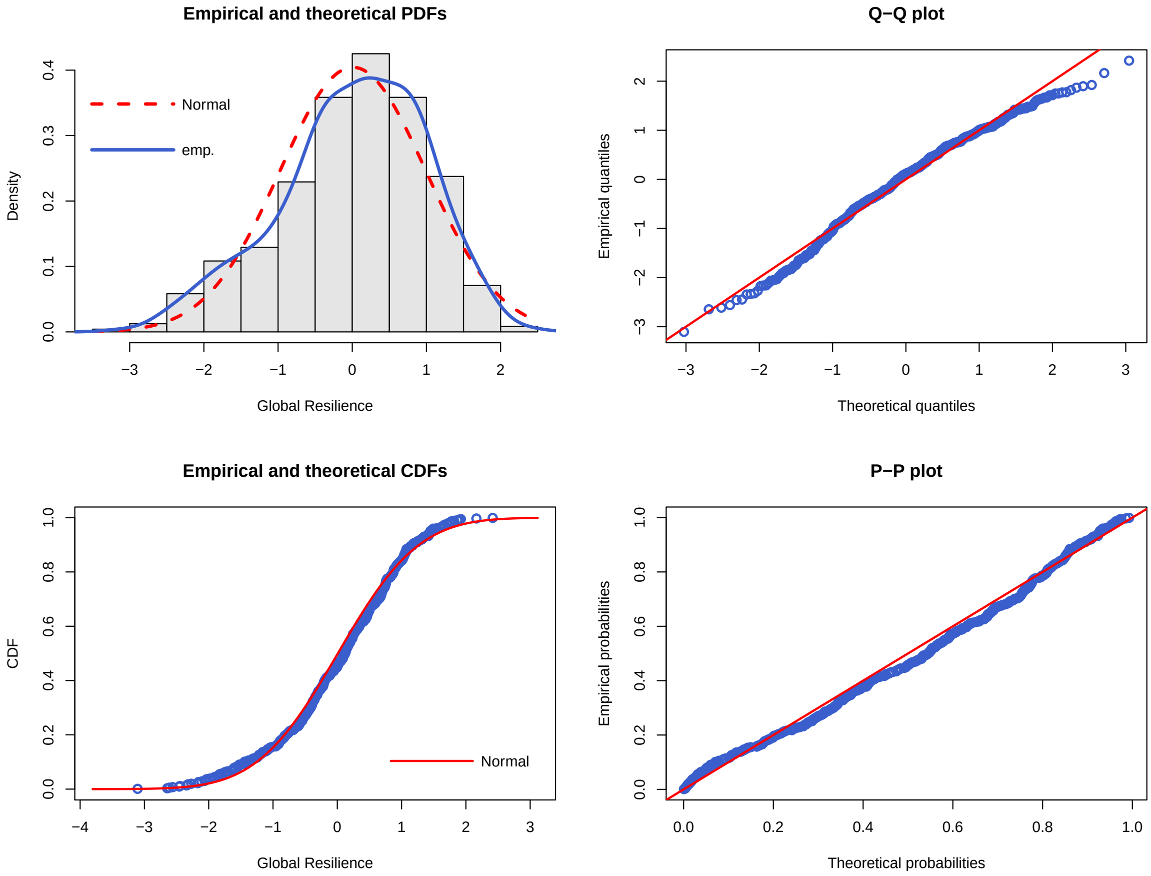
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5.2: Empirical distribution of Cognitive Reserve compared to Normal distribution



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5.3: Empirical distribution of Global Resilience compared to Normal distribution



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

5.3 Data analysis

We applied the OQMDR algorithm on the AD data set separately for each response. All possible 2-way and 3-way interactions between the genetic factors are examined along with their best risk patterns per the OCP procedure. Refer to tables 5.3 and 5.4 for the output of the analysis.

Table 5.3: Model selection of the three cognitive scores

	Best 2-way		Best 3-way		
Response	Model	t_k^*-score	Model	t_k^*-score	$\overline{\text{MSPE}}$
CRs	CL	0.1525	DJL	0.2373	1.0061
CRv	FI	-0.1713	JKL	0.2881	1.0238
GRs	EG	-0.9722	DJL	-0.0774	0.9948

As we can see from table 5.3, the selected 2-way and 3-way models for CRs are (NME8 and CASS4) and (EPHA1, SLC24A4, and CASS4), for CRv are (PICALM and PSEN1) and (SLC24A4, CD33, and CASS4), and for GRs are (CELF1 and SORL1) and (EPHA1, SLC24A4, and CASS4), respectively. Per the OQMDR algorithm, the best final models for all responses are the 3-way interaction (the one that maximizes the testing t-score). The risk patterns for each selected 2-way and 3-way interaction are demonstrated in figures 5.4 and 5.5, respectively. However, it's clear from table 5.4 that none of the proposed models are statistically significant.

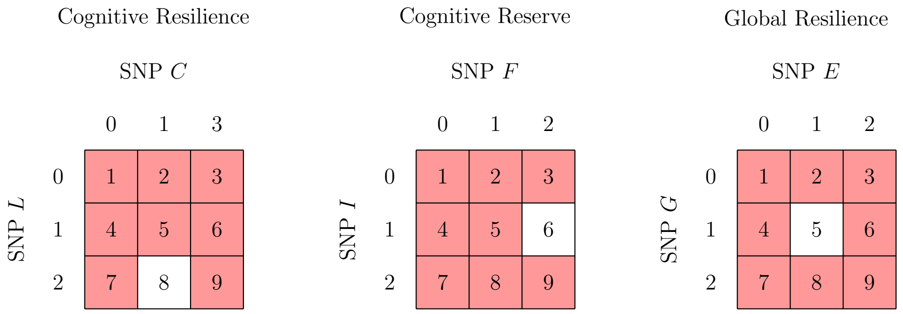
This could be attributed to one or more of the following issues: First, the sample size is relatively small (480) to make it possible to correctly spot a statistically significant interaction. Second, there is little or no true relationship between the responses and the considered factors from the current data set. Third, the eliminated variables might have heavy influence on the response variables, and are no longer accessible because of the elimination. Fourth, there could be a true relationship but the OQMDR algorithm is not able to catch it.

Table 5.4: Proposed model evaluation

	GEVD procedure									Permutation
Response	$t^{*(0)}_{k_{max}}$	$\hat{\mu}_{t^{*(0)}_{k_{max}}}$	$\hat{\sigma}_{t^{*(0)}_{k_{max}}}$	$\hat{\xi}_{t^{*(0)}_{k_{max}}}$	$p^{(0)}_{k_{max}}$	$\hat{\mu}_v$	$\hat{\sigma}_v$	$\hat{\xi}_v$	p_v	p-value
CRs	0.2373	0.2139	0.4841	-0.0612	0.6144	0.1716	0.9719	0.0930	0.5708	0.617
CRv	0.2881	0.2872	0.7864	-0.0984	0.6317	0.0412	0.9165	0.0078	0.6399	0.720
GRs	-0.0774	0.2282	0.5282	-0.2963	0.8483	-0.1221	0.9612	0.0544	0.8872	0.651

* Column headers are defined as in table 3.1.

Figure 5.4: Risk patterns of the proposed 2-way interactions for each response


Cognitive Resilience

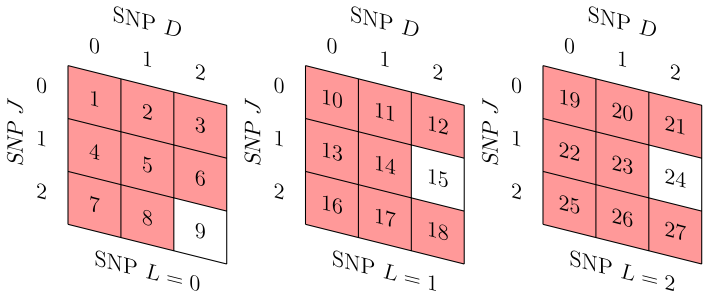
Cognitive Reserve

Global Resilience

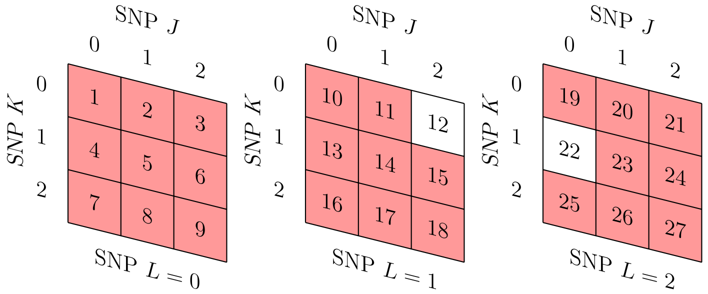
[*] Individuals in highlighted cells are at high risk.

Figure 5.5: Risk patterns of the proposed 3-way interactions for each response

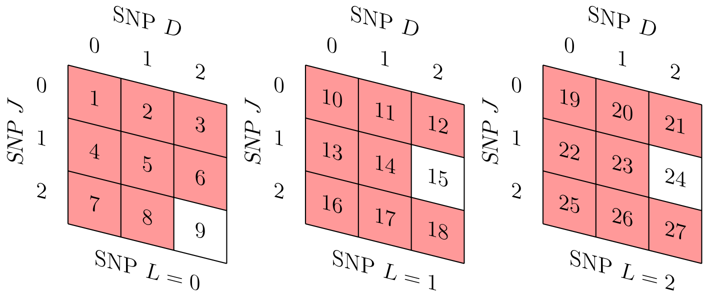
(a) Cognitive Resilience



(b) Cognitive Reserve



(c) Global Resilience



[*] Individuals in highlighted cells are at high risk.

On the other hand, the 3-way models that the OQMDR selected for the CRs and GRs response are the same (EPHA1, SLC24A4, and CASS4), with the exact same risk pattern (see figure 5.5). Also, two factors (SLC24A4, and CASS4) are chosen to form the 3rd-degree interaction for all three responses. This consistency could be a sign of true but weak interactive effect of the three factors on the cognitive impairment scores, especially because we know that (CASS4) is selected as the most important main effect factor (1-way) for all three scores.

Also, a Principal Components (PC) analysis on the three responses is carried out, and two of the PCs have been selected as alternative formulations of cognitive scores. Then, the OQMDR method is applied to the AD data set with the new responses from each selected PCs (See table 5.5). The third degree interaction (EPHA1, SLC24A4, and CASS4), with the same risk pattern that is selected for CRs and GRs (figures 5.5a and 5.5c), has been proposed again as the best final model but it failed to pass the significance assessment once again. Regardless of the insignificance, it's interesting to notice that the output from the PC analysis supports the importance of the selected 3-way model per the analyzed data set.

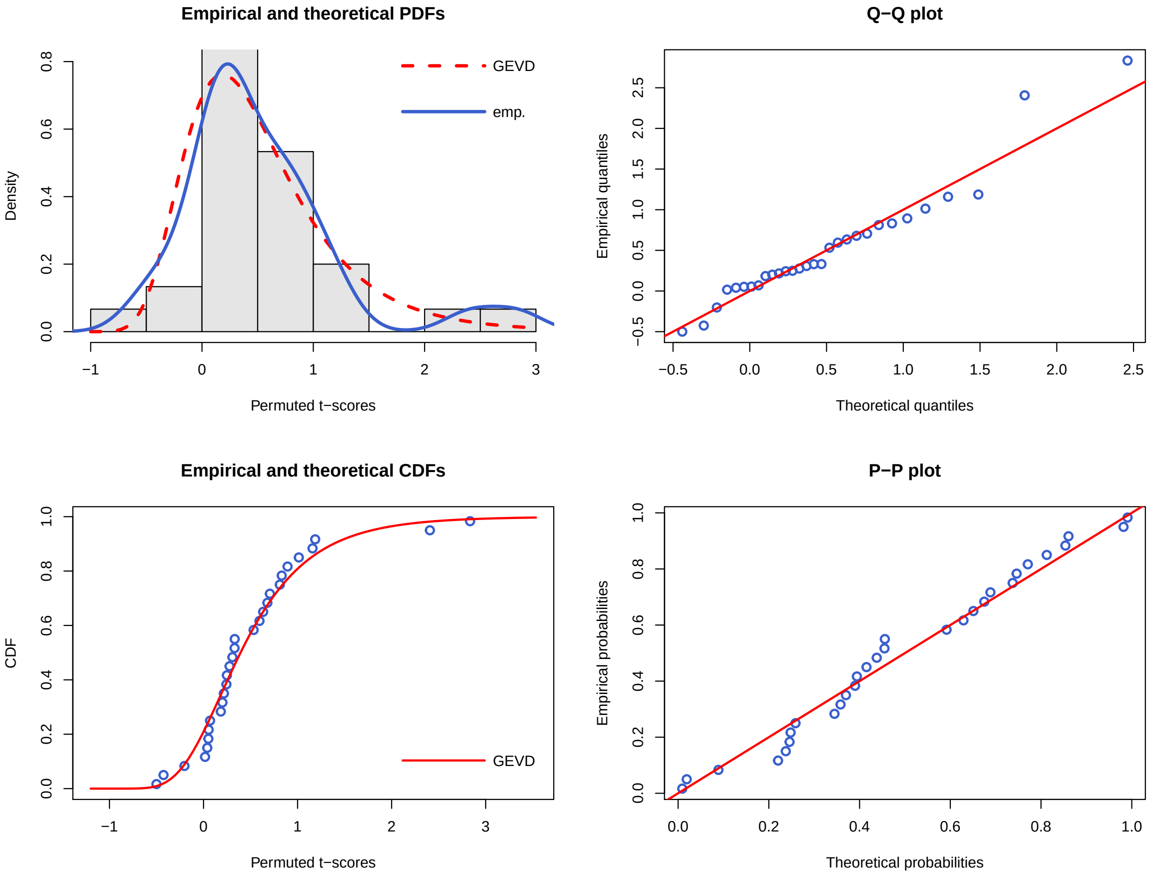
Table 5.5: Principal Components analysis

	CRs	CRv	GRs	SD	Model	t^*_k-score
PC1	0.5397	0.5278	0.6559	1.5046	DJL	0.1474
PC2	-0.6911	0.7227	-0.0129	0.8433	DJL	-0.1059
PC3	0.4808	0.4463	-0.7548	0.0008	—	—

* The rows in columns CRs, CRv, and GRv represent the eigenvectors of the covariance matrix of the matrix of responses.

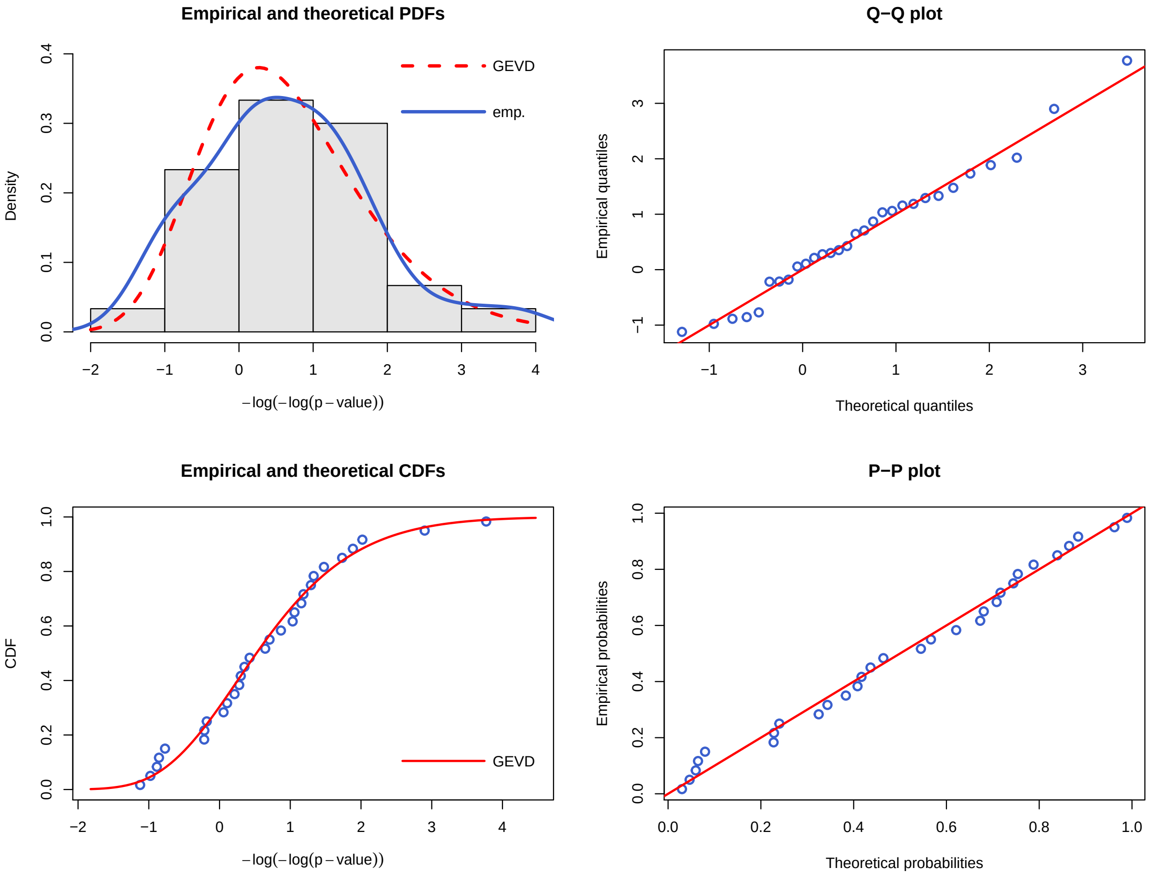
Technically, the OQMDR performed decently to examine all possible interactions. The evaluation process is done using $m = 30$ permuted test statistics and $m_1 = 30$ permuted p-values for each response. The MLE estimation procedure failed to diverge in one of the permuted cases of the CRv; therefore, we approximated the null distribution from $m_1 = 29$ permuted p-values only (see figures 5.8 and 5.9). The graphical presentations show that the GEVD well approximated the distribution of the testing score and the transformed p-value; yet, the transformed p-values and their p-values are close. This suggests that a uniform(0,1) distribution might be a valid assumption about the p-values of the test statistic. Both the GEVD and the permutation testings agree about the non-significance of the proposed models, which could explain why the p-values of the test statistics and their p-values are so close . The evaluations procedure was a bit faster for the GEVD procedure (about 16 hours for the GEVD compared to 17 hours for the regular permutations). Although it's not an enormous enhancement, it might be a positive indication on the GEVD side when there are many interacting factors in the study.

Figure 5.6: Cognitive Resilience; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



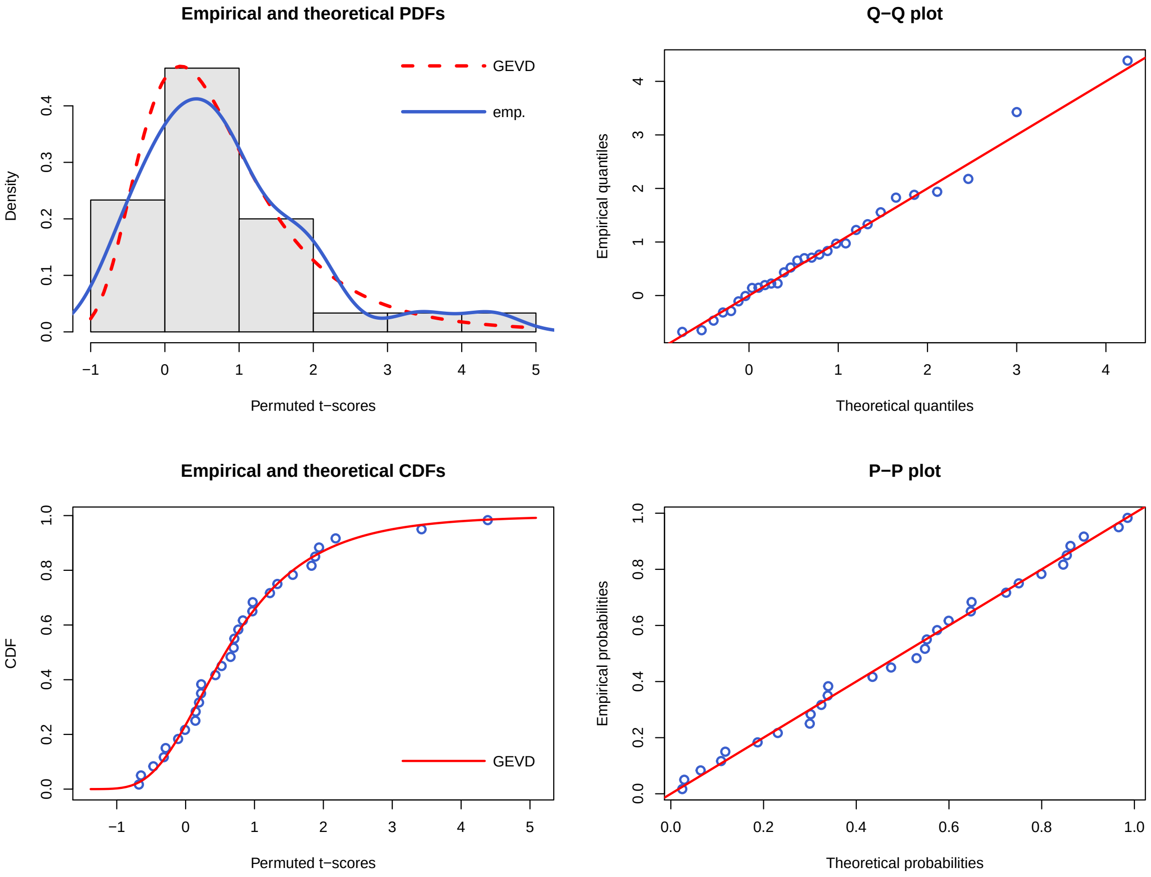
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5.7: Cognitive Resilience; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



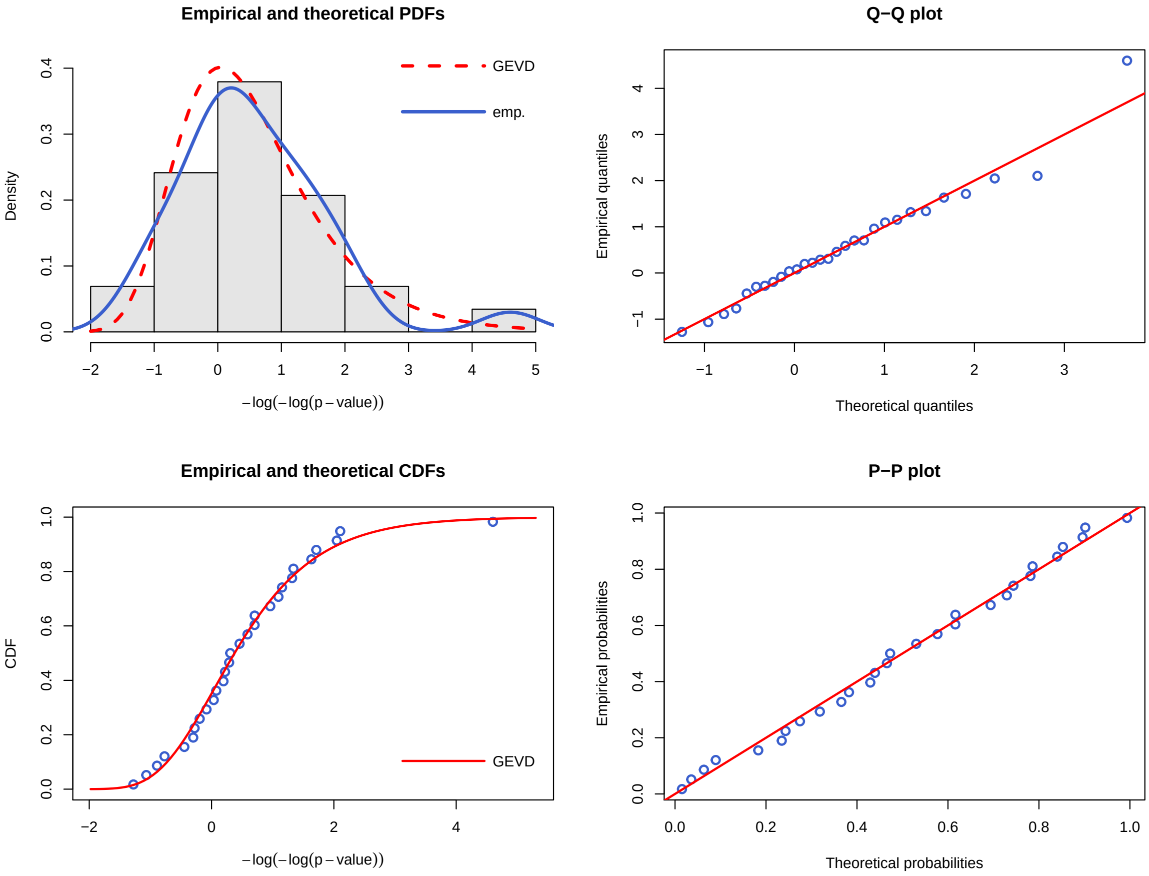
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5.8: Cognitive Reserve; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 29 permuted t-scores



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

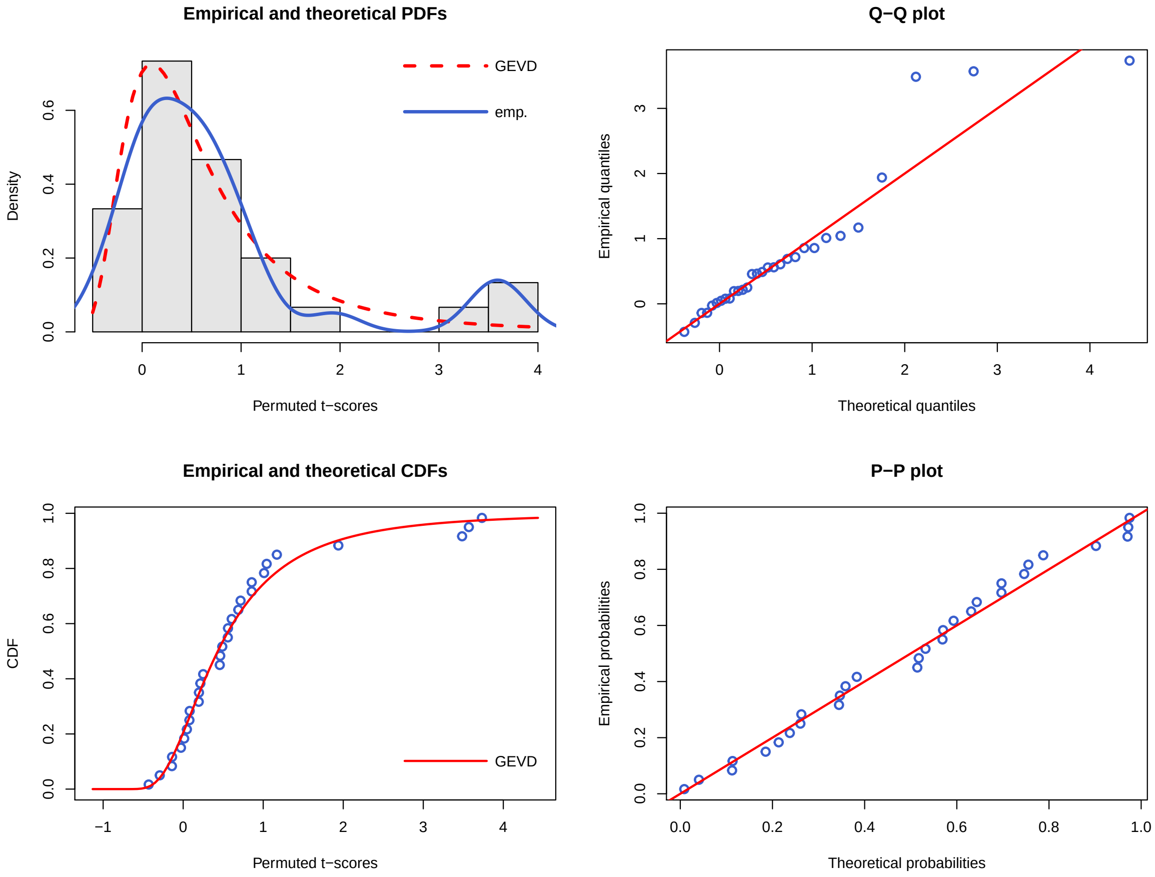
Figure 5.9: Cognitive Reserve; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 29 permuted p-values



---

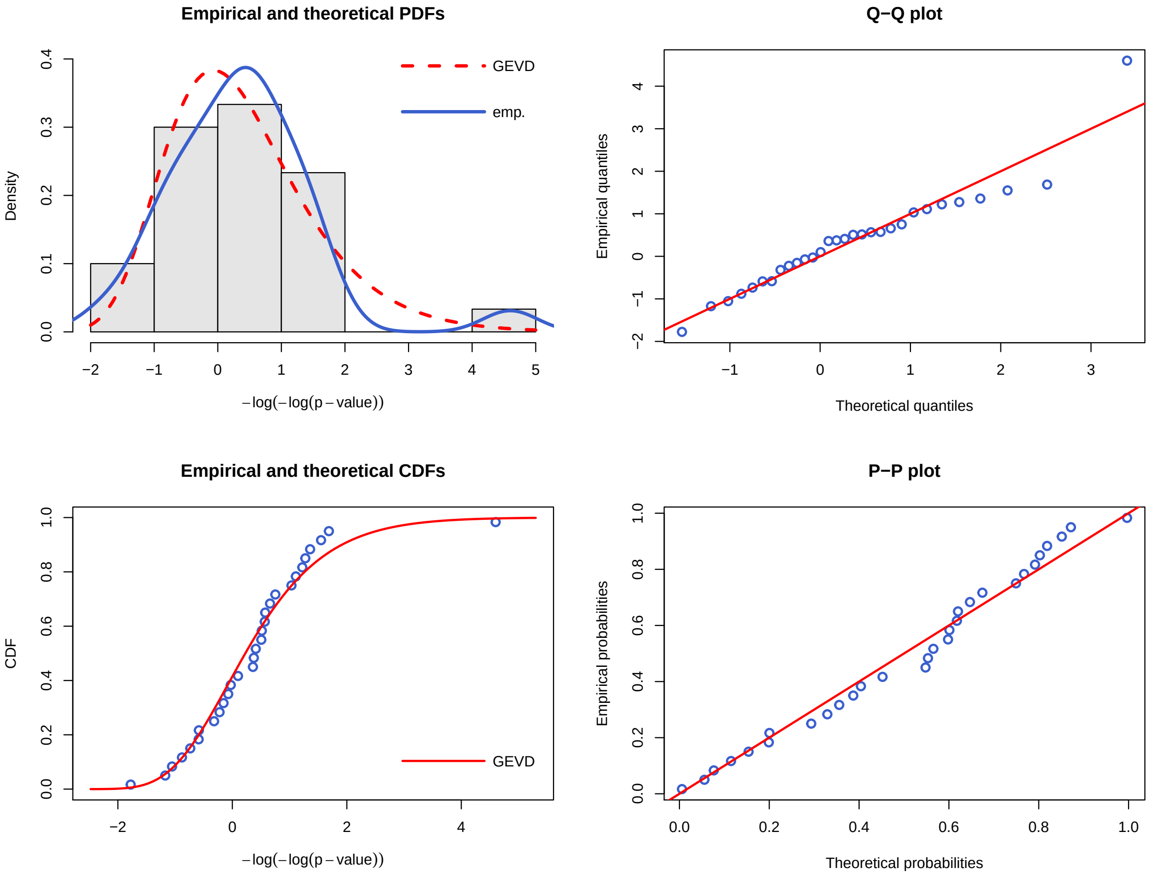
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5.10: Global Resilience; Graphical representation of the null distribution of $T^{*(0)}_{k_{max}}$ based on 30 permuted t-scores



[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5.11: Global Resilience; Graphical representation of the null distribution of $-\log(-\log(P^{(0)}_{k_{max}}))$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

5.4 Conclusion

Statistical analysis of the AD data set shows a consistent selection of the (SLC24A4, and CASS4) factors to explain the variation on the cognitive scores. The combination (EPHA1, SLC24A4, and CASS4) is chosen, as the best 3-way interaction, twice with the same risk pattern to model the relationship between genetic factors and patients' cognition. However, the contribution of these factors doesn't seem strong enough to approach statistical significance. Increasing the size of the data set might help for better recognizing the disease disposition. Model selection in OQMDR algorithm seems to work well choosing the most important interaction among all possible interactions; however, with such a small sample size, the sparsity of some multilocus allele combinations could substantially influence the selection mechanism. Regardless of weak significance of proposed models, the model evaluation component of the OQMDR method satisfactorily approximated the null distributions of the test statistic and the transformed p-value, which can be inferred by the graphical representation of the GEV and empirical distributions. Therefore we think that the GEV distribution is an ideal choice for assessing the validity of the interactions.

5.5 Further work

In this research, we proposed a new machine learning algorithm, the OQMDR, to handle genetic data sets with continuous trait response. The OQMDR is an adapted combination of the QMDR and the Optimal MDR algorithms [26, 30]. The modification was done by utilizing the concept of the Ordered Combinatorial Partitions (OCP) [46]. The new method shows a legitimate performance compared to QMDR in terms of selecting the most critical risk pattern that minimizes the prediction errors. The performance of the new method is presented in details in chapter 2. A comparison with the QMDR algorithm is carried out also in chapter 2. To enhance the efficiency and the accuracy of evaluation, the permutation testing for model assessment has been replaced with a parametric approach based on extreme value theory in chapter 3. Simulation studies in chapter 2 and chapter 3 exhibited an acceptable practical performance to capture the true models; however, there are some drawbacks of the OQMDR method that could be addressed in future works. One of the drawbacks is that the algorithm shows a poor performance with small size data sets, notably when high order interactions are examined due to the sparsity of information in some combinations. In addition, the OQMDR is vulnerable to missing information (NA), which is a pervasive issue with genetic data set [1]. Another weakness is the inability to analyze data sets with multiple responses simultaneously. For instance, we might have wished to do the analysis of the AD data set with all responses at once rather than performing three separate analysis. While this could be handled by doing a principal components (PC) analysis to aggregate all responses into one variable and apply the OQMDR on the new variable, a multivariate version of the OQMDR would be an interesting area to investigate in future research.

On the other hand, regardless of its complication, the modified GEVD evaluation component of the OQMDR has higher accuracy, compared to the regular permutation testings, in evaluating the significance of the proposed models and is more efficient under specific considerations. Despite, simplifying the theory-based approach could substantially benefit the efficiency of the algorithm. The simplification could involve revising the analytical MLE approach to lessen the required iterations to achieve convergence, or using a more efficient programming language.

Further, theoretical validation and power estimation studies would strengthen the findings of this research. In addition, utilizing the OCP approach on other MDR-based algorithms, where applicable, might benefit the performance of model selection and reduce the prediction error.

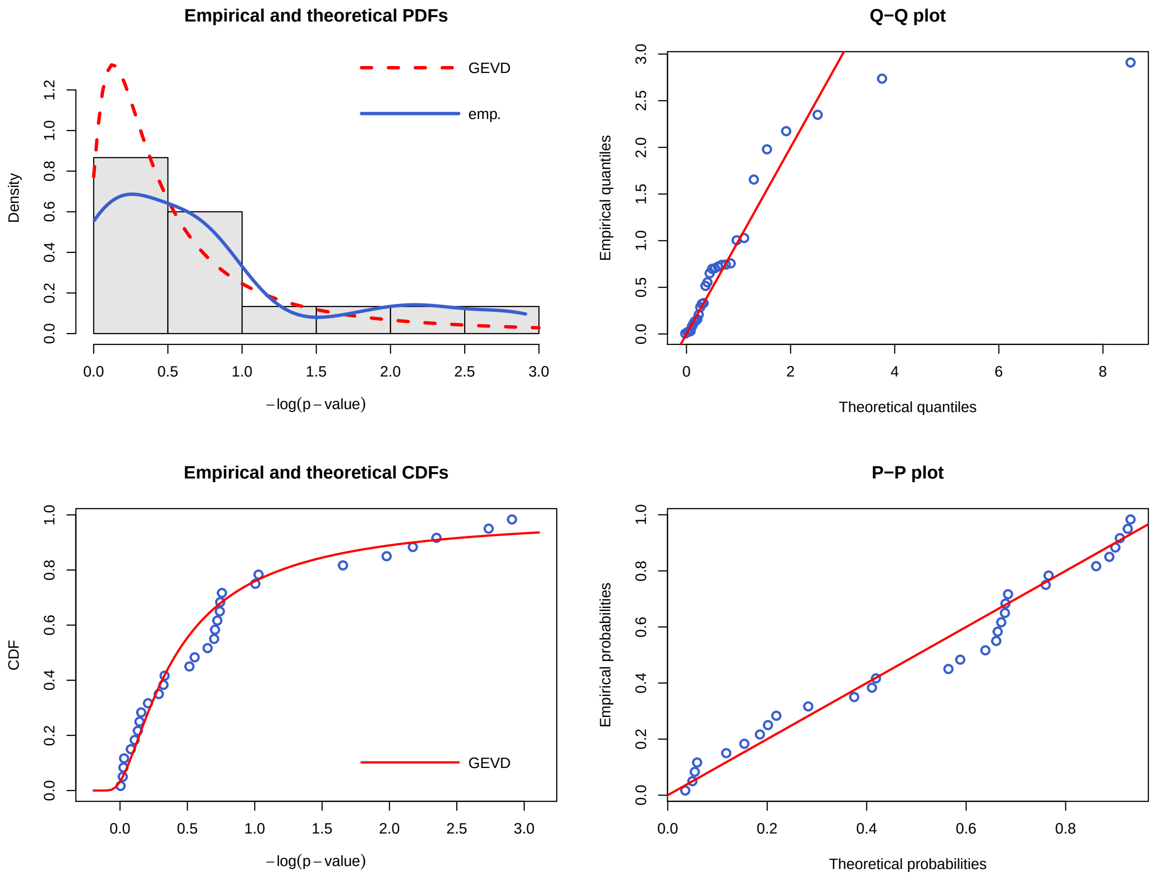
Copyright© Zaid T. Al-Khaledi, 2019.

Appendix

Presentation of other fitted distributions

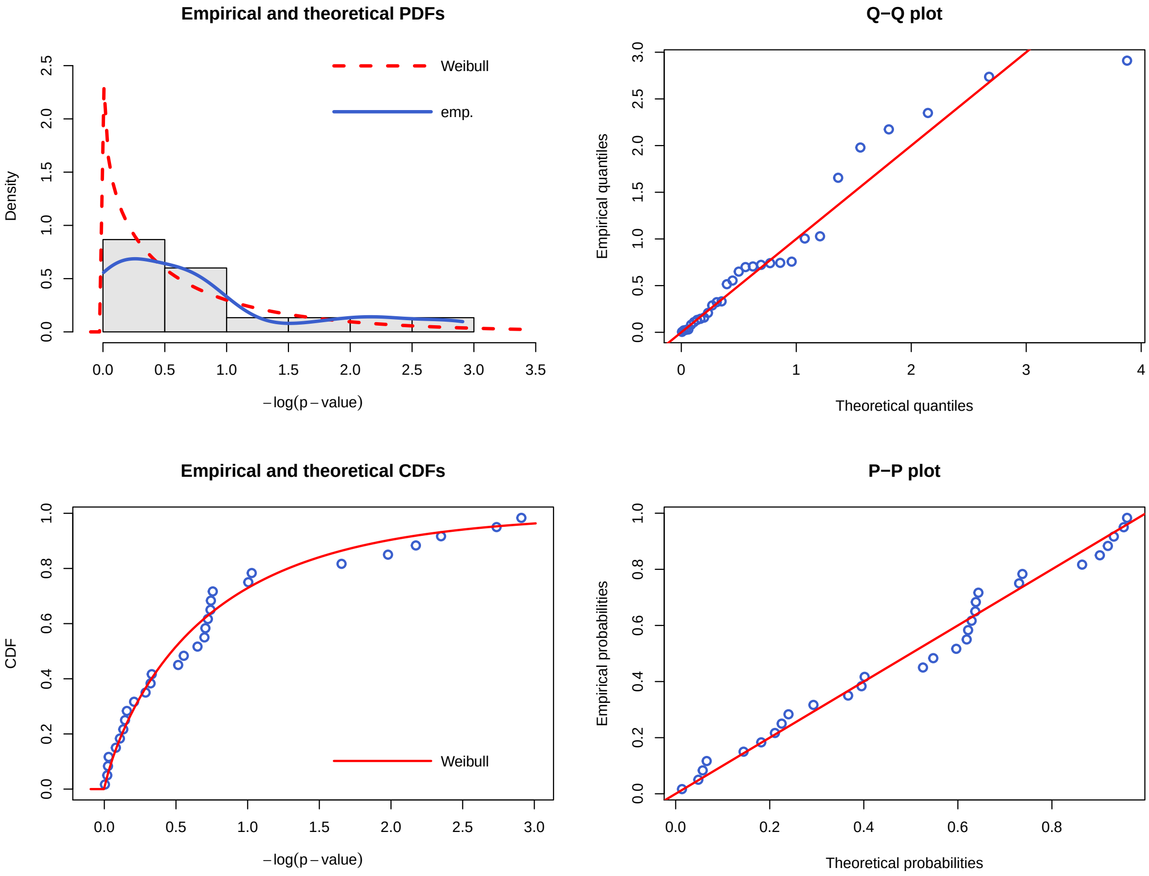
In this appendix, we presented some graphical results from chapter 3 for some fitted distributions and/or transformations besides the GEVD of the $-\log(-\log(P^{(0)}_{k_{max}}))$, which was demonstrated earlier in chapter 3. The case where the underlying interaction is AB with $n = 500$ is the only considered case in this appendix.

Figure 1: True model $= AB$, $n = 500$; Graphical representation of the null GEVD of the $-\log(P^{(0)}_{k_{max}})$ based on 30 permuted p-values



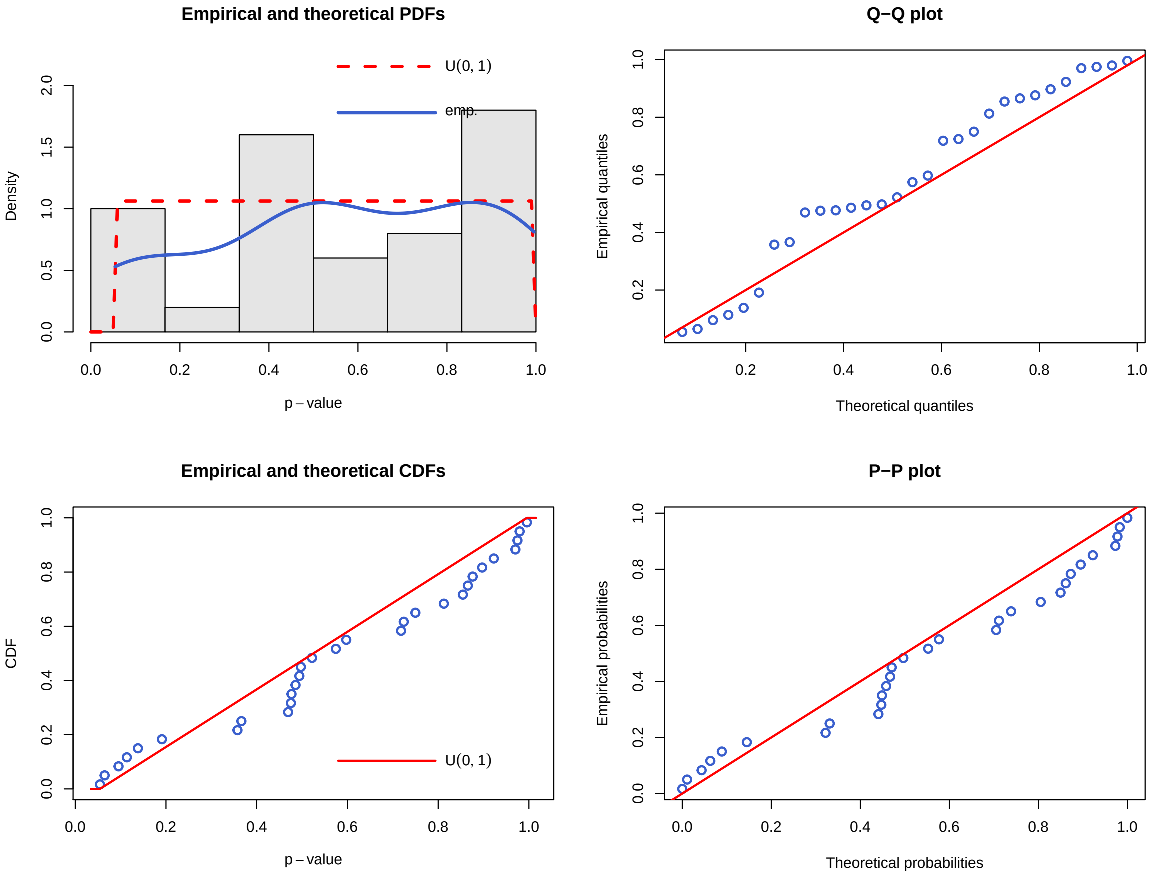
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 2: True model = AB, $n = 500$; Graphical representation of the null Weibull distribution of the $-\log(P^{(0)}_{k_{max}})$ based on 30 permuted p-values



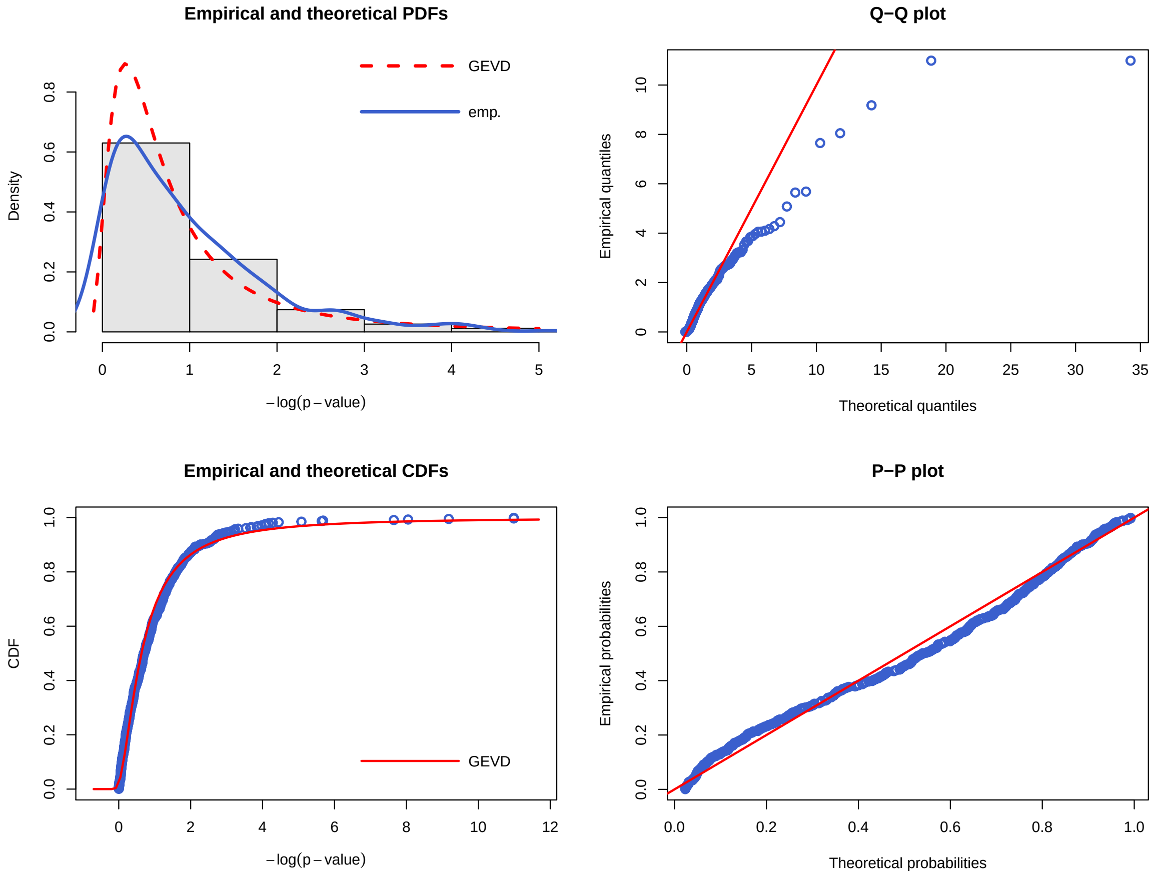
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 3: True model = AB, n = 500; Graphical representation of the null uniform(0,1) distribution of the $P^{(0)}_{k_{max}}$ based on 30 permuted p-values



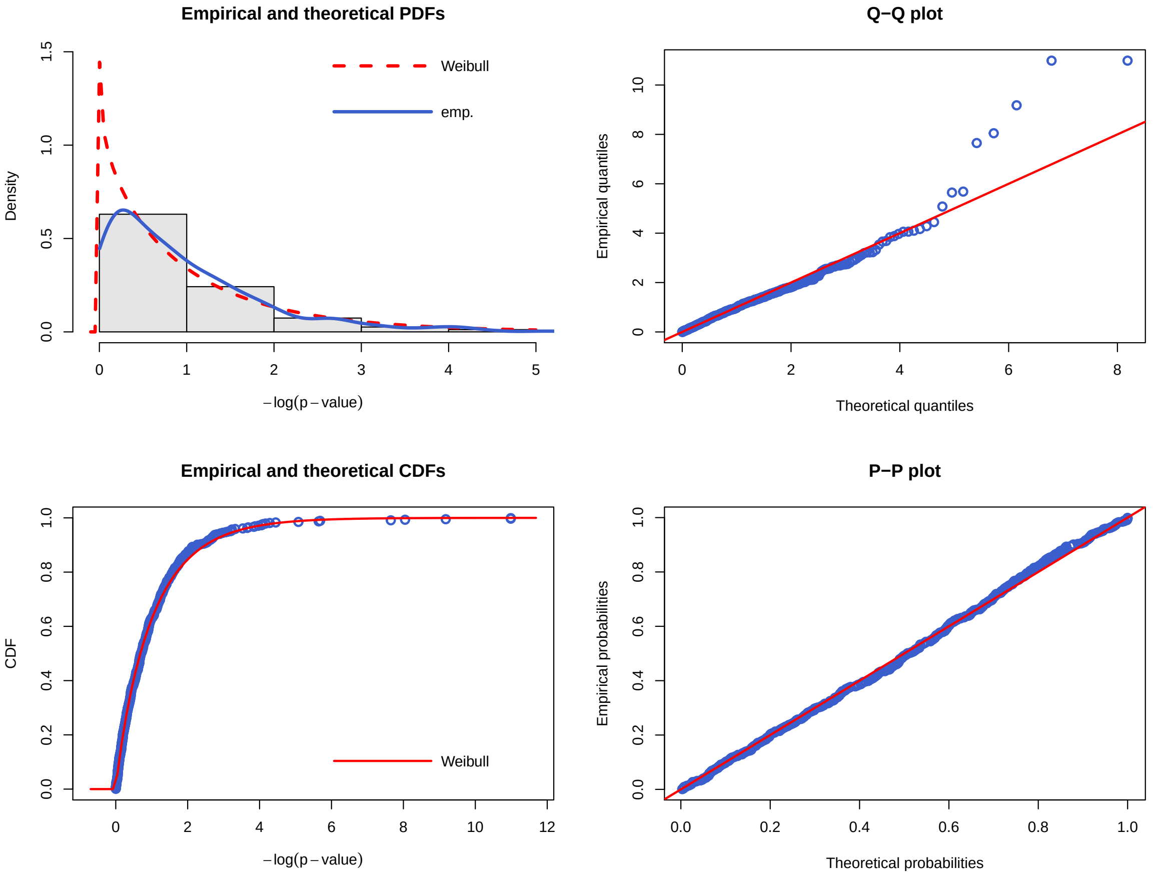
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 4: True model = AB, $n = 500$; Graphical representation of the null GEVD of the $-\log(P^{(0)}_{k_{max}})$ based on 500 permuted p-values



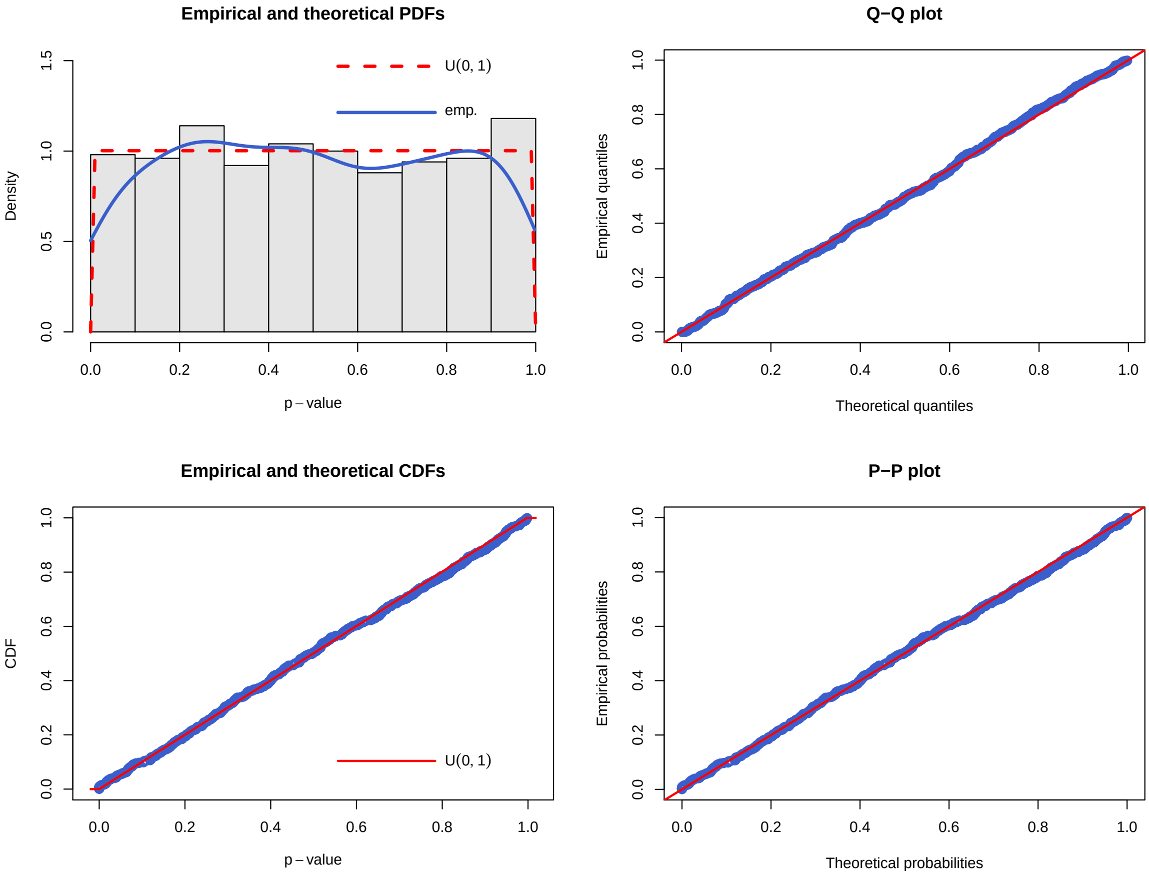
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 5: True model = AB, $n = 500$; Graphical representation of the null Weibull distribution of the $-\log(P^{(0)}_{k_{max}})$ based on 500 permuted p-values



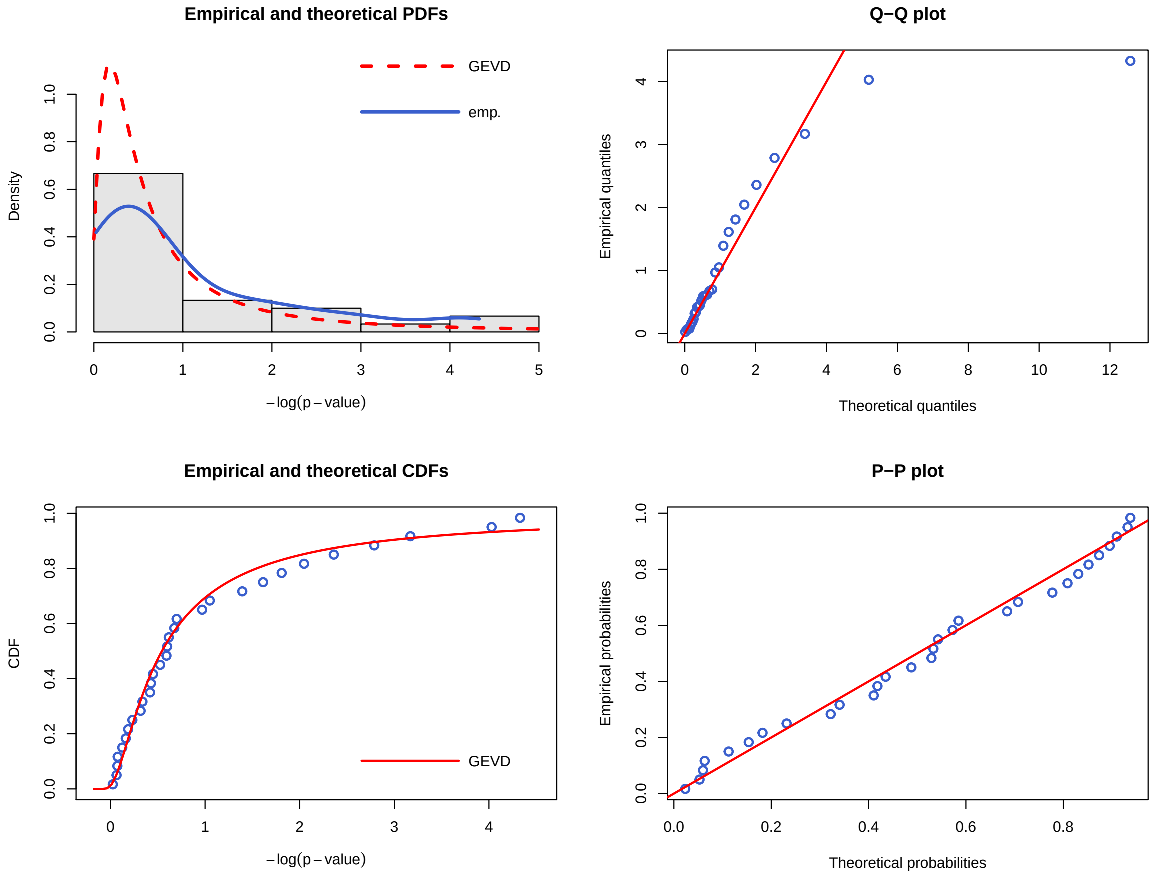
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 6: True model $= AB$, $n = 500$; Graphical representation of the null uniform(0,1) distribution of the $P^{(0)}_{k_{max}}$ based on 500 permuted p-values



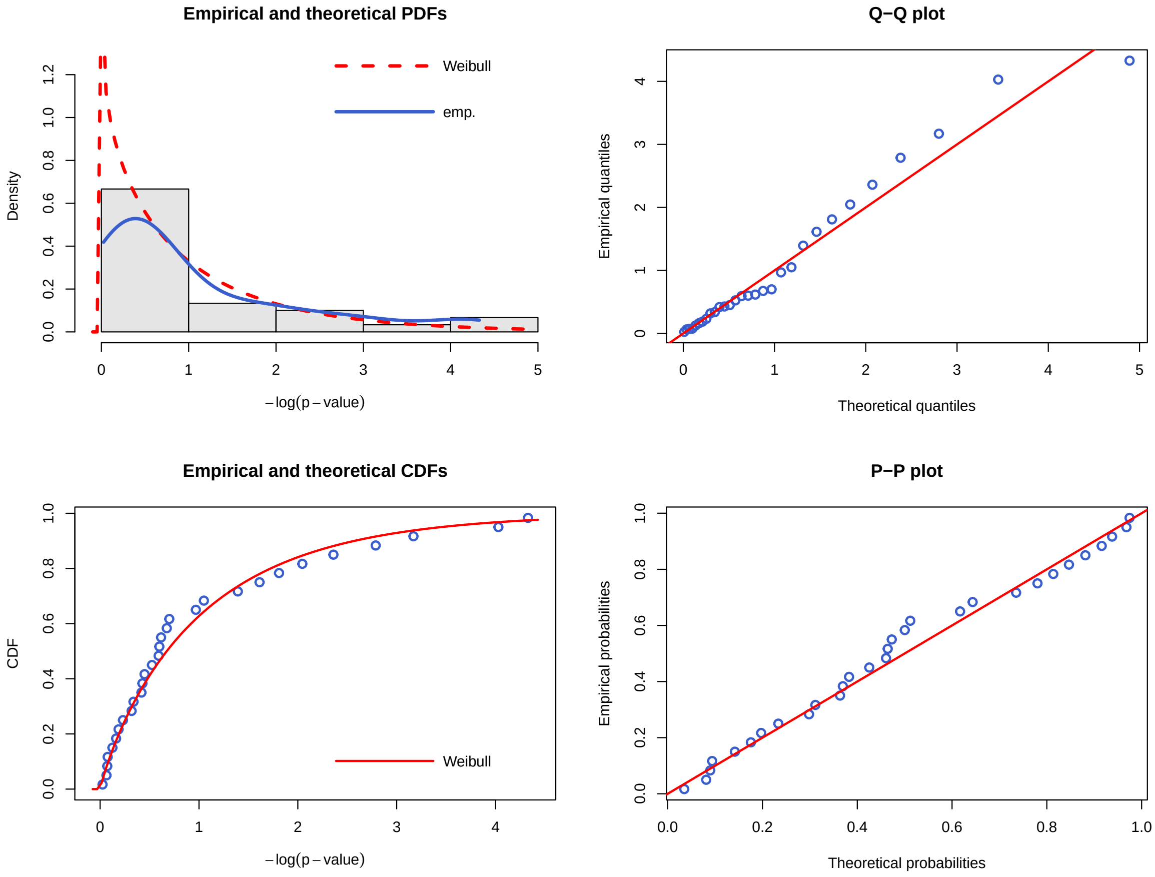
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 7: True model = AB, $n = 1000$; Graphical representation of the null GEVD of the $-\log(P^{(0)}_{k_{max}})$ based on 30 permuted p-values



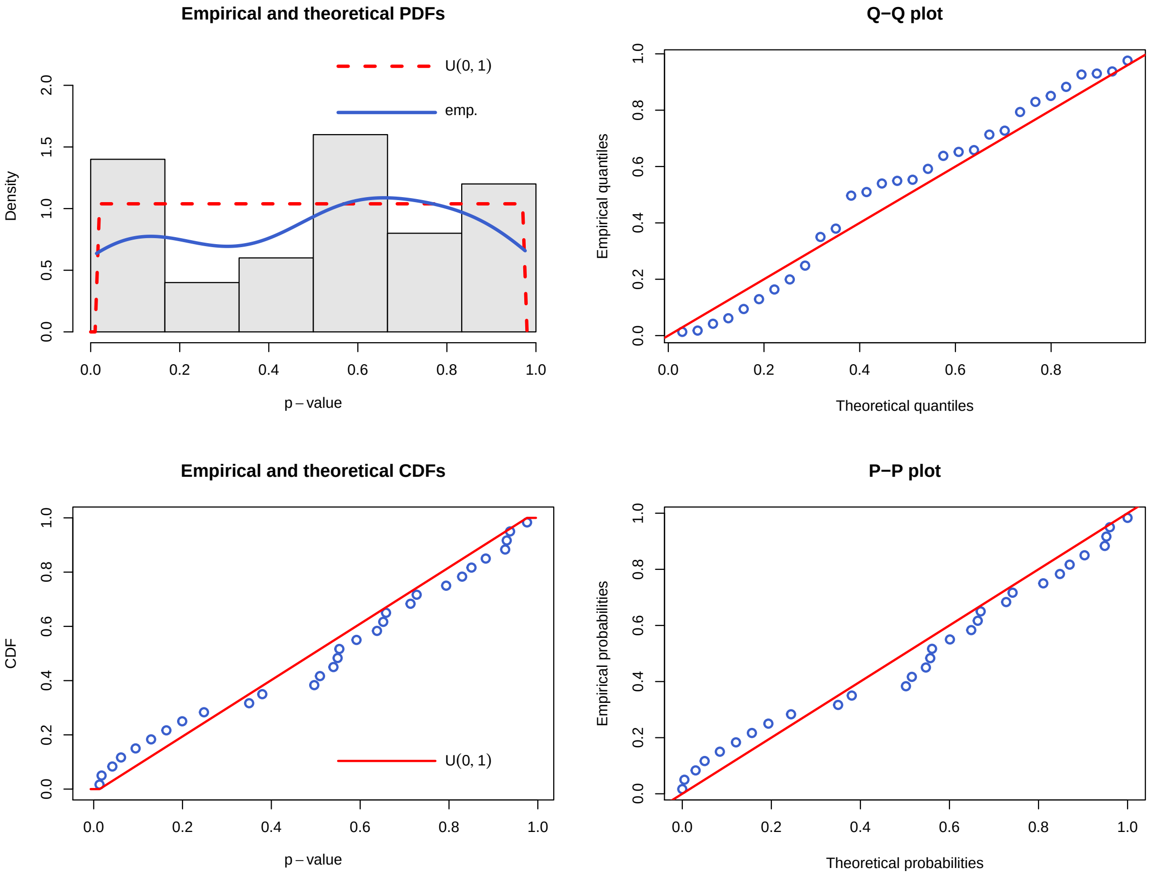
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 8: True model = AB, $n = 1000$; Graphical representation of the null Weibull distribution of the $-\log(P^{(0)}_{k_{max}})$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

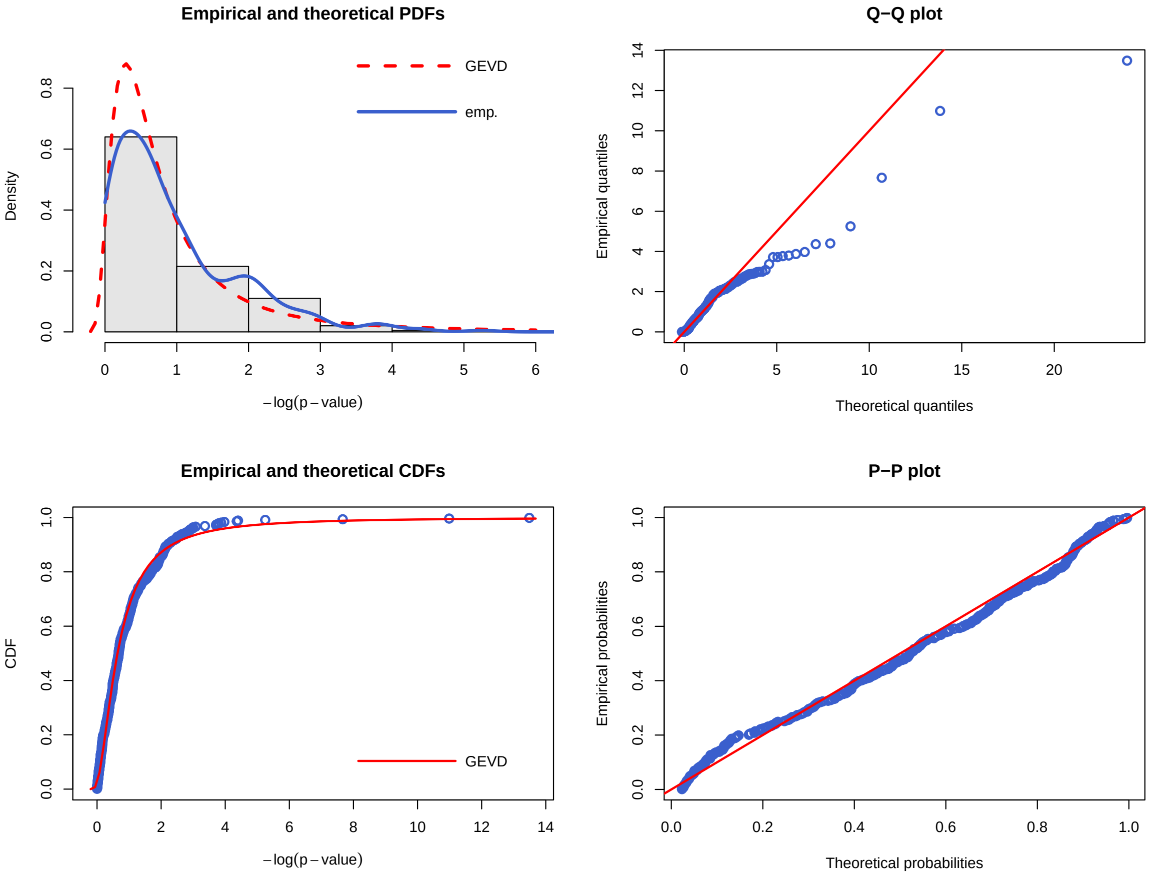
Figure 9: True model = AB, $n = 1000$; Graphical representation of the null uniform(0,1) distribution of the $P^{(0)}_{k_{max}}$ based on 30 permuted p-values



---

* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 10: True model = AB, $n = 1000$; Graphical representation of the null GEVD of the $-\log(P^{(0)}_{k_{max}})$ based on 400 permuted p-values


Empirical and theoretical PDFs

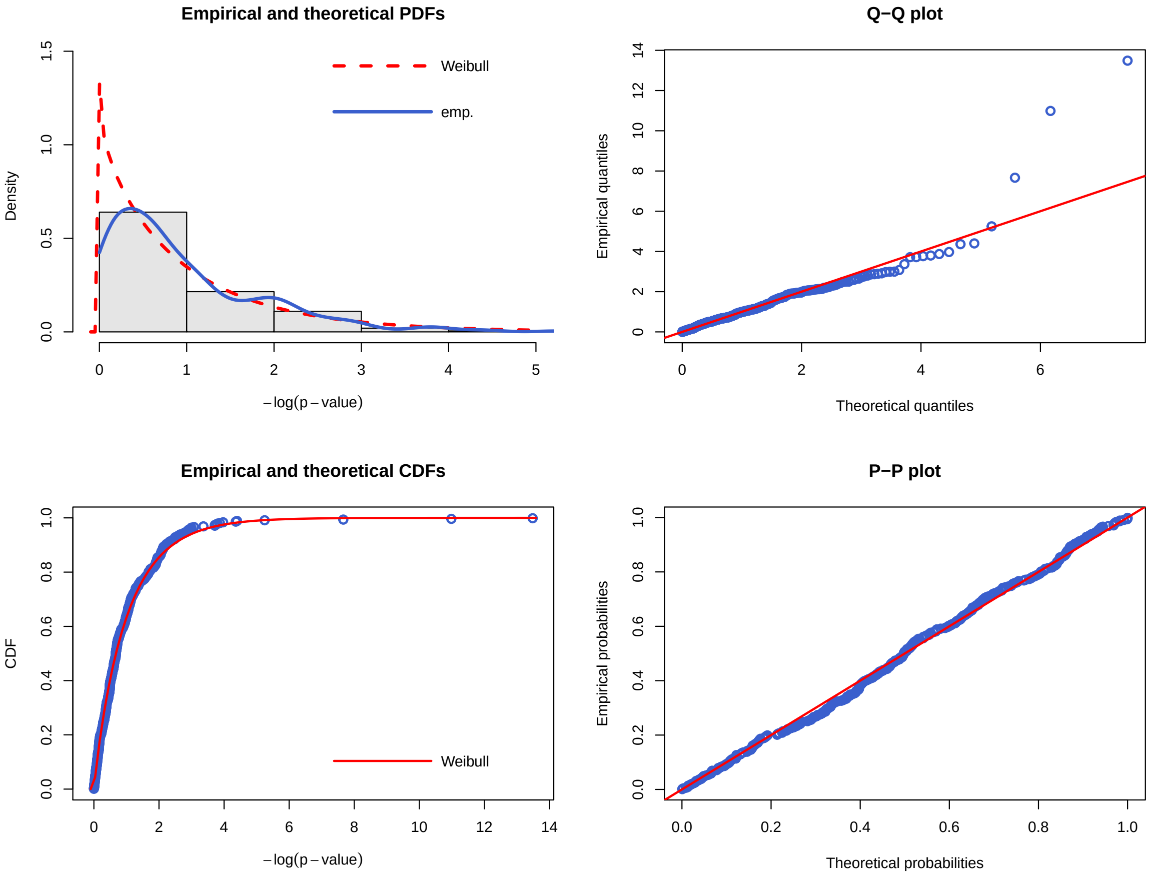
Q–Q plot

Empirical and theoretical CDFs

P–P plot

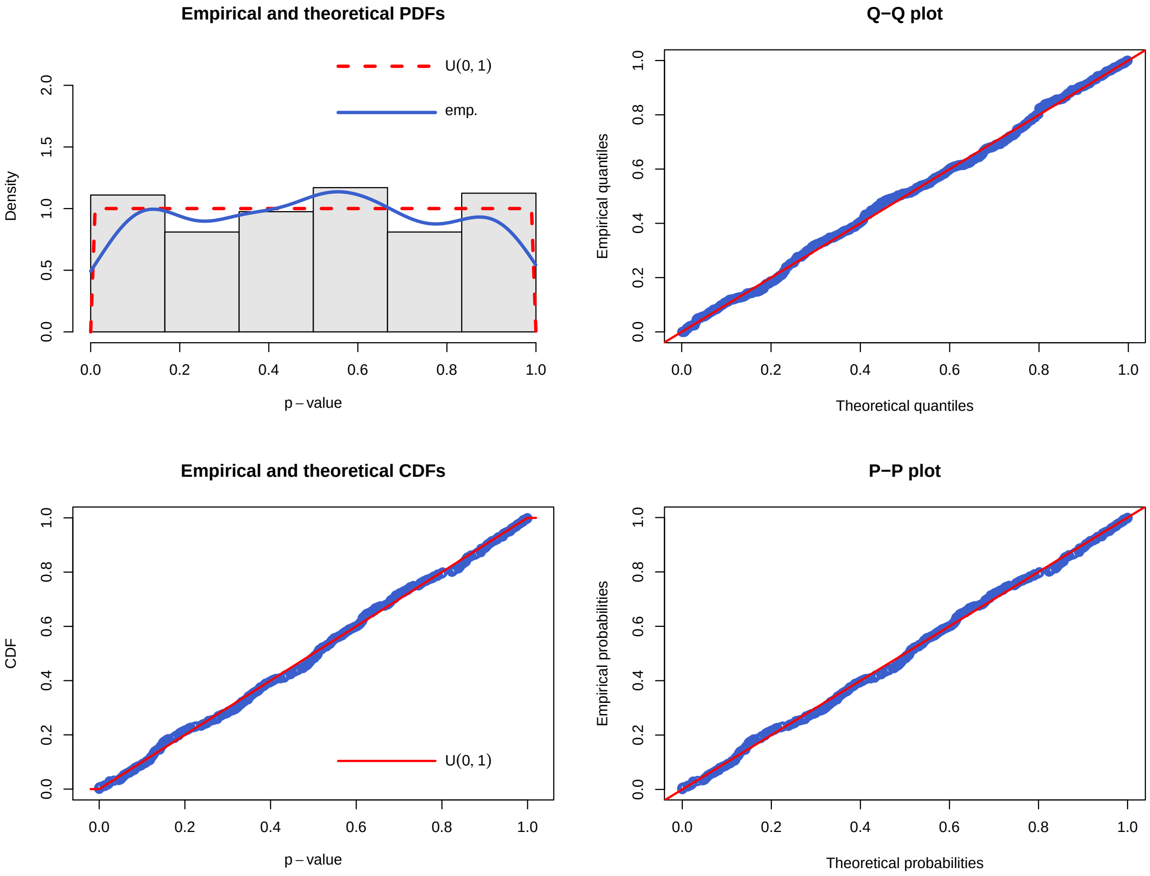
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 11: True model = AB, $n = 1000$; Graphical representation of the null Weibull distribution of the $-\log(P^{(0)}_{k_{max}})$ based on 400 permuted p-values



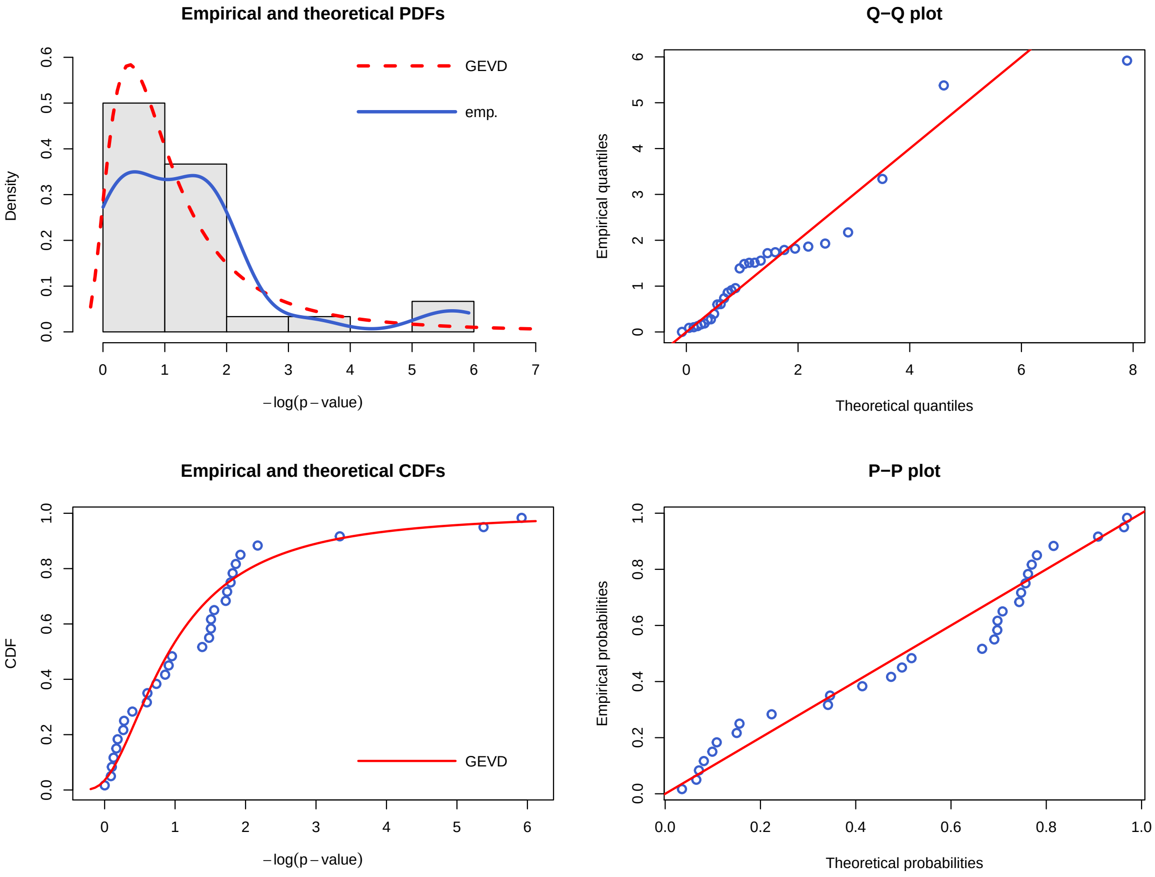
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 12: True model = AB, $n = 1000$; Graphical representation of the null uniform(0,1) distribution of the $P^{(0)}_{k_{max}}$ based on 400 permuted p-values



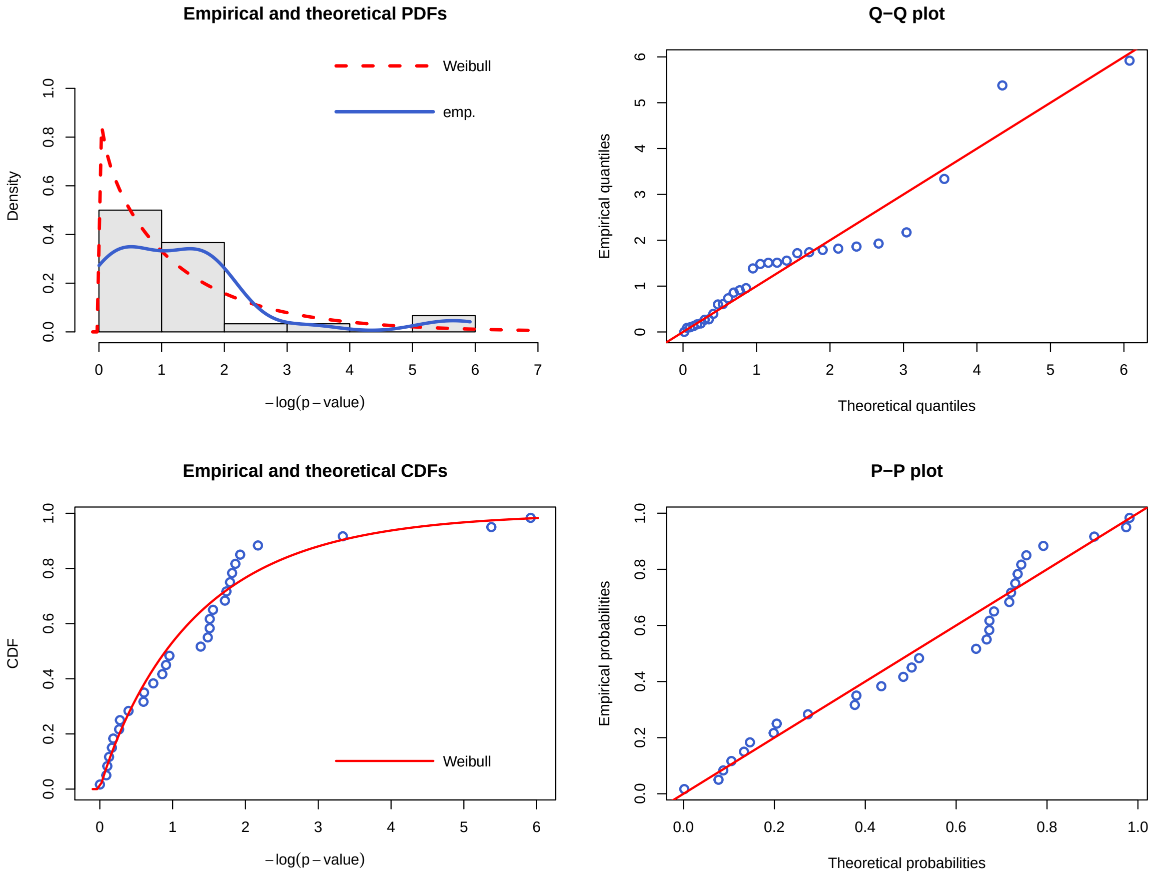
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 13: True model = AB, $n = 2000$; Graphical representation of the null GEVD of the $-\log(P^{(0)}_{k_{max}})$ based on 30 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 14: True model = AB, $n = 2000$; Graphical representation of the null Weibull distribution of the $-\log(P^{(0)}_{k_{max}})$ based on 30 permuted p-values


Empirical and theoretical PDFs

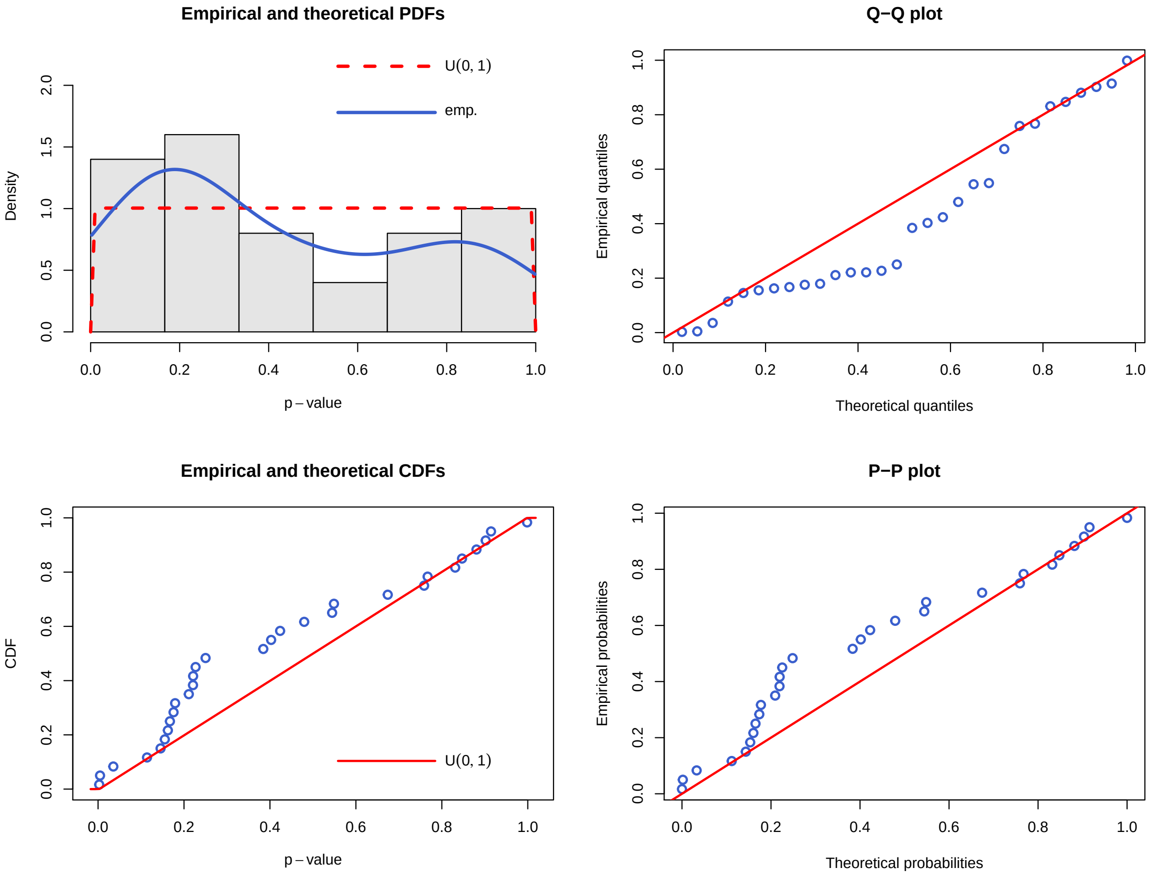
Q−Q plot

Empirical and theoretical CDFs

P−P plot

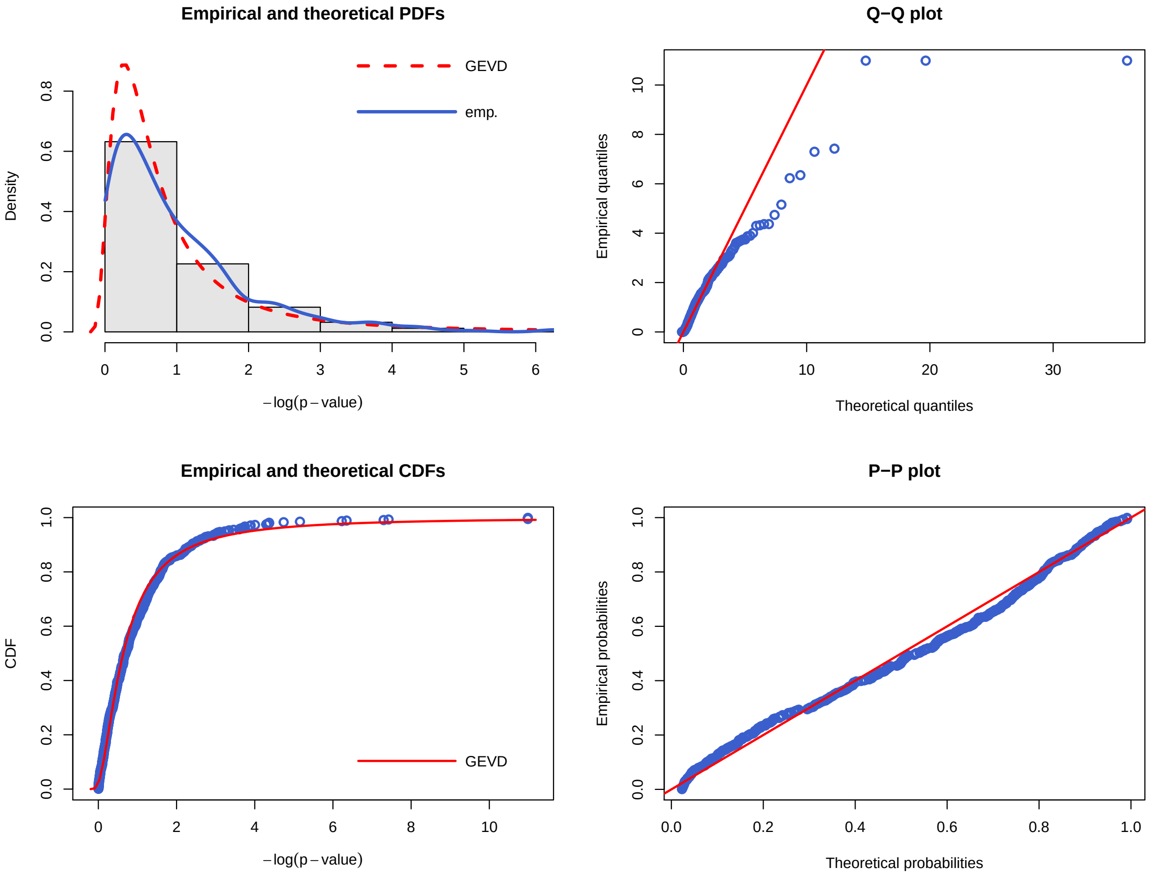
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 15: True model = AB, $n = 2000$; Graphical representation of the null uniform(0,1) distribution of the $P^{(0)}_{k_{max}}$ based on 30 permuted p-values



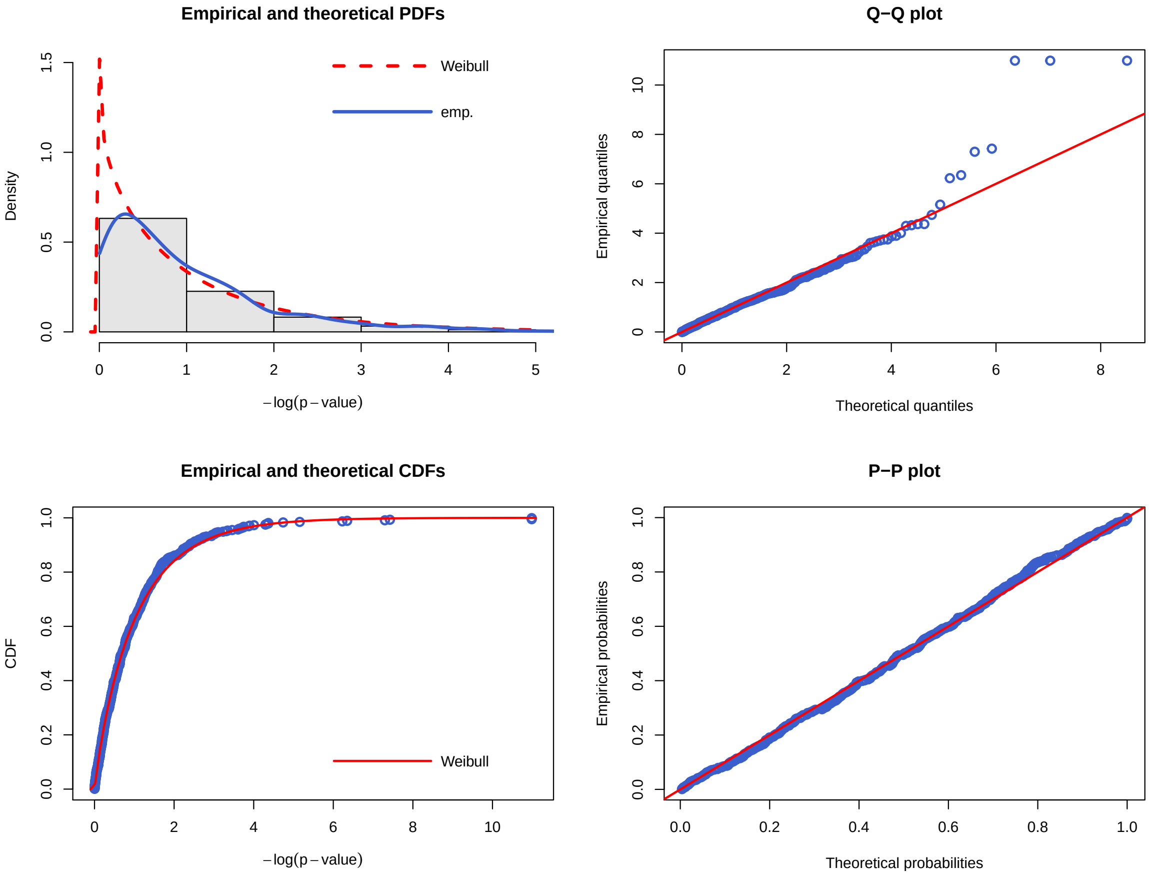
* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 16: True model = AB, $n = 2000$; Graphical representation of the null GEVD of the $-\log(P^{(0)}_{k_{max}})$ based on 500 permuted p-values



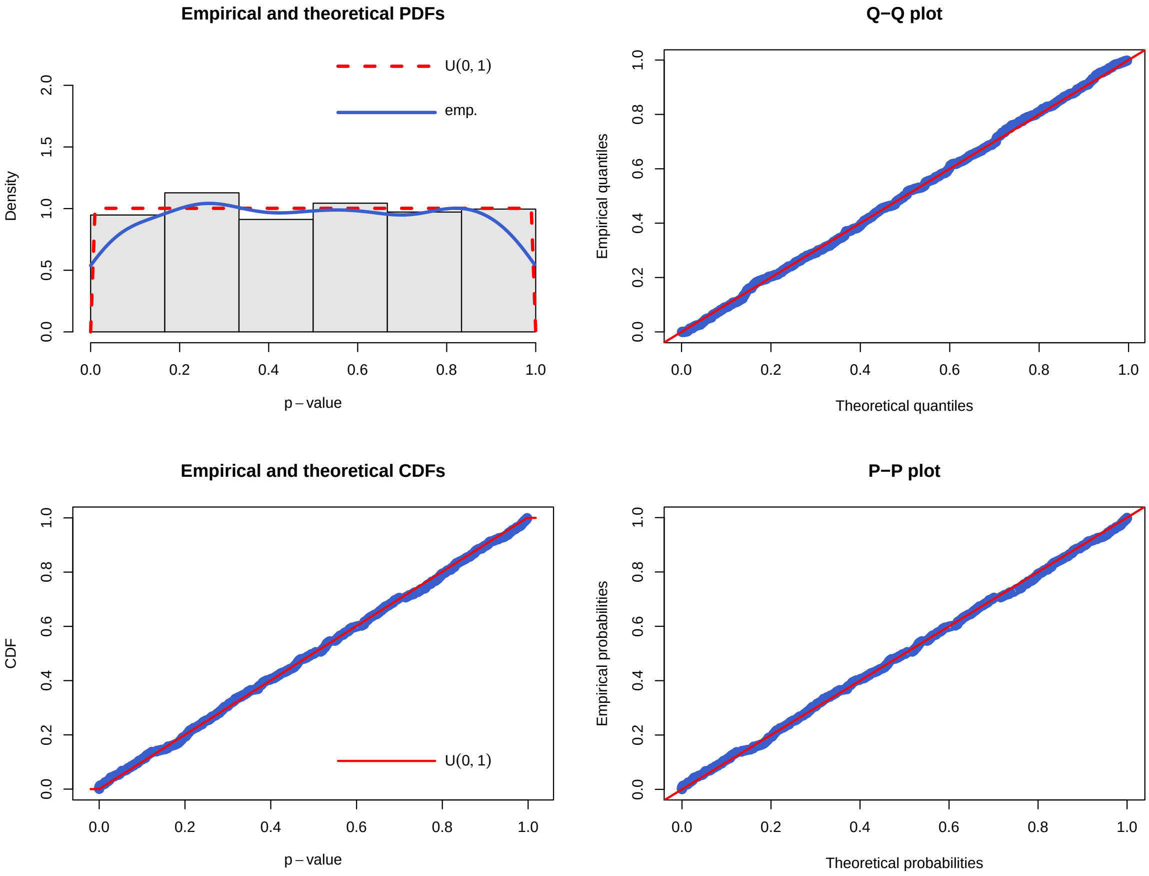
[*] The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 17: True model = AB, $n = 2000$; Graphical representation of the null Weibull distribution of the $-\log(P^{(0)}_{k_{max}})$ based on 500 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

Figure 18: True model $= AB$, $n = 2000$; Graphical representation of the null uniform(0,1) distribution of the $P^{(0)}_{k_{max}}$ based on 500 permuted p-values



* The four plots are produced using `R`. Refer to the second paragraph of section 3.5 and figure 3.1 for details.

References

[1] Pakeeza Akram and Li Liao. Prediction of missing common genes for disease pairs using network based module separation on incomplete human interactome. *BMC genomics*, 18(10):902, 2017.

[2] Alzheimer's Association. Alzheimer's and dementia: Prevalence, 2019. URL `https://www.alz.org/alzheimers-dementia/facts-figures`. [Online; accessed 20-April-2019].

[3] Alzheimer's Association. Alzheimer's and dementia: Treatments, 2019. URL `https://www.alz.org/alzheimers-dementia/treatments`. [Online; accessed 12-February-2019].

[4] Elizabeth Arias. United states life tables, 2004. *National vital statistics reports*, 56(9):1–40, 2007.

[5] Lorenzo Beretta, Alessandro Santaniello, Piet LCM van Riel, Marieke JH Coenen, and Raffaella Scorza. Survival dimensionality reduction (sdr): development and clinical application of an innovative approach to detect epistasis in presence of right-censored data. *BMC bioinformatics*, 11(1):416, 2010.

[6] Thomas D Bird. Alzheimer disease overview. In *GeneReviews®[Internet]*. University of Washington, Seattle, 2015.

[7] William S Bush, Todd L Edwards, Scott M Dudek, Brett A McKinney, and Marylyn D Ritchie. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *Bmc Bioinformatics*, 9 (1):238, 2008.

[8] ML Calle, V Urrea, G Vellalta, N Malats, and Kristel Van Steen. Model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data. Technical report, Department of Systems Biology, Universitat de Vic,, 2008.

[9] Elena Carapelle, Laura Serra, Sergio Modoni, Michele Falcone, Carlo Caltagirone, Marco Bozzali, Luigi Maria Specchio, and Carlo Avolio. How the cognitive reserve interacts with β-amyloid deposition in mitigating fdg metabolism: An observational study. *Medicine*, 96(16), 2017.

[10] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[11] Enrique Castillo, Ali S Hadi, Narayanaswamy Balakrishnan, and José-Mariá Sarabia. *Extreme value and related models with applications in engineering and science*. Wiley Hoboken, NJ, 2005.

[12] Bernard Cesarone. Resilience guide: A collection of resources on resilience in children and families. 1999.

[13] John M Chambers. *Graphical Methods for Data Analysis*. Chapman and Hall/CRC, 2017.

[14] Jiin Choi and Taesung Park. Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions. *BMC systems biology*, 7(Suppl 6):S15, 2013.

[15] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[16] James F Crow. Hardy, weinberg and language impediments. *Genetics*, 152(3): 821–825, 1999.

[17] Hongying Dai, Richard J Charnigo, Mara L Becker, J Steven Leeder, and Alison A Motsinger-Reif. Risk score modeling of multiple gene to gene interactions using aggregated-multifactor dimensionality reduction. *BioData mining*, 6(1):1, 2013.

[18] Rishika De, Shefali S Verma, Fotios Drenos, Emily R Holzinger, Michael V Holmes, Molly A Hall, David R Crosslin, David S Carrell, Hakon Hakonarson, Gail Jarvik, et al. Identifying gene-gene interactions that are highly associated with body mass index using quantitative multifactor dimensionality reduction (qmdr). *BioData mining*, 8(1):41, 2015.

[19] Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015. URL `http://www.jstatsoft.org/v64/i04/`.

[20] David W. Fardo. Private Communication, 2019.

[21] Norman Garmezy. Resiliency and vulnerability to adverse developmental outcomes associated with poverty. *American behavioral scientist*, 34(4):416–430, 1991.

[22] Generalized extreme value distribution. Generalized extreme value distribution - Wikipedia, the free encyclopedia, 2018. URL `https://en.wikipedia.org/wiki/Generalized_extreme_value_distribution`. [Online; accessed 10-February-2018].

[23] Genetic and Rare Diseases Information Center (GARD). Alzheimer disease, 2015. URL `https://rarediseases.info.nih.gov/diseases/10254/alzheimer-disease`. [Online; accessed 01-March-2019].

[24] Damian Gola, Jestinah M Mahachie John, Kristel Van Steen, and Inke R König. A roadmap to multifactor dimensionality reduction methods. *Briefings in bioinformatics*, 17(2):293–308, 2016.

[25] Jiang Gui, Jason H Moore, Karl T Kelsey, Carmen J Marsit, Margaret R Karagas, and Angeline S Andrew. A novel survival multifactor dimensionality reduction method for detecting gene–gene interactions with application to bladder cancer prognosis. *Human genetics*, 129(1):101–110, 2011.

[26] Jiang Gui, Jason H Moore, Scott M Williams, Peter Andrews, Hans L Hillege, Pim van der Harst, Gerjan Navis, Wiek H Van Gilst, Folkert W Asselbergs, and Diane Gilbert-Diamond. A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS One*, 8(6):e66545, 2013.

[27] Muralidhar L Hegde, Anil K Mantha, Tapas K Hazra, Kishor K Bhakat, Sankar Mitra, and Bartosz Szczesny. Oxidative genome damage and its repair: implications in aging and neurodegenerative diseases. *Mechanisms of ageing and development*, 133(4):157–168, 2012.

[28] Hanns Hippius and Gabriele Neundörfer. The discovery of alzheimer's disease. *Dialogues in clinical neuroscience*, 5(1):101, 2003.

[29] Jonathan RM Hosking, James R Wallis, and Eric F Wood. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261, 1985.

[30] Xing Hua, Han Zhang, Hong Zhang, Yaning Yang, and Anthony YC Kuk. Testing multiple gene interactions by the ordered combinatorial partitioning method in case–control studies. *Bioinformatics*, 26(15):1871–1878, 2010.

[31] Graeme D Hutcheson and Nick Sofroniou. *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage, 1999.

[32] Arthur F Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171, 1955.

[33] H. Joe. Estimation of quantiles of the maximum of n observations. Technical report, University of British Columbia, 1985.

[34] Harry Joe. Estimation of quantiles of the maximum of n observations. *Biometrika*, 74(2):347–354, 1987.

[35] PJ Kahle, M Jakowec, SJ Teipel, H Hampel, GM Petzinger, DA Di Monte, GD Silverberg, H-J Möller, JA Yesavage, JR Tinklenberg, et al. Combined assessment of tau and neuronal thread protein in alzheimer's disease csf. *Neurology*, 54(7):1498–1504, 2000.

[36] Kyunga Kim, Min-Seok Kwon, Sohee Oh, and Taesung Park. Identification of multiple gene-gene interactions for ordinal phenotypes. *BMC medical genomics*, 6(2):S9, 2013.

[37] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452, 2013.

[38] Kenneth Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

[39] Xiang-Yang Lou, Guo-Bo Chen, Lei Yan, Jennie Z Ma, Jun Zhu, Robert C Elston, and Ming D Li. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics*, 80(6):1125–1137, 2007.

[40] Richard Mayeux and Mary Sano. Treatment of alzheimer's disease. *New England Journal of Medicine*, 341(22):1670–1679, 1999.

[41] Hao Mei, Deqiong Ma, Allison Ashley-Koch, and Eden R Martin. Extension of multifactor dimensionality reduction for identifying multilocus effects in the gaw14 simulated data. *BMC genetics*, 6(1):S145, 2005.

[42] Steven P. Millard. *EnvStats: An R Package for Environmental Statistics*. Springer, New York, 2013. ISBN 978-1-4614-8455-4. URL `http://www.springer.com`.

[43] Alison A Motsinger-Reif. The effect of alternative permutation testing strategies on the performance of multifactor dimensionality reduction. *BMC research notes*, 1(1):139, 2008.

[44] Junghyun Namkung, Kyunga Kim, Sungon Yi, Wonil Chung, Min-Seok Kwon, and Taesung Park. New evaluation measures for multifactor dimensionality reduction classifiers in gene–gene interaction analysis. *Bioinformatics*, 25(3):338–345, 2009.

[45] National Institute of Aging. What are the signs of alzheimer's disease? - national institute of aging, 2017. URL `https://www.nia.nih.gov/health/what-are-signs-alzheimers-disease`. [Online; accessed 12-February-2019].

[46] MR Nelson, SLR Kardia, RE Ferrell, and CF Sing. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome research*, 11(3):458–470, 2001.

[47] A Otten and MAJ Van Montfort. Maximum-likelihood estimation of the general extreme-value distribution parameters. *Journal of Hydrology*, 47(1):187–192, 1980.

[48] Kristine A Pattin, Bill C White, Nate Barney, Jiang Gui, Heather H Nelson, Karl T Kelsey, Angeline S Andrew, Margaret R Karagas, and Jason H Moore. A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genetic epidemiology*, 33(1):87–94, 2009.

[49] P Prescott and AT Walden. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67(3):723–724, 1980.

[50] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[51] Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. 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, 2001.

[52] Dennis J Selkoe. Alzheimer's disease: genes, proteins, and therapy. *Physiological reviews*, 81(2):741–766, 2001.

[53] Digna R Velez, Bill C White, Alison A Motsinger, William S Bush, Marylyn D Ritchie, Scott M Williams, and Jason H Moore. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic epidemiology*, 31(4):306–315, 2007.

[54] Robert S Wilson, Carlos F Mendes De Leon, Lisa L Barnes, Julie A Schneider, Julia L Bienias, Denis A Evans, and David A Bennett. Participation in cognitively stimulating activities and risk of incident alzheimer disease. *Jama*, 287(6): 742–748, 2002.

Vita

Zaid Tariq Saleh Al-Khaledi
Birth Place: Mosul, Iraq.

Education

2015	University of Kentucky *Master of Science in Statistics, May 2015*
2005	University of Mosul *Master of Science in Statistics, July 2005*
2002	University of Mosul *Bachelor of Science in Statistics, June 2002*

Academic Positions

2016-2019	Teaching Assistant, University of Kentucky.
2008-2012	Assistant Lecturer, College of Computer Sciences and Mathematics - University of Mosul, Mosul, Iraq.
2005-2008	Assistant Lecturer, Computer and Internet Center - University of Mosul, Mosul, Iraq.

Other Experience

2010-2012	Internet Core Competency Certification (IC3) Administrator. The Graduate School, University of Mosul, Mosul, Iraq.
2005-2009	Assistant Director - Computer and Internet Center - University of Mosul, Mosul, Iraq.

Honors and Awards

2019	R.L. Anderson Outstanding Teaching Award, Department of Statistics, University of Kentucky.
2002	Superior Students in Iraq Award - 2nd rank in Statistics, Office of the President, Baghdad, Iraq.