Linkage fine-mapping on sequences from case-control studies and Goodness-of-fit tests based on empirical distribution function for general likelihood models

by

Payman Nickchi

M.Sc., University of Tehran, 2013
B.Sc., University of Tehran, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2024

Declaration of Committee

Name:	**Payman Nickchi**
Degree:	**Doctor of Philosophy**
Thesis title:	**Linkage fine-mapping on sequences from case-control studies and Goodness-of-fit tests based on empirical distribution function for general likelihood models**
Committee:	**Chair:** Liangliang Wang Associate Professor, Statistics and Actuarial Science
	Richard Lockhart Co-supervisor Professor, Statistics and Actuarial Science
	Jinko Graham Co-supervisor Professor, Statistics and Actuarial Science
	Brad McNeney Committee Member Associate Professor, Statistics and Actuarial Science
	Tim Swartz Examiner Professor, Statistics and Actuarial Science
	Reg Kulperger External Examiner Professor Department of Statistical and Actuarial Sciences Western University

Abstract

This thesis investigates two distinct projects: one in statistical genetics focusing on identifying rare causal variants using a sequence-relatedness approach, and another in goodness-of-fit test based on the empirical distribution function (EDF) for any general likelihood model. First, we investigate an association method based on sequence-relatedness for identifying causal variants in a genomic region. We focus on conducting linkage analysis by using sequences as the unit of observation rather than the traditional methods that relied on individuals. We introduce two sequence-relatedness approach to associate similarity in genetic relatedness with similarity in trait values. We compare them to two common genotypic-association methods. Based on a simulation study, we show the efficacy of sequence-relatedness methods in improving the localization and detection of rare causal variants in an allelically heterogeneous disease trait. In addition, a post-hoc labeling procedure based on the idea of genealogical nearest neighbors is introduced to identify potential carriers or non-carriers of causal variants among case sequences. Second, we introduce a goodness-of-fit test based on the EDF in the presence of parameter estimation, which can be applied to any general likelihood model. In summary, the computation of the P-value in goodness-of-fit tests based on EDF with parameter estimation depends on the limiting large-sample covariance function of a stochastic process. This function relies on key elements of the model, including the Fisher information matrix and the derivatives of the cumulative distribution function under the null hypothesis. Computing these elements is often not straightforward and can be computationally intensive or impractical in some cases. In this thesis, we review the theory and propose a new method to estimate the covariance function of the process directly from the sample instead of analytical calculation. We consider two broad cases: when the sample is independent and identically distributed, or when the expected value of the response variable depends on some covariates (e.g., linear model or generalized linear model). Through simulations, we demonstrate the reliability of the estimation method. Finally, we provide computational tools as an R package for practical implementation.

Keywords: linkage analysis; fine-mapping; sequence relatedness; goodness-of-fit test; empirical distribution function; general likelihood model

Dedication

To my parents and brother for their constant love and support.
To my lovely fiancée Miranti, the one who stood by me and made every step of this PhD adventure brighter and better.

"When the world says, "Give up" Hope whispers, "Try it one more time.""

- Unknown

Acknowledgements

I would like to start by thanking my supervisors for their support during my program. Thanks to Dr. Jinko Graham for introducing me to the field of statistical genetics and providing guidance during the initial phase of my thesis. I appreciate your time and energy at the outset of my research journey. I am deeply grateful to Dr. Richard Lockhart for his generous support during the second phase of my program. I have learned a great deal from you. Thank you for your patience, dedication, and unwavering support. Our conversations during our meetings were a source of inspiration for me. I cannot imagine completing this journey without your guidance. Thanks for believing in me and helping to rebuild my confidence. I truly appreciate you.

I would like to express my gratitude to the examining committee members: Dr. Liangliang Wang, Dr. Reg Kulperger, Dr. Tim Swartz, and Dr. Brad McNeney, for taking the time to read and review my thesis. Thank you all for your valuable and constructive comments.

Thanks to the Department of Statistics and Actuarial Sciences for giving me the opportunity. I learned a lot from our wonderful faculty members. A special thanks to Dr. Tim Swartz and Dr. Rachel Altman. I would like to thank our dedicated staff at the Department: Kelly, Sadika, Charlene, Anna, and Caitlin for making the administrative aspects of this journey effortless.

To my supervisors at the University of Tehran: Dr. Hamid Pezeshk and Dr. Ahmad Parsian, from whom I learned a lot. Special thanks to my mentor and friend, Dr. Mohieddin Jafari. I am grateful for the lessons I have learned from you. It is always a pleasure working with you.

Not to forget my friends who were constantly beside me during this difficult journey. I am sincerely grateful to Reza for our thoughtful conversations during our coffee breaks. I extend my thanks to Vahab for being a supportive companion even from a long distance. Thanks, William, for the engaging conversations, delightful tea gatherings, and your support during difficult times. My friendship with you all has been a source of great joy. My good fellow graduate students at the department, thanks for all the laughter and good times together: Alex, Angela, Charlie, Golara, Haoyao, Louis, Matt, Nirodha, Pulindu, Renny, Sashini, and Trevor.

I would like to thank my parents for their constant love and support. Thank you for never giving up on me and for encouraging me to pursue my dreams! To my wonderful brother, your support means everything. Thanks for ensuring I landed on my feet when I came to Vancouver. Last but certainly not least, I want to extend heartfelt thanks to the love of my life, Miranti. Your unwavering support and the sacrifices you made during this time have been nothing short of incredible. I would not be able to finish this journey without you. So let's start our new journey together!

Payman Nickchi
February 26, 2024

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis presents the theory and results of two distinct projects that have been undertaken as part of my thesis. The first part of the thesis discusses a project in the field of statistical genetics where we explore a linkage fine-mapping method to identify the causal variants in a genomic region. In the second part of the thesis, we review the theory of goodness-of-fit tests based on empirical distribution functions and propose a computational method for formal model evaluation in a general likelihood model.

Chapter 2 of the thesis presents a statistical genetics project focused on exploring the feasibility of linkage fine-mapping using sequences instead of individuals. The methods and results of the simulation are described, including the proposed method to associate sequence-relatedness with trait values in order to identify genomic regions containing causal variants. We compare the results of sequence-relatedness methods with genotypic association methods, such as Fisher's exact test and an optimal version of sequence kernel association test known as SKAT-O, to benchmark their performance. The simulation results demonstrate that sequence-relatedness methods improve the localization of rare causal variants and are comparable to genotypic-association methods in detecting them. Additionally, we introduce a post-hoc labeling approach to classify case sequences as potential carriers or non-carriers of causal variants once an association has been established. The results of this project are published in journal of *Genetic Epidemiology*, volume 47, 2023.

Moving on to the second part of the thesis in Chapter 3, we focus on goodness-of-fit tests based on empirical distribution functions. In this work, a computational framework is developed to apply these tests in order to verify the distribution or model assumptions. The chapter starts by reviewing some of the well-known goodness-of-fit test methods based on empirical distribution functions including Cramér-von-Mises and Anderson-Darling statistic. For the case of an i.i.d sample, we review the large sample theory to approximate the P-value for a simple null hypothesis where the distribution is fully specified and no parameter estimation is involved. We show that the problem can be reduced to studying a stochastic process based on the empirical distribution function. The covariance function of

this process is important since a limited number of the eigenvalues of that covariance can be used to approximate the P-value for the test.

In addition to the simple null hypothesis, a more challenging case is studied where a composite hypothesis is being tested. In a composite hypothesis, the model is not fully specified under the null hypothesis. The unknown parameters of the model must be estimated; as a result of estimation, the covariance function of the relevant stochastic process is altered. The exact form of the covariance depends on the estimator; since parameters are usually estimated by the maximum likelihood estimates obtained from the sample only maximum likelihood estimation is considered in this thesis. We review the limiting covariance function which will be used to approximate the P-value for the test. Later in the chapter, we broaden our view to the cases where the sample is not identically distributed and the expected value of the observations depends on some covariates. Some examples of this are linear models and generalized linear models.

Throughout the analysis of both simple null hypotheses and composite hypotheses, we demonstrate that the covariance function of the stochastic process relies on specific characteristics of the assumed model. Notably we need to know the Fisher information matrix and the partial derivatives of the cumulative distribution function with respect to the unknown parameters to compute the covariance function. However, obtaining these quantities can be computationally intensive or challenging in general likelihood models. To overcome this limitation, we propose an alternative method for estimating the covariance function of the stochastic process directly from the sample data rather than carrying out analytical calculations.

The results of a large scale simulation and application of the proposed goodness-of-fit tests are presented in Chapter 4 of the thesis where we evaluate the reliability of the estimation. In these simulations, the estimated version of the covariance function from the sample is used to compute the P-value. In summary, our simulations consists of four different cases, and we apply the goodness-of-fit test in each case. The first and second simulations involve an i.i.d. sample from a Normal distribution and a Gamma distribution, respectively. The third simulation investigates the behavior of the proposed method in a linear model. Lastly, the fourth simulation involves a generalized linear model with a Gamma-distributed response variable. Two different popular link functions (log and inverse function) are considered in this GLM case. We also apply the proposed goodness-of-fit test to third party motor insurance claims in Sweden in 1977. The chapter concludes with a discussion of the work and of possible future research direction. Finally, we have developed a computational package in R to implement the method of covariance function estimation for anyone who is interested to utilize this method in their research. In Chapter 5 of the thesis, a concise overview of the package's functionality is provided. The package, `gofedf`, is published on CRAN repository and is available for download.

Chapter 2

An exploration of linkage fine-mapping on sequences from case-control studies

2.1 Introduction

Linkage analysis is a classic tool to map genetic loci that contribute to a heritable trait. The basic idea is to look for genomic regions that have excess relatedness among individuals with similar trait values [1]. The approach therefore associates similarity in genetic relatedness with similarity in trait values (e.g. [2]; [3]). By contrast, genotypic-association analysis associates specific variants or aggregates of variants directly with trait values. Linkage analysis has traditionally been conducted on related individuals from families. However, the use of families for fine mapping requires many informative meioses [4], either through numerous small pedigrees or large extended pedigrees, and enrolling such families may be impractical.

An alternative that has been proposed for allelically heterogeneous traits is population-based linkage mapping, which gains meioses by adapting linkage analysis to readily-available population-based case-control, cohort or cross-sectional samples [5]. These methods scan individuals for excess ancestral sharing or identity by descent (IBD) at a locus, among individuals with similar trait values. The association between ancestral sharing and phenotypic similarity is assessed along the genome, and regions with high association are singled out for further study. Browning et. al. investigated the power of population-based linkage mapping to detect associations for complex diseases in case-control studies [6]. They contrasted rates of IBD in case/case and non-case/case pairs of individuals at each single-nucleotide variant (SNV), and showed that IBD-based mapping has higher power than genotypic-association mapping when there are multiple, rare causal variants. Their results confirm the expectation that linkage analysis can be more powerful than genotypic-association methods for allelically heterogeneous traits [7].

The linkage analyses reviewed so far consider individuals as the unit of observation. Here, we take a different approach and use sequences as the unit of observation. We use sequences rather than individuals because, at a given genomic location, the gene genealogy connecting the sampled sequences groups them according to their relatedness. We assume that sequences which carry the same rare causal variant descend from a common ancestral sequence. As a result, they are IBD around the variant and will cluster together on its local gene genealogy. Genomic regions with excess trait clustering on their local genealogy therefore indicate a causal locus.

In this chapter of thesis, we explore the feasibility of linkage fine-mapping on sequences, as an alternative to standard genotypic association mapping. In particular, we compare the ability of linkage and genotypic-association approaches to map an allelically heterogeneous disease of high penetrance. High-penetrance variants produce familial clusters that are easier to detect. Linkage methods are predicted to work well in such circumstances [7]. We consider two linkage or descent-based methods that associate similarity in relatedness of sequences with similarity in trait values. For comparison, we consider two genotypic-association methods, one which considers single variants and another which aggregates variants. Through a simulation study, we compare the ability of these methods to fine-map rare causal variants in a 2 million base-pair (Mbp) candidate region. We chose 2 Mbp because it is the approximate resolution of a moderate-sized linkage study in pedigrees. For example, a linkage analysis with 100 informative meioses is expected to map a disease locus to within 2 centiMorgans, or approximately 2 Mbp [4].

To illustrate ideas, we work through an example dataset as a case study. Following this, we use a coalescent simulation to evaluate the ability of these methods to *detect* and *localize* a disease locus. Specifically, we are interested in the ability to detect any association within the candidate genomic region being fine-mapped, and also in the ability to localize the association signal to the causal subregion within the candidate region. Having detected a disease locus, it is of interest to identify case sequences that may be carriers of a causal variant. We conclude by describing a *post hoc* labeling procedure to classify case sequences into carriers and non-carriers of causal variants, using estimated sequence relatedness.

2.2 Materials and Methods

In this section, we describe how we simulated the genetic data and disease phenotype given the genetic data. Next, we describe the association methods that we considered to *detect* and *localize* causal variants. Finally, we propose a method for *post-hoc* labeling of case sequences into carriers and non-carriers of causal variants, given that an association has been detected.

2.2.1 Genetic-data simulation

We used *msprime* to simulate the gene genealogy and sequences across a 2 Mbp genomic region for an entire population [8]. We applied a hybrid strategy in which a backwards Wright-Fisher model with recombination and mutation was run to 5000 generations before present, followed by a coalescent with recombination and mutation from 5000 generations back to the overall most recent common ancestor across the genomic region [9]. The hybrid strategy avoids inaccuracies in the coalescent approximation when the number of sampled sequences is large relative to the population effective size. The diploid population was of constant effective size, $N_e = 3100$, and consisted of 6200 sequences [10]. We used a recombination rate of 1×10^{-8} per base per generation and a mutation rate of 2×10^{-8} per base per generation to simulate 500 populations [11, 12]. Figure B.1 displays the distribution of variant allele frequencies in the simulated population from which the example dataset was drawn. This allele-frequency spectrum is similar across the 500 simulated populations. The spike in the lowest-frequency bin of the histogram (frequency ≤ 0.01) is consistent with an observed abundance of rare variants in real populations [13].

2.2.2 Disease-trait model

To mimic random mating in a diploid population, we randomly paired the population sequences into 3100 individuals. Case-control status was assigned to individuals in the population based on causal SNVs (cSNVs) randomly sampled from the middle 900-1100 kbp of the 2 Mbp candidate genomic region. For cSNVs, the risk of disease increases according to a logistic-regression model:

$$logit(P(D = 1|G)) = \beta_0 + \beta_1 \sum_{j=1}^{m} G_j,$$

where

- $logit(p) = log[\frac{p}{1-p}]$ for $0 < p < 1$,
- D is the disease status (D=1, case; D = 0, control),
- $G = (G_1, G_2, \cdots, G_m)$ is the multi-locus genotype of an individual at m causal SNVs, where G_j indicates the number of copies of the derived allele at the jth cSNV,
- β_0 is the intercept of the model and controls the sporadic-disease rate, i.e $P(D = 1|G = 0)$, and
- β_1 is the effect parameter which measures the influence of causal variants on the disease.

Simulations under the null hypothesis

We randomly assigned disease status to the 3100 individuals in the population. To ensure a disease prevalence of 5% in the population, 155 out of 3100 individuals were randomly assigned as disease-affected individuals. To form our case-control sample, we randomly sampled 50 cases from the 155 affected individuals and 50 controls from the 2945 unaffected individuals in the population.

Simulations under the alternative hypothesis

We used the disease-trait model above with parameter values set to ensure a high penetrance and low phenocopy rate consistent with genetic-linkage studies [7]. In particular, we set $\beta_0 = -10$ so that the phenocopy rate $f = P(D = 1|\sum_{j=1}^{m} G_j = 0) = 4.5 \times 10^{-5}$, and $\beta_1 = 16$ so that the genetic penetrance $g = P(D = 1|\sum_{j=1}^{m} G_j \geq 1) \approx P(D = 1|\sum_{j=1}^{m} G_j = 1) = 0.9975$. The penetrance ratio was therefore $g/f \approx .9975/4.5 \times 10^{-5} = 22167$. We aimed for an allelically heterogeneous disease with 15 rare cSNVs of roughly equal frequency in the population. To achieve the targeted disease prevalence, each cSNV had a population allele frequency around 0.16 percent (about 10 copies in the population of 6200 sequences). When necessary, additional very rare variants were chosen to be causal to attain the targeted 5% disease prevalence. Further details about the selection procedure for causal variants can be found in Appendix C. After assigning disease status to the 3100 individuals in the population, we randomly sampled 50 cases from the affected individuals and 50 controls from the unaffected individuals. We then extracted the SNV sequences of the case-control sample for analysis.

2.2.3 Genotypic association

We consider Fisher's exact test and an optimized sequence-kernel association test [14]. These methods test for association between the trait and the genotypes, either one-at-a-time or in aggregate, and do not consider the relatedness of sequences.

Fisher's exact test

We use a standard Fisher's exact test of disease association with genotype frequencies for each SNV implemented in the `stats` package in base R. Specifically, each of the SNV sites is tested for an association with the disease outcome using a 2×3 table to compare the genotype frequencies. At each SNV, our test statistic is the exact P-value, expressed in the negative, base-10-logarithmic scale.

SKAT-O

Single-variant association tests such as Fisher's exact test have limited power to detect rare variants [15]. To improve power, aggregation methods, such as the sequence-kernel

association test, collapse variants in a window of SNVs into a one-number summary that is then used to test for association. We consider the adaptive aggregation test SKAT-O [14] which finds the optimal linear combination of the burden test [16] and the sequence kernel association test in terms of power. We applied the SKAT-O test implemented in the SKAT R package [17] to each SNV, using a window size of 21 SNVs ($\sim 14-15$ kbp in the simulated datasets). The window includes the target SNV at the center and 10 SNVs to the right and left. Target SNVs at either edge of the candidate region had a smaller window size than 21. For example, the window centered at the first SNV has no SNVs to the left and 10 SNVs to the right and thus contains 11 SNVs in total. At each SNV, we record the P-value, expressed in the negative-base-10-log scale, as the test statistic.

2.2.4 Descent-based association

Rather than associating genotype frequencies with trait values, we propose instead a linkage analysis that associates similarity in sequence-relatedness with similarity in trait values. Since sequences carrying a causal variant tend to cluster on the gene genealogy around the variant, we expect sequence relatedness and trait similarity to be associated in genomic regions harbouring causal variants. As the true gene genealogy is unknown, we reconstruct sequence partitions on the genealogy from the sequence data and calculate distances on these partitions. We then calculate trait dissimilarities between sequences and use them to assess the association between the clustering of sequences and trait values.

Sequence partitions and their distances

To reconstruct sequence partitions, we apply the clustering methods implemented in the R package, `perfectphyloR` [18]. The package takes the sample sequences and returns a perfect phylogeny for a focal SNV. The perfect phylogeny is a rooted tree that recursively partitions DNA sequences [19]. These nested partitions provide insight into the relatedness of sequences around a focal SNV. Sequences descending from a common ancestral mutation tend to cluster together in a partition. We use the `reconstructPPregion()` function from `perfectphyloR` package with a minimum window size of 500 variants to reconstruct partitions across the 2-Mbp genomic region. We use a large window size to help resolve non-identical sequences in the reconstruction. Note that the sequence partitions provide no information on coalescence times or the ordering of non-nested coalescence events and so are not genealogical trees. They do however provide information on the nested structure of sequence clusters and therefore on sequence relationships.

At each SNV, we measure the scaled pairwise distances between sequences on the partitions as described in [20]. These distances measure how closely sequences are related around a focal SNV. To compute the pairwise distances, we apply the `perfectphyloR` function `rdistMatrix()`. Partitions may change along the genome due to recombination. As a re-

sult, the pairwise distances may differ for different focal SNVs. A small example of sequence distances on a partition is presented in the Appendix, Figure D.1.

Phenotypic distances are computed as described in [21]. These distances measure the trait dissimilarity of sequences. Briefly, the phenotypic distance between sequence i and j is defined to be $d_{ij} = 1 - s_{ij}$, where $s_{ij} = (y_i - \mu)(y_j - \mu)$ is the phenotypic similarity score between sequence i and j, y_i is the binary phenotype (0 for control or 1 for case), and μ is the disease prevalence in the population. For our disease prevalence of 5%, the phenotypic distances are essentially dichotomous. In one group, the distances between case sequences take on the same low values while, in the other group, the distances between control sequences or between case and control sequences take on similar high values.

Measures of association

We associate sequence and phenotypic distances in two ways: via the distance correlation as described in [22] or the Mantel coefficient [23]. The distance correlation measures non-linear dependence between two random vectors but can be expressed in terms of pairwise Euclidean distances [24]. In contrast, the Mantel coefficient measures linear dependence between elements of two distance matrices which do not necessarily have to be Euclidean. At each SNV, we record the distance correlation or Mantel coefficient as the test statistic.

2.2.5 Scoring detection

Through simulation, we compare the abilities of the two genotypic and two descent-based methods to both *detect* association and *localize* causal SNVs. For detection, we are interested in finding *any* association across the entire candidate region. For localization, we are interested in mapping the locus harboring causal variants. We describe a global test to detect association across the entire region and the empirical distribution function (EDF) to graphically compare the resulting global tests. We also describe how we compute the type-I error rate and power of the global tests.

Global tests

For each dataset, we use the maximum test statistic across all the SNVs to obtain a global test statistic across the candidate genomic region. We obtain the null distribution of this global test statistic by randomly permuting the case-control labels of the individuals 1000 times. The P-value for the global test is defined as the proportion of test statistics that are greater than or equal to the observed value. The nominal level of all tests is 5%.

Empirical distribution functions

To compare the distribution of p-values for each of the methods, we plot their empirical distribution functions or EDFs. The EDF at any point $x \in (0, 1)$ indicates the proportion

of simulated datasets with a p-value less than or equal to x. Therefore, any method with higher EDF at x has a larger proportion of simulated datasets with p-value less than or equal to x.

Type-I error rate and power

The estimated type-I error rate and power of each method is respectively the proportion of the 500 datasets simulated under the null or alternative hypothesis that are rejected at level 5%. Type-I error rate and power can be extracted from the EDF. For example, when datasets are simulated under the alternative hypothesis, any method with higher EDF at $x = 0.05$ appears to be more powerful at level 0.05. To assess whether the power of two methods differs, we apply McNemar's test to the EDFs evaluated at $x = .05$ [25]. We use McNemar's test to account for dependence in test results from the same dataset.

We are particularly interested in the type-I error rate of the Mantel test because it is known to be biased (i.e to have inflated type-I error rate) when the units being permuted are non-exchangeable under the null hypothesis [26]. In our context, the sample sequences are not exchangeable owing to their underlying ancestry. However, under the null hypothesis of no association, the case-control status of the individuals being permuted is exchangeable. We use a normal approximation to the binomial distribution to obtain an approximate 95% confidence interval for the type-I error rate.

2.2.6 Scoring localization

To evaluate the localization ability of each method, we calculate the distance of the maximum absolute association signal from the causal region, in base pairs. If more than one maximum is encountered, we take the average of all maxima. We then calculate the EDF of these average distances for the 500 simulated datasets. The EDF at any point $0 \leq x \leq 2000$ kbp gives the proportion of simulated datasets with peak association signal within x kbp of the causal region. Therefore, any method with higher values of the EDF at a given value x appears to localize better, within a distance of x kbp. To assess whether the localization ability of two methods differs, we apply McNemar's test to the EDFs evaluated at $x = 0$.

2.2.7 Post-hoc labeling of case sequences

We propose a procedure to label case sequences as potential carriers or non-carriers of causal variants. Our approach relies on the concept of a genealogical nearest neighbor or GNN [27]. GNNs arise from the topological properties of genealogical trees, as summarized by the sequence partitions. Case sequences that carry a given rare variant are descended from a common ancestral mutation that arose relatively recently back in time. Therefore, we expect these case sequences to cluster in the sequence partition as GNNs. The GNN proportion of a sequence for a given partition is the proportion of its nearest neighbors in the partition that

are case sequences. We then average this proportion over all sequence partitions along the genomic region to obtain an average GNN proportion. A worked example of the calculation of average GNN proportions is illustrated in Appendix E. Briefly, any case sequence whose average GNN proportion is consistent with the distribution of GNN proportions in controls is declared to be more closely related to controls than to cases. We group case sequences into carriers and non-carriers of causal variants according to their average GNN proportion. We consider the median of the distribution of GNN proportions in control sequences as our threshold. Specifically, any case sequence with an average GNN proportion less than the median of the average GNN proportion in control sequences is labelled as a non-carrier. We refer to this grouping as the *GNN labeling* of case sequences.

Since we have simulated the sequences and genealogies, we know the true carrier status of each sequence. Therefore, we can compare the accuracy of our GNN labeling to naive labeling, in which all case sequences are assumed to be carriers. Table 2.1 presents an example of a confusion matrix for the carrier status of $N = a+b+c+d$ case sequences in a simulated dataset. Referring to this confusion matrix, we see that the observed misclassification rate is $\frac{b+c}{N}$.

Table 2.1: Example confusion matrix of carrier status for N case sequences

		GNN-predicted status	
		Non-Carrier	Carrier
True status	Non-Carrier	a	b
	Carrier	c	d

N=a+b+c+d

2.3 Results

To start, we present an analysis of an example dataset to give insight into the association methods. We then present estimated type-I error rates and rates of *detection* and *localization* of the causal region. Finally, we present misclassification error rates for the proposed post-hoc labeling of case sequences.

2.3.1 Example dataset

We first summarize the causal variants in the population and sample. Next, we show profiles of various association statistics across the candidate genomic region and apply the proposed procedure for post-hoc labeling of the case sequences.

Population and sample summaries

The population of 3100 individuals (6200 sequences) has 4723 SNVs in a 2 Mbp genomic region. Among these SNVs, 2904 are segregating in the sample of 50 case and 50 control

individuals, including all 15 cSNVs. In the population, all sequences and all but one individual carry zero or one cSNV. The one individual with two carrier sequences is included in the sample as a case. Table 2.2 summarizes the causal variants in the sample and population. The column labeled "Position (kbp)" gives the physical position of cSNVs along the genome in kbp. The columns labeled "Population" and "Sample" count the number of case and control sequences that are carrying any causal variants in the population and sample, respectively. The column labeled "DAF" gives the derived allele frequency of causal variants in the population, expressed as a percentage. The causal variants are all rare with a maximum population DAF of 0.19%. A total of 154 and 51 case sequences carry a cSNV in the population and sample, respectively. None of the control sequences carry cSNVs.

Table 2.2: Summaries of causal variants in the sample and population.

		Population sequences			Sample sequences	
	Position	Case	Control		Case	Control
cSNV	(kbp)	(154)	(0)	DAF%	(100)	(100)
1	928.761	9	0	0.14	4	0
2	937.392	12	0	0.19	7	0
3	940.023	12	0	0.19	4	0
4	942.571	9	0	0.14	5	0
5	946.127	10	0	0.16	1	0
6	993.008	12	0	0.19	4	0
7	994.439	10	0	0.16	3	0
8	998.710	11	0	0.18	4	0
9	1002.568	9	0	0.14	2	0
10	1003.525	11	0	0.18	2	0
11	1016.514	9	0	0.14	1	0
12	1039.256	10	0	0.16	4	0
13	1045.524	11	0	0.18	3	0
14	1054.265	10	0	0.16	4	0
15	1082.301	9	0	0.14	3	0

Association profiles

The association profile is a scatter plot with genomic coordinates on the horizontal axis and SNV-specific measures of association on the vertical axis. Figure 2.1 presents the association profiles of different methods for the example dataset. The x-axes in all panels is the genomic position in kbp. The y-axes show either a transformed SNV-specific P-value (genotypic association methods), or an SNV-specific measure of association (descent-based association methods). The vertical red-dashed and horizontal red-dotted lines indicate, respectively, the

causal region from which the cSNVs were randomly selected and the 5% significant threshold for the global test of any association. The Mantel, SKAT-O and distance correlation tests detect significant association, but the Mantel test is the only method that correctly localizes the causal region. The profiles for Fisher's exact test and distance correlation in panels (a) and (c) appear similar, and the peak association signal of both methods occurs in approximately in the same genomic position. We will return to this point later when discussing the simulation results.

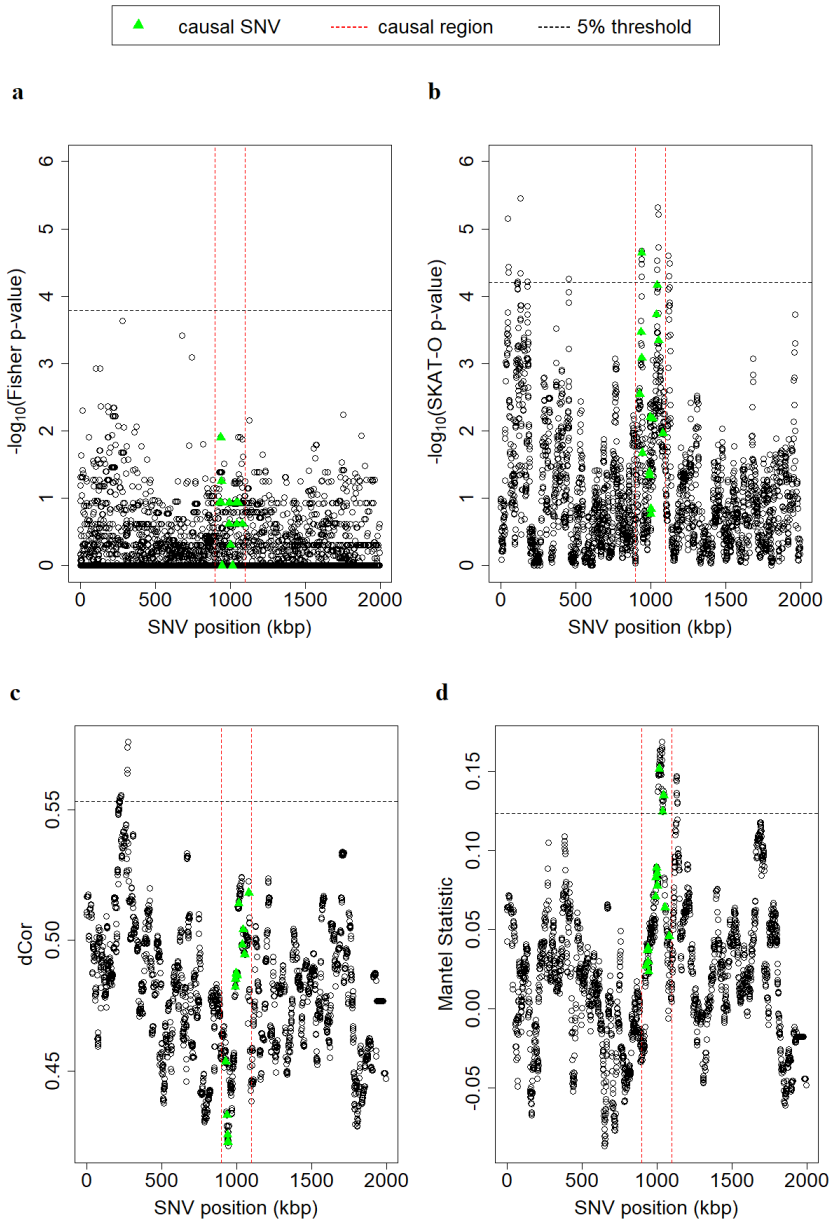


Figure 2.1: Association profiles, a) Fisher's exact test (FET), b) SKAT-O, c) distance correlation (dCor), and d) Mantel. The vertical red-dashed lines indicate the region from which the causal SNVs were selected. The green triangles represent the causal SNVs. The maximum value of SNV-specific statistics over the entire genomic region is used in a permutation test for the presence of any association. The horizontal dotted line represents the 5% significant threshold based on 1000 permutations of the individual disease phenotypes. The detection P-values for FET, SKAT-O, dCor, and Mantel are 0.089, 0.005, 0.008, and 0.002 respectively.

Post-hoc labeling of case sequences

We apply the proposed GNN-labeling procedure to classify case sequences in the example dataset. Figure 2.2 shows the boxplots of average GNN proportions of sequences grouped by their status as case carriers or case non-carriers of causal variants and controls. We use the median of average GNN proportion in control sequences to classify the case sequences into carriers and non-carriers. In the example dataset, all but three of the true carriers are correctly predicted by the GNN labeling.

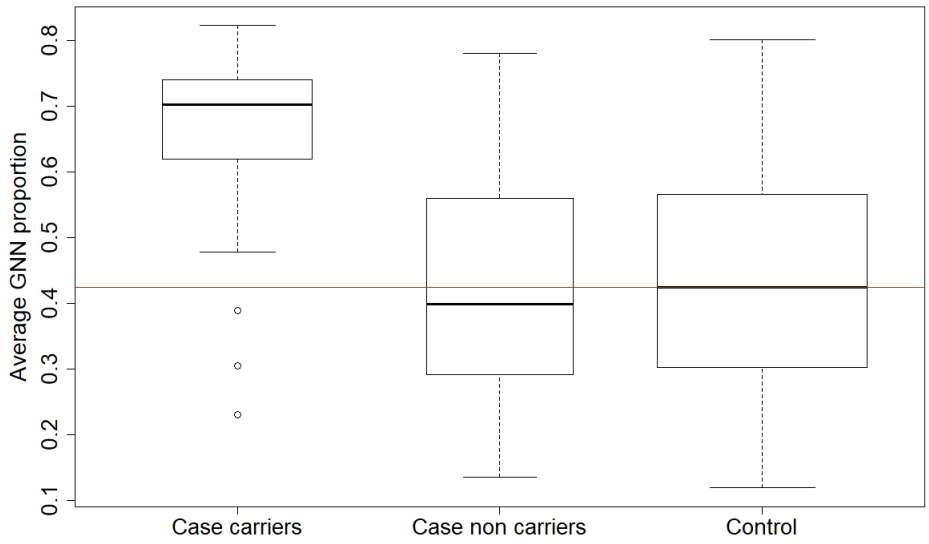


Figure 2.2: Average GNN proportions of sequences grouped by their status as case carriers of causal variants, case non-carriers of causal variants or controls. The horizontal red line is the median of the average GNN proportion in control sequences.

Tables 2.3 (a) and (b) show the confusion matrices for naive and GNN labeling, respectively. Naive labeling considers all 100 case sequences to be carriers of a cSNV. The observed misclassification rates for naive and GNN labeling are 49% and 25%, respectively. Post-hoc, GNN labeling therefore improves the identification of carriers of a cSNV among case sequences. Table 2.4 considers the 51 case sequences that carry a cSNV in the example dataset, and presents the number that are correctly predicted by GNN labeling. Three carrier case sequences are incorrectly predicted, corresponding to the cSNVs in highlighted rows of the table. As the case sequences carrying cSNVs 5 and 11 are sample singletons, we would not expect their genealogical nearest neighbours to be over-represented by case sequences. Thus, we would not expect the GNN labeling to correctly predict their carrier status.

Table 2.3: Carrier status for $N = 100$ case sequences in the example dataset using a) naive labeling and b) GNN labeling.

(a) Naive labelling

		Naive status	
		Non-Carrier	Carrier
True status	Non-Carrier	0	49
	Carrier	0	51

N=100

(b) GNN labelling

		GNN-predicted status	
		Non-Carrier	Carrier
True status	Non-Carrier	27	22
	Carrier	3	50

N=100

Table 2.4: Carrier case sequences for each cSNV and number predicted by GNN labeling.

	Sample sequences	
cSNV	Case	GNN predicted
1	4	4
2	7	7
3	4	4
4	5	5
5	1	0
6	4	4
7	3	3
8	4	4
9	2	2
10	2	2
11	1	0
12	4	4
13	3	3
14	4	3
15	3	3

2.3.2 Detection

We present the simulation results for type-I error rates first, then the results for power.

Type-I error rate

To estimate type-I error rates, we considered 500 datasets simulated under the null hypothesis of no association with cSNVs. Figure 2.3 shows the empirical distribution functions (EDFs) of the permutation P-values from a global test of association across the entire genomic region, for each of the association methods. In both panels, P-values are labeled in the natural scale but plotted in the log-10-scale. The y-axis in the left panel of the figure is shown up to 0.30. The right panel of the figure magnifies EDFs around the 5% significance level. Figure 2.4 presents point and approximate 95%-confidence interval estimates of the type-I error rates. The results of both figures suggest that the type-I error rates of all association methods are controlled at the nominal 5% level. Numerical values for the point and 95% confidence-interval estimates of the type-I error rates are reported in Table F.1.

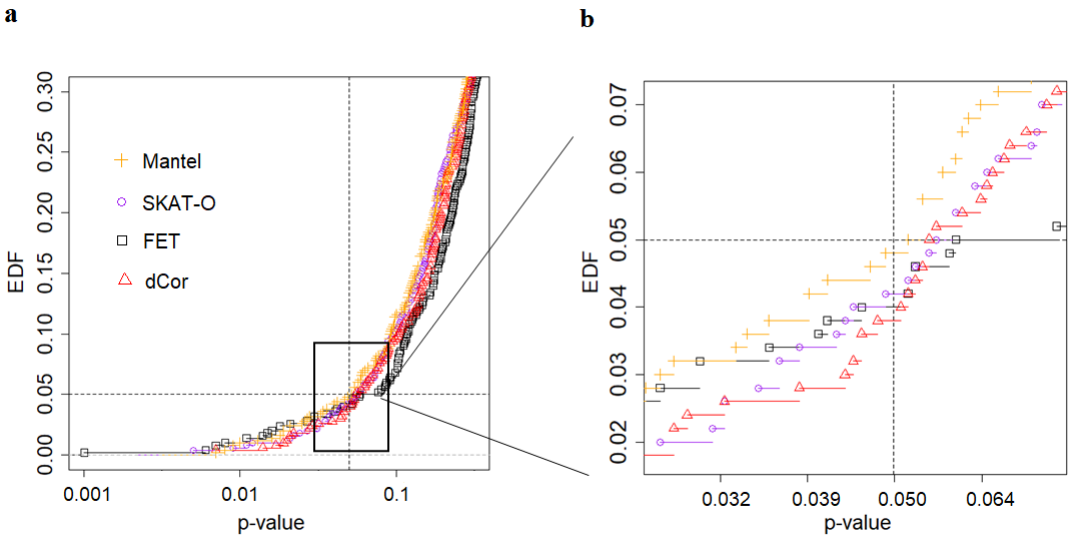


Figure 2.3: Empirical distribution functions (EDFs) of permutation P-values from a global test of association across the genomic region. Four methods are compared: Fisher's exact test (FET), SKAT-O, distance correlation (dCor) and Mantel. a) Original b) Zoomed version. On the x-axis, P-values are labeled in the natural scale but plotted in the log-10-scale. The vertical and horizontal dashed lines indicate the nominal 5% level.

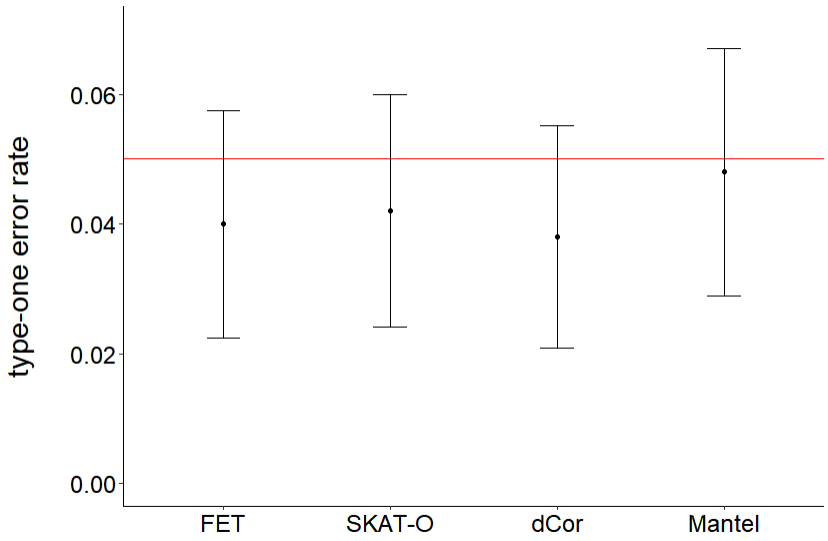


Figure 2.4: Point and approximate 95%-confidence interval estimates for type-I error rate in Fisher's exact (FET), SKAT-O, distance correlation (dCor), and Mantel tests. The horizontal dashed line is the nominal 5% level.

Power

The EDFs shown in Figure 2.5 are based on 500 datasets that have been simulated under the alternative hypothesis of association with causal SNVs. The EDFs suggest that the SKAT-O and Mantel tests outperform the other tests for detecting the association signal. The performance of the SKAT-O and Mantel tests at level 5% does not differ significantly (McNemar P-value = 0.60).

A scatter plot of detection P-values from the SKAT-O and Mantel tests is shown in Figure 2.6. The Pearson correlation between these two sets of P-values is 0.175 and differs significantly from zero ($p < 0.0001$). In 62 of the 500 datasets, the Mantel test detects the association signal but the SKAT-O test does not (fourth quadrant). In 68 datasets, the SKAT-O test detects the association signal but the Mantel test does not (second quadrant). Both methods detect the association signal in 345 datasets (third quadrant) and neither method detects the signal in 25 datasets (first quadrant). The observed discordance rate between the tests is 130/500, or 26%. These results suggest that the SKAT-O and Mantel tests are picking up on different aspects of the association signal.

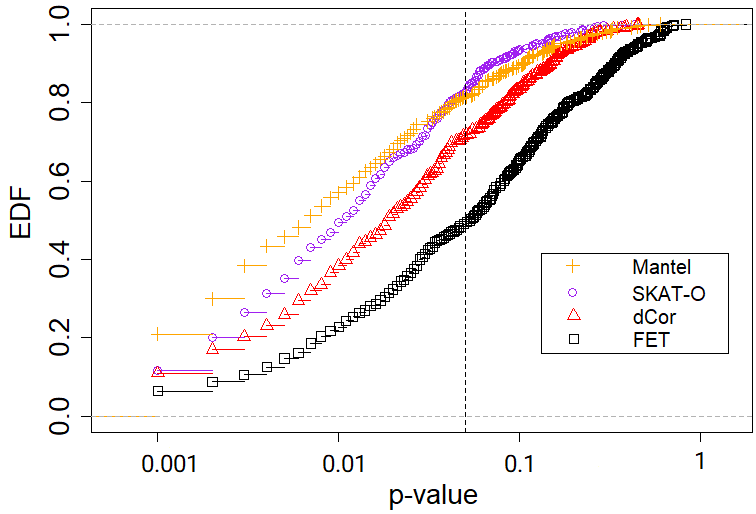


Figure 2.5: Empirical distribution functions (EDF) of permutation P-values from a global test of association across the genomic region. Four methods are compared: Fisher's exact test (FET), SKAT-O, distance correlation (dCor), and the Mantel statistic. On the x-axis, P-values are labeled in the natural scale but plotted in the log-10 scale. The vertical dashed line indicates a P-value of 0.05.

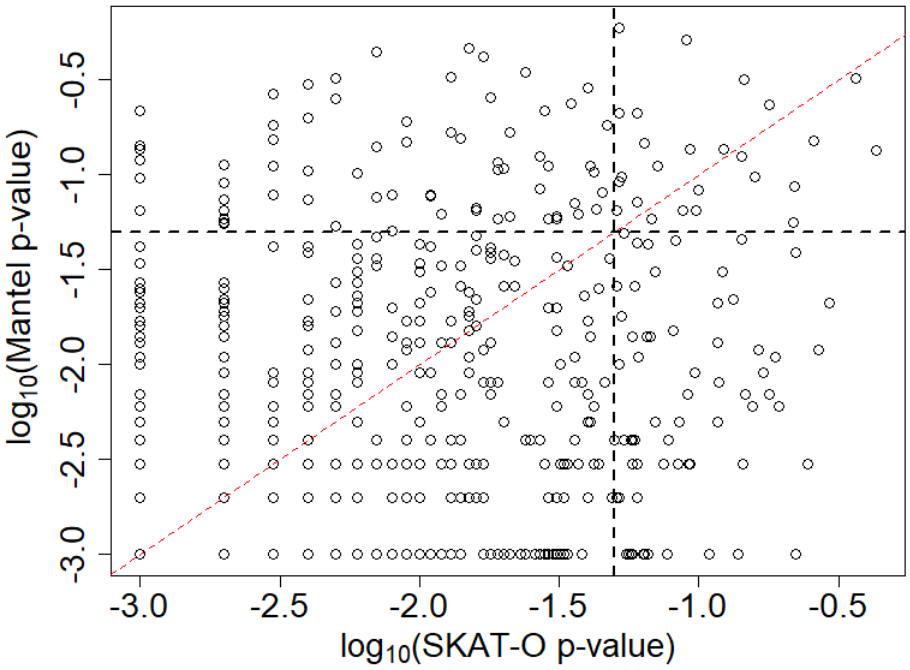


Figure 2.6: The relationship between P-values from the Mantel test and SKAT-O in the log-10 scale. The vertical and horizontal black-dashed lines show P-values of 0.05. The Pearson correlation between the transformed P-values is 0.175 (p <0.0001). The red-dashed line is y=x.

2.3.3 Localization

The EDFs of average distances from the causal region are shown in Figure 2.7, for the four association methods. The Mantel profile appears to localize the causal region far better than any of the others, followed by SKAT-O. In fact, the Mantel profile localizes significantly better than SKAT-O (McNemar P-value = 0.0042)

From the figure, we see that Fisher's exact test and the distance correlation localize about 10% of the 500 simulated datasets to the causal region. However, the causal region comprises 10% of the candidate region being fine-mapped, and so these two methods are localizing no better than random chance. In the example dataset, Fisher's exact test and the distance correlation had similar association profiles which localized the peak signal to roughly the same genomic position. To investigate the co-localization properties of the methods, we calculated the Pearson correlation of their average distances from the causal region. Amongst all pairs of methods, the maximum correlation of 0.30 (p-value $\approx$ 0) belongs to Fisher's exact test and the distance correlation. Our findings suggest that Fisher's exact test and distance correlation tend to co-localize the association signal more than any other pair of methods considered.

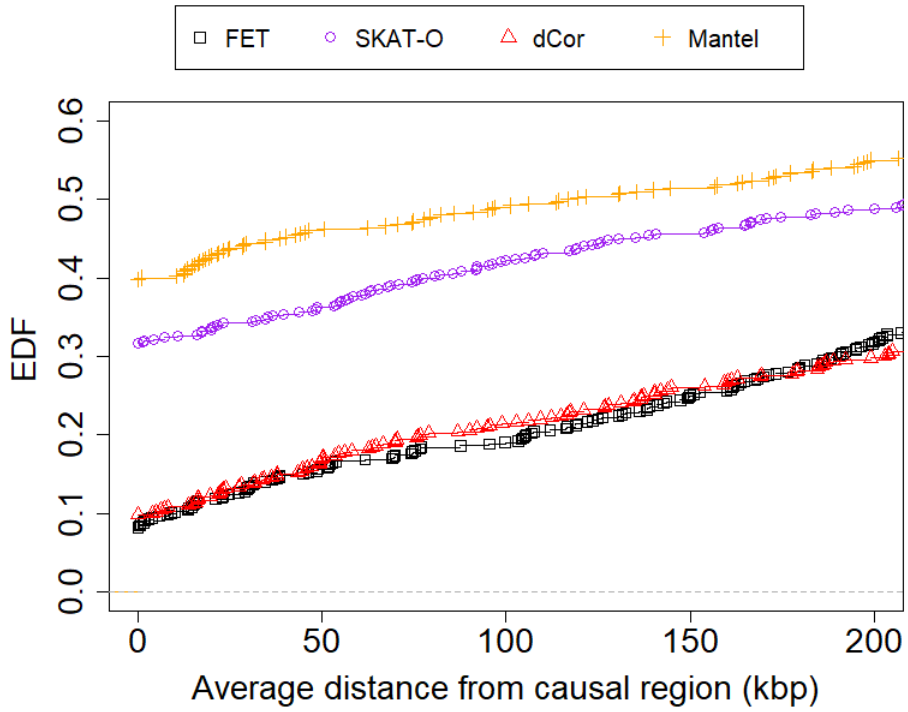


Figure 2.7: Empirical distribution functions (EDFs) for the average distance of the peak association signal from the causal region, for 500 datasets simulated under the alternative hypothesis of association. Four methods are compared: Fisher's exact test (FET), SKAT-O, distance correlation (dCor) and Mantel. To make the comparison easier and for better resolution, the x-axis is shown only for genomic distances less than 200 kbp.

2.3.4 Performance of case-sequence labeling

We use the 500 datasets simulated under the alternative hypothesis to compare the performance of GNN labeling of case sequences to a naive labeling scheme in which all case sequences are assumed to be carriers of cSNV. For the 100 sampled case sequences in each dataset, we compute the misclassification rates of the labeling procedures. Figure 2.8 shows the scatter plot of these rates for the 500 datasets, with the GNN rates on the vertical axis and the naive rates on the horizontal axis. The naive misclassification rates on the horizontal axis have only four values (0.48, 0.49, 0.50, and 0.51), which have been randomly perturbed for better viewing. From the red-dashed line indicating $y = x$, we can see that GNN labeling has a uniformly lower misclassification error rate than naive labeling, across all 500 datasets. Thus, post-hoc labeling of case sequences by the GNN procedure may be of practical use for predicting which case sequences carry causal variants.

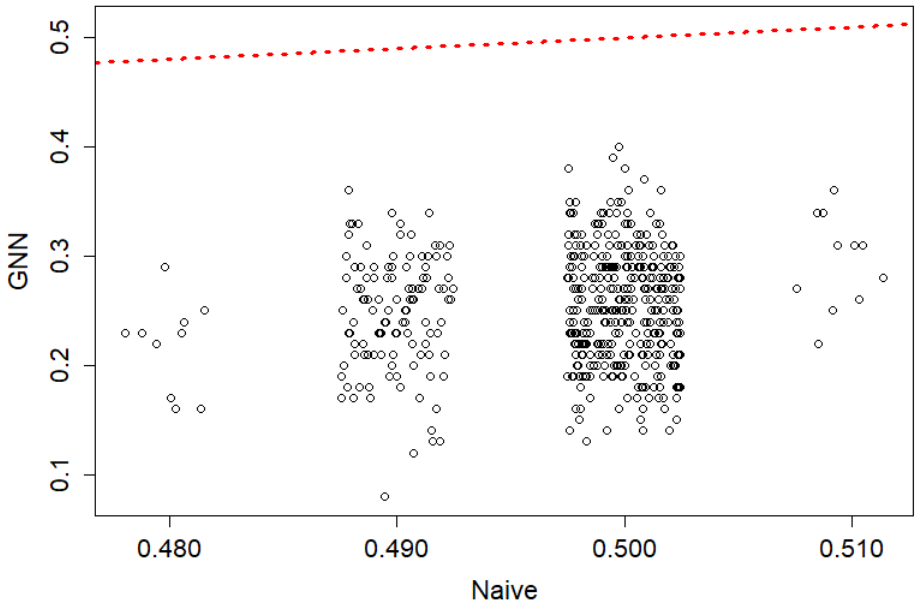


Figure 2.8: Misclassification error rate of cSNV carrier status in case sequences, for GNN versus Naive labeling across 500 simulated datasets. The red-dashed line is $y = x$.

2.4 Discussion and Conclusion

We have explored the feasibility of linkage fine-mapping on sequences for an allelically heterogeneous disease. In particular, we have compared the ability of linkage and genotypic-association approaches to detect and fine-map a causal locus in simulated datasets. The linkage methods we have considered use sequences rather than individuals as the unit of observation. These methods associate similarity in relatedness of sequences with similarity in trait values, using either the distance correlation or Mantel coefficient as a measure of association. For comparison, we include two genotypic-association methods, a single-variant Fisher's test and the SKAT-O test that aggregates variants. While our linkage methods use haploid sequences and the genotypic association methods use diploid individuals as the unit of observation, *both* classes of methods assume a diploid disease model. A consequence for the linkage methods is that some case individuals will have sequences that do not carry causal variants. We therefore introduce a *post-hoc* procedure to group case sequences into carriers and non-carriers of causal variants, inspired by the idea of genealogical nearest neighbors (GNN) described in [27]. We view this simulation investigation as a proof of principle illustrating the potential of sequence-based linkage approaches. A future direction of research would be to compare the linkage and SKAT-O approaches on real datasets from fine-mapping studies.

Our example data analysis and simulation study indicate that sequence-based linkage methods are useful for population-based fine-mapping of an allelically heterogeneous disease. In particular, we find that: (i) a linkage-based Mantel test detects rare causal variants as well as a state-of-the-art genotypic-association test, SKAT-O; (ii) the Mantel profiles best localize rare causal variants among all the methods; and (iii) GNN labeling of case sequences is helpful for removing sequences that do not carry causal variants. These findings suggest the following strategy for fine-mapping an heterogeneous disease in case-control samples. First, detect disease association with either the Mantel or SKAT-O test. Once a disease association has been detected, localize the causal region with the Mantel association profile and refine the case sequences by removing those labeled as non-carriers by the GNN procedure. The putative causal locus and putative carrier case sequences can then be searched for causal variants. Our work extends earlier investigations of sequence-based linkage mapping that relied on the known gene genealogies [20], [28]. Instead, we infer the topological structure of unknown genealogies from sequence data. Our sequence-based approaches are therefore practical for linkage fine-mapping with population-based data.

We evaluated the type-I error rate using datasets simulated under the null hypothesis of no association. Our results suggest that the type-I error rate of all the methods is well controlled at the 5% nominal level. The Mantel test was of particular interest because it has been criticized for inflated type-I error rates when the units being permuted are non-exchangeable [26]. In our context, however, the size of the Mantel test is maintained because the case and control status of individuals is exchangeable under the null hypothesis.

Fisher's exact test had the lowest power of all the methods, as expected for a single-variant method detecting rare variants. The Mantel test of linkage and the SKAT-O test of genotypic-association had the highest power of all methods. The estimated power of these tests at level 5% did not differ significantly (McNemar P-value = 0.60). The similar power of the Mantel and SKAT-O tests prompted us to look into the agreement of their p-values across the simulated datasets (Figure 2.6). The random pattern in the figure as well as the observed discordance rate of 26% between the two significance tests at level 5% suggests that the Mantel and SKAT-O tests detect different aspects of the association signal. The complementary nature of the tests indicates that a combined test, e.g., using Fisher's method of combining p-values (e.g. [29]), could be more powerful than either the Mantel or SKAT-O tests alone. Investigating the power of combined tests is an area for future work. The Mantel test had higher detection power than the distance-correlation test (Figure 2.5). We note that the Mantel test is well suited to a disease of low prevalence (5%) because the trait distances, between case/case pairs on one hand and case/non-case and non-case/non-case pairs on the other, are essentially binary (results not shown). The relationship between trait distances is then essentially a straight line and therefore well captured by the Pearson correlation coefficient. By contrast, the distance correlation assumes Euclidean distances [24] and so may be unsuitable for our partition distances between sequences.

The principal finding of our simulation study is that, when the penetrance model favours linkage analysis, the Mantel association profile localizes the causal region *significantly better* than the next-best SKAT-O (McNemar P-value = 0.0042). In contrast, Fisher's exact test and the distance correlation localized the causal region the worst, and in fact did no better than random guessing of the location. Interestingly, among all pairs of methods considered, the average distance of the peak association signal from the causal region was the most highly correlated for the Fisher's exact and distance-correlation methods ($r = 0.30$, p-value≈ 0). In some datasets, the peak association signals for Fisher's exact test arise from synthetic associations with common SNVs outside the causal region that happen to tag multiple causal SNVs by chance [30]. It would be interesting to investigate whether distance correlation is also vulnerable to synthetic associations, given its tendency to co-localize the association signal with Fisher's exact test.

Throughout, we have applied SKAT-O with a window size of 21 SNVs. This window size corresponds to a genomic region of roughly 14-15 kb, about the size of a typical human gene, in our simulated datasets. We also investigated the detection and localization properties of SKAT-O using windows of 11, 41, 61, and 81 SNVs (results not shown). We found that a SKAT-O window size of 61 SNVs yielded slightly greater detection power than the Mantel test. SKAT-O localization rates improved slightly, up to a window size of 61 SNVs, before falling off for larger window sizes. However, no SKAT-O window size achieved better localization than the Mantel method. Such tuning of the SKAT-O window size is not possible in practice, as it will depend on unknowns such as the presence and location of any causal variants. In addition, SKAT-O already involves optimizing over a linear combination of its constituent burden and variance-component tests, so that optimizing over window size would add extra computational burden. Development of practical and feasible procedures for tuning the SKAT-O window size would be an interesting avenue for further investigation.

Our results suggest that sequence-based linkage analysis is useful for fine-mapping allelically heterogeneous traits. To start the discussion, we have used simulation to explore disease traits in a case-control study design, under a genetic architecture that favors linkage methods. In particular, we simulated high-penetrance, low-frequency causal variants. Examples of diseases influenced by high-penetrance, low-frequency variants are familial breast and ovarian cancer [31], familial bipolar disorder [32], hearing impairment, familial goitres and familial hypertension [7]. In the absence of allelic heterogeneity, we do not expect sequence-based linkage methods to offer advantages over genotypic association methods such as SKAT-O, for either detection or localization. By analogy to family-based linkage analysis, lower penetrance ratios are expected to reduce the effectiveness of sequence-based linkage analysis in relation to association approaches such as SKAT-O. Further simulations under a larger variety of allele frequency and penetrance parameters are an important direction for future work.

A related area of future work is to expand the scope to other study designs and genetic architectures. For example, how is sequence-based fine-mapping affected as we vary the type of trait (i.e. binary disease versus quantitative), the type of population-based study design (e.g. case-control, cohort or cross-sectional sampling), and the number and frequency of causal variants within the trait locus? We would also like to investigate the impact of sequencing errors on the sequence-based linkage methods. Our initial thoughts are that sequencing errors will attenuate the association signal by misclassifying carriers of causal variants as non-carriers or *vice versa*, but is unlikely to create false-positive associations. Adjusting for confounding variables is another area of interest. One option to deal with confounding is a partial Mantel test of association between two distance matrices given a third [33]. This extension would allow testing for association between partition distances and phenotype distances given a third distance matrix based on the confounding variables.

In fine-mapping, fine-scale population structure is a confounding variable of particular concern because rare variants tend to cluster geographically due to their recent origin [34]. Adjusting for fine-scale population structure when fine-mapping rare variants is challenging, though recently proposed permutation approaches offer a potential way forward [35]. Our investigation of sequence-based linkage methods has focused on fine-mapping in a candidate region, but these methods also have the potential to scale up to genome-wide analysis as long as computational resources are available. For example, on a 2.1Ghz Intel processor, and with the sequence partition distances in hand, calculation of the Mantel association profile for the example dataset took 2.18 seconds, and calculation of 1000 further scans for the permutation distribution took about 18 minutes. Scaling from a 2MB region to the entire genome is expected to take about 54.5 minutes for a single scan, and about 27,000 minutes for the permutation replicates. However, permutations are easily parallelized across nodes of a compute cluster. Our study was focused on fine mapping, and so used the `perfectphyloR` R package for partition reconstruction. As `perfectphyloR`'s reconstruction does not scale to genome-wide data, we recommend alternative genome-wide reconstructors such as those implemented in `tsinfer` [36] or `Relate` [37]. For high-penetrance diseases influenced by multiple low-frequency variants, we expect sequence-based linkage analysis to have similar power to SKAT-O and better localization at the genome-wide scale, as in the fine-mapping results from the current investigation.

Chapter 3

Goodness-of-fit tests based on empirical processes

3.1 Introduction

A goodness-of-fit test is a statistical method which uses hypothesis testing to evaluate how well a theoretical distribution fits observed data. The process of testing whether sample data match a specific distribution with a cumulative distribution function (CDF) F begins by specifying a null hypothesis.

A simple null hypothesis arises when there are no unknown parameters, and a single CDF, F, fully describes the distribution of the data. However, in the much more common situations where some or all parameters of the distribution are unknown, a composite hypothesis is tested. Sometimes the observations in the sample are assumed to follow the same distribution but in more complex situations, the distribution of each observation in the sample may depend on some covariates. Important examples are linear and generalized linear models.

In its simplest form, a simple null hypothesis is being tested where F fully describes the distribution. To be more specific, let $Y_1, Y_2, \ldots, Y_n$ be a random sample of a continuous random variable (such as Y) from a population with the cumulative distribution function G. The basic problem of goodness-of-fit is to test the simple null hypothesis H_0 : For all y $G(y) = F(y)$, where F is a known continuous distribution, against the omnibus alternative that G is in the set of all CDFs which are not identically equal to F. The alternative hypothesis in goodness-of-fit test thus specifies no information about the distribution of the data and only indicates that the null hypothesis is false.

We review some of the suggestions in the literature for testing the simple null hypothesis when F has no unknown parameters and the distribution of sample data does not depend on covariates. The well-known Pearson's Chi-squared test was developed for the classic problem of testing goodness-of-fit [38]. To test the hypothesis, we partition the entire range of data into k different cells. The idea is to compare the number of observed values in any

cell range (denoted o_i) to the expected number of observations in the cell obtained from the distribution under the null hypothesis (denoted e_i). The sum of the squared differences between expected values and observed values divided by the expected values asymptotically follows a Chi-squared distribution; the degree of freedom, $k-1$ depends on the number of cells. Therefore, it is easy to compute the P-value for the test.

The Pearson Chi-squared test for the goodness-of-fit is an ideal case for discrete data but it lacks statistical power for the general alternative given above when the data is continuous. To overcome this, goodness-of-fit tests based on the empirical distribution function (EDF) were developed later [39]. The EDF is a step function calculated from sample data and is an estimate for the population distribution function [40]. Goodness-of-fit tests based on EDF give a non-parametric approach that does not make any assumption about the distribution of the data itself. Rather, the idea of these tests is to measure the discrepancy between the EDF obtained from the sample and the cumulative distribution function $F(y)$ specified by the null hypothesis.

A variety of methods have been proposed for goodness-of-fit tests based on the EDF. See for example Cramér[41], Smirnov [42], Kolmogorov [43], Anderson and Darling [44], and Lockhart and Stephens [45]. The asymptotic distribution of these statistics depends on whether or not any parameter estimation is involved, that is, on whether or not the null hypothesis is composite.

In this chapter of the thesis, we focus on goodness-of-fit tests based on the EDF. We mainly study the Cramér-Von-Mises (CvM) and Anderson-Darling (AD) statistics. We review the large sample theory for these EDF tests and show that each new model to be tested requires a substantial theoretical effort to derive the large sample distribution. That effort involves computing the covariance function of a suitable approximately Gaussian process and then solving a Fredholm integral equation to find the spectrum (eigenvalues) of the covariance function of that Gaussian process. We then show how to estimate, rather than compute analytically, the relevant covariance function and then estimate the required spectrum. Finally we review computation of approximate P-values of our statistic based on the computed or estimated spectrum.

The chapter is organized as follows. We begin in Section 3.2 by reviewing tests based on the empirical distribution function. We give the definitions and some properties of the well-known Cramér-von-Mises and Anderson-Darling statistics beginning with the simple null hypothesis. We present in Section 3.3 the well-known large sample theory that is required to study the asymptotic behavior of these statistics. When no parameters are estimated our statistics are the squared integral of a stochastic process which is a scaled average of independent and identically distributed terms. Well known large sample theory then shows that this stochastic process converges weakly, under the null hypothesis, to a Gaussian process with mean 0 and a covariance function which depends on whether we are discussing the Cramér-von-Mises statistic or the Anderson-Darling statistic. The properties of this

covariance function play a central role in calculating the P-value of the test; the limiting distribution of the test statistic is that of a linear combination of chi-squared variables. The weights are eigenvalues of the covariance; they solve a certain integral equation which we present.

We continue in Section 3.4 by studying the large sample theory when some or all parameters of the density model are not known. This case is important as it is more likely to happen in real-world applications of the goodness-of-fit test. In Section 3.5, we introduce a new method to estimate the covariance function of a stochastic process based on the sample which is helpful in applying goodness-of-fit tests based on EDF for any general likelihood model. To present our idea and emphasize the theory, we examine in Sections 3.6 and 3.7 the details of an i.i.d. sample from a Normal distribution and an i.i.d. sample from a Gamma distribution. The Gamma distribution has multiple standard parametrizations; motivated by this observation we show in Section 3.8 that our tests are parametrization invariant.

In Section 3.9 and 3.10 we broaden our view by considering linear models with a normal assumption for the residuals and then generalized linear models with an assumption of gamma distributed observations. In the generalized linear model case we focus on two popular links: log and inverse. Concluding remarks are in 3.11. To facilitate the application of our method and provide it to a wider audience, we have developed an R package to implement these methods in R statistical software; the package and its scope are described in Chapter 5.

3.2 Tests based on the empirical distribution function

In this section, we review the details of the goodness-of-fit tests based on the empirical distribution function (EDF). We describe two well-known statistics of this type which we will be using in this thesis. The simplest and classical problem of goodness-of-fit begins with an i.i.d. sample denoted by $Y_1, Y_2, \ldots, Y_n$ from a continuous random variable Y drawn from a population with some unknown cumulative distribution function (CDF) such as G, i.e $G(y) = Pr(Y \leq y)$. We would like to test the following hypothesis:

$$H_0 : \text{For all } y \; G(y) = F(y; \theta)$$
$$H_1 : \text{There is a } y \text{ such that } G(y) \neq F(y; \theta)$$

where F is the CDF of a fully known distribution. This means θ is a known value or vector. For example, F might be the CDF of a Normal distribution with mean zero and standard deviation of one. In this case θ is the vector $\theta = (\mu, \sigma) = (0, 1)$. Thus the precise form of the hypothesized distribution is known.

To test the above null hypothesis, one can calculate the empirical distribution function based on the random sample from the population. The empirical distribution function for

the sample is defined as:

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i \leq y\right). \tag{3.1}$$

Note that I is the indicator function with a value of one when $Y_i \leq y$ and zero otherwise. For all possible values of y, the quantity $F_n(y)$ simply calculates the proportion of observations in the sample that are less than or equal to y. If the null hypothesis is correct, we expect the proportions obtained from the sample to match very well to theoretical values obtained from $F(y;\theta)$. For example Pearson's chi-squared test assesses this expectation at certain points. The entire range of possible values is partitioned into different disjoint intervals. The points where intervals are defined are denoted by $C_0 = -\infty$, C_1, C_2, ...,C_{k-1}, $C_k = \infty$. The number of observations falling within the interval $[C_{i-1}, C_i]$ is $o_i = n\left(F_n(C_i) - F_n(C_{i-1})\right)$ where $F_n(C_i)$ and $F_n(C_{i-1})$ are the empirical cumulative distribution functions evaluated at C_i and C_{i-1}, respectively. On the other hand, the expected number of observations in each interval is $e_i = n\left(F(C_i) - F(C_{i-1})\right)$, where $F(C_i)$ and $F(C_{i-1})$ are the cumulative distribution functions of the theoretical distribution under consideration evaluated at C_i and C_{i-1}, respectively. Pearson's chi-squared statistic measures this discrepancy using the statistic $\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$ which asymptotically follows a Chi-square distribution with $k-1$ degrees of freedom. Note that the statistic essentially measures the discrepancy between the EDF and the hypothesized CDF at certain discrete data points.

More generally the idea of goodness-of-fit tests based on the EDF is to measure the discrepancy between the empirical distribution function obtained from the sample, $F_n(y)$, and the cumulative distribution function under the null hypothesis, $F(y;\theta)$, at all possible values of y. Goodness-of-fit tests based on the EDF summarize this distance over all possible values of y with a statistic. Generally speaking the distance between $F_n(y)$ and $F(y,\theta)$ can be calculated by two classes of statistics.

Supremum EDF statistics: The first class is the supremum norm of the weighted distance between the $F_n(y)$ and $F(y;\theta)$; the Kolmogorov-Smirnov (KS) statistic is a well-known example. It measures the discrepancy between $F_n(y)$ and $F(y;\theta)$ by calculating the absolute value of distance of the empirical distribution function from the cumulative distribution function for all possible values of y, weighted by square root of the sample size. The supremum of these differences is the KS statistic which is defined as follows:

$$K_n = \sup_{-\infty<y<\infty} \sqrt{n}\,|\,F_n(y) - F(y;\theta)\,|.$$

Quadratic EDF statistics: The second class of EDF-based statistics measures the discrepancy between the empirical distribution function obtained from the sample and theoretical values under the null hypothesis with a quadratic form of distance. This idea leads

to the family of Cramér-von-Mises statistics defined as follows:

$$Q = n \int_{-\infty}^{\infty} (\, F_n(y) - F(y;\theta)\,)^2 \; \zeta(y)\, dF(y;\theta). \tag{3.2}$$

Here $\zeta(y)$ controls the power of the test by assigning different weights to different parts of the distribution. If $\zeta(y) = 1$, the statistic is known as the Cramér-von-Mises statistic and is defined by:

$$W^2 = n \int_{-\infty}^{\infty} (\, F_n(y) - F(y;\theta)\,)^2 \; dF(y;\theta). \tag{3.3}$$

This is a distribution free statistic (that is, the distribution of W^2 does not depend on F when the null hypothesis holds) that compares how well the empirical distribution function of a sample matches to the theoretical distribution under the null hypothesis. Anderson and Darling [46] defined another measure of discrepancy by setting $\zeta(y) = [F(y;\theta)\,(1 - F(y;\theta)]^{-1}$ to define the statistic as follows:

$$A^2 = n \int_{-\infty}^{\infty} \frac{(\, F_n(y) - F(y;\theta)\,)^2}{F(y;\theta)(1 - F(y;\theta))} dF(y;\theta). \tag{3.4}$$

The Cramér-von Mises and Anderson-Darling statistics have easier alternative forms for computing purposes. The idea is to do the integral in 3.3 or 3.4 by integrating analytically between sorted probability integral transform values obtained from the sample. Suppose that $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ is the ordered sample from smallest to largest value. The Cramér-von Mises statistic can then be computed as [40]:

$$W^2 = \sum_{i=1}^{n} \left[U_{(i)} - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n} \tag{3.5}$$

where $U_{(i)} = F(Y_{(i)};\theta)$. Similarly, the Anderson-Darling statistic can be computed as [40]:

$$A^2 = \frac{-1}{n} \sum_{i=1}^{n} (2i-1) \left[\ln\left(U_{(i)}\right) + \ln\left(1 - U_{(n+1-i)}\right) \right] - n. \tag{3.6}$$

Both Cramér-von Mises and Anderson-Darling tests are well known in the literature and quite practical. Anderson-Darling statistic assigns more weight to the tails of the distribution than does the Cramér-Von Mises statistic (since $\zeta(y) = 1$ in CvM). This makes Anderson-Darling more sensitive to deviations from the theoretical distribution in the tails. Therefore using it against heavy tail alternative distributions often results in better power.

If the null hypothesis is not correct, we expect a large number for the discrepancy between the empirical distribution function obtained from sample, $F_n(y)$, and the cumulative distribution function under the null hypothesis, i.e $F(y;\theta)$. For example in the case of Cramér-von Mises statistic, this is reflected in the value of integrand, i.e $(F_n(y) - F(y;\theta))^2$.

As a result it is clear that the null hypothesis would be rejected when W^2 exceeds z for some value of z since this indicates a significant discrepancy between $F_n(y)$ and $F(y;\theta)$ from our assumption in the null hypothesis. This is true for both supremum and quadratic type statistics. In order to be able to work with any of these statistics, we need to obtain the distribution and the corresponding significance levels of these statistics. The asymptotic distribution of K_n is studied in the works of [43]. Massey [47] used combinatorial methods to obtain the exact distribution of K_n for sample sizes of $n \leq 35$. The asymptotic distribution of the Cramér-von Mises statistic is studied in the works of Smirnov (1936). The focus of this chapter is primarily on Cramér-von Mises statistics. In the next section, we review the required large sample theory to find the asymptotic distribution of quadratic statistics based on the EDF.

3.3 Large sample theory for the EDF

In this section, we provide the large sample theory required to perform the goodness-of-fit tests for the simple null hypothesis based on the empirical distribution function. As mentioned before, we primarily focus on the Cramér-von Mises statistic in this thesis but the idea can be generalized in a similar way to any quadratic EDF statistic for goodness-of-fit test. Again let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a continuous random variable Y with $G(y)$ being the unknown cumulative distribution function. The simple hypothesis test:

$$H_0 : \text{For all } y \; G(y) = F(y;\theta)$$
$$H_1 : \text{There is a } y \text{ such that } G(y) \neq F(y;\theta)$$

is of interest where F is the distribution in question and θ is a known vector of parameters that fully describes the distribution under the null hypothesis. Note that as Y is a continuous random variable, the probability integral transform theorem guarantees that the random variable $U = F(Y;\theta)$ has a Uniform distribution over the interval [0,1]. Therefore we can transform the Y_i sample into $U_i = F(Y_i;\theta)$ for $i = 1, 2, 3, \ldots, n$. If the null hypothesis is correct, the U_i's are a random sample drawn from a Uniform distribution over interval [0,1]. Therefore, the initial hypothesis testing problem can be reduced to determining whether these transformed samples follow a Uniform distribution over the interval [0, 1], with a cumulative distribution function of $F(u) = u$, for $0 \leq u \leq 1$. We can now write the Cramér-von Mises statistic in the following form based on the U_i sample:

$$W^2 = n \int_0^1 \left(F_n(u) - F(u) \right)^2 du = n \int_0^1 \left(F_n(u) - u \right)^2 du = \int_0^1 \left(\sqrt{n}(F_n(u) - u) \right)^2 du.$$

Note that we are now abusing notation by writing $F_n(u)$ for values of the empirical CDF of the U_i and $F_n(y)$ for values of the empirical CDF of the Y_i. That is we define for values

of $0 \leq u \leq 1$:

$$F_n(u) = \frac{1}{n}\sum_{i=1}^{n} I\left(F(Y_i;\theta) \leq u\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(U_i \leq u\right). \tag{3.7}$$

We finally define $W_n(u) = \sqrt{n}\left(F_n(u) - u\right)$ to write the CvM as follows:

$$W^2 = \int_0^1 \Big(W_n(u)\Big)^2 \, du. \tag{3.8}$$

The expected value of $F_n(u)$ is calculated as follows:

$$\begin{aligned} E\left[F_n(u)\right] &= E\left(\frac{1}{n}\sum_{i=1}^{n} I\left(F(Y_i;\theta) \leq u\right)\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(I\left(F(Y_i;\theta) \leq u\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} Pr\left(F(Y_i;\theta) \leq u\right) = \frac{1}{n}\sum_{i=1}^{n} Pr\left(F^{-1}(F(Y_i;\theta);\theta) \leq F^{-1}(u;\theta)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} Pr(Y_i \leq F^{-1}(u;\theta)) = \frac{1}{n}\sum_{i=1}^{n} F\left(F^{-1}(u;\theta);\theta\right) = \frac{1}{n}\sum_{i=1}^{n} u = u. \end{aligned}$$

The covariance between $F_n(s)$ and $F_n(t)$ for any values of $0 \leq s,t \leq 1$ is calculated as follows:

$$\begin{aligned} \mathrm{Cov}\left(F_n(s), F_n(t)\right) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} I\left(F(Y_i;\theta) \leq s\right), \frac{1}{n}\sum_{i=1}^{n} I\left(F(Y_i;\theta) \leq t\right)\right) \\ &= \frac{1}{n^2}\,\mathrm{Cov}\left(\sum_{i=1}^{n} I\left(U_i \leq s\right), \sum_{i=1}^{n} I\left(U_i \leq t\right)\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(I(U_i \leq s), I(U_j \leq t)\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{Cov}(I(U_i \leq s), I(U_i \leq t)) \;+\; \sum_{i \neq j} \mathrm{Cov}(I(U_i \leq s), I(U_j \leq t)\right). \end{aligned}$$

Note that each term in the second summation in the parentheses is 0; it is the covariance between two independent random variables – since Y_i and Y_j are independent we see that U_i and U_j are independent too. In the first term we must compute for $0 \leq s,t \leq 1$:

$$\mathrm{Cov}(I(U_i \leq s), I(U_i \leq t) = E(I(U_i \leq s)I(U_i \leq t)) - st.$$

Since $I(U_i \leq s)I(U_i \leq t) = I(U_i \leq \min(s,t))$ we find

$$\operatorname{Cov}\left(F_n(s), F_n(t)\right) = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Cov}(I(U_i \leq s), I(U_i \leq t) \tag{3.9}$$

$$= \frac{1}{n} \left(\min(s,t) - st\right). \tag{3.10}$$

We now describe the limiting distribution of $W^2 = \int_0^1 (W_n(u))^2\, du$. For values of $0 \leq u \leq 1$, $W_n(u) = \sqrt{n}\left(F_n(u) - u\right)$ is an empirical process. The limiting behaviour of this process is of interest. For fixed values of $u_1, u_2, \ldots, u_k$ the joint distribution of $W_n(u_1), W_n(u_2), \ldots, W_n(u_k)$ converges to a Gaussian process with mean zero and a certain covariance function, as $n \to \infty$ [48]. We here highlight the properties of the Gaussian process in the case of Cramér-von Mises statistic. The mean and covariance function of the sequence of stochastic process is calculated as follows:

$$E\left[W_n(u)\right] = E\left[\sqrt{n}(F_n(u) - u)\right] = \sqrt{n}\left[E(F_n(u)) - u\right] = \sqrt{n}(u-u) = 0$$

For values of $0 \leq s, t \leq 1$ the covariance function of the stochastic process is:

$$\begin{aligned} \operatorname{Cov}(W_n(s), W_n(t)) &= n(\operatorname{Cov}(F_n(s), F_n(t))) \\ &= \min(s,t) - st \end{aligned}$$

We use the notation:

$$\rho(s,t) = \min(s,t) - st \tag{3.11}$$

to refer to the covariance function of the stochastic process $W_n(u)$. First note that we can use the Karhunen-Loève theorem to expand the covariance function $\rho(s,t)$ in terms of its eigenvalues (spectrum) and its orthonormal eigenfunctions as follows [48]:

$$\rho(s,t) = \sum_{i=1}^{\infty} \lambda_j f_j(s) f_j(t)$$

On the other hand, since $W_n(u)$ is a random process in L_2 such that $\int_0^1 W_n^2(u)\, du < \infty$, it can be shown that the expansion of $W_n(u)$ is [48]:

$$W_n(u) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} Z_{n,j} f_j(u)$$

where $f_j(t)$ are orthonormal eigenfunctions, λ_j are the eigenvalues of covariance function $\rho(s,t)$ and:

$$Z_{n,j} = \int_0^1 W_n(u) f_j(u)\, du.$$

Note that:

$$\Big(W_n(u)\Big)^2 = \Big(\sum_{j=1}^{\infty}\sqrt{\lambda_j}Z_{n,j}f_j(u)\Big)^2 = \sum_{j=1}^{\infty}\lambda_j Z_{n,j}^2 f_j^2(u) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sqrt{\lambda_j}Z_{n,j}f_j(u)\,\sqrt{\lambda_k}Z_{n,k}f_k(u)$$

Now we can rewrite Cramér-von-Mises statistics defined in formula 3.8 as follows:

$$\begin{aligned}
W^2 &= \int_0^1 \Big(W_n(u)\Big)^2\,du = \int_0^1 \Big(\sum_{j=1}^{\infty}\lambda_j Z_{n,j}^2 f_j^2(u) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sqrt{\lambda_j}Z_{n,j}f_j(u)\,\sqrt{\lambda_k}Z_{n,k}f_k(u)\Big)\,du \\
&= \int_0^1 \sum_{j=1}^{\infty}\lambda_j Z_{n,j}^2 f_j^2(u)\,du \;+\; \int_0^1 \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sqrt{\lambda_j}Z_{n,j}f_j(u)\sqrt{\lambda_k}Z_{n,k}f_k(u)\,du \\
&= \sum_{j=1}^{\infty}\lambda_j Z_j^2\int_0^1 f_j^2(u)\,du \;+\; \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sqrt{\lambda_j}Z_{n,j}\sqrt{\lambda_k}Z_{n,k}\int_0^1 f_j(u)f_k(u)\,du \\
&= \sum_{j=1}^{\infty}\lambda_j Z_{n,j}^2.
\end{aligned}$$

The last line holds since $f_j(u)$ are orthonormal functions (i.e. $\int_0^1 f_k^2(t)dt = 1$ and for values of $i \neq j$ we have $\int_0^1 f_j(t)f_k(t)dt = 0$). Replacing $W_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(U_i \leq u) - u\right)$ in the formula for $Z_{n,j}$ we can write:

$$Z_{n,j} = \int_0^1 \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(U_i \leq u) - u\right) f_j(u)\,du.$$

The $Z_{n,j}$ have mean zero, standard deviation one, and for values of $i \neq j$, $\mathrm{Cov}(Z_{n,j}, Z_{n,k}) = 0$. It can be shown that $Z_{n,j}$ converges in distribution to an i.i.d. sequence of variables Z_j with standard normal distributions. Therefore the Cramér-von-Mises defined in formula 3.8 converges in distribution to a linear combination of Chi-squared random variables with a quadratic form such as

$$Q = \sum_{j=1}^{\infty}\lambda_j Z_j^2$$

where the Z_j's are i.i.d. random variables from a standard normal distribution and the λ_j's are eigenvalues of the covariance function of the stochastic process defined in formula 3.11. For a rigorous proof of the convergence see page 210 on Shorack and Wellner [48]. As a result, there are normalized eigenfunctions denoted $f_k(t)$ that satisfy the following integral equation [49]:

$$\int_0^1 \rho(s,t)f_k(t)dt = \lambda_k f_k(s) \qquad k = 1, 2, 3, \ldots \tag{3.12}$$

Note that the value of k varies from 1 to infinity as there are infinitely many eigenvalues. For the basic Cramér-von Mises statistic defined in formula 3.8, solving the integral equation gives the following eigenvalues and eigenfunctions [49]:

$$\lambda_k = \frac{1}{\pi^2 k^2} \qquad f_k(t) = \sqrt{2}\ \sin(\sqrt{\lambda_k}t) \quad k = 1, 2, 3, \ldots . \tag{3.13}$$

More general statistics such as the Anderson-Darling statistic incorporate a weight function ζ in the integral defining the test statistic. In this case the covariance function ρ is replaced by ρ_ζ defined by

$$\rho_\zeta(s,t) = \sqrt{\zeta(s)\zeta(t)}\rho(s,t).$$

A similar strategy can be incorporated to calculate eigenvalues and eigenfunctions in the case of Anderson-Darling and prove the convergence.

As we will see later in this chapter, a more complex form of the covariance function, as defined in formula 3.11, emerges in examples where the model contains unknown parameters. In such cases, an analytical solution for the integral equation in Formula 3.12 rarely exists. A common approach to solve for λ numerically is to discretize the integral over the interval [0,1]. For any given covariance function $\rho(s,t)$, the eigenvalues can be approximated by solving the following system of equations:

$$\sum_{j=1}^{m} w_j \rho(s_i, s_j) f(s_j) = \lambda_i f(s_i) \qquad i = 1, 2, 3, \ldots, m \tag{3.14}$$

where m is the number of knots s_j being used to discretize the integral over [0,1] and w_j are quadrature weights [50]. Specifically, the eigenvalues can be approximated by computing the non-zero eigenvalues of an m by m matrix M, where the elements of this matrix are:

$$M_{i,j} = w_j \rho(s_i, s_j)$$

The elements are calculated by evaluating the covariance function at values of s_i and s_j where $0 \leq s_i, s_j \leq 1$. In this thesis, these values are either selected to be equally spaced over the [0,1] interval or obtained as probability integral transformed values from the sample. We have also tried two sets of quadrature weights: either $w_j = 1/m$ (uniform weight over m different knots s_j) or $w_j = (U_{(j+1)} - U_{(j-1)})/2$ where $U_{(j)} = F(Y_{(j)}; \theta)$ is the j-th ordered probability integral transformed value of sample.

Returning to the covariance function in formula 3.11 and its eigenvalues in formula 3.13 for CvM, we use only a finite number, m, of these eigenvalues to numerically compute the P-value by computing the tail probability of the distribution of a linear combination of chi-squared variables,

$$Q_m = \sum_{i=1}^{m} \lambda_i Z_i^2.$$

We do this computation using either the method of Imhof [51] (numerical inversion of the characteristic function) or the method of Farebrother [52] (a more complex infinite expansion). In this thesis and in the package we developed, we worked with the first $m = 100$ eigenvalues in the case considered in this section: the simple null hypothesis tested using the Cramér-von Mises statistic. The P-value is $Pr(Q_m > x)$ where x is the Cramér-von Mises statistic calculated from the sample based on formula 3.5. The package *CompQuadForm* in R statistical software can be used to compute the P-value based on either Imhof's or Farebrother's method.

In summary, the following are the steps to calculate the P-value to test the null hypothesis when there is no unknown parameter in the model:

Step 1. For a sample of size n, obtain $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ where $Y_{(i)}$ is the i-th sorted value in the sample.

Step 2. Compute the probability integral transformed values of each $Y_{(i)}$ by applying $U_{(i)} = F(Y_{(i)}; \theta)$

Step 3. Compute the Cramer-von Mises statistic, W^2, using computation formula in 3.5.

Step 4. Compute a finite set of m eigenvalues as defined in formula 3.13.

Step 5. Numerically compute the P-value by either Imhof's or Farebrother's method.

In this section, we reviewed the goodness-of-fit test based on EDF for a simple hypothesis testing setting where θ is known, and $F(Y; \theta)$ fully describes the distribution under the null hypothesis. As a result there is no parameter estimation involved. In the next section, we will review the necessary theory for a more challenging and practical case. Specifically, we will explore cases where some or all elements of the vector parameter θ are unknown and need to be estimated from the sample. We will review the large sample theory in this case and how it alters the covariance function.

3.4 Empirical distribution function test for a composite hypothesis

We now consider a more challenging situation where $F(y; \theta)$ is not fully specified under the null hypothesis since some or all elements of θ, vector of parameter, are unknown. To be more specific, assume $Y_1, Y_2, \ldots, Y_n$ is an independent and identically distributed random sample from a continuous distribution. We denote the unknown cumulative distribution function with $G(y)$. We would like to test the following null hypothesis vs the alternative:

$$H_0 : \text{There is a } \theta \in \Theta \text{ such that } G(y) = F(y, \theta) \text{ for all } y,$$
$$H_1 : \text{For every } \theta \in \Theta \text{ there is a } y \text{ such that } G(y) \neq F(y, \theta).$$

In this hypothesis, $F(y;\theta)$ is the distribution of interest. For example, it can be a Normal distribution with unknown parameters of μ and σ^2. In this case, the vector of parameters is $\theta = (\mu, \sigma^2) \in \Theta$ where $\Theta = \{(\mu, \sigma^2); \mu \in R, \sigma > 0\}$, or it could be a Gamma distribution with unknown shape and scale parameters. In this case $\theta = (\alpha, \lambda) \in \Theta$ where $\Theta = \{(\alpha, \lambda), \alpha, \lambda > 0\}$. The large sample theory for composite hypotheses is similar to the large sample theory that we discussed in section 3.3 but with some important changes. Now θ is not known and needs to be estimated to compute the estimated distribution function which will be compared to the empirical distribution function. As a result of parameter estimation, the covariance function of the corresponding stochastic process is changed. This case has attracted more attention in the literature, particularly in the work of Michael Stephens [49], as it is more relevant to real-life applications. Here, we review the theory and summarize the steps for conducting the hypothesis tests.

The first step is to estimate the unknown parameter from the sample using a reasonable estimator. Maximum Likelihood Estimators (MLEs) are known for their efficiency and well-behaved asymptotic properties. MLEs can be easily computed for some known distributions from the sample data. In more complex likelihood models, when possible, numerical methods, such as the Newton-Raphson algorithm, can be employed to compute the MLE for θ numerically. We use the conventional notation $\hat{\theta}$ to represent the MLE of θ. In this thesis, we only consider MLE estimation of θ. Once the unknown parameter is estimated, the second step involves transforming the original data, $Y_1, Y_2, \ldots, Y_n$, into $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n$ by applying the probability integral transform $\hat{U}_i = F(Y_i; \hat{\theta})$. The estimated empirical distribution function is then calculated for all possible values of $0 \leq u \leq 1$ by computing:

$$\hat{F}_n(u) = \frac{1}{n} \sum_{i=1}^{n} I\left(F(Y_i; \hat{\theta}) \leq u\right) = \frac{1}{n} \sum_{i=1}^{n} I\left(\hat{U}_i \leq u\right). \tag{3.15}$$

Note that this is an estimate for the empirical distribution function defined in formula 3.7. Recall that the Cramér-von Mises statistic measures the discrepancy between the empirical distribution function and the cumulative distribution function under the null hypothesis. The same principle applies in the context of parameter estimation but the theoretical distribution function is replaced with the distribution function based on the parameter estimates. The Cramér-von Mises statistic is then used to quantify the goodness-of-fit between the EDF and the CDF estimated under the null hypothesis. It is worth noting that for relatively large sample sizes, we can expect the MLE of θ to be sufficiently close to the true value of θ. As a result, under the null hypothesis we can expect the estimated values $\hat{U}_i = F(Y_i; \hat{\theta})$ to be very close to the true values $U_i = F(Y_i; \theta)$, which means that we can reasonably expect $\hat{F}_n(u)$ to be close to $F_n(u)$. Therefore, we can reduce the problem to checking whether the transformed values of $\hat{U}_i$ follow a uniform distribution over the interval [0,1]. Thus for testing our composite null hypothesis we define the Cramér-von Mises statistic in the following

form:

$$W^2 = n\int_0^1 \left(\hat{F}_n(u) - u\right)^2 du$$
$$= \int_0^1 \left(\sqrt{n}(\hat{F}_n(u) - u)\right)^2 du = \int_0^1 \left(\hat{W}_n(u)\right)^2 du,$$

where now $\hat{W}_n(u) = \sqrt{n}\left(\hat{F}_n(u) - u\right)$ which can be thought of as an estimate for $W_n(u) = \sqrt{n}\,(F_n(u) - u)$ (defined in section 3.3). The limiting distribution of W^2 is still a quadratic form such as $\sum_k \lambda_k Z_k^2$ but the estimation of θ changes the covariance function of $\hat{W}_n(u)$ process compared to the covariance function of $W_n(u)$ process, [49]. We now review the steps needed to calculate the limiting covariance function of the $\hat{W}_n(u)$ process when the parameters of the model need to be estimated. (Our presentation does not give rigorous proofs that the various remainder terms are uniformly negligible; see [53], chapter 5, section 5 on page 228 for more details.) The idea is to expand $\hat{W}_n(u)$ in terms of $W_n(u)$ since we know the covariance function of $W_n(u)$. Using formula 3.15, we start by writing:

$$\begin{aligned}\hat{W}_n(u) &= \sqrt{n}\left(\hat{F}_n(u) - u\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} I\left(F(Y_i,\hat{\theta}) \le u\right) - u\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} I\left(F(Y_i,\hat{\theta}) \le u\right) - \frac{nu}{n}\right) \\ &= \frac{\sqrt{n}}{n}\left(\sum_{i=1}^{n} I\left(F(Y_i,\hat{\theta}) \le u\right) - nu\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I\left(F(Y_i,\hat{\theta}) \le u\right) - u\right).\end{aligned} \tag{3.16}$$

We define the inverse of the cumulative distribution function of $F(Y;\hat{\theta})$ using the notation $Q(Y;\hat{\theta}) = F^{-1}(Y;\hat{\theta})$. Applying the inverse function to both sides of the inequality in the indicator function in formula 3.16, resulting in the following expression:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(Y_i \le Q(u;\hat{\theta})) - u\right).$$

Since $F(Y;\theta)$ is a non-decreasing function of Y, we can write:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(F(Y_i;\theta) \le F(Q(u;\hat{\theta});\theta)) - u\right).$$

It is easy to verify that adding and subtracting $F(Q(u;\hat{\theta});\theta)$ results in:

$$\begin{aligned}\hat{W}_n(u) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I(F(Y_i;\theta)\le F(Q(u;\hat{\theta});\theta)) - F(Q(u;\hat{\theta});\theta) + F(Q(u;\hat{\theta});\theta) - u\Big)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I(F(Y_i;\theta)\le F(Q(u;\hat{\theta});\theta)) - F(Q(u;\hat{\theta});\theta)\Big)\\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(F(Q(u;\hat{\theta});\theta) - u\Big)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I(U_i\le F(Q(u;\hat{\theta});\theta)) - F(Q(u;\hat{\theta});\theta)\Big) + \sqrt{n}\Big(F(Q(u;\hat{\theta});\theta) - u\Big).\end{aligned} \tag{3.17}$$

We approximate the first of the two terms on last line by $W_n(u)$ to write the expression in formula 3.17 as follows [53]:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{I(U_i\le u) - u\} + \sqrt{n}\{F(Q(u;\hat{\theta});\theta) - u\} + R_n(u)$$

where $R_n(u)$ is the remainder term given by

$$R_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I(U_i\le F(Q(u;\hat{\theta});\theta)) - F(Q(u;\hat{\theta});\theta)\Big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{I(U_i\le u) - u\}.$$

In [53] it is shown under reasonable conditions on the null hypothesis model that

$$\sup_{0\le u\le 1}|R_n(u)| \to 0$$

in probability; we therefore drop the remainder term from our formulas in the discussion which follows. Recall that $W_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{I(U_i\le u) - u\}$; therefore we can write:

$$\hat{W}_n(u) = W_n(u) + \sqrt{n}\{F(Q(u;\hat{\theta});\theta) - u\} \tag{3.18}$$

The next step is to approximate $F(Q(u;\hat{\theta});\theta)$ by deriving the Taylor expansion of $F(Q(u;\hat{\theta}))$ around the value of θ. Before that, we will introduce some notation and review some derivatives. Since Q is the inverse of cumulative distribution function, we have $F(Q(u;\theta);\theta) = u$. Taking the derivative of both sides with respect to θ and applying the chain rule gives:

$$\frac{\partial}{\partial\theta}F(Q(u;\theta);\theta) = \frac{\partial}{\partial\theta}u$$

$$\frac{\partial F(Q(u;\theta);\theta)}{\partial Q(u,\theta)} \times \frac{\partial Q(u,\theta)}{\partial\theta} + \frac{\partial F(Q(u;\theta);\theta)}{\partial\theta} = 0$$

$$f(Q(u;\theta);\theta) \times \frac{\partial Q(u;\theta)}{\partial \theta} + \frac{\partial F(Q(u;\theta);\theta)}{\partial \theta} = 0$$

which leads to:

$$\frac{\partial Q(u;\theta)}{\partial \theta} = \frac{-D_2 F(Q(u;\theta),\theta)}{f(Q(u;\theta);\theta)}$$

where $D_2 F(Q(u;\theta),\theta)$ is the partial derivative of $F(Q(u;\theta),\theta)$ with respect to the second parameter, i.e θ. Using this formula, we can evaluate $\frac{\partial Q(u;\theta)}{\partial \theta}$ at any specific value of θ. For instance, we can calculate it at $\theta = \hat{\theta}$ to write:

$$\frac{\partial Q(u;\hat{\theta})}{\partial \hat{\theta}} = \frac{-D_2 F(Q(u;\hat{\theta}),\hat{\theta})}{f(Q(u;\hat{\theta});\hat{\theta})} = \frac{-\frac{\partial}{\partial \hat{\theta}} F(Q(u;\hat{\theta}),\hat{\theta})}{f(Q(u;\hat{\theta});\hat{\theta})}. \tag{3.19}$$

Note that in this thesis, we assume that partial derivatives of $F(y;\theta)$ exists and $f(y;\hat{\theta})$ evaluated at $y = Q(u;\hat{\theta})$ is not zero ($0 \leq u \leq 1$). Our simulations are based on these assumptions. Next we write the Taylor expansion of $F(Q(u;\hat{\theta});\theta)$ around the true value of θ as follows:

$$\begin{aligned} F(Q(u;\hat{\theta});\theta) &= F(Q(u;\theta);\theta) + \frac{\partial}{\partial \hat{\theta}} F(Q(u;\hat{\theta});\theta)\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2] \\ &= u + \frac{\partial}{\partial \hat{\theta}} F(Q(u;\hat{\theta});\theta)\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2] \\ &= u + f(Q(u;\hat{\theta});\theta)\frac{\partial Q(u;\hat{\theta})}{\partial \hat{\theta}}\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2] \end{aligned} \tag{3.20}$$

and plug in the value of $\frac{\partial Q(u;\hat{\theta})}{\partial \hat{\theta}}$ from 3.19 to obtain:

$$F(Q(u;\hat{\theta});\theta) = u - f(Q(u;\hat{\theta});\theta)\frac{\frac{\partial}{\partial \hat{\theta}} F(Q(u;\hat{\theta}),\hat{\theta})}{f(Q(u;\hat{\theta});\hat{\theta})}\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2]. \tag{3.21}$$

In the limit, i.e as $n \to \infty$, we expect $\hat{\theta} \to \theta$. In formula 3.21, we evaluate the second term at $\hat{\theta} = \theta$ to continue:

$$\begin{aligned} F(Q(u;\hat{\theta});\theta) &= u - f(Q(u;\theta);\theta)\frac{\frac{\partial}{\partial \theta} F(Q(u;\theta),\theta)}{f(Q(u;\theta);\theta)}\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2] \\ &= u - \frac{\partial}{\partial \theta} F(Q(u;\theta),\theta)\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2] \\ &= u - \Psi^T(u)\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2] \end{aligned} \tag{3.22}$$

Note in the formula $\Psi(u)$ is a column vector. The elements of this vector are the partial derivatives of the cumulative distribution function with respect to the unknown parameter θ. If θ contains p unknown parameters then the $\Psi(u)$ vector has p elements of the form

$\psi_i(u) = \frac{\partial}{\partial \theta_i} F(y;\theta)$, for $i = 1, 2, 3, \ldots, p$, and all elements are evaluated at $y = Q(u;\theta)$, the inverse of cumulative distribution function. Returning to formula 3.18, we replace the value of $F(Q(u;\hat{\theta});\theta)$ from the formula in 3.22 to write:

$$\begin{aligned}
\hat{W}_n(u) &= W_n(u) + \sqrt{n}\{F(Q(u;\hat{\theta});\theta) - u\} \\
&= W_n(u) + \sqrt{n}\{u - \Psi^T(u)\,(\hat{\theta} - \theta) + O_p[(\hat{\theta} - \theta)^2] - u\} \\
&= W_n(u) - \sqrt{n}\{\Psi^T(u)\,(\hat{\theta} - \theta) + O_p[(\hat{\theta} - \theta)^2]\} \\
&= W_n(u) - \Psi^T(u)\,\sqrt{n}(\hat{\theta} - \theta) - \sqrt{n} O_p[(\hat{\theta} - \theta)^2].
\end{aligned} \tag{3.23}$$

Since $\hat{\theta} - \theta = O_p(\frac{1}{\sqrt{n}})$ we conclude that $(\hat{\theta} - \theta)^2 = O_p(\frac{1}{n})$ and the remainder term converges to zero as $n \to \infty$. Conditions under which the remainder term, which depends on u, converges uniformly to 0 at the rate $1/\sqrt{n}$ are in [48]. We can write:

$$\hat{W}_n(u) = W_n(u) - \Psi^T(u)\,\sqrt{n}(\hat{\theta} - \theta).$$

We now use the standard asymptotic properties of maximum likelihood estimator to replace $\sqrt{n}(\hat{\theta} - \theta)$ in the expansion. We will see the usual score function and the Fisher information matrix, but we also need an $n \times p$ matrix, here denoted by $S(\theta)$ (the score matrix) where the i-th row and the j-th column of $S(\theta)$ is:

$$S_{ij}(\theta) = \frac{\partial \log f(Y_i, \theta)}{\partial \theta_j}$$

Note that the partial derivative is with respect to the j-th element of θ vector. Each row of this matrix contains the partial derivative of the log-likelihood function with respect to θ. For instance, the i-th row of the matrix is given by $S_i(\theta) = \frac{\partial}{\partial \theta} \log f(Y_i;\theta)$. We use the notation of $S_i(\theta)$ to denote the i-th row in score matrix. Then we can write the usual score function as a row vector of length p given by:

$$\mathbf{1}^T S(\theta) = \sum_{i=1}^{n} S_i(\theta)$$

where $\mathbf{1}$ is a column vector of length n with all values one. The usual asymptotic properties of MLE results in:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}}\,\mathbf{1}^T S(\theta) I^{-1}(\theta) + O_P(\frac{1}{\sqrt{n}}).$$

In this formula, $I(\theta)$ represents the Fisher information matrix based on one observation and is a p by p matrix (corresponding to the number of unknown parameters in θ) and $S(\theta)$ is the score function, a matrix with n rows (sample size) and p columns (number of

parameters), as defined before. Replacing $\sqrt{n}(\hat{\theta}-\theta)$, we can write $\hat{W}_n(u)$ as follows:

$$\hat{W}_n(u) = W_n(u) - \frac{1}{\sqrt{n}}\mathbf{1}^T S(\theta) I^{-1}(\theta)\Psi(u) + O_P(\frac{1}{\sqrt{n}}).$$

The remainder term converges to zero as $n \to \infty$. We can finally write $\hat{W}_n(u)$ according to $S_i(\theta)$, the score of each observation as follows:

$$\begin{aligned}
\hat{W}_n(u) &= W_n(u) - \frac{1}{\sqrt{n}}\mathbf{1}^T S(\theta) I^{-1}(\theta)\Psi(u) + O_P(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{I\left(F(Y_i,\theta)\le u\right) - u\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\theta) I^{-1}(\theta)\Psi(u) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I\left(F(Y_i,\theta)\le u\right) - u - S_i(\theta) I^{-1}(\theta)\Psi(u)\Big).
\end{aligned}$$

It can be shown that $\hat{W}_n(u)$ converges weakly to a stochastic process with mean zero and a certain covariance [48]. Note that the mean is zero since $E(S_i(\theta)) = 0$. To calculate the covariance, we take the following steps. Start by defining $Z_i(u) = I\left(F(Y_i,\theta)\le u\right) - u - S_i(\theta) I^{-1}(\theta)\Psi(u)$ to rewrite $\hat{W}_n(u)$ as:

$$\hat{W}_n(u) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i(u).$$

The covariance function of $\hat{W}_n(u)$ for values of $0 \le s,t \le 1$ is now:

$$\begin{aligned}
\rho(s,t) &= \mathrm{Cov}\left(\,\hat{W}_n(s)\,,\,\hat{W}_n(t)\,\right) \\
&\approx \mathrm{Cov}\left(\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i(s)\,,\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i(t)\,\right) \\
&= \frac{1}{n}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(Z_i(s), Z_j(t))\Big) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathrm{Cov}(Z_i(s), Z_i(t)) + \sum_{i\ne j}^{n}\mathrm{Cov}(Z_i(s), Z_j(t))\right).
\end{aligned}$$

Note that if $i \ne j$, then $\mathrm{Cov}(Z_i(s), Z_j(t))$ is zero since Y_i and Y_j are independent. Therefore, we only need to calculate the covariance when $i = j$ as follows:

$$\begin{aligned}
\rho(s,t) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(Z_i(s), Z_i(t)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}\Big(I(U_i\le s) - s - S_i(\theta) I^{-1}(\theta)\Psi(s), I(U_i\le t) - t - S_i(\theta) I^{-1}(\theta)\Psi(t)\Big).
\end{aligned}$$

We expand the covariance and calculate each term separately. The calculation of

$$\mathrm{Cov}\left(I(U_i \le s) - s, I(U_i \le t) - t\right) = \min(s,t) - st$$

remains the same as in formula 3.11. To compute the remaining terms define y by $F(y,\theta) = s$. Since $E(S_i(\theta)) = 0$, it is easy to verify that:

$$\begin{aligned}
&\mathrm{Cov}(I(U_i \le s)\,,\, S_i(\theta)I^{-1}(\theta)\Psi(t)) = E\Bigg[I(U_i \le s)S_i(\theta)I^{-1}(\theta)\Psi(t)\Bigg] \\
&= E\Bigg[I(Y_i \le y)S_i(\theta)I^{-1}(\theta)\Psi(t)\Bigg] = \left(\int_{-\infty}^{y} f(y';\theta)\, S_i(\theta)\, dy'\right) I^{-1}(\theta)\Psi(t) \\
&= \left(\int_{-\infty}^{y} f(y';\theta)\, \frac{\partial \log f(y';\theta)}{\partial\theta}\, dy'\right) I^{-1}(\theta)\Psi(t) \\
&= \left(\int_{-\infty}^{y} f(y';\theta)\, \frac{\frac{\partial}{\partial\theta} f(y';\theta)}{f(y';\theta)}\, dy'\right) I^{-1}(\theta)\Psi(t) \\
&= \left(\int_{-\infty}^{y} \frac{\partial}{\partial\theta} f(y';\theta)\, dy'\right) I^{-1}(\theta)\Psi(t) = \frac{\partial}{\partial\theta}\left(\int_{-\infty}^{y} f(y';\theta)\, dy'\right) I^{-1}(\theta)\Psi(t) \\
&= \frac{\partial}{\partial\theta} F(y,\theta) I^{-1}(\theta)\Psi(t) \\
&= \Psi^T(s) I^{-1}(\theta)\Psi(t).
\end{aligned}$$

Switch the roles of s and t to get:

$$\mathrm{Cov}(S_i(\theta)I^{-1}(\theta)\Psi(s),\, I(U_i \le t)) = \Psi^T(t) I^{-1}(\theta)\Psi(s).$$

Finally, the last term in the covariance calculation is:

$$\begin{aligned}
&\mathrm{Cov}\Big(S_i(\theta)I^{-1}(\theta)\Psi(s)\,,\, S_i(\theta)I^{-1}(\theta)\Psi(t)\Big) \\
&= \Psi^T(s) I^{-1}(\theta)\, \mathrm{Cov}\Big(S_i(\theta)\,,\, S_i(\theta)\Big) I^{-1}(\theta)\Psi(t) = \Psi^T(s) I^{-1}(\theta)\, I(\theta) I^{-1}(\theta)\Psi(t) \\
&= \Psi^T(s) I^{-1}(\theta)\Psi(t).
\end{aligned}$$

Therefore, the approximate covariance function of the stochastic process $\hat{W}_n(u)$ is:

$$\rho(s,t) = \min(s,t) - st - \Psi^T(t) I^{-1}(\theta)\Psi(s) \tag{3.24}$$

for values of $0 \le s,t, \le 1$. As mentioned before, $\Psi(u)$ is a column vector with the same length as the number of parameters in θ. The elements of the vector are the partial derivatives of $F(y;\theta)$ with respect to θ which are evaluated at $y = Q(u;\theta)$, for values of $0 \le u \le 1$. Note that in formula 3.24, the covariance function depends on θ since both Ψ and I depend

on θ. We replace θ by the maximum likelihood estimator of θ to estimate both the Fisher information matrix and Ψ in the formula.

Now we summarize the steps to compute the P-value when θ is unknown and is estimated from the sample.

Step 1. Estimate the unknown parameter using the computing the maximum likelihood estimate of θ, i.e., $\hat{\theta}$.

Step 2. Sort the data to obtain $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ where $Y_{(i)}$ is the i-th sorted value in the sample.

Step 3. Compute the (estimated) probability integral transforms, $\hat{U}_{(i)} = F(Y_{(i)}, \hat{\theta})$.

Step 4. Compute the Cramér-von Mises statistic from these $\hat{U}_{(i)}$ as in formula 3.5:

$$W^2 = \sum_{i=1}^{n} \left[\hat{U}_{(i)} - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n}.$$

Step 5. Calculate the elements of the vector $\hat{\Psi}(y)$ and evaluate them at $y = Q(u; \hat{\theta})$ where Q is the inverse of the cumulative distribution function under the null hypothesis and $0 \leq u \leq 1$.

Step 6. Calculate the Fisher information matrix and replace the unknown parameters by $\theta = \hat{\theta}$.

Step 7. Using the estimates from Steps 5 and 6, estimate the covariance function defined in 3.24.

Step 8. Approximate the eigenvalues of the covariance function calculated in Step 7 by computing the eigenvalues of matrix M. Detailed suggestions for the choice of M are given in 3.14.

Step 9. Discard any zero eigenvalues (if any) and compute the P-value by using the Imhof or Farebrother method.

It is evident that knowledge of the covariance function is crucial for calculating the P-value. This necessitates the availability of the $\Psi(y)$ vector and $I(\theta)$. As we will see in the following sections, these components may not be readily available for all likelihood models. In the next section, we propose a method to estimate the covariance function from sample data without the need for analytical element calculations.

3.5 Estimation of the covariance function

Understanding the covariance function of the stochastic process $\hat{W}_n(u)$ is crucial for computing the P-value for the goodness-of-fit test. Some distributions, such as the Normal and Gamma, allow a relatively straightforward computation of the covariance function, but computing the covariance function can be a daunting or time consuming task for a general likelihood model. The problem occurs when computing the partial derivatives of CDF with

respect to parameters, i.e. the vector $\Psi(u)$, or the inverse of the Fisher information matrix. Consequently computing $\rho(s,t)$ may not always be a straightforward process if these quantities are too difficult or time consuming to compute analytically. This problem can limit our ability to compute the P-value for the goodness-of-fit test. To address this challenge, this section proposes an alternative method of estimating $\rho(s,t)$ using sample data rather than calculating it analytically from the model. The proposal needs only algorithms to find the MLE, $\hat{\theta}$, evaluate the score function components, $S_i(\hat{\theta})$, and compute the probability integral transforms $F(Y_i,\hat{\theta})$; it is not necessary to have compute the expectations defining the Fisher information matrix, nor to compute the derivative of $F(\cdot,\theta)$ with respect to θ.

The idea starts by finding an alternative expression for the elements of $\Psi(t)$ based on $S_i(\theta)$ and $I(F(Y_i;\theta)\leq t)$. Note that under conditions where we can change the order of integral and derivative, for any value of $0\leq t\leq 1$ we can write $\Psi(t)$ as follows:

$$\begin{aligned}
\Psi(t) &= \frac{\partial}{\partial\theta}F(y;\theta)\bigg|_{y=Q(t;\theta)} \\
&= \int_{-\infty}^{y}\frac{\partial f(y';\theta)}{\partial\theta}\,dy'|_{y=Q(t;\theta)} \\
&= \int_{-\infty}^{y}\frac{\frac{\partial f(y';\theta)}{\partial\theta}}{f(y';\theta)}\,f(y';\theta)\,dy'|_{y=Q(t;\theta)} = \int_{-\infty}^{y}\frac{\partial}{\partial\theta}log(f(y;\theta))\,f(y';\theta)\,dy'|_{y=Q(t;\theta)} \\
&= E\left[\frac{\partial}{\partial\theta}log(f(Y_i;\theta))\,I(F(Y_i;\theta)\leq t)\right] = E\left[S_i(\theta)I(F(Y_i;\theta)\leq t)\right] \\
&= \mathrm{Cov}\bigg(S_i(\theta),I(F(Y_i;\theta)\leq t)\bigg).
\end{aligned}$$

Note that the last line results since $E\left(S_i(\theta)\right)=0$. The calculation presented above brings two important notes. First, the random variable $S_i(\theta)I(F(Y_i;\theta)\leq t)$ has expected value $\Psi(t)$. Second, $\Psi(t)$ can also be computed using an alternative approach by calculating the covariance between $S_i(\theta)$ and $I(F(Y_i;\theta)\leq t)$.

Having these points in mind, note that for $i=1,2,3,\ldots,n$ the sequence of random variables $S_i(\theta)I(F(Y_i;\theta)\leq t)$ are mutually independent with the same expected value. Therefore by the weak law of large numbers the average of these random variables converges in probability to $\Psi(t)$. In other words:

$$\frac{1}{n}\sum_{i=1}^{n}S_i(\theta)I(F(Y_i;\theta)\leq t)\xrightarrow{P}\Psi(t).$$

Now suppose θ_0 is the true value of θ. Define:

$$H(\theta,\theta_0)=E_{\theta_0}\left[S_i(\theta)I(F(Y_i;\theta)\leq t)\right]$$

which is a continuous function of θ. The function can be evaluated at $\hat{\theta}$, the MLE of θ to get $H(\hat{\theta}, \theta_0)$. Since $\hat{\theta}$ is consistent, continuity implies that $H(\hat{\theta}, \theta_0)$ converges in probability to $H(\theta, \theta_0)$. Hence:

$$\frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\theta}) I(F(Y_i; \hat{\theta}) \leq t) \xrightarrow{P} \Psi(t).$$

Therefore, for a sample of size n, we estimate $\Psi(t)$ by the covariance between the score function and indicator function computed from the sample, i.e:

$$\hat{\Psi}(t) = \frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\theta}) I(F(Y_i; \hat{\theta}) \leq t).$$

Similarly, the Fisher information matrix can be estimated by the sample covariance of the score function. We use the notation of $\hat{I} = \hat{I}(\hat{\theta})$ for the estimate of Fisher information to define the estimate by:

$$\hat{I}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\theta})^T S_i(\hat{\theta}) = \frac{S(\hat{\theta})^T S(\hat{\theta})}{n}$$

where $S_i(\hat{\theta}) = \frac{\partial}{\partial \theta} \log f(Y_i; \hat{\theta})$ is the i-th row of $S(\hat{\theta})$ matrix.

Our proposal for testing the fit of a general model for independent observations is to estimate θ by maximum likelihood, compute probability integral transforms using this estimate, then estimate a covariance function as described above. Discretize the integral equation as described at 3.14 and compute the eigenvalues of the resulting matrix Q (this matrix is defined in the steps below). Use these eigenvalues to approximate an infinite linear combination of chi-squares by a finite linear combination and compute tail probabilities of the approximation to get a P-value.

In summary, the steps for a goodness-of-fit test in a composite hypothesis for a general likelihood model are as follows:

Step 1. Sort the sample $Y_1, Y_2, \ldots, Y_n$ to obtain $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$.

Step 2. Compute $\hat{\theta}$, the MLE of θ.

Step 3. Calculate the Cramér-von Mises statistic by:

$$W^2 = \sum_{i=1}^{n} \left[\hat{U}_{(i)} - \frac{2i-1}{2n}\right]^2 + \frac{1}{12n}$$

where $\hat{U}_{(i)} = F(Y_{(i)}; \hat{\theta})$.

Step 4. Calculate the score function, $S(\hat{\theta})$, which is an $n \times p$ matrix where n is the sample size and p is the number of parameters. Note that the ith row of S presents the contribution to the score for the ith observation and that $\hat{\theta}$ is plugged-in for θ whenever necessary.

Step 5. Estimate the Fisher information matrix by computing the $p \times p$ matrix $\hat{I}(\hat{\theta})$ which is the sample variance of $S(\hat{\theta})$. That is,

$$\hat{I}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\theta})^T S_i(\hat{\theta}) = \frac{S(\hat{\theta})^T S(\hat{\theta})}{n}$$

and compute the inverse of $\hat{I}(\hat{\theta})$.

Step 6. Estimate the function $\Psi(u)$ by the sample covariance of the score function and an indicator function:

$$\hat{\Psi}(u) = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{U}_i \leq u) S_i(\hat{\theta}).$$

In this thesis, we evaluated this estimate at m values of u within the interval [0,1] to obtain a matrix with dimensions $m \times p$. We tried using either the sorted estimated values of the probability integral transform obtained from the sample as the grid, denoted as $\hat{U}_{(1)}, \hat{U}_{(2)}, \ldots, \hat{U}_{(n)}$ or a grid of m equally spaced points, such as $u_i = \frac{i}{m}$, within the interval $[0-\epsilon, 1+\epsilon]$, where we chose ϵ to be 10^{-5}. This choice of ϵ was made to ensure that the first and last columns of the matrix Q (as defined below) do not consist solely of zeros and ones. In our developed R package, we decided to evaluate $\hat{\Psi}(u)$ at PIT values since the simulation results clearly showed the advantage of PIT over an equally spaced grid.

Step 7. Compute the $n \times m$ matrix $\Delta = S(\hat{\theta}) \left(\hat{I}(\hat{\theta})\right)^{-1} \hat{\Psi}^T(u)$ using values from step 5 and step 6. Then compute matrix Q with elements defined as:

$$q_{ij} = I(\hat{U}_i \leq u_j) - \delta_{ij}$$

where δ_{ij} is the i,j-th entry in the matrix Δ. Here u_j are the chosen grid of values over interval [0,1]. Note that Q is an $n \times m$ matrix, where i ranges from 1 to n (sample size) and j ranges from 1 to m (number of points in the grid).

Step 8. Compute the m by m matrix Λ whose i,jth entry is the sample covariance between the ith and jth columns of matrix Q; then multiply Λ by $\frac{n-1}{n-p-1}$. In the R code in our package, we use function *var* to compute this variance-covariance matrix; our choice of scaling is motivated by our simulation studies in Chapter 4.

Step 9. Calculate the eigenvalues of this covariance matrix, drop all those which are numerically 0, and obtain the P-value using either Imhof's or Farebrother's method. In our package we use *imhof* function from *CompQuadForm* package, unless an error is detected in which case *farebrother* is substituted.

Our experience shows that when computing the probabilities in the extreme tail of quadratic forms in normal variables, sometimes the P-value is inaccurate or outright incorrect. For example, we noticed that if the P-value is greater than 10^{-7}, the *Imhof* method has no numerical issue with computation. If the P-value is between 10^{-10} and 10^{-7}, *Imhof*

fails to generate a correct value but *Farebrother* computes the P-value with a good accuracy. For P-values less than 10^{-10}, both methods have difficulty and produce incorrect values. While P-values this small have little importance it is important in writing code to be clear about the accuracy of all presented results; in particular we must avoid reporting negative P-values. To overcome this, in our R package, we have developed procedures to numerically compute a lower bound (LB) and an upper bound (UB) for the P-value. Then, depending on the values of LB and UB, we compute the exact P-value using either *Imhof*'s or *Farebrother*'s method. In the case when both methods fail, we simply return LB and UB values as an interval for the P-value.

In the next two sections, we review the theory behind computing the covariance function in two examples. The first example is an i.i.d. sample from a Normal distribution. The second example reviews the theory for a Gamma distribution. In both examples, we derive the limiting covariance function both analytically and with the estimation method described in this section.

3.6 Example 1: Normal distribution

The simplest example to start with is the Normal distribution. In this section we derive the analytic formulas from Section 3.4 and show the details of the steps presented in Section 3.5 to estimate the covariance and eigenvalues. For this distribution the analytic formulas for $\Psi(u)$ are relatively easy but we also present the precise formulas for the method of Section 3.5 for comparison.

Suppose we have a random sample $Y_1, Y_2, \ldots, Y_n$ from a population. We would like to test if this sample follows a Normal distribution. In other words $H_0 : F \in \{G(\cdot, \theta); \theta \in \Theta\}$ where $G(\cdot, \theta)$ is the CDF of a Normal distribution, $\theta = (\mu, \sigma)$ and $\Theta = \{(\mu, \sigma);\ \mu \in R\ ,\ \sigma > 0\}$. Both parameters are unknown. For a sample of size n, the likelihood function is

$$L(\mu, \sigma) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-1}{2\sigma^2}(Y_i-\mu)^2} = (2\pi\sigma^2)^{\frac{-n}{2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2}$$

and the log-likelihood function is:

$$l(\mu, \sigma) = \left(\frac{-n}{2}\right) \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2.$$

We take the following steps to calculate the limiting covariance function of the stochastic process $\hat{W}_n(u)$. The first step is to obtain the Fisher information matrix by calculating the partial derivatives of the log-likelihood function with respect to parameters:

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = \left(\frac{-1}{2\sigma^2}\right) \sum_{i=1}^{n} (-2)(Y_i - \mu) = \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma^2}, \quad \frac{\partial^2 l(\mu, \sigma)}{\partial \mu^2} = \frac{-n}{\sigma^2}$$

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu)^2, \quad \frac{\partial^2 l(\mu,\sigma)}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(Y_i-\mu)^2$$

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma \partial \mu} = \frac{\partial l(\mu,\sigma)}{\partial \mu \partial \sigma} = \frac{-2}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu).$$

The Fisher information matrix is thus:

$$I_n(\theta) = -E\begin{pmatrix} \frac{\partial^2}{\partial \mu^2} l(\mu,\sigma) & \frac{\partial}{\partial \mu \partial \sigma} l(\mu,\sigma) \\ \frac{\partial}{\partial \sigma \partial \mu} l(\mu,\sigma) & \frac{\partial^2}{\partial \sigma^2} l(\mu,\sigma) \end{pmatrix} = -E\begin{pmatrix} \frac{-n}{\sigma^2} & \frac{-2}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu) \\ \frac{-2}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu) & \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(Y_i-\mu)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}$$

and the Fisher information matrix based on one observation is:

$$I(\theta) = \frac{1}{n} I_n(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

The inverse of the Fisher information matrix (based on one observation) is:

$$I^{-1}(\theta) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}.$$

The next step is to calculate the column vector $\Psi(y)$, partial derivatives of cumulative distribution function $F(y;\theta)$ with respect to θ evaluated at $y = F^{-1}(u;\theta)$. In the Normal example, $\Psi(u)$ is a column vector with two elements as follows:

$$\Psi(u) = \begin{pmatrix} \psi_1(u) \\ \psi_2(u) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \mu} F(y;\theta)\Big|_{y=F^{-1}(u;\theta)} \\ \frac{\partial}{\partial \sigma} F(y;\theta)\Big|_{y=F^{-1}(u;\theta)} \end{pmatrix}.$$

It is easy to verify that:

$$\frac{\partial}{\partial \mu} F(y;\theta) = \frac{\partial}{\partial \mu}\Phi(\frac{y-\mu}{\sigma}) = \left(\frac{-1}{\sigma}\right)\phi(\frac{y-\mu}{\sigma})\Big|_{y=F^{-1}(u;\theta)} = \left(\frac{-1}{\sigma}\right)\phi(\Phi^{-1}(u))$$

and:

$$\frac{\partial}{\partial\sigma}F(y;\theta)=\frac{\partial}{\partial\sigma}\Phi(\frac{y-\mu}{\sigma})=\left(\frac{-1}{\sigma^2}\right)(y-\mu)\phi(\frac{y-\mu}{\sigma})\Big|_{y=F^{-1}(u;\theta)}=\left(\frac{-1}{\sigma}\right)\Phi^{-1}(u)\phi(\Phi^{-1}(u))$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the standard normal distribution, respectively. Therefore:

$$\Psi(u)=\left(\frac{-1}{\sigma}\right)\begin{pmatrix}\phi(\Phi^{-1}(u))\\ \\ \Phi^{-1}(u)\phi(\Phi^{-1}(u))\end{pmatrix}.$$

Thus the limiting covariance of $\hat{W}_n(u)$ process is

$$\rho(s,t)=\min(s,t)-st-\Psi^T(t)I^{-1}(\theta)\Psi(s)$$

Finally, we can write the covariance function of $\hat{W}_n(u)$ process for the case of an i.i.d. normal sample as follows:

$$\begin{aligned}\Psi^T(t)I^{-1}(\theta)&=\left(\frac{-1}{\sigma}\right)\begin{pmatrix}\phi(\Phi^{-1}(t)) & \Phi^{-1}(t)\phi(\Phi^{-1}(t))\end{pmatrix}\begin{pmatrix}\sigma^2 & 0\\ 0 & \frac{\sigma^2}{2}\end{pmatrix}\\&=(-\sigma)\begin{pmatrix}\phi(\Phi^{-1}(t)) & \frac{1}{2}\Phi^{-1}(t)\phi(\Phi^{-1}(t))\end{pmatrix}.\end{aligned}$$

So

$$\begin{aligned}&\Psi^T(t)I^{-1}(\theta)\Psi(s)=\\&\qquad(-\sigma)\begin{pmatrix}\phi(\Phi^{-1}(t)) & \frac{1}{2}\Phi^{-1}(t)\phi(\Phi^{-1}(t))\end{pmatrix}(-\frac{1}{\sigma})\begin{pmatrix}\phi(\Phi^{-1}(s))\\ \\ \Phi^{-1}(s)\phi(\Phi^{-1}(s))\end{pmatrix}\\&\qquad=\phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s))+\frac{1}{2}\Phi^{-1}(t)\phi(\Phi^{-1}(t))\Phi^{-1}(s)\phi(\Phi^{-1}(s)).\end{aligned}$$

Thus,

$$\rho(s,t)=\min(s,t)-st-\phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s))-\frac{1}{2}\Phi^{-1}(t)\phi(\Phi^{-1}(t))\Phi^{-1}(s)\phi(\Phi^{-1}(s)).$$

In the case of an i.i.d. sample from Normal distribution, it is worth noting that the covariance function of $\hat{W}_n(u)$ is independent of any unknown parameters. Instead, it solely relies on the values of $0\le s,t\le 1$. This is to be expected because μ and σ are location and scale parameters. If our sample values $Y_1,Y_2,\ldots,Y_n$ are drawn from a $\mathrm{N}(\mu,\sigma^2)$ distribution then the variates $Z_1,Z_2,\ldots,Z_n$ where $Z_i=(Y_i-\mu)/\sigma$ are iid standard normal. We can express

the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ in terms of the Z's as follows:

$$\hat{\mu} = \mu + \sigma\bar{Z}$$

and

$$\hat{\sigma} = \sigma\sqrt{\frac{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}{n}}.$$

Define

$$s_Z = \sqrt{\frac{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}{n}}$$

to be the sample standard deviation of the Z values. The probability integral transforms for $i = 1, 2, \cdots, n$ are then

$$\hat{U}_i = \Phi\left(\frac{Y_i - \hat{\mu}}{\hat{\sigma}}\right) = \Phi\left(\frac{Z_i - \bar{Z}}{s_Z}\right).$$

This formula shows that goodness-of-fit statistics which depend only on the $\hat{U}_i$ have distributions which do not depend on μ or σ.

By employing the method described in Section 3.4, one can use the covariance function defined in Formula 3.25, discretize the integral equation, and approximate the eigenvalues of this covariance function to compute the P-value. As demonstrated, deriving this covariance function needs theoretical calculations. Instead of analytically computing the limiting covariance, we can estimate the covariance function directly from the sample using the approach detailed in Section 3.5. For a Normal distribution, first, the calculations depend on the matrix $S(\hat{\theta})$ whose i-th row presents the contribution to the score for the i-th observation in the score function with the following values:

$$S_i(\hat{\theta}) = \left[\begin{array}{cc} \frac{Y_i - \bar{Y}}{\hat{\sigma}^2} & \frac{(Y_i - \bar{Y})^2}{\hat{\sigma}^3} - \frac{1}{\hat{\sigma}} \end{array}\right].$$

The Fisher information matrix is estimated by the variance of the score function, i.e:

$$\hat{I}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\theta})^T S_i(\hat{\theta}) = \frac{S(\hat{\theta})^T S(\hat{\theta})}{n} = \frac{1}{\hat{\sigma}^2}\begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\hat{\sigma}}\right)^3 \\ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\hat{\sigma}}\right)^3 & \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\hat{\sigma}}\right)^4 - 1 \end{pmatrix}.$$

Second, the elements of the column vector $\Psi(u)$ are estimated by computing the covariance between the score function and an indicator function as described in Step 6 in Section 3.5. For each value of the $0 \leq u \leq 1$ in the grid being used, we estimate $\Psi(u)$ by:

$$\hat{\Psi}(u) = \frac{1}{n}\sum_{i=1}^{n} I(\hat{U}_i \leq u)\left[\begin{array}{cc} \frac{Y_i - \bar{Y}}{\hat{\sigma}^2} & \frac{(Y_i - \bar{Y})^2}{\hat{\sigma}^3} - \frac{1}{\hat{\sigma}} \end{array}\right].$$

We compute the matrix Δ as described in Section 3.5 step 7, and follow step 8 and step 9 in that section to compute the P-value.

3.7 Example 2: Gamma distribution

In this section, we show how to calculate the covariance function of the stochastic process defined in Section 3.24 for the case of an i.i.d. sample from Gamma distribution. Suppose we have an i.i.d. sample such as $Y_1, Y_2, \ldots, Y_n$ from Gamma(α, λ) distribution with the following probability distribution function:

$$f(Y; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\ Y^{\alpha-1}\ e^{-\lambda Y}$$

where $\alpha > 0$ is the shape, $\lambda > 0$ is the inverse scale parameter and $\Gamma(\alpha)$ is the Gamma function defined as $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$. Therefore we can write the likelihood function as:

$$\begin{aligned} L(\alpha, \lambda) = \prod_{i=1}^{n} f(Y_i; \alpha, \lambda) &= \prod_{i=1}^{n} \frac{\lambda^\alpha}{\Gamma(\alpha)}\ Y_i^{\alpha-1}\ e^{-\lambda Y_i} = \Big(\frac{\lambda^\alpha}{\Gamma(\alpha)}\Big)^n \Big(\prod_{i=1}^{n} Y_i\Big)^{\alpha-1} e^{-\lambda \sum_{i=1}^n Y_i} \\ &= \frac{\lambda^{n\alpha}}{\Big(\Gamma(\alpha)\Big)^n}\ \Big(\prod_{i=1}^{n} Y_i\Big)^{\alpha-1}\ e^{-\lambda \sum_{i=1}^n Y_i}. \end{aligned}$$

We now proceed by reparametrizing the model in terms of α and the mean of Y denoted by μ. This will help connect this example to work in later sections when we study generalized linear models. Note that in Gamma distribution $\mu = E(Y_i) = \frac{\alpha}{\lambda}$. We plug in the value of λ and rewrite the likelihood with the new set of parameters, i.e (α, μ) as follows:

$$L(\alpha, \mu) = \frac{(\frac{\alpha}{\mu})^{n\alpha}}{\Big(\Gamma(\alpha)\Big)^n}\Big(\prod_{i=1}^{n} Y_i\Big)^{\alpha-1}\ e^{-\frac{\alpha}{\mu}\sum_{i=1}^n Y_i}.$$

The log-likelihood follows:

$$\begin{aligned} l(\alpha, \mu) &= (n\alpha)\ln\Big(\frac{\alpha}{\mu}\Big) - n\ln\Big(\Gamma(\alpha)\Big) + \Big(\alpha - 1\Big)\sum_{i=1}^{n}\ln(Y_i) - \Big(\frac{\alpha}{\mu}\Big)\sum_{i=1}^{n} Y_i \\ &= (n\alpha)\ln(\alpha) - (n\alpha)\ln(\mu) - n\ln\Big(\Gamma(\alpha)\Big) + (\alpha - 1)\sum_{i=1}^{n}\ln(Y_i) - \Big(\frac{\alpha}{\mu}\Big)\sum_{i=1}^{n} Y_i. \end{aligned}$$

We take the following steps to calculate the covariance function of the stochastic process $\hat{W}_n(u)$. The first step is to obtain the Fisher information matrix by calculating the partial

derivatives of the log-likelihood function with respect to the parameters as follows:

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= n\ln(\alpha) + (n\alpha)\frac{1}{\alpha} - n\ln(\mu) - n\frac{\Gamma^{'}(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\ln(Y_i) - \frac{1}{\mu}\sum_{i=1}^{n}Y_i \\
&= n\ln(\alpha) + n - n\ln(\mu) - nD(\alpha) + \sum_{i=1}^{n}\ln(Y_i) - \frac{1}{\mu}\sum_{i=1}^{n}Y_i \\
\frac{\partial^2 l}{\partial \alpha^2} &= \frac{n}{\alpha} - nD^{'}(\alpha), \quad \frac{\partial l}{\partial \mu} = \frac{-n\alpha}{\mu} + \alpha\sum_{i=1}^{n}\frac{Y_i}{\mu^2}, \quad \frac{\partial^2 l}{\partial \mu^2} = \frac{n\alpha}{\mu^2} - \frac{2\alpha}{\mu^3}\sum_{i=1}^{n}Y_i \\
\frac{\partial l}{\partial \mu \partial \alpha} &= \frac{-n}{\mu} + \sum_{i=1}^{n}\frac{Y_i}{\mu^2}
\end{aligned}$$

where $D(\alpha) = \frac{\Gamma^{'}(\alpha)}{\Gamma(\alpha)}$ is the *digamma* function whose derivative $D^{'}(\alpha) = \frac{\partial}{\partial \alpha}D(\alpha)$ is the *trigamma* function. The Fisher information matrix is:

$$\begin{aligned}
I_n(\theta) &= -E\begin{pmatrix} \frac{\partial^2}{\partial \alpha^2}l(\alpha,\mu) & \frac{\partial}{\partial \alpha \partial \mu}l(\alpha,\mu) \\ \frac{\partial}{\partial \mu \partial \alpha}l(\alpha,\mu) & \frac{\partial^2}{\partial \mu^2}l(\alpha,\mu) \end{pmatrix} = -E\begin{pmatrix} \frac{n}{\alpha} - nD^{'}(\alpha) & \frac{-n}{\mu} + \frac{\sum_{i=1}^{n}Y_i}{\mu^2} \\ \frac{-n}{\mu} + \frac{\sum_{i=1}^{n}Y_i}{\mu^2} & \frac{n\alpha}{\mu^2} - \frac{2\alpha}{\mu^3}\sum_{i=1}^{n}Y_i \end{pmatrix} \\
&= \begin{pmatrix} nD^{'}(\alpha) - \frac{n}{\alpha} & 0 \\ 0 & \frac{-n\alpha}{\mu^2} \end{pmatrix}
\end{aligned}$$

and the Fisher information matrix based on one observation in the sample is:

$$I(\theta) = \begin{pmatrix} D^{'}(\alpha) - \frac{1}{\alpha} & 0 \\ 0 & \frac{-\alpha}{\mu^2} \end{pmatrix}$$

and the inverse of Fisher information matrix is:

$$I^{-1}(\theta) = \begin{pmatrix} \frac{-\alpha}{1-\alpha D^{'}(\alpha)} & 0 \\ 0 & \frac{-\mu^2}{\alpha} \end{pmatrix}.$$

The next step is to obtain the vector $\Psi(y)$, the partial derivatives of the cumulative distribution function with respect to the parameters and evaluate it at $y = F^{-1}(u)$. The cumulative distribution function of a Gamma distribution at any value of y is calculated as:

$$G(y;\alpha,\mu) = \int_0^y \frac{\alpha^\alpha}{\mu^\alpha\ \Gamma(\alpha)}\ t^{\alpha-1}\ e^{-\frac{\alpha}{\mu}t}dt.$$

We need to calculate $\frac{\partial}{\partial\alpha}G(y;\alpha,\mu)$ and $\frac{\partial}{\partial\mu}G(y;\alpha,\mu)$ to obtain $\Psi(u)$ vector. In order to do so, we apply the change of variable as follows:

$$z = \frac{\alpha}{\mu}t \Rightarrow dz = \frac{\alpha}{\mu}dt \Rightarrow dt = \frac{\mu}{\alpha}dz$$

to write the cumulative distribution function as:

$$\begin{aligned} G(y;\alpha,\mu) &= \int_0^{\frac{\alpha y}{\mu}} \frac{\alpha^\alpha}{\mu^\alpha\ \Gamma(\alpha)}\left(\frac{\mu z}{\alpha}\right)^{\alpha-1} e^{-z}\ \frac{\mu}{\alpha}\ dz = \frac{\alpha^\alpha}{\mu^\alpha\ \Gamma(\alpha)}\ \frac{\mu^{\alpha-1}}{\alpha^{\alpha-1}}\ \frac{\mu}{\alpha}\int_0^{\frac{\alpha y}{\mu}} z^{\alpha-1}\ e^{-z}\ dz \\ &= \int_0^{\frac{\alpha y}{\mu}} \frac{1}{\Gamma(\alpha)}\ z^{\alpha-1}\ e^{-z}\ dz. \end{aligned}$$

Now the partial derivative with respect to α is:

$$\begin{aligned} \frac{\partial}{\partial\alpha}G(y;\alpha,\mu) &= \frac{\partial}{\partial\alpha}\ \left(\int_0^{\frac{\alpha y}{\mu}} \frac{1}{\Gamma(\alpha)}\ z^{\alpha-1}\ e^{-z}\ dz\right) = \frac{\partial}{\partial\alpha}\ \left(\int_0^{\frac{\alpha y}{\mu}}\ g(z;\alpha,1)\ dz\right) \\ &= \frac{\partial}{\partial\alpha}\left(\frac{\alpha y}{\mu}\right)\ g(\frac{\alpha y}{\mu};\alpha,1)\ +\ \int_0^{\frac{\alpha y}{\mu}} \frac{\partial}{\partial\alpha}\ g(z;\alpha,1)\ dz \\ &= \frac{y}{\mu}\ g(\frac{\alpha y}{\mu};\alpha,1)\ +\ \int_0^{\frac{\alpha y}{\mu}} \frac{\partial \ln(g(z;\alpha,1))}{\partial\alpha}\ g(z;\alpha,1)\ dz \\ &= \frac{y}{\mu}\ g(\frac{\alpha y}{\mu};\alpha,1)\ +\ \int_0^{\frac{\alpha y}{\mu}} \frac{\partial}{\partial\alpha}\left((\alpha-1)\ln(z)\ -\ z\ -\ \ln(\Gamma(\alpha))\right)\ g(z;\alpha,1)\ dz \\ &= \frac{y}{\mu}\ g(\frac{\alpha y}{\mu};\alpha,1)\ +\ \int_0^{\frac{\alpha y}{\mu}} \left(\ln(z)\ -\ D(\alpha)\right)\ g(z;\alpha,1)\ dz \\ &= \frac{y}{\mu}\ g(\frac{\alpha y}{\mu};\alpha,1)\ +\ \int_0^{\frac{\alpha y}{\mu}}\ \ln(z)\ g(z;\alpha,1)\ dz\ -\ \int_0^{\frac{\alpha y}{\mu}}\ D(\alpha)\ g(z;\alpha,1)\ dz \end{aligned}$$

and the partial derivative with respect to μ is:

$$\begin{aligned} \frac{\partial}{\partial\mu}G(y;\alpha,\mu) &= \frac{\partial}{\partial\mu}\ \left(\int_0^{\frac{\alpha y}{\mu}} \frac{1}{\Gamma(\alpha)}\ z^{\alpha-1}\ e^{-z}\ dz\right) \\ &= \frac{\partial}{\partial\mu}\ \left(\int_0^{\frac{\alpha y}{\mu}}\ g(z;\alpha,1)\ dz\right) \\ &= \frac{\partial}{\partial\mu}\left(\frac{\alpha y}{\mu}\right)\ g(\frac{\alpha y}{\mu};\alpha,1)\ +\ \int_0^{\frac{\alpha y}{\mu}} \frac{\partial}{\partial\mu}\ g(z;\alpha,1)\ dz \\ &= \frac{-\alpha\ y}{\mu^2}\ g(\frac{\alpha y}{\mu};\alpha,1) + \int_0^{\frac{\alpha y}{\mu}}\ \frac{\partial}{\partial\mu}g(z;\alpha,1)\ dz \\ &= \frac{-\alpha\ y}{\mu^2}\ g(\frac{\alpha y}{\mu};\alpha,1) + \int_0^{\frac{\alpha y}{\mu}}\ \frac{\partial}{\partial\alpha}g(z;\alpha,1)\frac{\partial\alpha}{\partial\mu}\ dz \end{aligned}$$

$$= \frac{-\alpha\ y}{\mu^2}\ g(\frac{\alpha y}{\mu};\alpha,1) + \int_0^{\frac{\alpha y}{\mu}} \Big[\ln(z) - D(\alpha)\Big] g(z;\alpha,1)\lambda\ dz$$
$$= \frac{-\alpha\ y}{\mu^2}\ g(\frac{\alpha y}{\mu};\alpha,1) + \lambda\int_0^{\frac{\alpha y}{\mu}} \ln(z)\ g(z;\alpha,1)\ dz - \lambda D(\alpha)\int_0^{\frac{\alpha y}{\mu}} g(z;\alpha,1)\ dz.$$

We evaluate both terms at $y = Q(u;\alpha,\mu) = \frac{\mu}{\alpha}Q(u;\alpha,1)$ where Q is the inverse of cumulative distribution function of a Gamma distribution with shape parameter of α and scale of one. Therefore we can write the elements of $\Psi(u)$ vector as follows:

$$\frac{\partial}{\partial\alpha}G(y;\alpha,\mu)\big|_{y=\frac{\mu}{\alpha}Q(u;\alpha,1)} = \frac{1}{\alpha}Q(u;\alpha,1)g(Q(u;\alpha,1);\alpha,1) + \int_0^{Q(u;\alpha,1)} \ln(z)g(z;\alpha,1)\ dz$$
$$- \int_0^{Q(u;\alpha,1)} D(\alpha)g(z;\alpha,1)$$

$$\frac{\partial}{\partial\mu}G(y;\alpha,\mu)\big|_{y=\frac{\mu}{\alpha}Q(u;\alpha,1)} = \frac{-1}{\mu}Q(u;\alpha,1)g(Q(u;\alpha,1);\alpha,1) + \lambda\int_0^{Q(u;\alpha,1)} \ln(z)g(z;\alpha,1)\ dz$$
$$- \lambda D(\alpha)\int_0^{Q(u;\alpha,1)} g(z;\alpha,1)\ dz.$$

It is possible to compute the Fisher information matrix and the elements of the $\Psi(u)$ vector numerically for a given sample. For example, the functions *digamma* and *trigamma* function from the *MASS* package in the R statistical software can be used to calculate $D(\alpha)$ and its derivative $D'(\alpha)$. The maximum likelihood estimate of α and λ can be computed using a procedure to find the root of the score function with any desired precision. Finally, a simple routine, in combination with the *qgamma* function from the *stats* package in R, can be applied to compute $\int_0^{Q(u;\alpha,1)} \log(z)g(z;\alpha,1)dz$ accurately. Thus the limiting covariance function of $\hat{W}_n(u)$ process will be:

$$\rho(s,t) = \min(s,t) - st - \Psi^T(t)I^{-1}(\theta)\Psi(s)$$
$$= \min(s,t) - st - \Psi^T(t)\begin{pmatrix} \frac{-\alpha}{1-\alpha D'(\alpha)} & 0 \\ 0 & \frac{-\mu^2}{\alpha} \end{pmatrix}\Psi(s).$$

In the case of an i.i.d. sample from the Gamma distribution, the covariance function of $\hat{W}_n(u)$ not only depends on s and t but also depends on the shape parameter in the Gamma distribution. The analytic calculation of the covariance function is long and tedious. We will not show the details here but similar to the Normal example described in Section 3.6, we can instead estimate the covariance function and eigenvalues. We follow the steps as described in Section 3.5.

The likelihood function of a Gamma distribution can be expressed in two forms. In the example we discussed here, we examined one form of the likelihood and derived the

covariance function based on it. It is natural to wonder which likelihood parametrization one should choose. In the next section, we will demonstrate the invariance property of the covariance function of the $\hat{W}_n(u)$ process we have been discussing thus far. We will show that both the covariance function and the statistic itself are independent of the parametrization of the likelihood. This applies to the CvM statistic, the estimated eigenvalues, and the final P-value.

3.8 Parametrization invariance of the statistic and the covariance function

In this section, we show the invariance of our statistics and of the covariance function of the stochastic process $\hat{W}_n(u)$ to the parametrization of θ in the likelihood function. Consider an i.i.d. sample $Y_1, Y_2, \ldots, Y_n$ from a parametric model $F = \{f_\theta(y); \theta \in \Theta\}$ where Θ is the parameter space and θ contains p unknown parameters. A new parametrization of the parametric model is obtained by defining a one-to-one function which maps θ into a new set of parameters such as $\phi \in \Phi$ where Φ is the new parameter space resulted from the mapping. Note that the map is one-to-one therefore the two class of parametric models are essentially the same, i.e $\{f_\theta(y); \theta \in \Theta\} = \{f_\phi(y); \phi \in \Phi\}$. One example of this is the Gamma distribution where one can write the likelihood in terms of (α, λ) or (α, β) where α is the shape parameter and $\beta = \frac{1}{\lambda}$ is the inverse of scale parameter λ. The goal of this section is to calculate the covariance function in new parameter space and show the invariance property of such a function.

Note first that it is well known that the mle of ϕ is simply $\hat{\phi} = \phi(\hat{\theta})$. As a result the estimated probability integral transforms do not depend on the parametrization so the goodness-of-fit statistics do not depend on the parametrization. Once we show that the estimated covariance functions are also parametrization invariant we will see that the p-values themselves do not depend on the parametrization.

To start with, we define the score function based on the original parametrization of the likelihood function as $S(\theta) = S(Y; \theta)$ which is a matrix with n rows and p columns. We define the elements of this matrix as follows:

$$S(Y_i; \theta_j) = \frac{\partial}{\partial \theta_j} log\Big(f(Y_i; \theta)\Big) \quad i = 1, 2, 3, \ldots, n \quad j = 1, 2, 3, .., p.$$

The score function based on the new parametrization is $S(\phi) = S(Y; \phi)$, a matrix with n rows and p columns with the following elements:

$$S(Y_i; \phi_k) = \frac{\partial}{\partial \phi_k} log\left(f(Y_i; \phi)\right) \quad i = 1, 2, 3, \ldots, n \quad k = 1, 2, 3, .., p.$$

It is easy to verify that for a fixed value of k, the derivative of log-likelihood in the new parametrization with respect to the k-element of ϕ is:

$$\frac{\partial}{\partial \phi_k} log\left(f(Y_i;\phi)\right) = \sum_{j=1}^{p} \frac{\partial}{\partial \theta_j} log\left(f(Y_i;\phi)\right) \frac{\partial \theta_j}{\partial \phi_k} = \sum_{j=1}^{p} \frac{\partial}{\partial \theta_j} log\left(f(Y_i;\phi(\theta))\right) \frac{\partial \theta_j}{\partial \phi_k}$$
$$= \sum_{j=1}^{p} S(Y_i;\phi(\theta)) \frac{\partial \theta_j}{\partial \phi_k}.$$

Note that for each value of k, $\frac{\partial}{\partial \phi_k} log\left(f(y_i;\phi)\right)$ produces a column vector of size n indexed by i. For each value of i it produces a row vector with p entries indexed by k. As a result, the score matrix based on the new parametrization is $n \times p$ and is given by $S(\phi) = S(\theta)D$ where D is the $p \times p$ matrix with i, jth entry $\frac{\partial \theta_j}{\partial \phi_k}$ in which both j and k vary from 1 to p.

We also need to calculate the elements of the column vector $\Psi_\phi(u)$ in the new presentation of likelihood. Again we can use the chain rule to compute the column vector $\Psi_\phi(u)$ from the column vector $\Psi_\theta(u)$ as follows:

$$\Psi_\phi(u) = \frac{\partial}{\partial \phi} F(y;\phi) = \frac{\partial}{\partial \phi} \int_{-\infty}^{y} f(t;\phi)\ dt = \int_{-\infty}^{y} \frac{\partial}{\partial \phi} f(t;\phi)\ dt = \int_{-\infty}^{y} \frac{\partial \theta}{\partial \phi}^T \frac{\partial}{\partial \theta} f(t;\phi)\ dt$$
$$= \frac{\partial \theta}{\partial \phi}^T \left(\int_{-\infty}^{y} \frac{\partial}{\partial \theta} f(t;\phi)\ dt \right) = \frac{\partial \theta}{\partial \phi}^T \left(\frac{\partial}{\partial \theta} \int_{-\infty}^{y} f(t;\phi)\ dt \right) = \frac{\partial \theta}{\partial \phi}^T \left(\frac{\partial}{\partial \theta} \int_{-\infty}^{y} f(t;\phi(\theta))\ dt \right)$$
$$= \frac{\partial \theta}{\partial \phi}^T \frac{\partial}{\partial \theta} F\left(y;\phi(\theta)\right) = \frac{\partial \theta}{\partial \phi}^T \Psi_\theta(u) = D^T \Psi_\theta(u).$$

To be clear, the k-th element of this column vector is:

$$\frac{\partial}{\partial \phi_k} F(y;\phi) = \frac{\partial}{\partial \phi_k} F(y;\phi(\theta)) = \sum_{j=1}^{p} \frac{\partial}{\partial \theta_j} F(y;\phi(\theta)) \frac{\partial \theta_j}{\partial \phi_k} = \sum_{j=1}^{p} \Psi_{\theta_j} \frac{\partial \theta_j}{\partial \phi_k}$$

and Ψ_{θ_j} is the j-th element of $\Psi_\theta(u)$ vector. The Fisher information matrix based on the original parametrization is:

$$I(\theta) = Var\left(S(\theta)\right) = E\left[S(Y;\theta)^T S(Y;\theta)\right].$$

The Fisher information matrix according to the new parametrization is:

$$I(\phi) = Var\left(S(\phi)\right) = Var\left(S(\theta)D\right) = D^T Var(S(\theta))D = D^T I(\theta) D$$

and the inverse of the Fisher information is:

$$I^{-1}(\phi) = \left(D^T I(\theta) D\right)^{-1} = D^{-1} I^{-1}(\theta)(D^T)^{-1}.$$

The stochastic process based on the new parametrization is calculated as follows:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(I(F(Y_i;\phi) \leq u) - u - S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(u) \Big).$$

Now we show that the covariance of this process based on new parameter of ϕ is the same as the covariance of the $\hat{W}_n(u)$ process based on θ. To start, for any values of $0 \leq u \leq 1$ we define:

$$Z_i(u) = I(F(Y_i;\phi) \leq u) - u - S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(u)$$

therefore we can write:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(u).$$

We use the notation of $\rho_\phi(s,t)$ to denote the limiting covariance function in the new parametrization and calculate as follows:

$$\begin{aligned}
\rho_\phi(s,t) = \mathrm{Cov}(\hat{W}_n(s), \hat{W}_n(t)) &= \mathrm{Cov}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(s), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(t) \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{n} \mathrm{Cov}(Z_i(s), Z_i(t)) + \sum_{i \neq j}^{n} \mathrm{Cov}(Z_i(s), Z_j(t)) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Cov}\left(Z_i(s), Z_i(t) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Cov}\Big(I(F(Y_i;\phi) \leq s) - s - S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(s), \\
&\qquad I(F(Y_i;\phi) \leq t) - t - S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(t) \Big) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big\{ \mathrm{Cov}\Big(I(F(Y_i;\phi) \leq s) - s, I(F(Y_i;\phi) \leq t) - t \Big) \\
&\qquad - \mathrm{Cov}\Big(I(F(Y_i;\phi) \leq s) - s, S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(t) \Big) \\
&\qquad - \mathrm{Cov}\Big(S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(s), I(F(Y_i;\phi) \leq t) - t \Big) \\
&\qquad + \mathrm{Cov}\Big(S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(s), S(Y_i;\phi) I^{-1}(\phi) \Psi_\phi(t)) \Big) \Big\}.
\end{aligned}$$

Note that the summation applies over all four terms. The calculation continues as follows. For the first term, note that since $F(Y_i;\phi) = F(Y_i;\theta)$, we can write:

$$\frac{1}{n}\sum_{i=1}^{n}\text{Cov}(I(F(Y_i;\theta)\leq s)-s, I(F(Y_i;\theta)\leq t)-t) = \min(s,t)-st.$$

For the second term, we can write (note that we apply the fact that $E\left(S(Y_i;\phi)\right)=0$):

$$\begin{aligned}
\text{Cov}\Big(I(F(Y_i;\phi)\leq s), S(Y_i;\phi)I^{-1}(\phi)\Psi_\phi(t)\Big) &= E\Big[I(F(Y_i;\phi)\leq s)S(Y_i;\phi)I^{-1}(\phi)\Psi_\phi(t)\Big] \\
&= \Psi_\phi^T(t)I^{-1}(\phi)E\Big[I(F(Y_i;\phi)\leq s)S(Y_i;\phi)\Big] \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\int_0^s g(u;\phi)S(Y_i;\phi)du \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\int_0^s g(u;\phi)\frac{\partial}{\partial\phi}log(g(u;\phi))du \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\int_0^s g(u;\phi)\frac{\frac{\partial}{\partial\phi}g(u;\phi)}{g(u;\phi)}du \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\int_0^s \frac{\partial}{\partial\phi}g(u;\phi)du \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\frac{\partial}{\partial\phi}\int_0^s g(u;\phi)du \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\frac{\partial}{\partial\phi}\left(G(s)-G(0)\right) \\
&= \Psi_\phi^T(t)I^{-1}(\phi)\Psi_\phi(s).
\end{aligned}$$

For the third term, switch the roles of s and t to check that:

$$\text{Cov}\Big(I(F(Y_i;\phi)\leq t), S(Y_i;\phi)I^{-1}(\phi)\Psi_\phi(s)\Big) = \Psi_\phi^T(s)I^{-1}(\phi)\Psi_\phi(t).$$

Finally we calculate the last term in the summation as follows:

$$\begin{aligned}
&\text{Cov}\Big(S(Y_i;\phi)I^{-1}(\phi)\Psi_\phi(s), S(Y_i;\phi)I^{-1}(\phi)\Psi_\phi(t)\Big) \\
&= \Psi_\phi^T(s)I^{-1}(\phi)\text{Cov}\Big(S(Y_i;\phi), S(Y_i;\phi)\Big)I^{-1}(\phi)\Psi_\phi(t) \\
&= \Psi_\phi^T(s)I^{-1}(\phi)Var(S(Y_i;\phi))I^{-1}(\phi)\Psi_\phi(t) \\
&= \Psi_\phi^T(s)I^{-1}(\phi)I(\phi)I^{-1}(\phi)\Psi_\phi(t) \\
&= \Psi_\phi^T(s)I^{-1}(\phi)\Psi_\phi(t).
\end{aligned}$$

This shows that the limiting covariance function of the stochastic process $\hat{W}_n(u)$ based on the new parametrization is:

$$\rho_\phi(s,t) = \min(s,t) - st - \Psi_\phi^T(t) I^{-1}(\phi) \Psi_\phi(s)$$

for any $0 \leq s,t \leq 1$. To show the equivalence of covariance function with the original parametrization, we note that $\Psi_\phi(u) = D^T \Psi_\theta(u)$ and $I^{-1}(\phi) = D^{-1} I^{-1}(\theta)(D^T)^{-1}$. Plug these values back in the formula for $\rho_\phi(s,t)$, we conclude:

$$\begin{aligned}
\rho_\phi(s,t) &= \min(s,t) - st - \left(D^T \Psi_\theta(t)\right)^T \left(D^{-1} I^{-1}(\theta)(D^T)^{-1}\right) D^T \Psi_\theta(s) \\
&= \min(s,t) - st - \Psi_\theta^T(t)\, D\, D^{-1} I^{-1}(\theta)(D^T)^{-1}\, D^T \Psi_\theta(s) \\
&= \min(s,t) - st - \Psi_\theta^T(t) I^{-1}(\theta) \Psi_\theta(s) = \rho(s,t).
\end{aligned}$$

As we can see, the covariance function does not change as a result of the new parametrization. Consequently, the estimated eigenvalues, as discussed in Section 3.5, remain the same. As previously noted, since $F(Y_i;\phi) = F(Y_i;\theta)$, the value of the CvM statistic also remains unchanged. This leads to the conclusion that the parametrization of the likelihood model does not alter the p-value of the goodness-of-fit test. In the next section, we review the theory and steps for applying goodness-of-fit tests based on empirical distribution functions in a linear model.

3.9 Example 3: Linear model

In sections 3.6 and 3.7, we examined the large sample theory for independently and identically distributed (i.i.d) samples from the Normal and Gamma distributions. Specifically, we showed the steps to calculate the covariance function of the stochastic process. In both examples, Y_i was assumed to be derived from the same $F(Y;\theta)$ distribution. In particular, we assumed that the expected value of the response variable was the same for all Y_i. However, the expected value of each response variable could depend on some covariates. In this section, we review the theory for linear models where the response variables for different observations are still independent but their expected values depend on some covariates. The goal is to test the null hypothesis that the usual assumptions about the distribution of the response variable are correct.

To start with, consider a linear regression model for the relationship between k explanatory variables and a response variable. An independent sample of n observations is available in the form of $(Y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$. Each Y_i is a response value and the x_{ij} is the value of the jth explanatory variable for the i-th observation. For $i = 1, 2, 3, \ldots, n$, we define the covariate vector as $x_i^T = (x_{1i}, x_{2i}, \ldots, x_{ki})$. The Y_i's are independent from each other but their expected value, $E[Y_i]$, depends on the values of the covariate vector, x_i^T; as usual in linear

regression our analysis is conditional on the covariates – they are treated as non-random constants. Therefore, the linear regression model is written as $Y = X\beta + \epsilon$, where Y and ϵ are both column vectors with n elements. The matrix X has n rows and $p = k + 1$ columns to include the intercept in the model. The column vector β has p elements. The usual assumptions apply for error terms where means are assumed to be zero and the variance σ^2 is constant.

Here we review the theory when the response is assumed to follow a normal distribution but the idea remains the same for other continuous distributions. We start by writing the likelihood function of the response variable:

$$L(\theta) = \left(2\pi\sigma^2\right)^{\frac{-n}{2}} e^{\frac{-1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}$$

and the log-likelihood is:

$$l(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta).$$

For convenience, we define θ which contains both β and σ. The next step is to calculate partial derivatives:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= \frac{-1}{2\sigma^2}\left(2X^TX\beta - 2X^TY\right) = \frac{1}{\sigma^2}X^T(Y - X\beta) \\
\frac{\partial^2 l}{\partial \beta^2} &= \frac{-1}{\sigma^2}X^TX \\
\frac{\partial l}{\partial \sigma} &= \frac{-n}{\sigma} + \frac{1}{\sigma^3}(Y - X\beta)^T \\
\frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}(Y - X\beta)^T(Y - X\beta) \\
\frac{\partial l}{\partial \sigma \partial \beta} &= \frac{2}{\sigma^3}X^T(X\beta - Y).
\end{aligned}$$

The Fisher information matrix is:

$$I(\theta) = -E\left[\frac{\partial^2 l}{\partial \theta^2}\right] = \begin{pmatrix} \frac{X^TX}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} X^TX & 0 \\ 0 & 2n \end{pmatrix}.$$

For a linear model with intercept, we can write X^TX in the following form. Note that for this model the first column in X has all entries equal to 1. Therefore:

$$X^TX = [1, X]^T[1, X] = \begin{pmatrix} n & n\bar{X}^T \\ n\bar{X} & X^TX \end{pmatrix} = n\begin{pmatrix} 1 & \bar{X}^T \\ \bar{X} & \frac{X^TX}{n} \end{pmatrix} = n\begin{pmatrix} 1 & \bar{X}^T \\ \bar{X} & \frac{1}{n}\sum_{i=1}^n x_i x_i^T \end{pmatrix}$$

where $\bar{X}^T = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ is the vector of averages of explanatory variables. We can thus write the Fisher information matrix as follows:

$$I_n(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} 1 & \bar{X}^T & 0 \\ \bar{X} & \frac{1}{n}\sum_{i=1}^n x_i x_i^T & 0 \\ 0 & 0 & 2 \end{pmatrix}. \tag{3.25}$$

The average Fisher information matrix based on one observation is:

$$I(\theta) = \frac{1}{n} I_n(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & \bar{X}^T & 0 \\ \bar{X} & \frac{1}{n}\sum_{i=1}^n x_i x_i^T & 0 \\ 0 & 0 & 2 \end{pmatrix}. \tag{3.26}$$

Before we calculate the covariance function of the stochastic process $\hat{W}_n(u)$ in the linear model case, we would like to point out some important changes from the i.i.d. case. The distribution of Y_i depends on covariate x_i^T values and as a result the calculations of $\hat{W}_n$ in 3.16 will change slightly. First, $F(Y;\theta)$ depends on covariates now. Therefore we rewrite 3.16 in the following way:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(I\left(F_i(Y_i, \hat{\theta}) \leq u\right) - u \right). \tag{3.27}$$

Note that the subscript in F shows the dependency of the CDF on i-th observation. Similar to calculations in Section 3.4, we define $Q_i(Y;\hat{\theta}) = F_i^{-1}(Y;\hat{\theta})$ as the inverse of the cumulative distribution function $F_i(Y;\hat{\theta})$ and apply this inverse transform on both sides of the inequality in 3.27 to obtain the following expression:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(I\left(Y_i \leq Q_i(u;\hat{\theta})\right) - u \right).$$

Using the fact that $F_i(Y_i;\theta)$ is a non-decreasing function of Y_i, we continue:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(I\left(F_i(Y_i;\theta) \leq F_i(Q_i(u;\hat{\theta});\theta)\right) - u \right).$$

Following the ideas leading to formula 3.17, the expansion of $\hat{W}_n(u)$ in this case is:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(I(U_i \leq F_i(Q_i(u;\hat{\theta});\theta)) - F_i(Q_i(u;\hat{\theta});\theta) \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(F_i(Q_i(u;\hat{\theta});\theta) - u \right).$$

We approximate the first of the two terms on the last line by $W_n(u)$ to write the expression as follows [53]:

$$\begin{aligned}\hat{W}_n(u) &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I(U_i \le u) - u\Big) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(F_i(Q_i(u;\hat{\theta});\theta) - u\Big) \\ &= W_n(u) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(F_i(Q_i(u;\hat{\theta});\theta) - u\Big).\end{aligned} \tag{3.28}$$

As in 3.21, we write the Taylor expansion of $F_i(Q_i(u;\hat{\theta});\theta)$ around θ in the form:

$$F_i(Q_i(u;\hat{\theta});\theta) = u - \frac{\partial}{\partial\hat{\theta}}F_i(Q_i(u;\hat{\theta});\theta)\,(\hat{\theta}-\theta) + O_p[(\hat{\theta}-\theta)^2]. \tag{3.29}$$

Note that the remainder term is negligible for relatively large values of n. Replacing the value of $F_i(Q_i(u;\hat{\theta});\theta)$ from 3.29 in 3.28 and evaluating at $\hat{\theta} = \theta$ we can write $\hat{W}_n(u)$ as:

$$\begin{aligned}\hat{W}_n(u) &= W_n(u) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}F_i(Q_i(u;\theta);\theta)\,(\hat{\theta}-\theta) \\ &= W_n(u) - \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}F_i(Q_i(u;\theta);\theta)(\hat{\theta}-\theta) \\ &= W_n(u) - \sqrt{n}\Psi_n^T(u)(\hat{\theta}-\theta).\end{aligned}$$

In this expansion, $\Psi_n(u)$ remains a column vector with elements representing the partial derivatives of the cumulative distribution function with respect to θ. However, the values are averaged over the sample since the distribution of each Y_i depends on covariates. In other words:

$$\Psi_n(u) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}F_i(Q_i(u;\theta);\theta).$$

Following the same approach as for an i.i.d. sample, we can finally write $\hat{W}_n(u)$ in terms of the score function based on each observation as follows:

$$\hat{W}_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(I\left(F_i(Y_i,\theta) \le u\right) - u - S_i(\theta)I^{-1}(\theta)\Psi_n(u)\Big).$$

The covariance function of $\hat{W}_n(u)$ remains the same as the i.i.d. case except the $\Psi(u)$ is replaced by $\Psi_n(u)$, as follows:

$$\rho(s,t) = \min(s,t) - st - \Psi_n^T(t)I^{-1}(\theta)\Psi_n(s). \tag{3.30}$$

In this example for linear model with normal residuals, there are $k+1$ estimated values for the coefficients (k explanatory variables and one intercept) and one estimate for σ. In

total there are $(k+1)+1 = p+1$ parameters. Hence the vector $\Psi_n(u)$ contains $p+1$ elements. For $j = 1, 2, 3, \ldots, p$, the j-th element of $\Psi_n(u)$ evaluated at the inverse of the CDF is:

$$\begin{aligned}\Psi_j(u) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta_j}F_i(Y_i;\beta,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta_j}\Phi(\frac{Y_i - x_{ij}\beta_j}{\sigma}) = \frac{1}{n}\sum_{i=1}^{n}(\frac{-x_{ij}}{\sigma})\phi(\frac{Y_i - x_{ij}\beta_j}{\sigma}) \\ &= \frac{-1}{n\sigma}\sum_{i=1}^{n}x_{ij}\,\phi(\Phi^{-1}(u)) = \frac{-\phi(\Phi^{-1}(u))}{n\sigma}\sum_{i=1}^{n}x_{ij} = \frac{-\phi(\Phi^{-1}(u))}{\sigma}\,\bar{x_j}.\end{aligned}$$

Similarly, for $j = p+1$ the j-th element of $\Psi_n(u)$ vector evaluated at the inverse of CDF is:

$$\begin{aligned}\Psi_j(u) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\sigma}F_i(Y_i;\beta,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\sigma}\Phi(\frac{Y_i - x_{ij}\beta_j}{\sigma}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\frac{-1}{\sigma^2})(Y_i - x_{ij}\beta_j)\phi(\frac{Y_i - x_{ij}\beta_j}{\sigma}) \\ &= (\frac{-1}{n\sigma})\sum_{i=1}^{n}(\frac{Y_i - x_{ij}\beta_j}{\sigma})\phi(\frac{Y_i - x_{ij}\beta_j}{\sigma}) \\ &= (\frac{-1}{n\sigma})\sum_{i=1}^{n}\Phi^{-1}(u)\phi(\Phi^{-1}(u)) = \left(\frac{-1}{\sigma}\right)\Phi^{-1}(u)\,\phi(\Phi^{-1}(u)).\end{aligned}$$

Therefore $\Psi_n(u)$ is a column vector with elements as follows:

$$\Psi_n(u) = \left(\frac{-1}{\sigma}\right)\,\phi(\Phi^{-1}(u))\begin{pmatrix}1\\ \bar{X_1}\\ \bar{X_2}\\ .\\ \bar{X_k}\\ \Phi^{-1}(u)\end{pmatrix} = \left(\frac{-1}{\sigma}\right)\,\phi(\Phi^{-1}(u))\begin{pmatrix}1\\ \bar{\mathbf{X}}\\ \Phi^{-1}(u)\end{pmatrix}.$$

We start with the calculation of the term $\Psi_n^T(t)I^{-1}(\theta)\Psi_n(s)$ in the covariance function $\rho(s,t)$. Using the Fisher information matrix obtained in 3.26 we get:

$$(\frac{-1}{\sigma})\phi(\Phi^{-1}(t))\begin{pmatrix}1 & \bar{X}^T & \Phi^{-1}(t)\end{pmatrix}\sigma^2\begin{pmatrix}1 & \bar{X}^T & 0\\ \bar{X} & \frac{1}{n}\sum_{i=1}^{n}x_ix_i^T & 0\\ 0 & 0 & 2\end{pmatrix}^{-1}(\frac{-1}{\sigma})\phi(\Phi^{-1}(s))\begin{pmatrix}1\\ \bar{X}\\ \Phi^{-1}(s)\end{pmatrix}$$

$$= \phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s))\Big[\frac{1}{2}\Phi^{-1}(t)\Phi^{-1}(s) + \begin{pmatrix}1 & \bar{X}^T\end{pmatrix}\begin{pmatrix}1 & \bar{X}^T\\ \bar{X} & \frac{1}{n}\sum_{i=1}^{n}x_ix_i^T\end{pmatrix}^{-1}\begin{pmatrix}1\\ \bar{X}\end{pmatrix}\Big].$$

It is easy to verify that:

$$\begin{pmatrix} 1 & \bar{X}^T \\ \bar{X} & \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \bar{X} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

which results in :

$$\begin{aligned} \Psi_n^T(t) I^{-1}(\theta) \Psi_n(s) &= \phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s)) \Big[\frac{1}{2}\Phi^{-1}(t)\Phi^{-1}(s) + \begin{pmatrix} 1 & \bar{X}^T \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Big] \\ &= \phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s)) \Big[\frac{1}{2}\Phi^{-1}(t)\Phi^{-1}(s) + 1 \Big] \\ &= \frac{1}{2}\phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s))\Phi^{-1}(t)\Phi^{-1}(s) + \phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s)). \end{aligned}$$

This gives the covariance function for the case of linear model as:

$$\rho(s,t) = \min(s,t) - st - \frac{1}{2}\phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s))\Phi^{-1}(t)\Phi^{-1}(s) + \phi(\Phi^{-1}(t))\phi(\Phi^{-1}(s)).$$

Note that in a linear model with an intercept, the covariance function only depends on the values of $0 \leq s, t \leq 1$; it does not depend on the values of explanatory variables. Instead of computing the limiting covariance analytically, we can estimate it using a similar approach to that we used in Section 3.6 and Section 3.7. In the next section, we will review the theory and steps for applying goodness-of-fit tests based on empirical distribution functions in a generalized linear model with a Gamma-distributed response variable.

The arguments in this section have involved non-identically distribute summands in the process whose limiting distribution must be computed. In [54] it is shown that all the remainder terms discarded along the way are uniformly negligible for any sequence of full rank designs with an intercept term.

3.10 Example 4: Generalized linear model

Generalized linear models (GLMs) are also used to model the relationship between a response variable and one or more covariates. GLMs extend the linear regression model by relaxing the assumption of normally distributed errors and allowing for a broader range of response distributions. In a GLM, the expected value of the response variable is related to a linear combination of the covariates through a link function. Specifically, for each observation $i = 1, 2, 3, \cdots, n$, we have $E(Y_i) = \mu_i$, where Y_i represents the response variable, and μ_i is the expected value, which is modeled as a function of the covariates. In a GLM, the relationship between the mean μ_i and the linear predictor η_i is captured by $g(\cdot)$, a link function. Specifically, we assume that $g(\mu_i) = \eta_i$ where $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$, and $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients to be estimated, and $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the covariates for observation i.

The choice of link function allows us to model flexibly the relationship between μ_i and η_i. For example, in a linear regression model, the function $g(\cdot)$ connects η_i and μ_i through an identity link, $g(x) = x$. By specifying an appropriate link function and response distribution, GLMs can handle a wide range of data types and non-linear relationship between the response variable and covariates [55].

This section reviews the theory required to apply a goodness-of-fit test based on the empirical distribution function for checking the model assumption regarding the distribution of the response variable in a GLM. To show the required theory, we specifically consider a GLM with a log link function. However, the concept remains the same for any link function. Consider a GLM with Y_i as the response variable, where $i = 1, 2, 3, \cdots, n$. In this model, the expected value of Y_i, denoted as $E(Y_i)$, depends on certain covariates. Specifically, we express $E(Y_i)$ as μ_i, where $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$. To simplify the notation, we can rewrite this equation as $\mu_i = \exp(x_i^T \beta)$, where x_i^T represents a row vector with p elements, and β represents the column vector of coefficients with p elements. We assume Y_i's are drawn from a population with a Gamma distribution with the following probability distribution function (PDF):

$$f(Y_i; \alpha, \theta_i) = \frac{1}{\theta_i \Gamma(\alpha)} \left(\frac{Y_i}{\theta_i}\right)^{\alpha-1} e^{\frac{-Y_i}{\theta_i}}$$

where α is the shape and θ_i is the scale parameter. Since $E(Y_i) = \mu_i = \alpha\theta_i$, we can substitute $\theta_i = \frac{\mu_i}{\alpha}$ in the PDF as follows:

$$f(Y_i; \alpha, \beta) = \frac{1}{\frac{\mu_i}{\alpha}\Gamma(\alpha)} \left(\frac{Y_i}{\frac{\mu_i}{\alpha}}\right)^{\alpha-1} e^{-\frac{\alpha}{\mu_i}Y_i} = \frac{\alpha}{\Gamma(\alpha)} \frac{1}{\mu_i} \alpha^{\alpha-1} \left(\frac{Y_i}{\mu_i}\right)^{\alpha-1} e^{-\frac{\alpha}{\mu_i}Y_i}.$$

Note that in this particular example, the log link function results in $\ln(\mu_i) = x_i^T \beta$, or alternatively, $\mu_i = \exp(x_i^T \beta)$. In addition, the PDF is now a function of α and β. We can now write the likelihood function for the sample as follows:

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} f(Y_i; \alpha, \beta) = \prod_{i=1}^{n} \frac{\alpha}{\Gamma(\alpha)} \frac{1}{\mu_i} \alpha^{\alpha-1} \left(\frac{Y_i}{\mu_i}\right)^{\alpha-1} e^{-\frac{\alpha}{\mu_i}Y_i} \\ &= \left(\frac{\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} \frac{1}{\mu_i}\right) \alpha^{n(\alpha-1)} \prod_{i=1}^{n} \left(\frac{Y_i}{\mu_i}\right)^{\alpha-1} e^{-\alpha \sum_{i=1}^{n} \frac{Y_i}{\mu_i}} \end{aligned}$$

and log-likelihood as:

$$l(\alpha, \beta) = n \ln(\frac{\alpha}{\Gamma(\alpha)}) + \sum_{i=1}^{n} \ln(\frac{1}{\mu_i}) + n(\alpha - 1)\ln(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \ln(\frac{Y_i}{\mu_i}) - \alpha \sum_{i=1}^{n} \frac{Y_i}{\mu_i}$$

$$= n\ln(\alpha) - n\ln(\Gamma(\alpha)) - \sum_{i=1}^{n}\ln(\mu_i) + n(\alpha-1)\ln(\alpha) + (\alpha-1)\sum_{i=1}^{n}\ln(Y_i)$$
$$- (\alpha-1)\sum_{i=1}^{n}\ln(\mu_i) - \alpha\sum_{i=1}^{n}\frac{Y_i}{\mu_i}.$$

To calculate the Fisher information matrix, we derive the partial derivatives of the log-likelihood function with respect to parameters as follows:

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \frac{n}{\alpha} - nD(\alpha) + n\left(\ln(\alpha) + (\alpha-1)\frac{1}{\alpha}\right) + \sum_{i=1}^{n}\ln(Y_i) - \sum_{i=1}^{n}\ln(\mu_i) - \sum_{i=1}^{n}\frac{Y_i}{\mu_i} \\
&= \frac{n}{\alpha} - nD(\alpha) + n\ln(\alpha) + n - \frac{n}{\alpha} + \sum_{i=1}^{n}\ln(Y_i) - \sum_{i=1}^{n}\ln(\mu_i) - \sum_{i=1}^{n}\frac{Y_i}{\mu_i} \\
&= -nD(\alpha) + n\ln(\alpha) + n + \sum_{i=1}^{n}\ln(Y_i) - \sum_{i=1}^{n}\ln(\mu_i) - \sum_{i=1}^{n}\frac{Y_i}{\mu_i} \\
\frac{\partial^2 l}{\partial \alpha^2} &= -nD^{'}(\alpha) + \frac{n}{\alpha}
\end{aligned}$$

where $D(\alpha) = \frac{\Gamma^{'}(\alpha)}{\Gamma(\alpha)}$. The partial derivatives with respect to β is:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= -\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\ln(\mu_i) - (\alpha-1)\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\ln(\mu_i) - \alpha\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\left(\frac{y_i}{\mu_i}\right) \\
&= -\alpha\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\ln(\mu_i) - \alpha\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\left(\frac{y_i}{\mu_i}\right) \\
&= -\alpha\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\ln(\mu_i) - \alpha\sum_{i=1}^{n}y_i(-1)(\mu_i)^{-2}\frac{\partial\mu_i}{\partial\beta} \\
&= -\alpha\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\ln(\mu_i) + \alpha\sum_{i=1}^{n}\frac{y_i}{\mu_i}\frac{\frac{\partial\mu_i}{\partial\beta}}{\mu_i} \\
&= -\alpha\sum_{i=1}^{n}x_i^T + \alpha\sum_{i=1}^{n}\frac{y_i}{\mu_i}x_i^T \\
\frac{\partial^2 l}{\partial\beta^2} &= \alpha\sum_{i=1}^{n}y_i\ \frac{\partial}{\partial\beta}(\frac{1}{\mu_i})\ x_i^T = \alpha\sum_{i=1}^{n}y_i(\frac{-1}{\mu_i^2})\frac{\partial}{\partial\beta}\mu_i\ x_i^T \\
&= -\alpha\sum_{i=1}^{n}y_i\frac{1}{\mu_i}\frac{\frac{\partial}{\partial\beta}\mu_i}{\mu_i}x_i^T = -\alpha\sum_{i=1}^{n}\frac{y_i x_i x_i^T}{\mu_i}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l}{\partial\alpha\partial\beta} &= -\sum_{i=1}^{n}\frac{\partial}{\partial\beta}ln(\mu_i) - \sum_{i=1}^{n}y_i\frac{\partial}{\partial\beta}\left(\frac{1}{\mu_i}\right) = -\sum_{i=1}^{n}x_i^T - \sum_{i=1}^{n}(-1)y_i\mu_i^{-2}\frac{\partial}{\partial\beta}\mu_i \\
&= -\sum_{i=1}^{n}x_i^T + \sum_{i=1}^{n}\frac{y_i}{\mu_i}\frac{\frac{\partial\mu_i}{\partial\beta}}{\mu_i} = -\sum_{i=1}^{n}x_i^T + \sum_{i=1}^{n}\frac{y_i}{\mu_i}x_i^T.
\end{aligned}$$

Therefore the Fisher information matrix is:

$$\begin{aligned} I_n(\theta) &= -E\begin{pmatrix} \frac{\partial^2}{\partial\alpha^2}l(\alpha,\beta) & \frac{\partial}{\partial\alpha\partial\beta}l(\alpha,\beta) \\ \frac{\partial}{\partial\alpha\partial\beta}l(\alpha,\beta) & \frac{\partial^2}{\partial\beta^2}l(\alpha,\beta) \end{pmatrix} \\ &= -E\begin{pmatrix} -nD^{'}(\alpha)+\frac{n}{\alpha} & -\sum_{i=1}^n x_i^T+\sum_{i=1}^n \frac{y_i}{\mu_i}x_i^T \\ -\sum_{i=1}^n x_i^T+\sum_{i=1}^n \frac{y_i}{\mu_i}x_i^T & -\alpha\sum_{i=1}^n \frac{y_i x_i x_i^T}{\mu_i} \end{pmatrix} \\ &= \begin{pmatrix} nD^{'}(\alpha)-\frac{n}{\alpha} & 0 \\ 0 & \alpha\sum_{i=1}^n x_i x_i^T \end{pmatrix} \end{aligned}$$

and the Fisher information matrix based on one observation is:

$$I(\theta) = \frac{1}{n}I_n(\theta) = \begin{pmatrix} D^{'}(\alpha)-\frac{1}{\alpha} & 0 \\ 0 & \frac{\alpha}{n}\sum_{i=1}^n x_i x_i^T \end{pmatrix}.$$

In the GLM case, the covariance function of $\hat{W}_n(u)$ remains the same as in the i.i.d. case, except the $\Psi(s)$ vector values are averaged out, similar to the calculations we discussed to obtain the covariance function in formula 3.30. The covariance function is:

$$\rho(s,t) = \min(s,t) - st - \Psi_n^T(t)I^{-1}(\theta)\Psi_n(s). \tag{3.31}$$

The elements of $\Psi_n(u)$ are the partial derivative terms with respect to the parameters, evaluated at the inverse of the CDF, and then averaged. For example, in a GLM-Gamma model with k explanatory variables, there are $p = k+2$ parameters (with k for coefficients, one for the intercept, and one for the shape parameter α). Therefore, $\Psi_n(u)$ is a vector with p elements. The next step is to calculate these elements in the column vector. If the response variable, Y_i, follows a Gamma distribution, we can compute the cumulative distribution function as follows:

$$F_i(Y_i;\alpha,\beta) = \int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ \frac{t^{\alpha-1}}{\mu^\alpha}\ e^{-\left(\frac{\alpha t}{\mu}\right)}\ dt.$$

The partial derivative of $F_i(Y_i;\alpha,\beta)$ with respect to β is:

$$\begin{aligned} \frac{\partial}{\partial\beta}F_i(Y_i;\alpha,\beta) &= \frac{\partial}{\partial\beta}\left(\int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ \frac{t^{\alpha-1}}{\mu_i^\alpha}\ e^{-\left(\alpha\frac{t}{\mu_i}\right)}\ dt\right) = \int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ t^{\alpha-1}\frac{\partial}{\partial\beta}\left(\frac{e^{-\left(\alpha\frac{t}{\mu_i}\right)}}{\mu_i^\alpha}\right)dt \\ &= \int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ t^{\alpha-1}\left\{\frac{\mu_i^\alpha\frac{\partial}{\partial\beta}\left(e^{-\left(\alpha\frac{t}{\mu_i}\right)}\right) - e^{\frac{-\alpha t}{\mu_i}}\frac{\partial}{\partial\beta}(\mu_i^\alpha)}{\mu_i^{2\alpha}}\right\}dt \\ &= \int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ t^{\alpha-1}\left\{\frac{\left(\mu_i^\alpha\frac{\alpha t}{\mu_i^2}e^{-\frac{\alpha t}{\mu_i}} - e^{-\frac{\alpha t}{\mu_i}}\alpha\mu_i^{\alpha-1}\right)\frac{\partial}{\partial\beta}\mu_i}{\mu_i^{2\alpha}}\right\}dt \end{aligned}$$

$$
\begin{aligned}
&= \int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ t^{\alpha-1}\ \left\{\alpha t \mu_i^{\alpha-2-2\alpha}\ e^{-\frac{\alpha}{\mu_i}t}\ \left(\frac{\partial}{\partial\beta}\mu_i\right)\ dt\right\} \\
&= \int_0^{Y_i} \frac{\alpha^{\alpha+1}}{\Gamma(\alpha)}\ t^{\alpha}\ \mu_i^{-\alpha-2}\ e^{-\frac{\alpha}{\mu_i}t}\ \left(\frac{\partial}{\partial\beta}\mu_i\right)\ dt \\
&- \int_0^{Y_i} \frac{\alpha^{\alpha+1}}{\Gamma(\alpha)}\ t^{\alpha-1}\ \mu_i^{-\alpha-1}\ e^{-\frac{\alpha}{\mu_i}t}\ \left(\frac{\partial}{\partial\beta}\mu_i\right)\ dt \\
&= \int_0^{Y_i} \frac{\alpha}{\Gamma(\alpha)}\ \left(\frac{\alpha t}{\mu_i}\right)^{\alpha}\ e^{-\frac{\alpha}{\mu_i}t}\ \frac{1}{\mu_i}\ x_i^T\ dt \\
&- \int_0^{Y_i} \frac{\alpha^2}{\Gamma(\alpha)}\ \left(\frac{\alpha t}{\mu_i}\right)^{\alpha-1}\ \frac{1}{\mu_i}\ e^{-\frac{\alpha}{\mu_i}t}\ x_i^T\ dt.
\end{aligned}
$$

We apply the change of variable $v = \frac{\alpha t}{\mu_i}$ to continue:

$$
\begin{aligned}
\frac{\partial}{\partial\beta}F_i(Y_i;\alpha,\beta) &= \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{\alpha}{\Gamma(\alpha)}\ v^{\alpha}\ e^{-v}\ \frac{1}{\mu_i}\ x_i^T\ \frac{\mu_i}{\alpha}\ dv - \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{\alpha^2}{\Gamma(\alpha)}\ v^{\alpha-1}\ \frac{1}{\mu}\ e^{-v}\ x_i^T\ \frac{\mu_i}{\alpha}\ dv \\
&= \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)}\ v^{\alpha}\ e^{-v}\ x_i^T\ dv - \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{\alpha}{\Gamma(\alpha)}\ v^{\alpha-1}\ e^{-v}\ x_i^T\ dv \\
&= x_i^T \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha+1)}\ v^{(\alpha+1)-1}\ e^{-v}\ dv \\
&- \alpha x_i^T \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)}\ v^{\alpha-1}\ e^{-v} dv \\
&= \alpha x_i^T \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha+1)}\ v^{(\alpha+1)-1}\ e^{-v} dv \\
&- \alpha x_i^T \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)}\ v^{\alpha-1}\ e^{-v}\ dv.
\end{aligned}
$$

The partial derivative of $F_i(Y_i;\alpha,\beta)$ with respect to α is:

$$
\begin{aligned}
\frac{\partial}{\partial\alpha}F_i(Y_i;\alpha,\beta) &= \frac{\partial}{\partial\alpha}\ \int_0^{Y_i} \frac{\alpha^\alpha}{\Gamma(\alpha)}\ \frac{t^{\alpha-1}}{\mu_i^\alpha}\ e^{(-\alpha\frac{t}{\mu_i})}\ dt \\
&= \frac{\partial}{\partial\alpha}\ \int_0^{Y_i} \frac{1}{\Gamma(\alpha)}\ \left(\frac{\alpha t}{\mu_i}\right)^{\alpha}\ e^{(-\alpha\frac{t}{\mu_i})}\ t^{-1}\ dt \\
&= \frac{\partial}{\partial\alpha}\ \int_0^{Y_i} \frac{1}{\Gamma(\alpha)}\ e^{ln\left\{(\frac{\alpha t}{\mu_i})^\alpha\right\}}\ e^{-\alpha\frac{t}{\mu_i}}\ t^{-1}\ dt \\
&= \frac{\partial}{\partial\alpha} \int_0^{Y_i} \frac{1}{\Gamma(\alpha)} e^{-\alpha\left\{\frac{t}{\mu_i} - ln(\frac{\alpha t}{\mu_i})\right\}}\ t^{-1}\ dt \\
&= \int_0^{Y_i} \frac{\partial}{\partial\alpha}\left(\frac{1}{\Gamma(\alpha)} e^{-\alpha\left\{\frac{t}{\mu_i} - ln(\frac{\alpha t}{\mu_i})\right\}}\ t^{-1}\ dt\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \int_0^{Y_i} \frac{\partial}{\partial\alpha}\left(\frac{1}{\Gamma(\alpha)}\right) e^{-\alpha\left\{\frac{t}{\mu_i}-ln(\frac{\alpha t}{\mu_i})\right\}}\, dt \\
&+ \int_0^{Y_i} \frac{1}{\Gamma(\alpha)}\, \frac{\partial}{\partial\alpha}\left(e^{-\frac{\alpha t}{\mu_i}+\alpha ln(\frac{\alpha t}{\mu_i})}\right) t^{-1}\, dt \\
&= \int_0^{Y_i} \frac{-\Gamma'(\alpha)}{\Gamma(\alpha)}\, e^{-\alpha\left\{\frac{t}{\mu_i}-ln(\frac{\alpha t}{\mu_i})\right\}}\, t^{-1} dt \\
&+ \int_0^{Y_i} \frac{1}{\Gamma(\alpha)} \left(\frac{-t}{\mu_i} + ln(\frac{\alpha t}{\mu_i}) + 1\right) e^{-\alpha\left\{\frac{t}{\mu_i}-ln(\frac{\alpha t}{\mu_i})\right\}}\, dt.
\end{aligned}
$$

We apply the change of variable $v = \frac{\alpha t}{\mu_i}$ to continue:

$$
\begin{aligned}
\frac{\partial}{\partial\alpha} F_i(Y_i;\alpha,\beta) &= \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{-\Gamma'(\alpha)}{\Gamma(\alpha)}\, e^{-v}\, e^{\ln(v^\alpha)}\, \frac{\alpha}{\mu_i v}\, \frac{\mu_i}{\alpha}\, dv \\
&+ \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)} \left(\frac{-v}{\alpha} + \ln(v) + 1\right) e^{-v+\alpha\ln(v)}\, \frac{\mu_i}{\alpha}\, dv \\
&= \int_0^{\frac{\alpha Y_i}{\mu_i}} -D(\alpha)\, e^{-v}\, v^{\alpha-1}\, dv + \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)} \left(\frac{-v}{\alpha} + \ln(v) + 1\right) e^{-v}\, v^\alpha\, dv \\
&= \int_0^{\frac{\alpha Y_i}{\mu_i}} -D(\alpha)\, e^{-v}\, v^{\alpha-1}\, dv - \frac{1}{\alpha}\int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)} v^{\alpha-1}\, e^{-v}\, dv \\
&+ \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)} \ln(v)\, v^\alpha\, e^{-v}\, dv + \int_0^{\frac{\alpha Y_i}{\mu_i}} \frac{1}{\Gamma(\alpha)}\, v^\alpha\, e^{-v}\, dv.
\end{aligned}
$$

We evaluate both terms at $Y_i = Q(u;\alpha,\mu_i) = \frac{\mu_i}{\alpha} Q(u;\alpha,1)$ where Q is the inverse of cumulative distribution function of a Gamma distribution with shape parameter of α and scale parameter of one. Therefore we can write the elements of the vector $\Psi_n(u)$ as the average over i of:

$$
\begin{aligned}
\frac{\partial}{\partial\beta} F_i(Y_i;\alpha,\beta)\big|_{Y_i=\frac{\mu_i}{\alpha}Q(u;\alpha,1)} &= \alpha x_i^T \int_0^{Q(u;\alpha,1)} \frac{1}{\Gamma(\alpha+1)}\, v^\alpha\, e^{-v}\, dv \\
&- \alpha x_i^T \int_0^{Q(u;\alpha,1)} \frac{1}{\Gamma(\alpha)} v^{\alpha-1}\, e^{-v}\, dv
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial\alpha} F_i(Y_i;\alpha,\beta)\big|_{Y_i=\frac{\mu_i}{\alpha}Q(u;\alpha,1)} &= \int_0^{Q(u;\alpha,1)} -D(\alpha)\, e^{-v}\, v^{\alpha-1}\, dv \\
&- \frac{1}{\alpha}\int_0^{Q(u;\alpha,1)} \frac{1}{\Gamma(\alpha)} v^{\alpha-1}\, e^{-v}\, dv \\
&+ \int_0^{Q(u;\alpha,1)} \frac{1}{\Gamma(\alpha)} \ln(v)\, v^\alpha\, e^{-v}\, dv \\
&+ \int_0^{Q(u;\alpha,1)} \frac{1}{\Gamma(\alpha)}\, v^\alpha\, e^{-v}\, dv.
\end{aligned}
$$

The vector $\Psi_n(u)$ has p elements, denoted by $\Psi_{n,j}(u)$, and can be computed as follows. For j values ranging from 2 to p, $\Psi_n(u)$ includes the following values:

$$\Psi_{n,j}(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta_{j-1}} F_i(Y_i; \alpha, \beta)$$

and for $j = 1$:

$$\Psi_j(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \alpha} F_i(Y_i; \alpha, \beta).$$

It is possible to compute both elements of the $\Psi_n(u)$ vector numerically for a given sample. For example, the *digamma* function from the *MASS* package in the R statistical software can be used to calculate $D(\alpha)$. The maximum likelihood estimate of α can be computed using a procedure to find the root of the score function with any desired precision. Finally, a simple routine, in combination with the *qgamma* function from the *stats* package in R, can be applied to compute $\int_0^{Q(u;\alpha,1)} \ln(z) g(z; \alpha, 1) dz$ accurately. Thus the approximating covariance function of the process $\hat{W}_n(u)$ will be:

$$\begin{aligned}
\rho(s,t) &= \min(s,t) - st - \Psi_n^{'}(t) I^{-1}(\theta) \Psi_n(s) \\
&= \min(s,t) - st - \Psi_n^{'}(t) \begin{pmatrix} D^{'}(\alpha) - \frac{1}{\alpha} & 0 \\ 0 & \frac{\alpha}{n}\sum_{i=1}^{n} x_i x_i^T \end{pmatrix}^{-1} \Psi_n(s).
\end{aligned}$$

This covariance function depends not only on s and t but also on the shape parameter α. As described in Section 3.4, one can estimate the eigenvalues and compute the P-value. Instead of computing the limiting covariance analytically, we can estimate it with a similar approach to what we show in Section 3.6 and Section 3.7.

3.11 Concluding remarks

Throughout the previous sections of this chapter, we have discussed various examples, including i.i.d. cases, linear models, and generalized linear models. We have covered the underlying theory necessary for computing the covariance function of the $\hat{W}_n(u)$ process analytically, in each of these examples. We started our exploration with a simple i.i.d. case and gradually considered more complex models. As will be evident to the reader, the complexity of obtaining the essential components within the covariance function increases notably as we transition towards working with general likelihood models.

In Section 3.5, we introduced our main contribution: an alternative approach to estimating the covariance function. Instead of relying on theoretical calculations, we presented a method for directly estimating the covariance function from sample data. This method replaces the need for complex computations and opens the door to more practical applica-

tions. In the upcoming chapter, we will investigate the reliability of this estimation through various simulations, where we will apply this estimation method to derive the covariance function, estimate the eigenvalues, and compute the P-value.

Chapter 4

Simulation results and real data

4.1 Overview

In this Chapter, we conduct four large-scale simulations in which we estimate the covariance function directly from the sample using the method described in Section 3.5, rather than relying on analytical calculations of the covariance function. In each simulation, the data is randomly generated from a model under the null hypothesis which we will explain. We use the simulations to evaluate the quality of the asymptotic approximations and to guide specific choices for estimation methods and for the approximate quadrature method.

For the first and second simulations, presented in Sections 4.2 and 4.3, we consider generating data from univariate Normal and Gamma distributions, respectively. In the third simulation, presented in Section 4.4, we briefly examine a linear regression model where the error terms follow a Normal distribution. In the last simulation, presented in Section 4.5, we address a common problem in generalized linear models (GLM) where the responses follow a Gamma distribution, and the link function is either the logarithm or the inverse of the linear predictor.

Throughout the chapter we use the Cramér-von Mises goodness-of-fit test and compute P-values. Under the null hypothesis, we expect P-values to follow a uniform distribution. We investigate this by examining the quantile-quantile plot (QQ-plot) of P-values to detect any discrepancy between quantiles of computed P-values and the theoretical quantiles of the uniform distribution. Additionally, we are interested to check if the distribution of small P-values is close to uniform, since the common significance level for statistical testing is 0.01 or 0.05 most of the times. For this, we assess the uniformity of P-values less than or equal to 0.10 by examining their QQ-plot.

The estimated type-one error rate is also computed to check the rate at which the test rejects a correct null hypothesis. This is estimated by computing the proportion of tests that are rejected at 0.01 or 0.05 significant levels.

We investigate the effect of sample size and suggest a minimum sample size in each simulation setting. To approximate the eigenvalues of the covariance function using matrix

Q (defined in Section 3.5), we employed two different approaches. We either used a grid of n equally spaced points over the interval [0,1] (i.e $m = n$ points), or we utilized the probability integral transformed values. We also considered $m = 2n$ and $m = 3n$ equally spaced points over the interval [0,1], but we did not observe any significant differences compared to using $m = n$ data points. Therefore we are not presenting those results here. We use uniform weights w_j throughout but a brief discussion of alternatives is presented in the Conclusion of the chapter in Section 4.7.

For the Normal and Gamma simulations, we run four sub-simulations in each case. We consider two estimates of the Fisher information matrix: calculating the sample variance of the score function or computing the negative of the Hessian matrix evaluated at the MLE. Those results led us to estimate the Fisher information matrix based on the sample variance of the score in the simulations with covariates.

After the simulations, Section 4.6 presents the results of our proposed goodness-of-fit test in a real-world example where the response variable is assumed to follow a Gamma distribution. We finish the chapter with Section 4.7 which presents concluding remarks and some ideas for future research.

4.2 Normal distribution

We show the results of four large scale simulations in the case of an i.i.d sample from the Normal distribution. In all of these simulations, a random sample of size n was generated from $N(0, 1)$ distribution. We only considered a standard normal distribution, i.e $\mu = 0$ and $\sigma = 1$. This setting is sufficient because if Z_i follows a standard normal distribution, then $Y_i = \mu + \sigma Z_i$ follows a Normal distribution with a mean of μ and a standard deviation of σ. It is easy to verify that $\Phi\left(\frac{Y_i - \bar{Y}}{s_Y}\right) = \Phi\left(\frac{Z_i - \bar{Z}}{s_Z}\right)$ and as a result, the probability integral transformed (PIT) values are the same, and the CvM statistic remains unchanged. Additionally, in Section 3.6, we demonstrated that the limiting covariance function for the Normal distribution does not depend on the choice of μ and σ.

We varied the sample size to assess its effect by including values of $n = 50$, $n = 100$, and $n = 250$. To calculate the matrix Q, we either divided the interval [0,1] into $m = n$ equally spaced points depending on the value of n or considered the probability integral transformed values, as described in the methods section. The Fisher information matrix was estimated either by the variance of the score function (estimated from the sample) or by the negative of the observed Hessian matrix evaluated at the MLE. In each simulation setting, 10,000 Monte Carlo samples were simulated.

The four sub-simulations in this section are as follows. The Fisher information matrix is estimated by the variance of the score in simulation 1 and simulation 3, and in simulation 2 and simulation 4 is estimated by the negative of the observed Hessian matrix evaluated at the MLE. In simulation 1 and simulation 2, we used $m = n$ equally spaced data points

over interval [0,1] to compute the Q matrix and estimate the eigenvalues (we show the computation of matrix Q in Section 3.5) while simulation 3 and simulation 4, used probability integral transformed values as a grid over interval [0,1]. In all these simulations we used uniform weights $1/n$.

4.2.1 Simulation 1

In this simulation, we estimate the Fisher information matrix using the sample variance of the score function. To compute the matrix Q and estimate eigenvalues, we used $m = n$ equally spaced points over [0,1] grid as described in Section 3.5. Quadrature weights were $w_i = 1/n$. Figure 4.1 shows the Quantile-Quantile plot or Q-Q plot of the P-values obtained from our goodness-of-fit test in each simulation setting. The x and y-axes in all panels are the theoretical quantiles of the Uniform distribution and quantiles of the P-value obtained from the sample, respectively. The figure is arranged in a 1×3 gird, where each column presents a sample size. In other words, the results related to sample sizes of $n = 50$, $n = 100$, and $n = 250$ are arranged in first, second, and third column, respectively. For sample size 50 there is some departure from the straight line but it is still acceptable considering the small sample size. Also we can see that the sample quantiles correspond very well with the corresponding theoretical values for sample sizes of 100 and 250.

It is of interest to investigate how well theoretical and sample quantiles correspond not only for all P-values but also focusing on smaller P-values. Figure 4.2 shows the Q-Q plot of P-values less than or equal to 0.10 (multiplied by 10 to run from 0 to 1). The x and y-axes in all panels are the theoretical quantiles of the Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as Figure 4.1. As we can see, regardless of the sample size, in all panels, the theoretical values correspond very well with sample quantiles.

We estimate the type-one error rate of the test at two nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$ in Table 4.1, for each sample size. The estimated rate is the proportion of all P-values among 10000 Monte Carlo samples that are less than the nominal level. We can see that the estimated type-one error rate is well controlled at level 0.01 for all simulation settings. It is clear that increasing the sample size from $n = 50$ to $n = 250$ results in a better controlled type one error rate. The estimated type-one error rate is inflated at level $\alpha = 0.05$ for sample sizes of $n = 50$ and $n = 100$. However, increasing sample size to $n = 250$ controls the type-one error rate at the desired level.

Figure 4.1: Normal distribution, simulation 1, using the variance of the score and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, and $n = 250$ respectively.

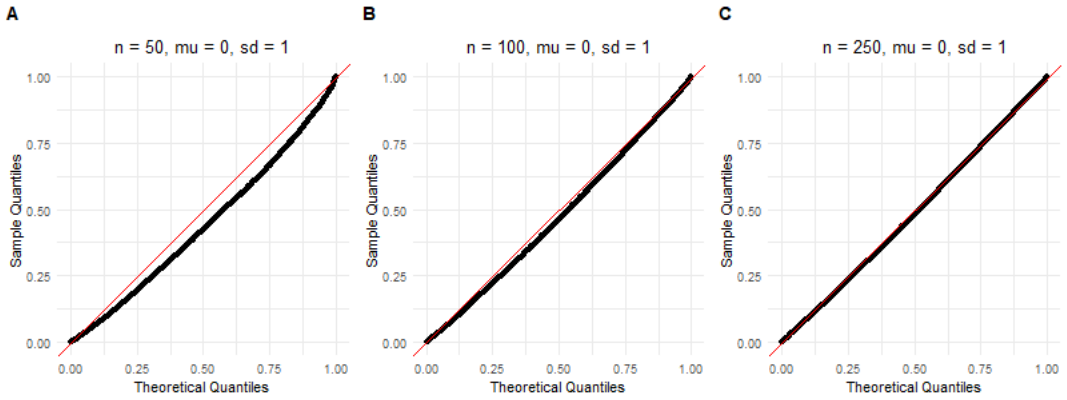


Figure 4.2: Normal distribution, simulation 1, using the variance of the score and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, and $n = 250$, respectively.

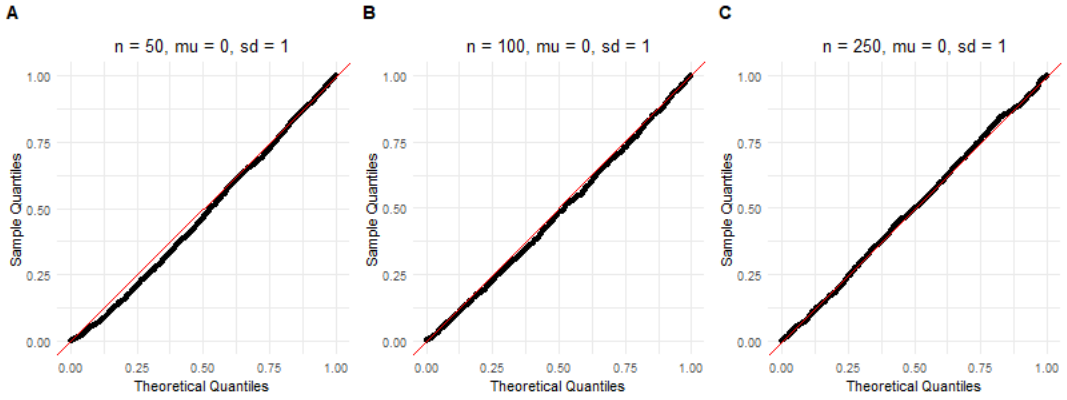


Table 4.1: Normal distribution, simulation 1, using the variance of the score and an evenly spaced grid. The estimated type one error rate at level $\alpha = 0.01$ and level $\alpha = 0.05$. Rows are the level of the test and columns are the sample sizes.

	Sample size		
	n = 50	n = 100	n = 250
$\alpha = 0.01$	0.0199	0.0136	0.0122
$\alpha = 0.05$	0.0745	0.0635	0.0553

4.2.2 Simulation 2

In this simulation, we estimate the Fisher information matrix by the negative of the observed Hessian matrix, where the required parameters are estimated from the data. To compute the matrix Q and estimate eigenvalues, we used $m = n$ equally spaced points over [0,1] grid as described in Section 3.5. Figure 4.3 shows the Q-Q plot of all P-values and Figure 4.4 presents the Q-Q plot for P-values that are less than or equal to 0.10.

Looking at the Q-Q plot of all P-values, it seems that estimating Fisher information with the observed Hessian matrix rather than the variance of score, produces unacceptably conservative P-values. This can be visually verified in all panels for any sample size that we considered in Figure 4.3. The Q-Q plot of P-values less than or equal to 0.10 in Figure 4.4 suggests that the sample quantiles are larger than the expected theoretical values.

The estimated type-one error rate at two nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$ are given in table 4.2. It is worth noting that by estimating the Fisher information matrix with Hessian matrix, the estimated type one error rate is well controlled in both nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$ but the test is too conservative.

Figure 4.3: Normal distribution, simulation 2, using the negative Hessian and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample sizes $n = 50$, $n = 100$, and $n = 250$, respectively.

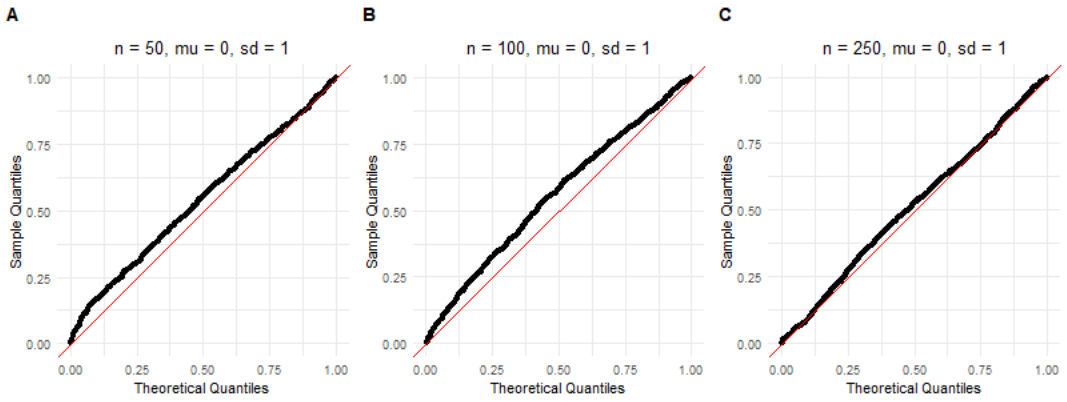


Figure 4.4: Normal distribution, simulation 2, using the negative Hessian and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample sizes $n = 50$, $n = 100$, and $n = 250$, respectively.

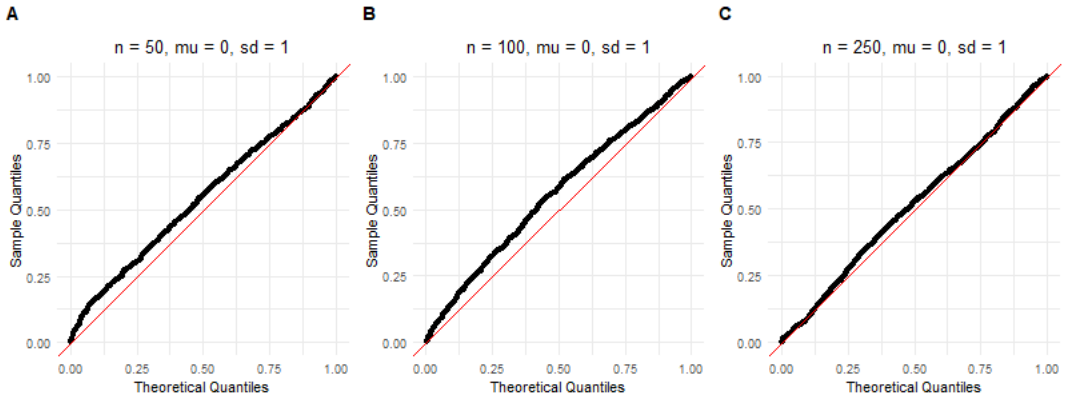


Table 4.2: Normal distribution, simulation 2, using the negative Hessian and an evenly spaced grid. The estimated type one error rate at level $\alpha = 0.01$ and level $\alpha = 0.05$. Rows are the level of the test and columns are the sample sizes.

	Sample size		
	n = 50	n = 100	n = 250
$\alpha = 0.01$	0.0037	0.0055	0.0114
$\alpha = 0.05$	0.0363	0.0367	0.0517

4.2.3 Simulation 3

In this simulation, we estimate the Fisher information matrix by the variance of the score function. In order to compute the matrix Q and estimate eigenvalues, we used the probability integral transformed values for grid, as described in Section 3.5. Figure 4.5 shows the Q-Q plot of all P-values and Figure 4.6 shows the Q-Q plot of P-values less than or equal to 0.10 only. It is clear from both plots that estimating the covariance function of the empirical process by probability integral transformed values results in a more uniform distribution of P-values under the null hypothesis. We estimate the type-one error rate at two nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$. These estimates are given in table 4.3. As we can see, the estimated type one error rate is well controlled at nominal level of $\alpha = 0.01$. The estimated type-one error rate at level $\alpha = 0.05$ is a bit higher than 0.05 but not importantly so.

Figure 4.5: Normal distribution, simulation 3, using the variance of the score and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, and $n = 100$, respectively.

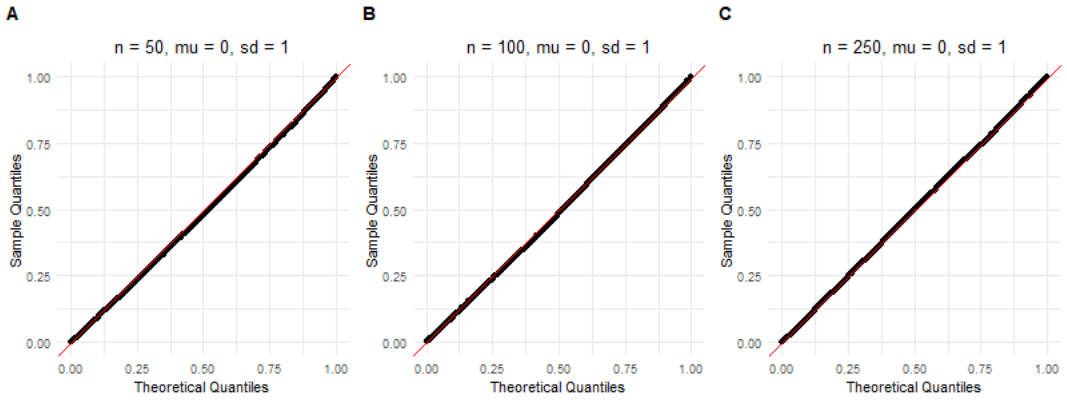


Figure 4.6: Normal distribution, simulation 3, using the variance of the score and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, and $n = 250$, respectively.

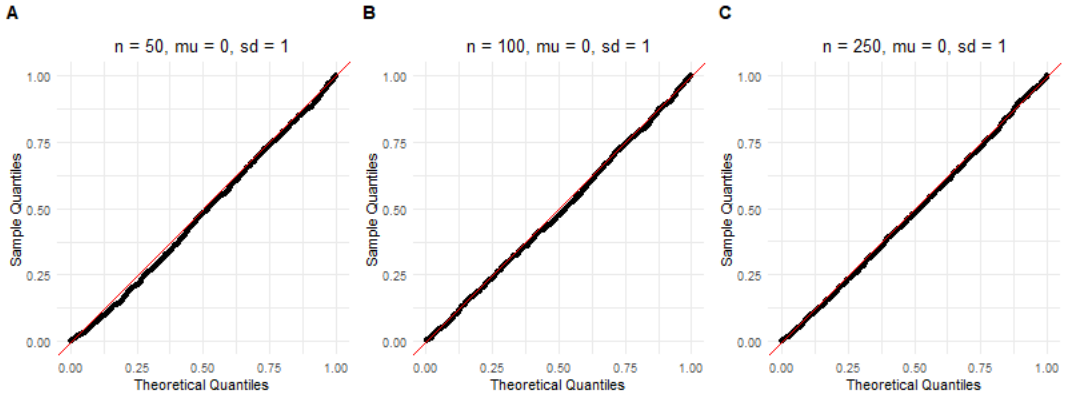


Table 4.3: Normal distribution, simulation 3, using the variance of the score and the PITs for the grid. The estimated type one error rate at level $\alpha = 0.01$ and level $\alpha = 0.05$. Rows are the level of the test and columns are the sample sizes.

	Sample size		
	n = 50	n = 100	n = 250
$\alpha = 0.01$	0.0147	0.0117	0.0119
$\alpha = 0.05$	0.0570	0.0560	0.0548

4.2.4 Simulation 4

In this simulation, we estimate the Fisher information matrix by the negative observed Hessian matrix evaluated at the maximum likelihood estimates. To compute the matrix Q and estimate eigenvalues, we used the probability integral transformed values, as described in Section 3.5. Figure 4.7 shows the Q-Q plot of all P-values and Figure 4.8 shows the Q-Q plot of P-values less than or equal to 0.10 only. For sample size of $n = 50$, there is a visible departure from the straight line. But we can see that increasing the sample size help resolving this issue. The estimated type-one error rate at two nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$ are given in table 4.4. The estimated type one error rate is well controlled at nominal level of $\alpha = 0.01$ and $\alpha = 0.05$. However, the results of the test are too conservative.

Figure 4.7: Normal distribution, simulation 4, using the negative Hessian and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, and $n = 250$, respectively.

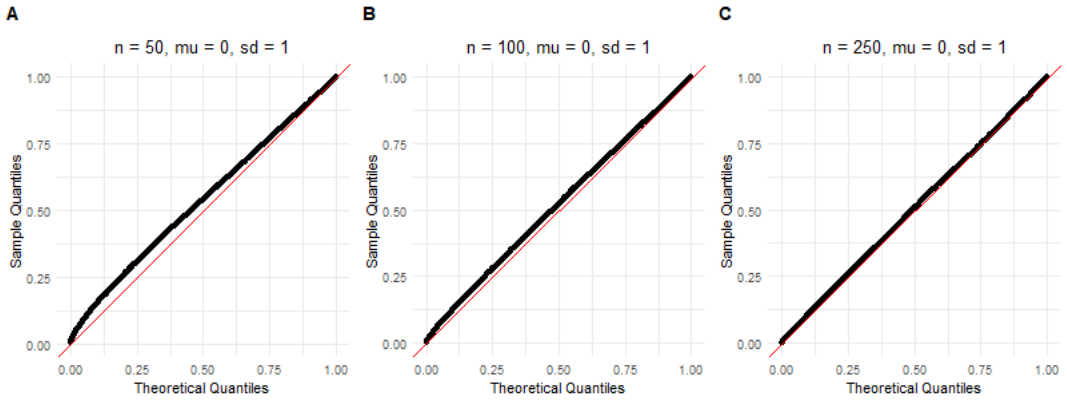


Figure 4.8: Normal distribution, simulation 4, using the negative Hessian and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different sample sizes in each panel: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, and $n = 250$, respectively.

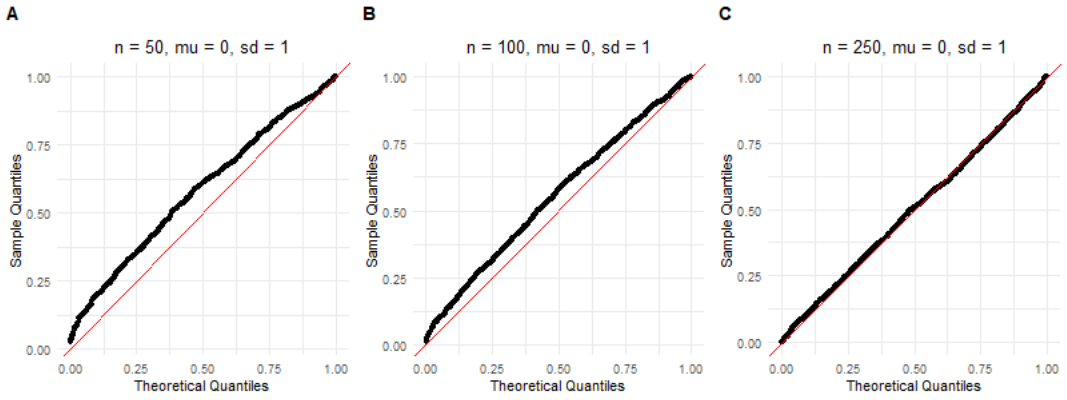


Table 4.4: Normal distribution, simulation 4, using the negative Hessian and the PITs for the grid. The estimated type one error rate at level $\alpha = 0.01$ and level $\alpha = 0.05$. Rows are the level of the test and columns are the sample sizes.

	Sample size		
	n = 50	n = 100	n = 250
$\alpha = 0.01$	0.0018	0.0041	0.0085
$\alpha = 0.05$	0.0222	0.0325	0.0458

The results of the simulation in the Normal case suggests the following. We recommend to use the PIT values as a grid over interval [0,1] with quadrature weights of $w_i = \frac{1}{n}$. In addition, our findings suggest that estimating the Fisher information matrix by the variance of score is more reliable. The test demonstrate an excellent performance for sample sizes as low as $n = 50$ at both significance levels of $\alpha = 0.01$ and $\alpha = 0.05$.

4.3 Gamma distribution

In this section, we present the results of four large scale simulations in the case of the Gamma distribution. In all these simulations, a random sample of size n is generated from Gamma$(\alpha, 1)$ distribution for a wide range of values for the shape parameters, including $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$. As shown in 3.7, the limiting covariance function does not depend on the choice for the scale parameter, as a result we set $\lambda = 1$. We consider different values of $n = 50$, $n = 100$, and $n = 250$ for the sample size. To compute the matrix Q and estimate the eigenvalues, we either divided the interval [0,1] into $m = n$ equally spaced data points,

depending on the value of n, or considered the probability integral transformed values as described in Section 3.5 section to estimate the covariance function of the empirical process. The Fisher information matrix was estimated either by the variance of the score function or by the negative value of the observed Hessian matrix. In each simulation setting, 10,000 Monte Carlo samples were simulated.

The four sub-simulations in this section follow the same pattern as in the Normal case. The Fisher information matrix, is estimated by the variance of score in simulation 1 and simulation 3, and in simulation 2 and simulation 4 is estimated by the negative of the observed Hessian matrix evaluated at MLE. In simulation 1 and simulation 2, we used $m = n$ equally spaced data points over interval [0,1] to compute the Q matrix and estimate the eigenvalues (we show the computation of matrix Q in Section 3.5) while simulation 3 and simulation 4, used probability integral transformed values as a grid over interval [0,1]. In all these simulations we used uniform weights $1/n$.

4.3.1 Simulation 1

In this simulation, the Fisher information matrix is estimated by the variance of the score function. We used a grid with $m = n$ equally spaced data points over [0,1] interval to compute the matrix Q and estimate the eigenvalues, as described in Section 3.5. Figure 4.9 shows the Q-Q plot of the obtained P-value from goodness-of-fit test in each simulation setting. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the computed P-value from the test, respectively. The figure is arranged so that each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter ($\alpha = 1$, $\alpha = 7$, and $\alpha = 50$) are arranged in first, second, and third columns, respectively. There are some visible discrepancies between theoretical and sample quantiles when $n = 50$ but a sample of size $n = 100$ seems to resolve the issue. In general, we see that the sample quantiles matches well with the corresponding theoretical values as sample size increases from $n = 50$ to $n = 250$.

We also look at Q-Q plots of smaller P-values to get a better insight into the performance of the method. Figure 4.10 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the same as 4.9 and panels are arranged in the same way. As we can see, theoretical and sample quantiles matches very well for all parameter settings. The estimated type one error rate at nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$ are given in table 4.5 and 4.6, respectively. At level $\alpha = 0.05$, it seems that the type one error rate is slightly inflated specially for small sample sizes. Increasing the sample size helps control the type one error rate at the desired level.

Figure 4.9: Gamma distribution, simulation 1, using the variance of the score and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

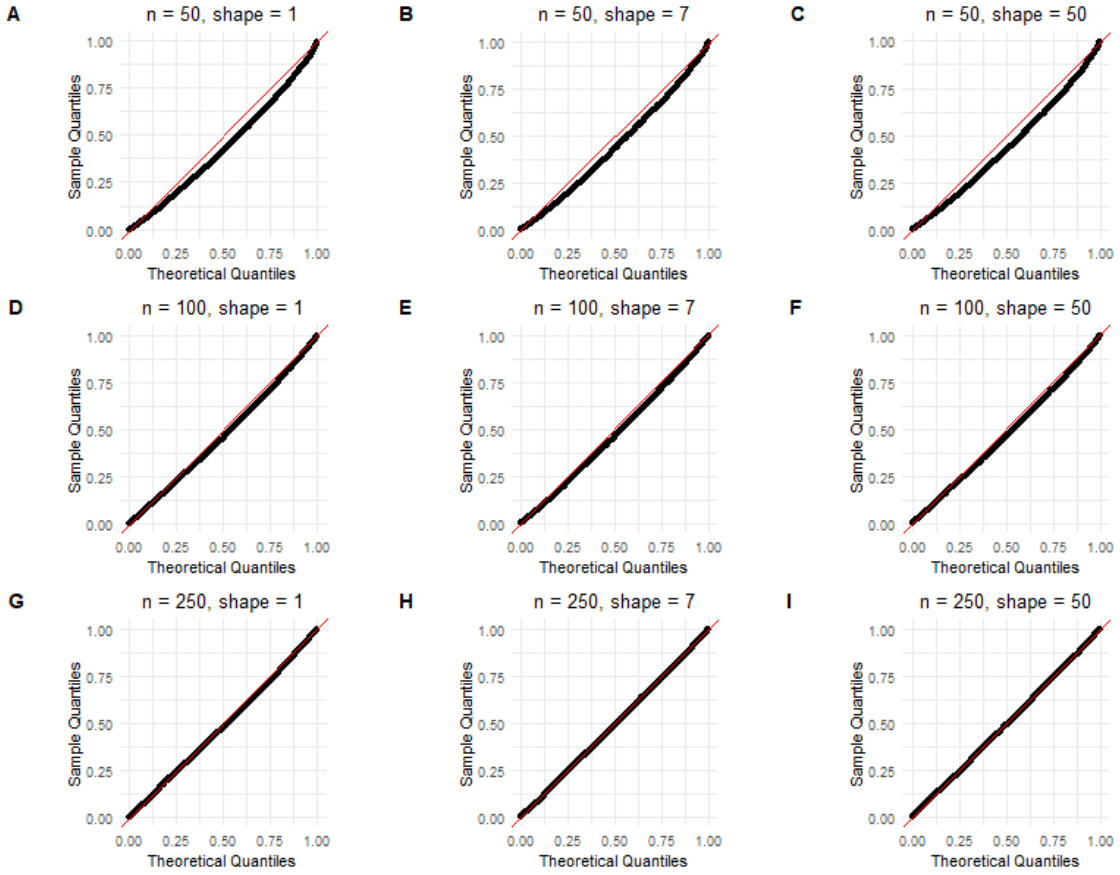


Figure 4.10: Gamma distribution, simulation 1, using the variance of the score and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

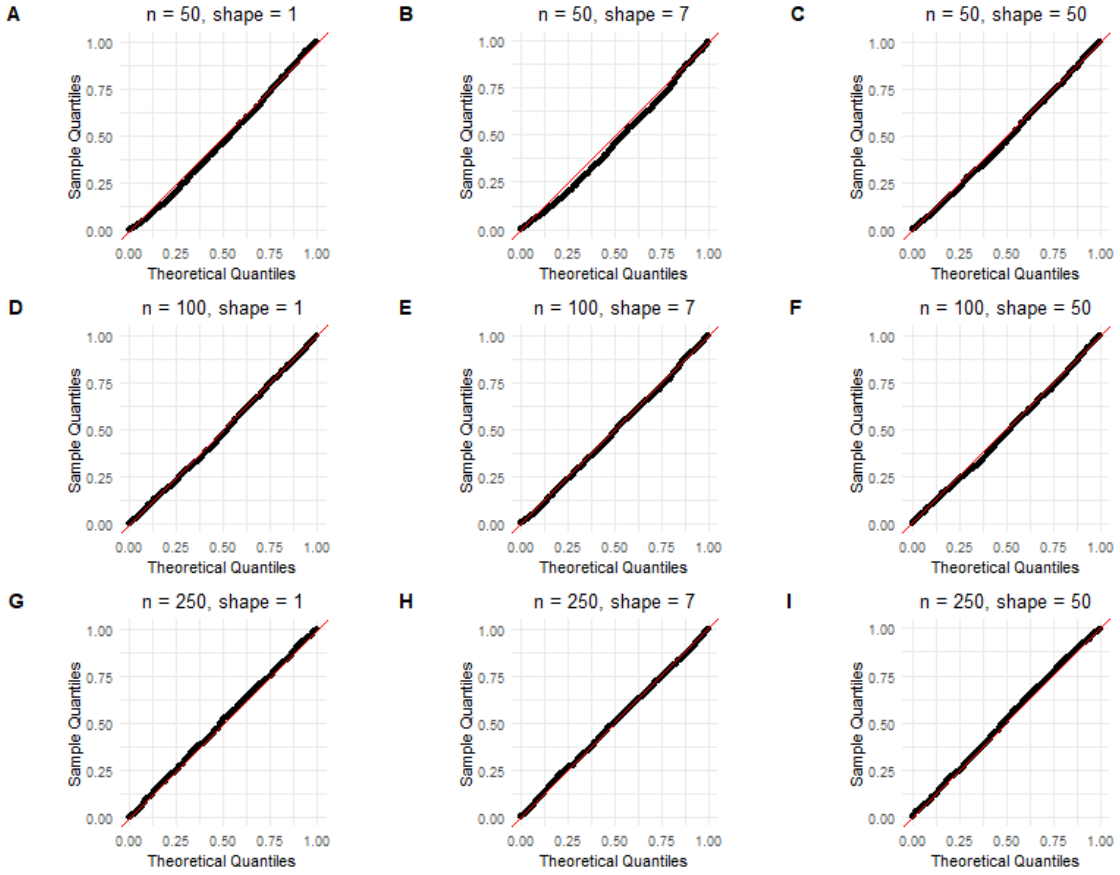


Table 4.5: Gamma distribution, simulation 1, using the variance of the score and an evenly spaced grid. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0195	0.0192	0.0175
n = 100	0.0124	0.0147	0.0123
n = 250	0.0112	0.0108	0.0107

Table 4.6: Gamma distribution, simulation 1, using the variance of the score and an evenly spaced grid. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0783	0.0764	0.0751
n = 100	0.0605	0.0624	0.0601
n = 250	0.0553	0.0557	0.0492

4.3.2 Simulation 2

In this simulation, the Fisher information matrix is estimated by the negative of the observed Hessian matrix. The covariance function of the empirical process is estimated over a grid of n equally spaced points over [0,1] (i.e $m = n$) interval as described in Section 3.5. Figure 4.11 shows the Q-Q plot of the obtained P-value from goodness-of-fit test in each simulation setting. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and sample quantiles of P-values, respectively. The figure is arranged so that each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively. We can clearly see that the sample quantiles matches very well with the theoretical ones. Figure 4.12 depicts the Q-Q plot of P-values less than or equal 0.10. For these P-values, the sample quantiles are larger than the theoretical values, as there is an obvious curve in the case of $n = 50$. This seems to improve a bit as sample size increases to $n = 100$ and much better as sample size reaches $n = 250$. Table 4.7 shows the estimated type one error rate at level 0.01 for each of the parameter settings. For all simulation settings, the type one error rate is below the desired level. The conclusion remains the same at level 0.05 as shown in Table 4.8. This observation persuaded us not to use the Hessian matrix to estimate Fisher information matrix.

Figure 4.11: Gamma distribution, simulation 2. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

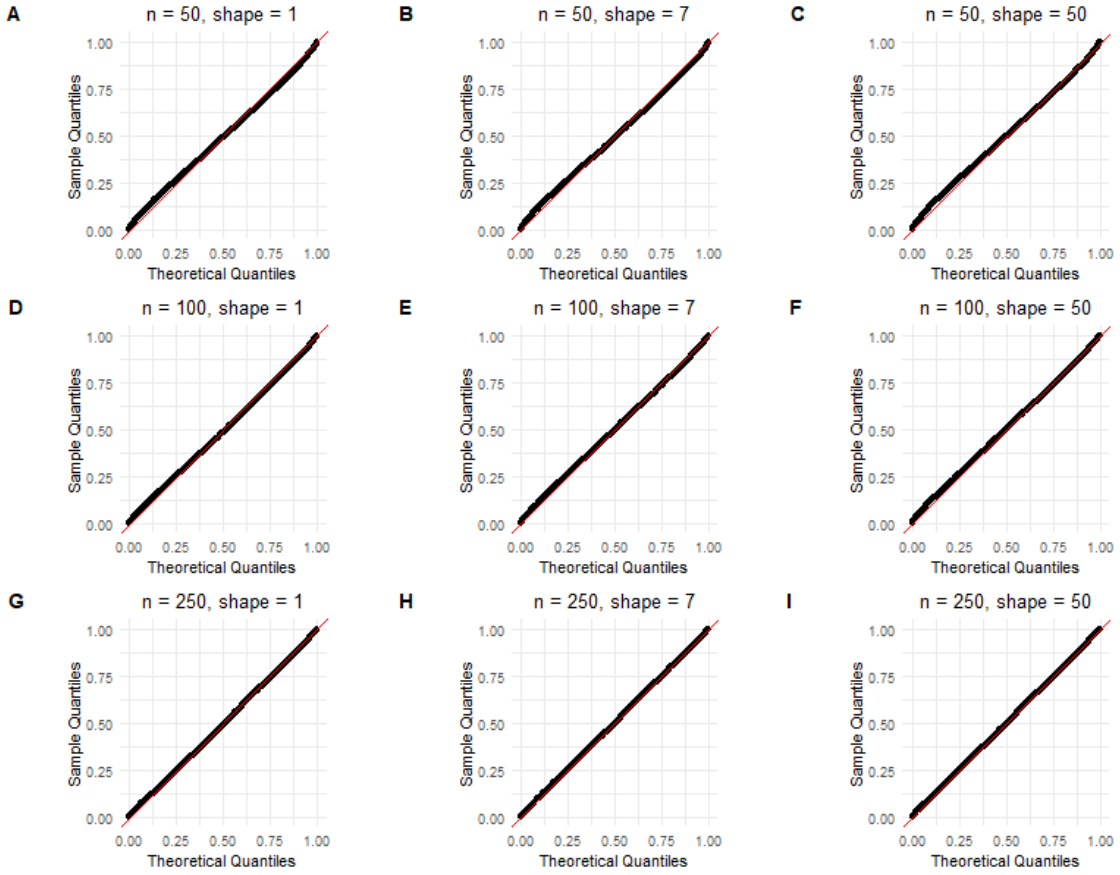


Figure 4.12: Gamma distribution, simulation 2, using the negative Hessian and an evenly spaced grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

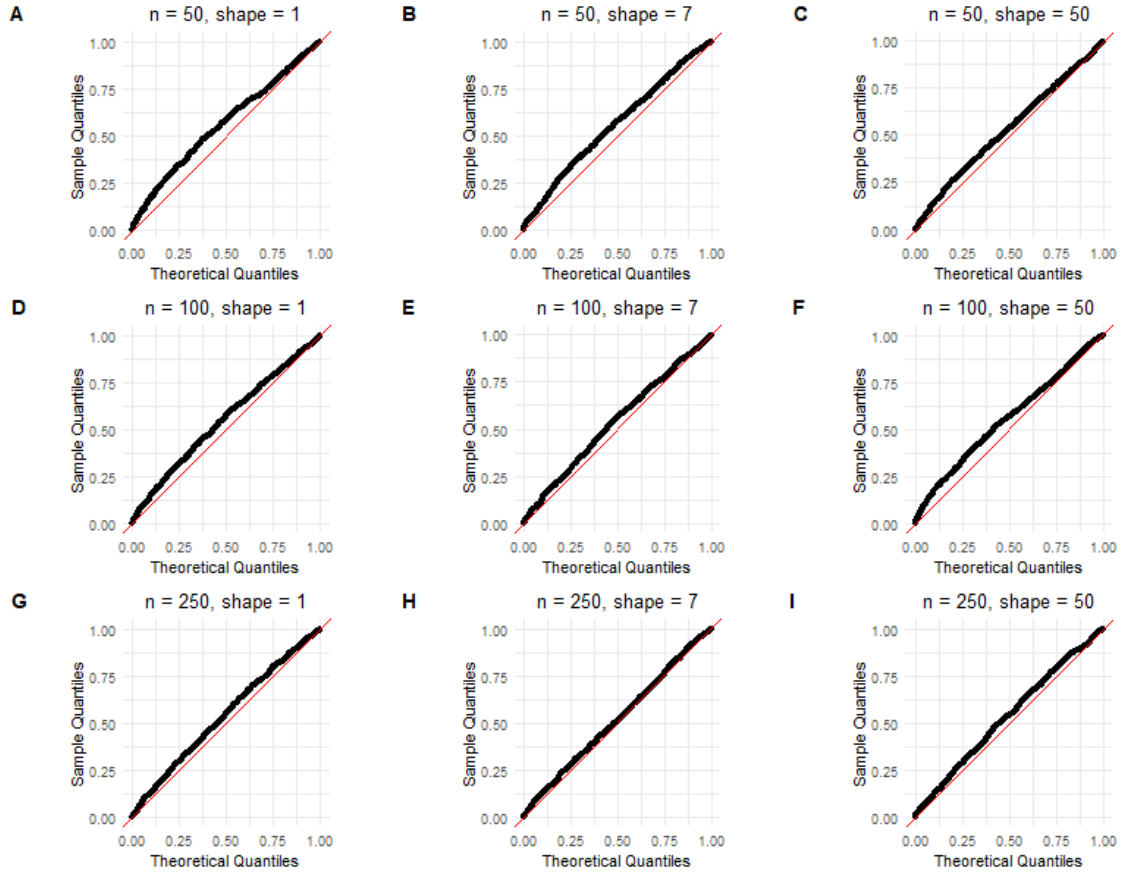


Table 4.7: Gamma distribution, simulation 2, using the negative Hessian and an evenly spaced grid. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameter.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0043	0.0055	0.006
n = 100	0.0061	0.0065	0.0042
n = 250	0.0060	0.0068	0.0085

Table 4.8: Gamma distribution, simulation 2, using the negative Hessian and an evenly spaced grid. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameter.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0325	0.0329	0.0348
n = 100	0.0393	0.0386	0.0341
n = 250	0.0436	0.0461	0.0461

4.3.3 Simulation 3

In this simulation, the Fisher information matrix is estimated by the variance of the score function. The covariance of the empirical process is calculated by probability integral transformed (PIT) values as described in Section 3.5. Figure 4.13 shows the Q-Q plot of all P-values. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged so that each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively. As we can see, the sample quantile matches very well with theoretical quantiles for any simulation settings. The conclusion remains the same for P-values that are less or equal than 0.10 as shown in figure 4.14. Table 4.9 shows the estimated type one error rate at level 0.01 for each of the parameter settings. It seems that the type of error rate is well controlled at level $\alpha = 0.01$. The estimated type one error rate at level 0.05 seems to be a bit inflated for $n = 50$ as we can see in table 4.10. It seems that increasing sample size to $n = 100$ and $n = 250$ controls the type one rate error at the desired level.

Figure 4.13: Gamma distribution, simulation 3. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

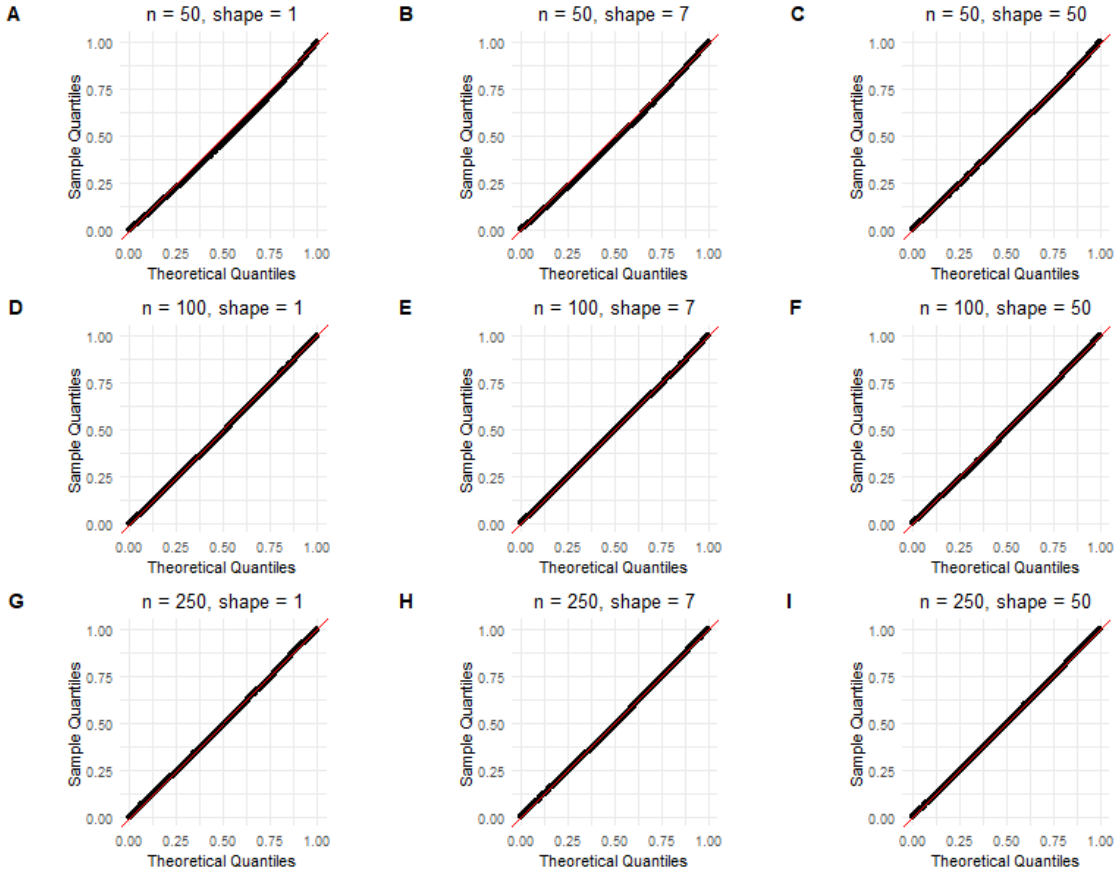


Figure 4.14: Gamma distribution, simulation 3. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

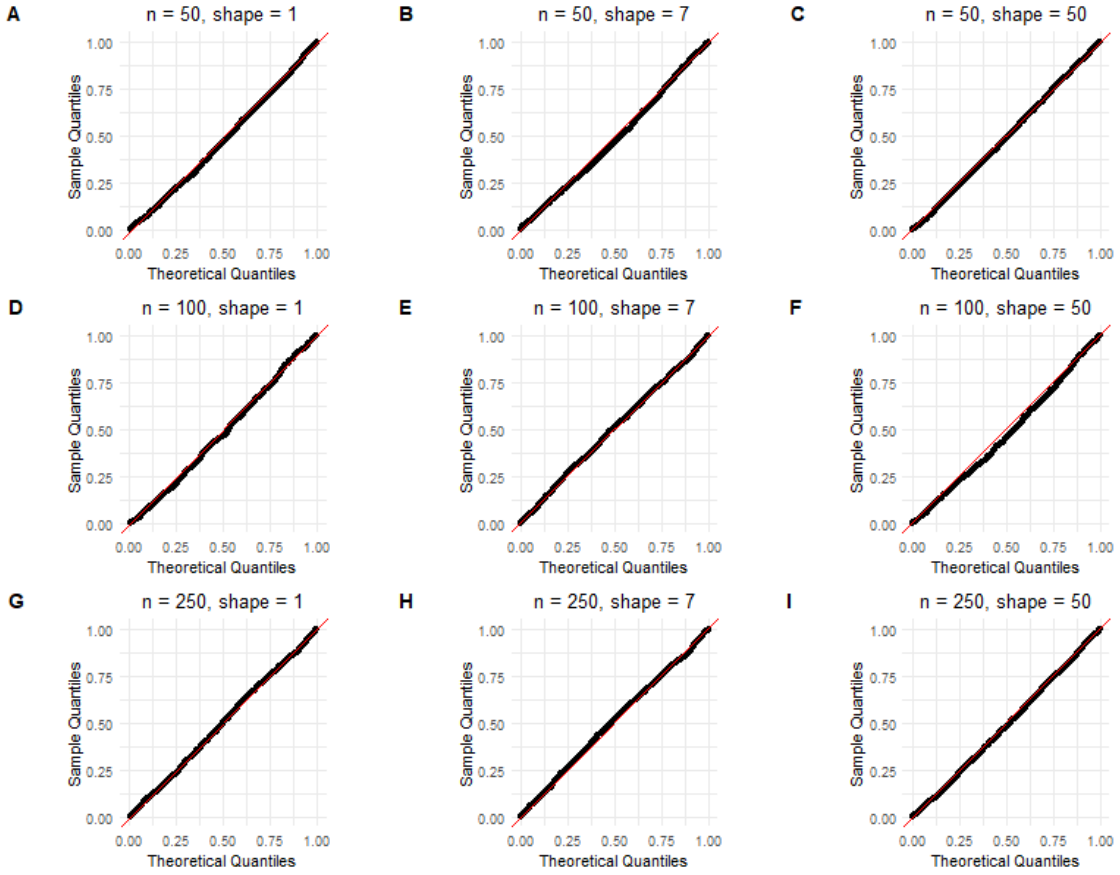


Table 4.9: Gamma distribution, simulation 3. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0144	0.0117	0.0130
n = 100	0.0137	0.0107	0.0131
n = 250	0.0103	0.0103	0.0125

Table 4.10: Gamma distribution, simulation 3. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0618	0.0619	0.0559
n = 100	0.0588	0.0529	0.0610
n = 250	0.0505	0.0492	0.0549

4.3.4 Simulation 4

In this simulation, the Fisher information matrix is estimated by the observed Hessian matrix. To compute the matrix Q and estimate eigenvalues, we used the probability integral transformed (PIT) values as described in Section 3.5. Figure 4.15 shows the Q-Q plot of all P-values and 4.16 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. Both figures are arranged so that each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively. The sample quantiles matches very well with the theoretical quantiles and the match gets better as the sample size increases from $n = 50$ to $n = 250$. Tables 4.11 and 4.12 presents the estimated type-one error rates at nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$, respectively. The type one error rate seems to be controlled at the desired levels.

Figure 4.15: Gamma distribution, simulation 4, using the negative Hessian and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

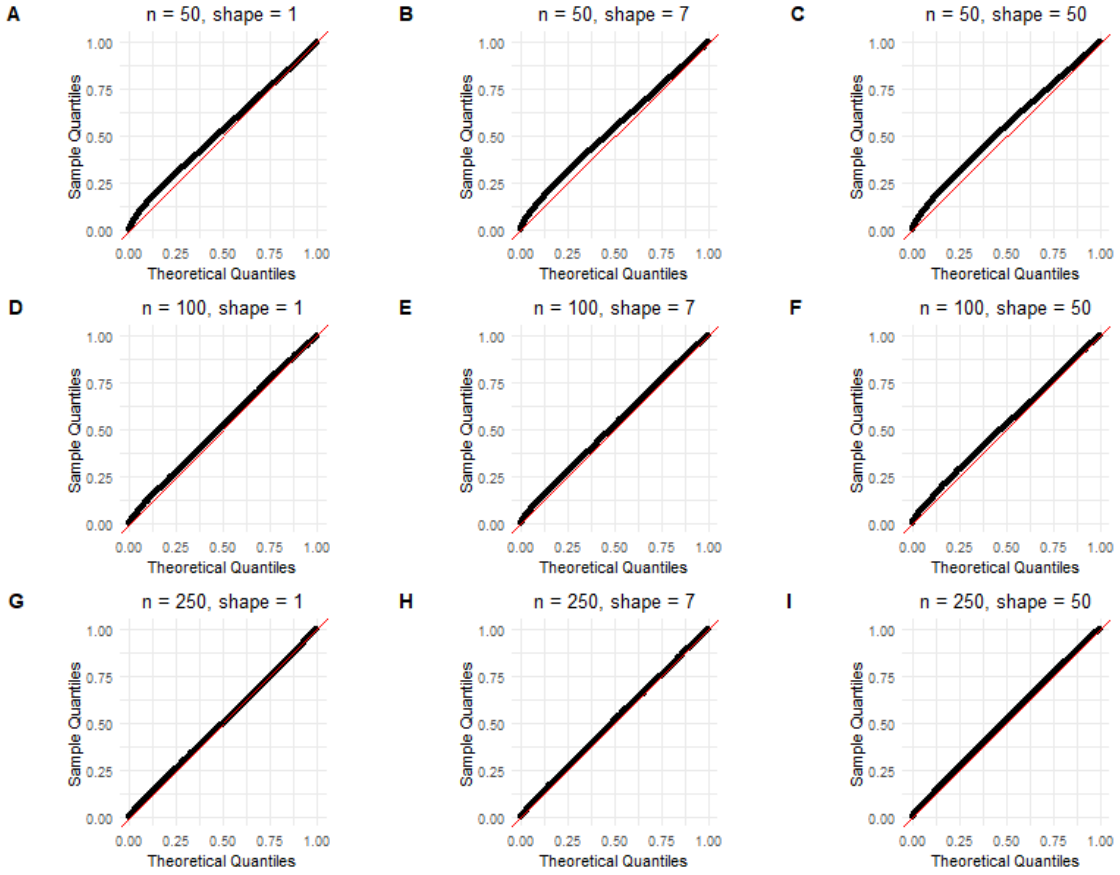


Figure 4.16: Gamma distribution, simulation 4, using the negative Hessian and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

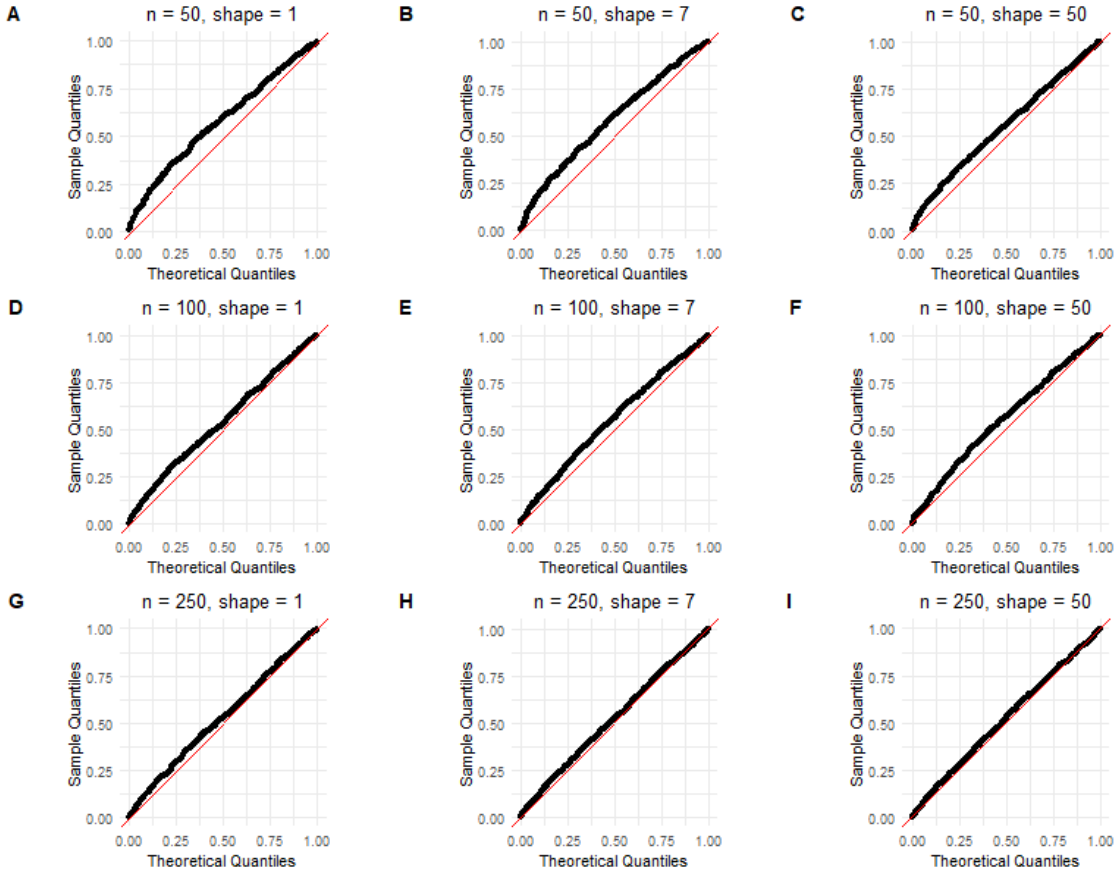


Table 4.11: Gamma distribution, simulation 4, using the negative Hessian and the PITs for the grid. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0020	0.0022	0.0029
n = 100	0.0048	0.0047	0.0056
n = 250	0.0063	0.0075	0.0071

Table 4.12: Gamma distribution, simulation 4, using the negative Hessian and the PITs for the grid. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0221	0.0226	0.0266
n = 100	0.0354	0.0330	0.0320
n = 250	0.0417	0.0429	0.0441

The results of the simulations for Gamma example suggest a similar strategy to the Normal example: We recommend using the PIT values as a grid over interval [0,1] with quadrature weights of $w_i = \frac{1}{n}$, and estimating the Fisher information matrix by the variance of score. The test demonstrate an excellent performance for sample sizes as low as $n = 50$ at both significance levels of $\alpha = 0.01$ and $\alpha = 0.05$.

4.4 Linear models

In this section, we show the results of one simulation for a linear model with a normal assumption for error terms. We considered the data generating model $y = X\beta + e$ where $\beta^T = (0.5, -1.34)$ is the vector of coefficients (including intercept), X is a design matrix with n rows and 2 columns, and e is the error terms. (As discussed below the null distribution of our test statistic does not depend on the actual parameter values.) In each Monte Carlo sample, the values of the design matrix and error terms are randomly generated from a standard normal distribution with mean zero and standard deviation of one but the vector of coefficients is kept constant between samples.

In view of the results obtained in the i.i.d. sampling examples we did not use a uniform grid; instead we used the probability integral transformed (PIT) values as a grid to compute the Q matrix, as described in Section 3.5.

As we described in Section 3.9, the limiting covariance function does not depend on the values of X, β, and on the choice of mean and standard deviation for the error terms. We applied the goodness-of-fit test to assess the normality of residuals in this model. We considered sample sizes $n = 50$, $n = 100$, and $n = 250$. Relying again on the univariate results the Fisher information matrix is here estimated by the variance of the score function obtained from the sample. We generated 10,000 Monte Carlo samples.

Figure 4.17 shows the theoretical quantiles vs the sample quantiles of all P-values resulted from goodness-of-fit test. For sample size of $n = 50$, the P-values seem to be uniformly distributed under the null hypothesis. There is a bit of departure from the straight line but the general performance seems fine. As the sample sizes increases, we can clearly see that the points get closer to the straight line. Figure 4.18 shows the theoretical quantiles vs the

sample quantiles of P-values that are less than or equal to 0.10, resulted from goodness-of-fit test. The results clearly suggests that, regardless of sample size, both theoretical and sample quantiles matches very well.

Table 4.13 presents the estimated type one error rate based on 10,000 Monte Carlo sample studies at two nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$. The colums of this table shows the estimated values for each sample size. As we can see, for sample size $n = 50$ the values are a bit inflated specially for $\alpha = 0.05$. However, increasing the sample size seems to properly control the type one error rate at the desired level.

Figure 4.17: Simulation studies, linear models. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different sample sizes: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, $n = 250$, respectively.

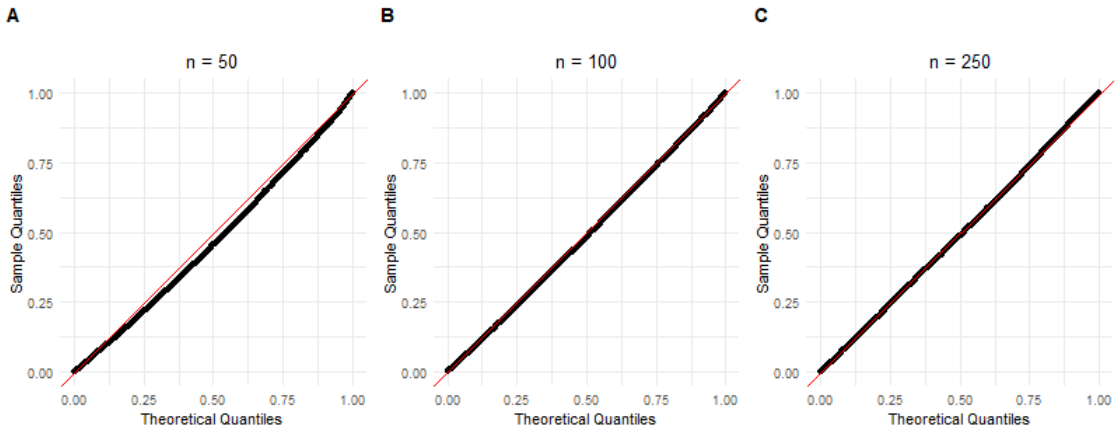


Figure 4.18: Simulation studies, linear models. Theoretical quantiles vs sample quantiles of P-values less than or equal to 0.10 from goodness-of-fit test for different sample sizes: panels (A), (B), and (C) are for sample size $n = 50$, $n = 100$, $n = 250$, respectively. The plot shows P-values less than 0.10 only.

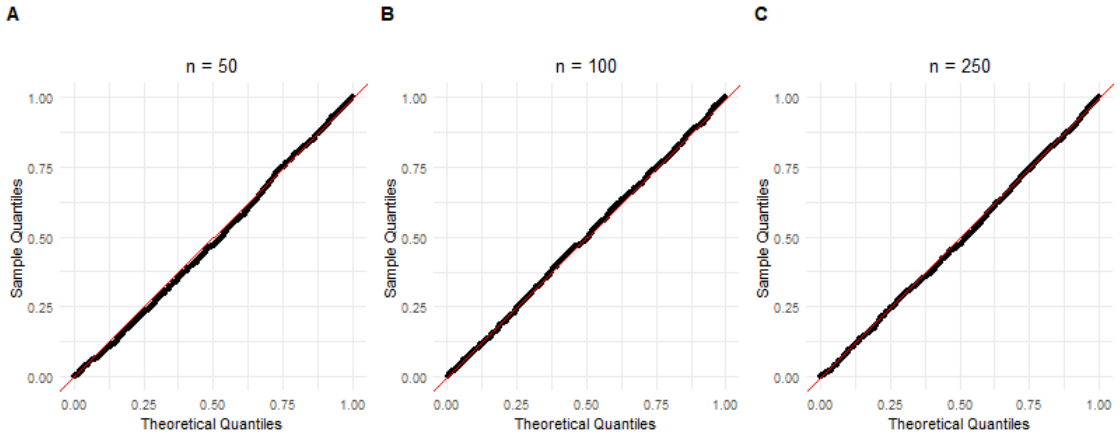


Table 4.13: Simulation studies, linear models. The estimated type one error rate at level $\alpha = 0.01$ and $\alpha = 0.05$ for different sample sizes. The rows and the columns are nominal type-one error rate (1 and 5 percent) and sample sizes, respectively. The values in the cells are the estimated type one error rates based on 10,000 Monte Carlo samples.

	n= 50	n =100	n = 250
$\alpha = 0.01$	0.0141	0.0118	0.0110
$\alpha = 0.05$	0.0634	0.0557	0.0543

4.5 Generalized linear model

In this section, we extend our simulations to a more general case and consider Gamma regression in the class of generalized linear models with two popular link functions, 1) log and 2) inverse. We test the null hypothesis that the response variable follows a Gamma distribution. We show the results from five different simulations.

In the first four simulations, we consider a generalized linear model with a log link function and one explanatory variable, i.e $log(\mu) = \beta_0 + \beta_1 X$ where $\mu = E(Y|X)$. We keep the value of the intercept constant at $\beta_0 = 0.56$ in all four simulations. Each simulation uses a different value os β_1 to investigate any effect of the slope. The value of X was generated from a Uniform distribution over the interval [0,1]. The response variable was generated by $Y = e^{\beta_0+\beta_1 X} \times e$, where e is the error term that follows a Gamma distribution with shape parameter α and scale value of one.

In the last simulation, we consider a generalized linear model with more explanatory variables. In this setting, we consider an inverse link function, i.e. $\mu = \frac{1}{X\beta}$. The values of β vector are generated from a Uniform distribution over the interval [0.5,1.5]. The values of the X matrix are generated from a normal distribution with a mean of 2 and standard deviation of 0.1 to ensure the values of μ are positive. The error terms of the model are generated as before.

In all simulations, we considered different values for the shape parameter, $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, and different sample sizes, $n = 50$, $n = 100$, and $n = 250$. We generated 10,000 Monte Carlo samples for each parameter setting. The Fisher information matrix was estimated by the score function. We estimated the covariance function of the empirical process by probability inverse transformed values. In all of the simulations below, we used the probability integral transformed (PIT) values to compute matrix Q, as described in Section 3.5.

4.5.1 Simulation 1

In this simulation, we set the intercept $\beta_0 = 0.56$ and slope $\beta_1 = -1.3$. Figure 4.19 shows the Q-Q plot of the P-values obtained from our goodness-of-fit test in each simulation setting. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

We can see that the sample quantiles correspond very well with the corresponding theoretical values as sample size increases from $n = 50$ to $n = 250$. We also look at Q-Q plot of smaller P-value to get a better insight into the performance of the method. Figure 4.20 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as previous figures. As we can see, theoretical and sample quantiles corresponds very well for all parameter settings.

We estimate the type-one error rate at levels of $\alpha = 0.01$ and $\alpha = 0.05$ in Table 4.14 and Table 4.15, respectively. The level of the test is well controlled at level $\alpha = 0.01$ but inflated at level of $\alpha = 0.05$. But increasing the sample size from $n = 50$ to $n = 250$ seems to control the level of the test. Note that when comparing this to the case of i.i.d. Gamma, the type-one error rate in that simulation is controlled, regardless of the sample size.

Figure 4.19: Generalized linear model, simulation 1. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

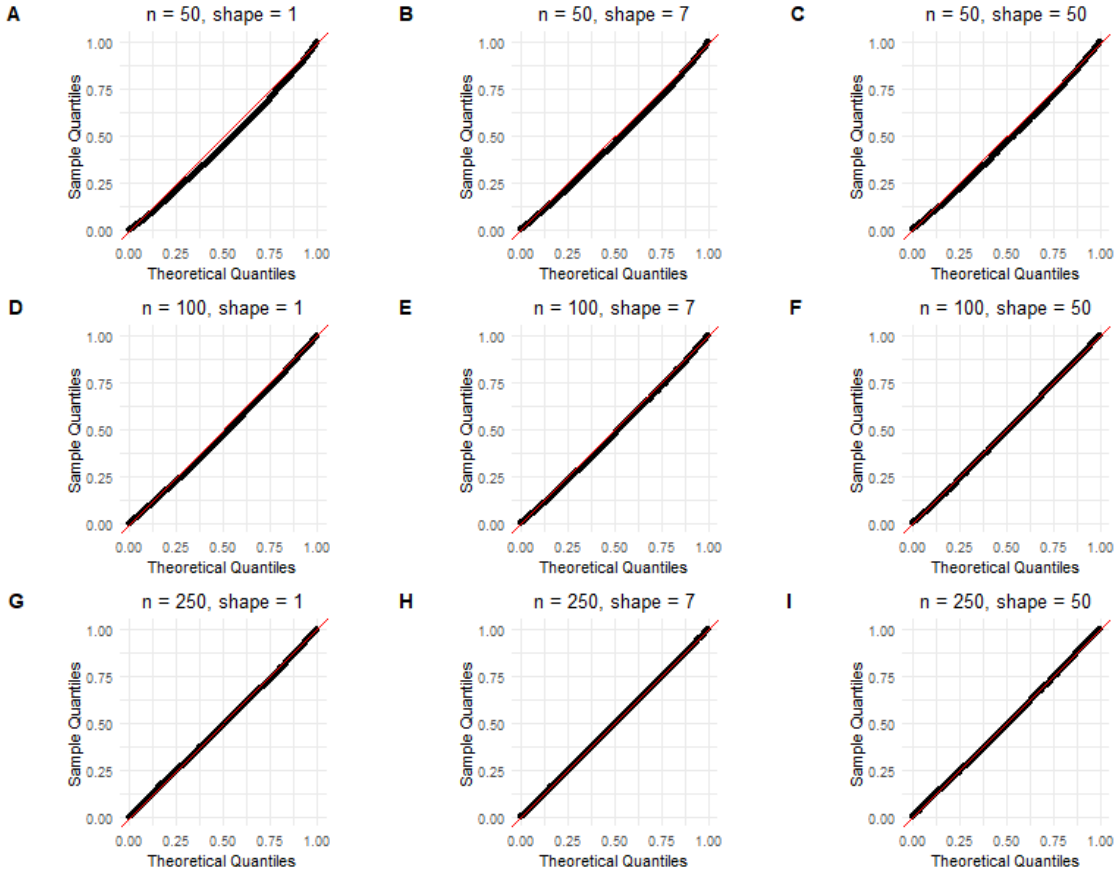


Figure 4.20: Generalized linear model, simulation 1. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

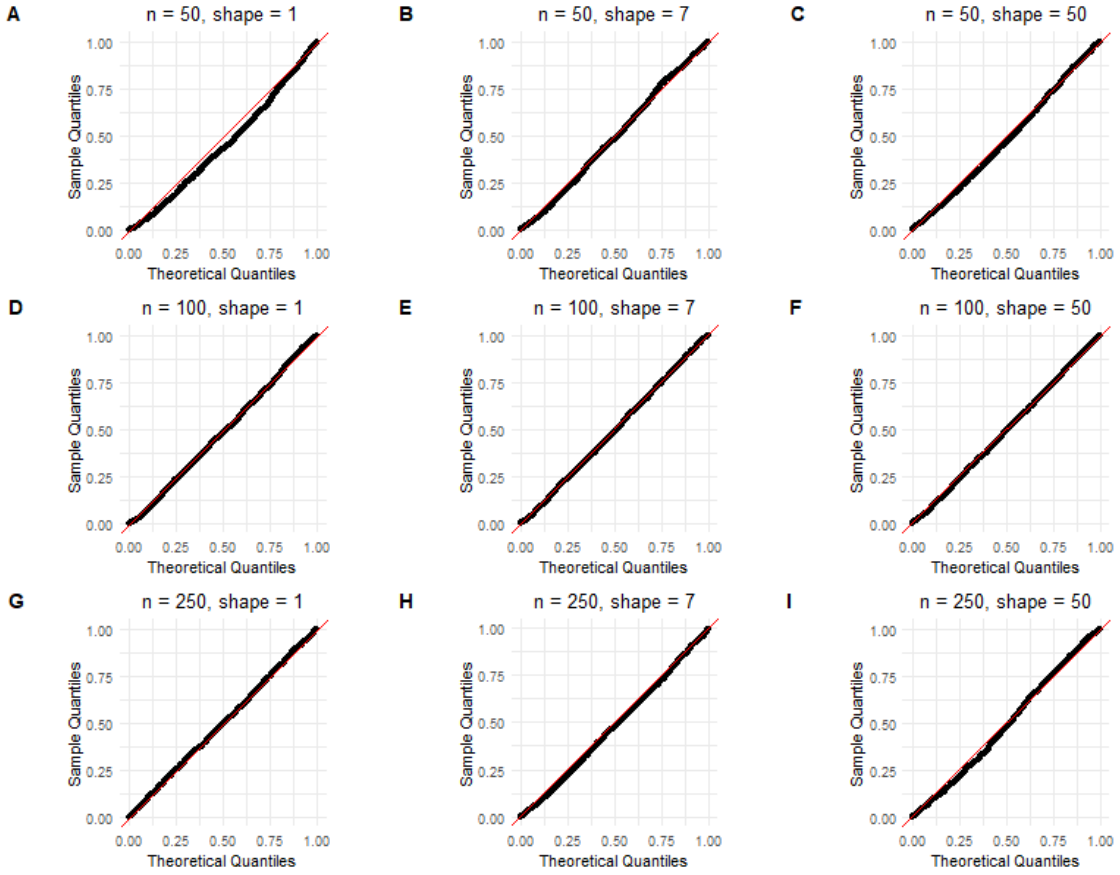


Table 4.14: Generalized linear model, simulation 1. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0192	0.0164	0.0150
n = 100	0.0154	0.0142	0.0140
n = 250	0.0104	0.0127	0.0124

Table 4.15: Generalized linear model, simulation 1. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0767	0.0640	0.0669
n = 100	0.0615	0.0607	0.0604
n = 250	0.0520	0.0552	0.0539

4.5.2 Simulation 2

We keep the value of the intercept $\beta_0 = 0.56$ – the same as in simulation 1. We increase the slope to $\beta_1 = 2.1$. Figure 4.21 shows the Q-Q plot of the all obtained P-value from goodness-of-fit test in each simulation setting. Figure 4.22 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

We can see that the sample quantiles correspond very well with the corresponding theoretical values as sample size increases from $n = 50$ to $n = 250$. Table 4.16 and Table 4.17 shows the estimated type-one error rate. The conclusion remains the same as simulation 1. The level of the test is well controlled at $\alpha = 0.01$ and a bit inflated at level $\alpha = 0.05$ and improves by increasing the sample size.

Figure 4.21: Generalized linear model, simulation 2. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

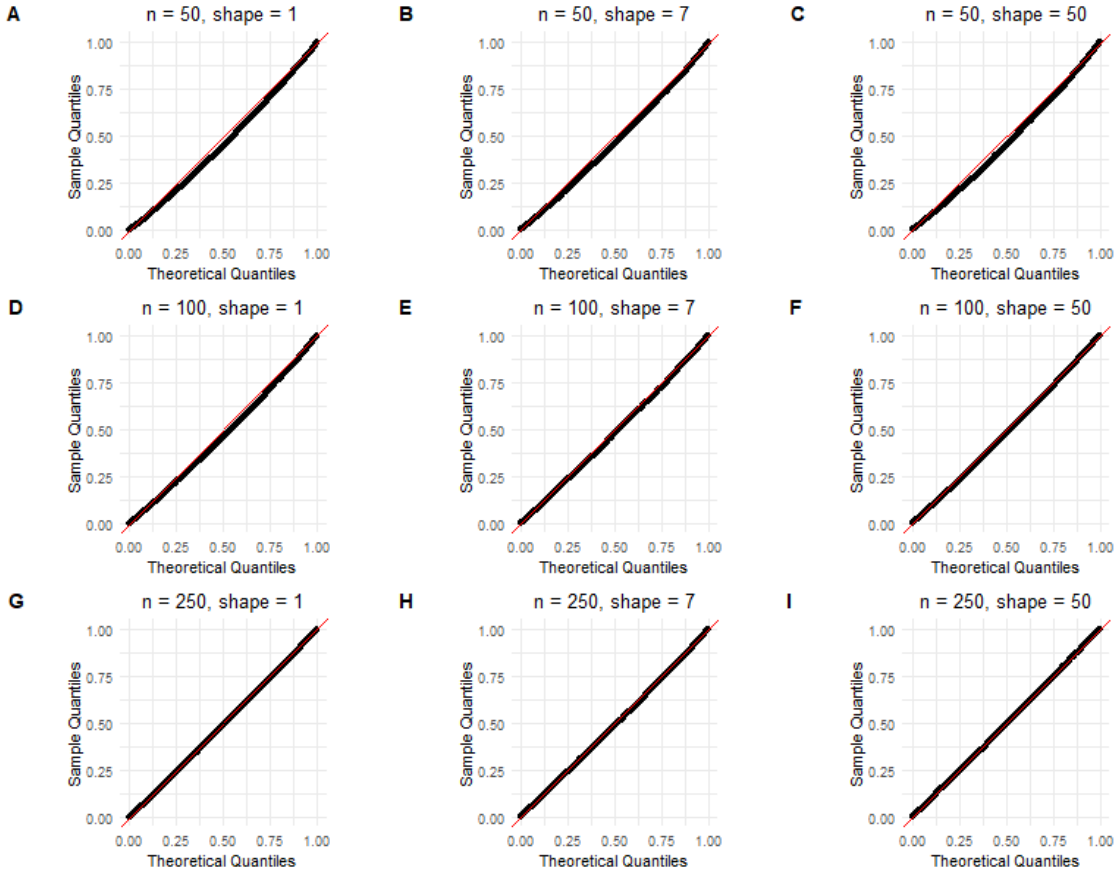


Figure 4.22: Generalized linear model, simulation 2. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

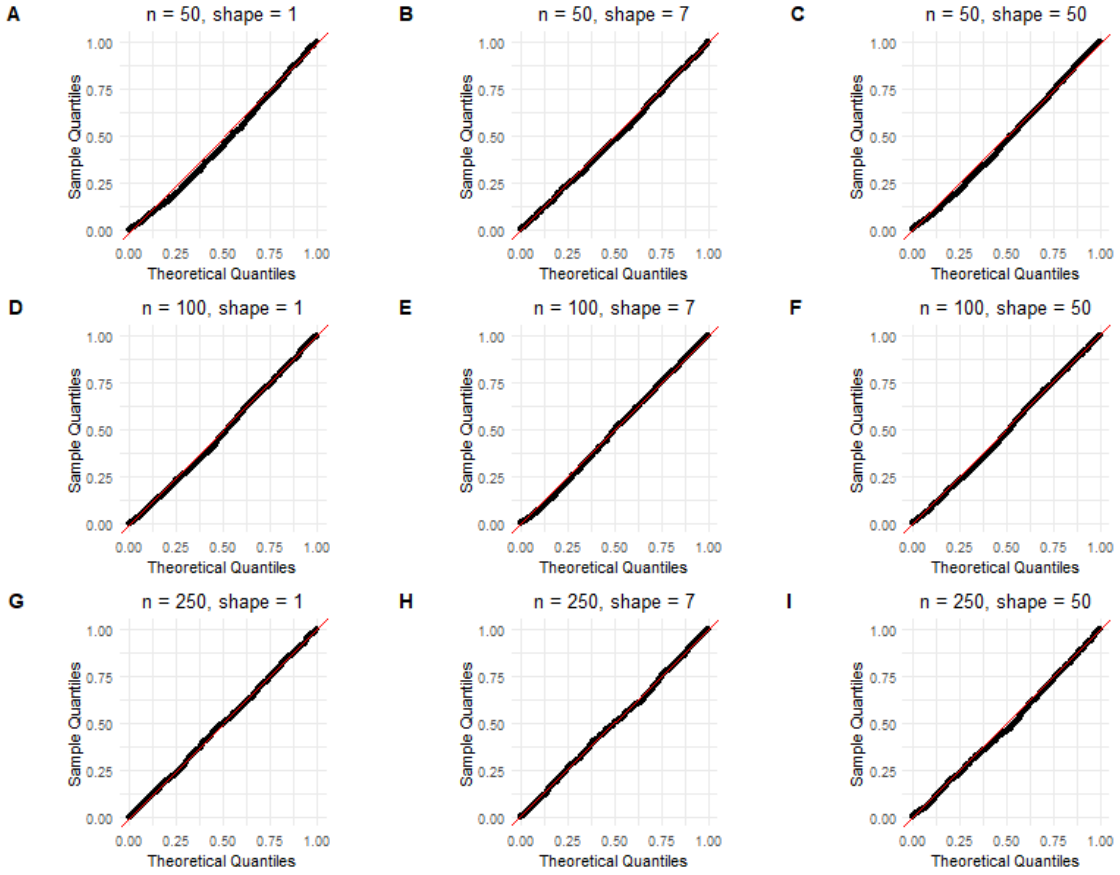


Table 4.16: Generalized linear model, simulation 2. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0177	0.0136	0.0173
n = 100	0.0149	0.0152	0.0131
n = 250	0.0116	0.0109	0.0124

Table 4.17: Generalized linear model, simulation 2. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0718	0.0654	0.0689
n = 100	0.0640	0.0580	0.0590
n = 250	0.0560	0.0523	0.0578

4.5.3 Simulation 3

For this simulation, we keep the value of intercept ($\beta_0 = 0.56$) the same as the previous simulations. We change the value of the slope to $\beta_1 = -2.6$. Figure 4.23 shows the Q-Q plot of all P-value from goodness-of-fit test in each simulation setting. Figure 4.24 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

We can see that the sample quantiles correspond very well with the corresponding theoretical values as sample size increases from $n = 50$ to $n = 250$. Table 4.18 and Table 4.19 shows the estimated type-one error rate at different levels of $\alpha = 0.01$ and $\alpha = 0.05$, respectively.

Figure 4.23: Generalized linear model, simulation 3, using the variance of the score and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

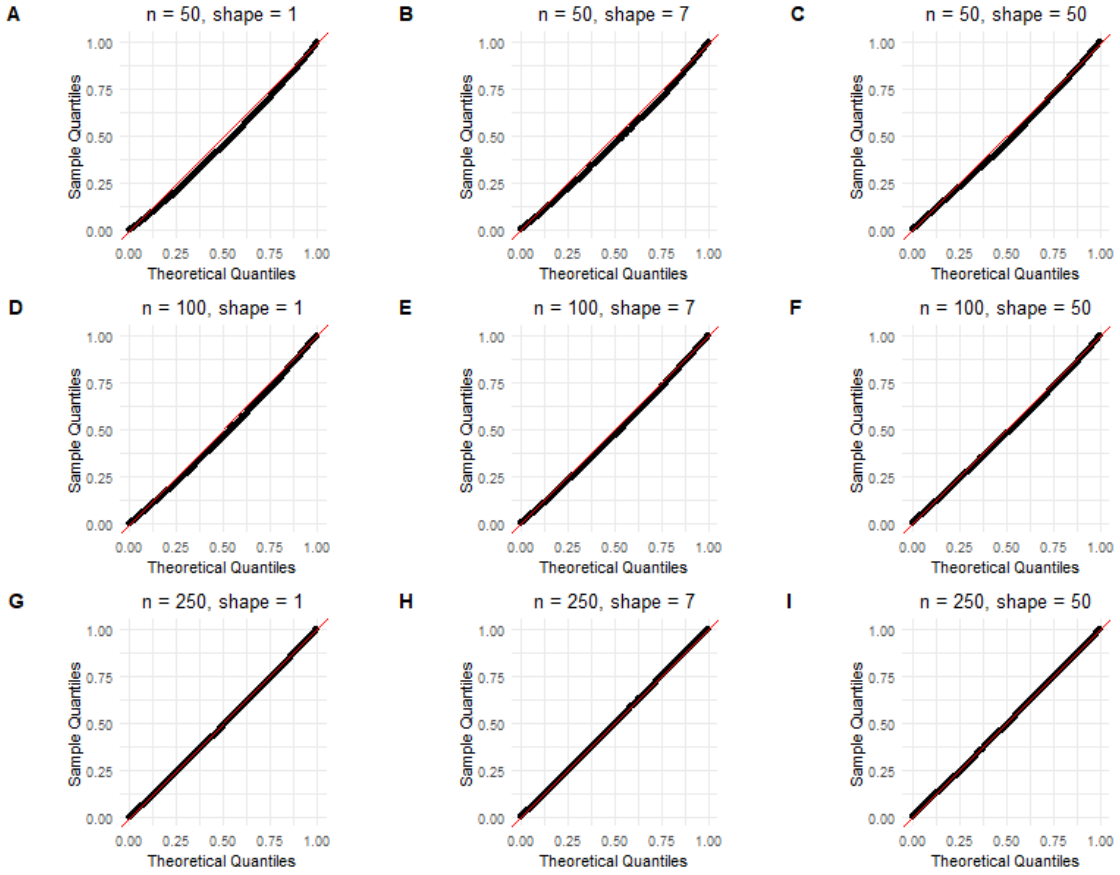


Figure 4.24: Generalized linear model, simulation 3, using the variance of the score and the PITs for the grid. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

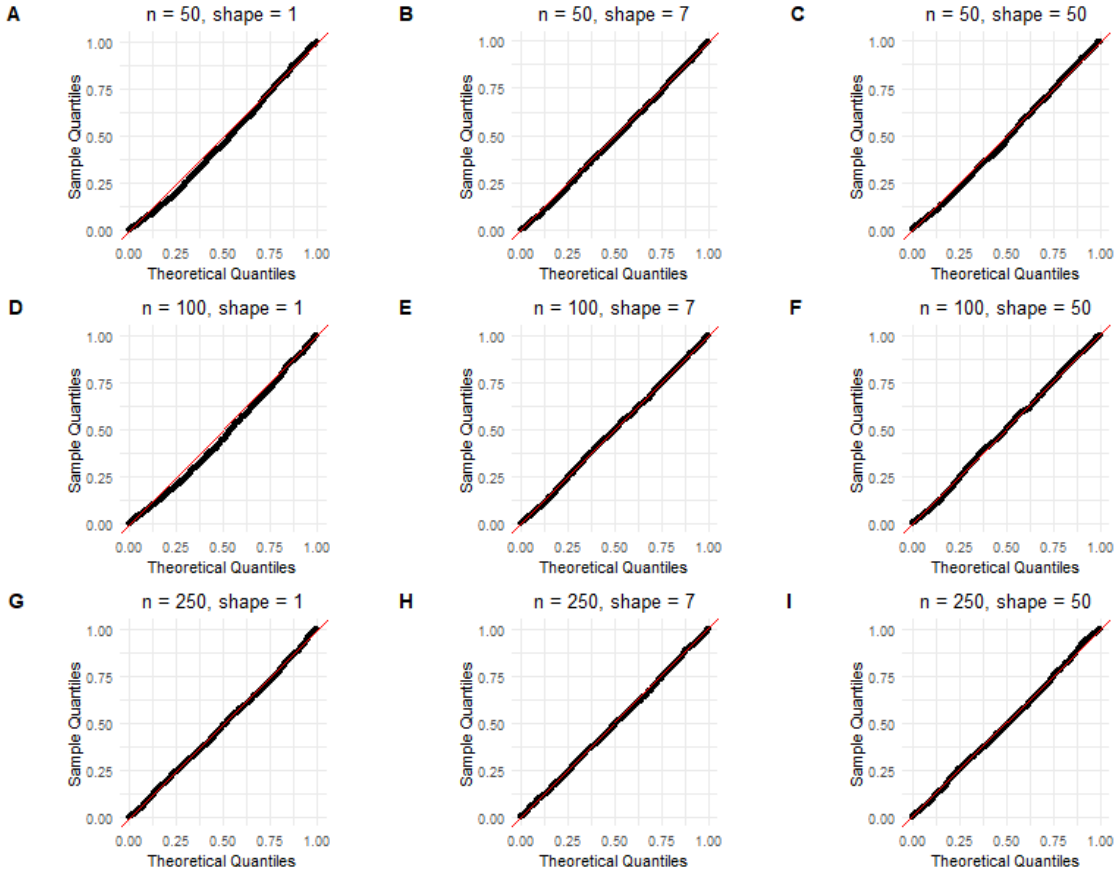


Table 4.18: Generalized linear model, simulation 3, using the variance of the score and the PITs for the grid. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0193	0.0154	0.0149
n = 100	0.0156	0.0142	0.0141
n = 250	0.0123	0.0102	0.0114

Table 4.19: Generalized linear model, simulation 3. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0728	0.0640	0.0604
n = 100	0.0649	0.0597	0.0559
n = 250	0.0561	0.0524	0.0529

4.5.4 Simulation 4

In this simulation, we kept the value of intercept the same as $\beta_0 = 0.56$ and considered a more negative slope, $\beta_1 = -5$. Figure 4.25 shows the Q-Q plot of all P-values from the goodness-of-fit test in each simulation setting. Figure 4.26 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

We can see that the sample quantiles correspond very well with the corresponding theoretical values as sample size increases from $n = 50$ to $n = 250$. Table 4.20 and Table 4.21 shows the estimated type-one error rate at different levels of $\alpha = 0.01$ and $\alpha = 0.05$, respectively.

Figure 4.25: Generalized linear model, simulation 4. Theoretical quantiles vs sample quantiles of P-values obtained from the goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

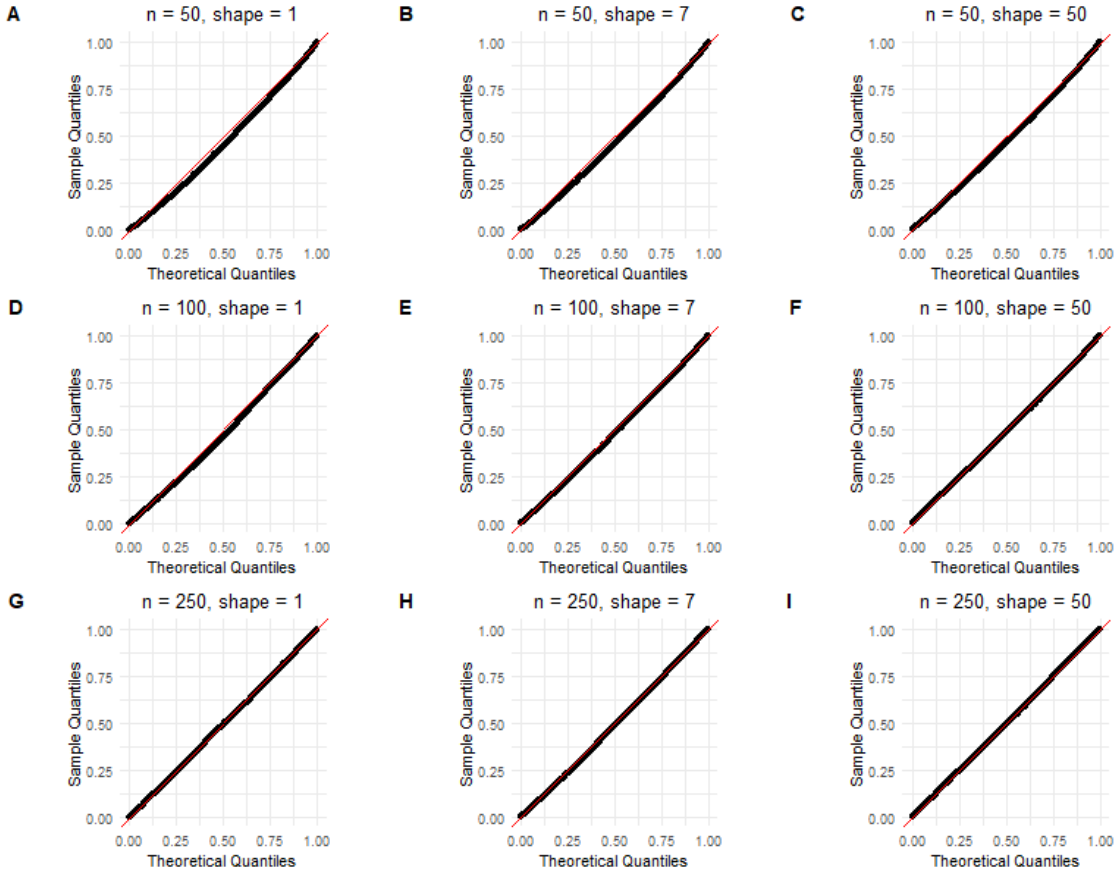


Figure 4.26: Generalized linear model, simulation 4. Theoretical quantiles vs sample quantiles of P-values less than or equal to 0.1 from the goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

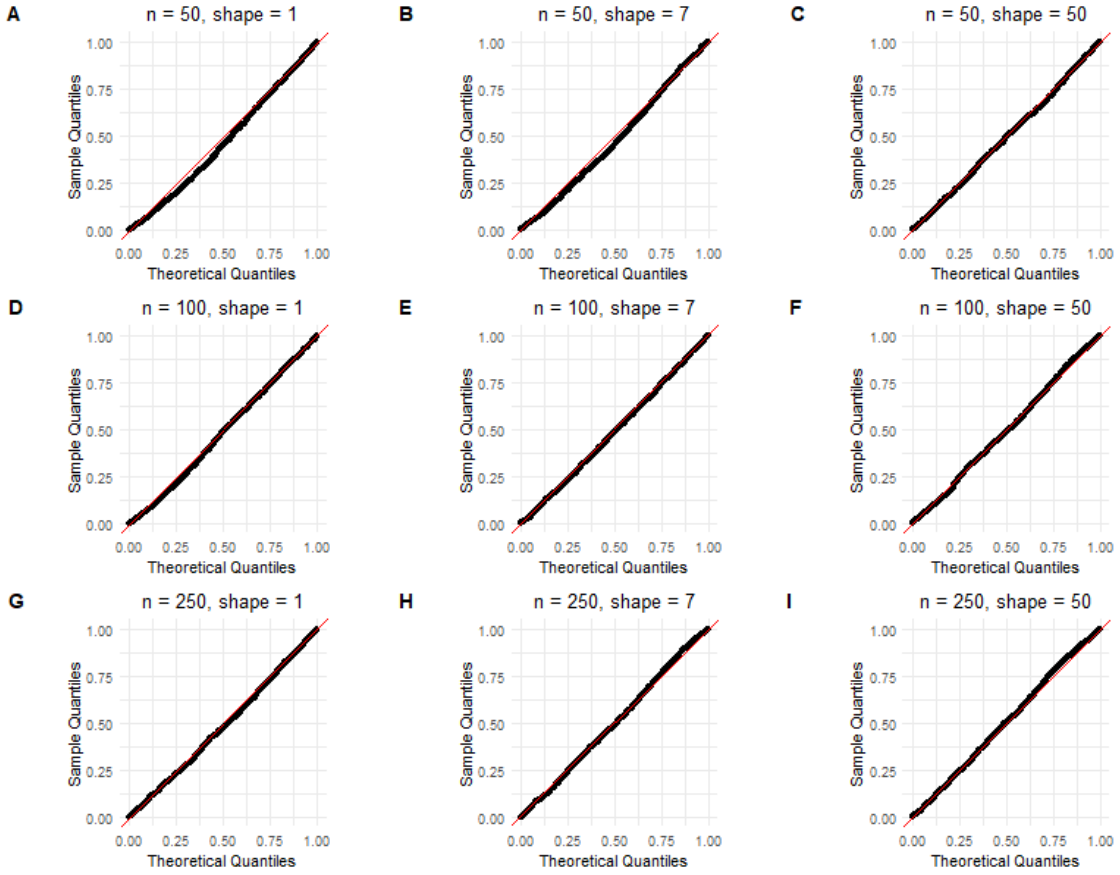


Table 4.20: Generalized linear model, simulation 4. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0183	0.0170	0.0144
n = 100	0.0160	0.0127	0.0125
n = 250	0.0116	0.0129	0.0121

Table 4.21: Generalized linear model, simulation 4. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0730	0.0708	0.0620
n = 100	0.0611	0.0538	0.0542
n = 250	0.0563	0.0580	0.0521

4.5.5 Simulation 5

In this simulation, we extend our model to include more explanatory variables. As noted at the beginning of the section, we generated the data based on a generalized linear model with an inverse link function. The values of β and design matrix X were chosen to ensure the value of $\mu = \frac{1}{X\beta}$ is positive. The number of explanatory variables in this simulation was $k = 2, 3, 4, 5, 10$ but we only show the results for $k = 2, 5, 10$. However, the conclusion remains the same.

Number of explanatory variables = 2

Figure 4.27 shows the Q-Q plot of all P-value from goodness-of-fit test in each simulation setting. Figure 4.28 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

Figure 4.27: Generalized linear model, simulation 5, k=2. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

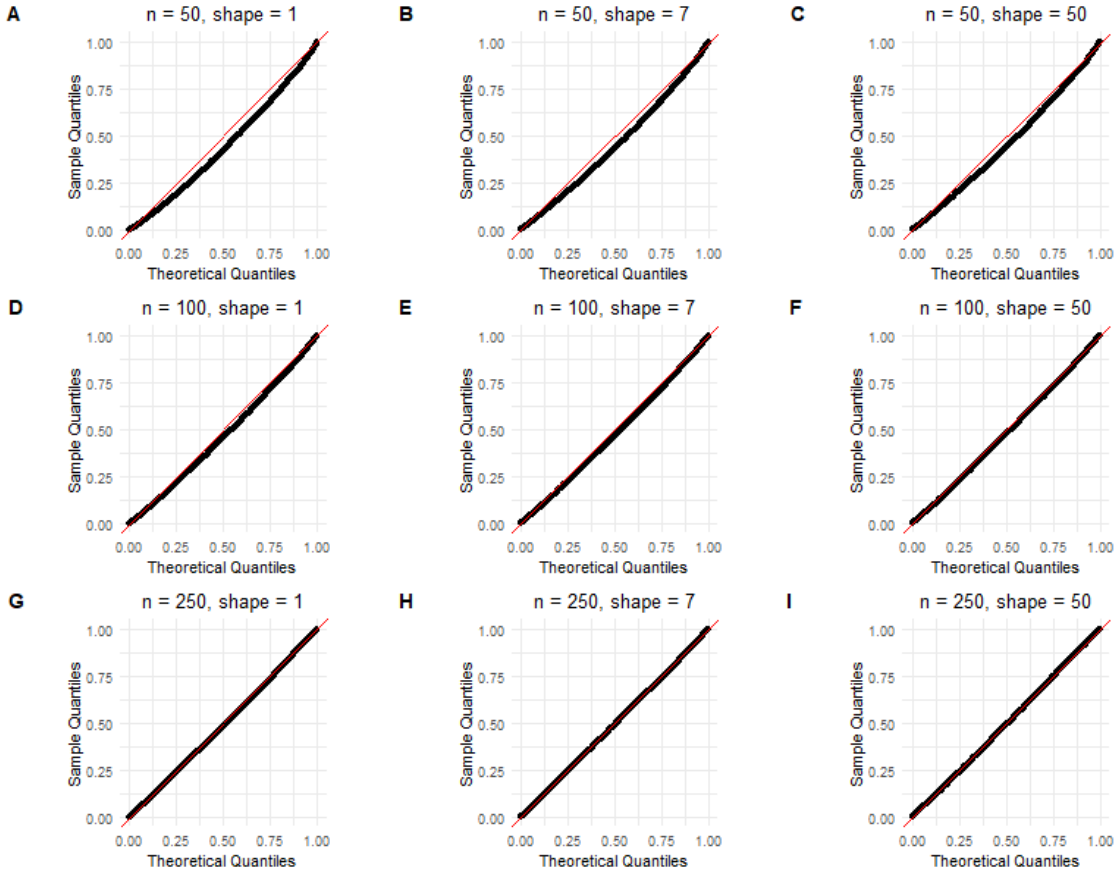


Figure 4.28: Generalized linear model, simulation 5, k=2. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

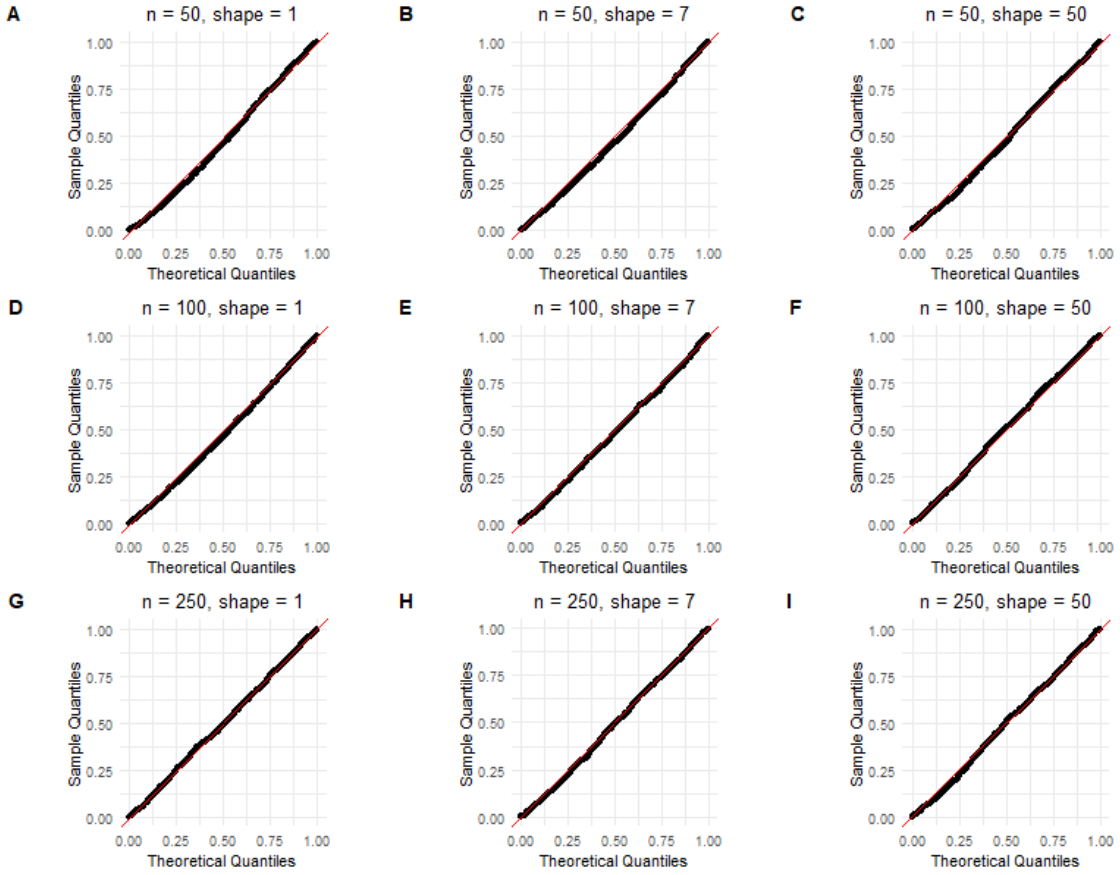


Table 4.22: Generalized linear model, simulation 5, k = 2. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0207	0.0165	0.0165
n = 100	0.0155	0.0149	0.0144
n = 250	0.0124	0.0129	0.0144

Table 4.23: Generalized linear model, simulation 5, k = 2. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0808	0.0716	0.0711
n = 100	0.0679	0.0620	0.0598
n = 250	0.0558	0.0540	0.0534

Number of explanatory variables = 5

Figure 4.29 shows the Q-Q plot of all P-value from goodness-of-fit test in each simulation setting. Figure 4.30 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

Figure 4.29: Generalized linear model, simulation 5, k=5. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

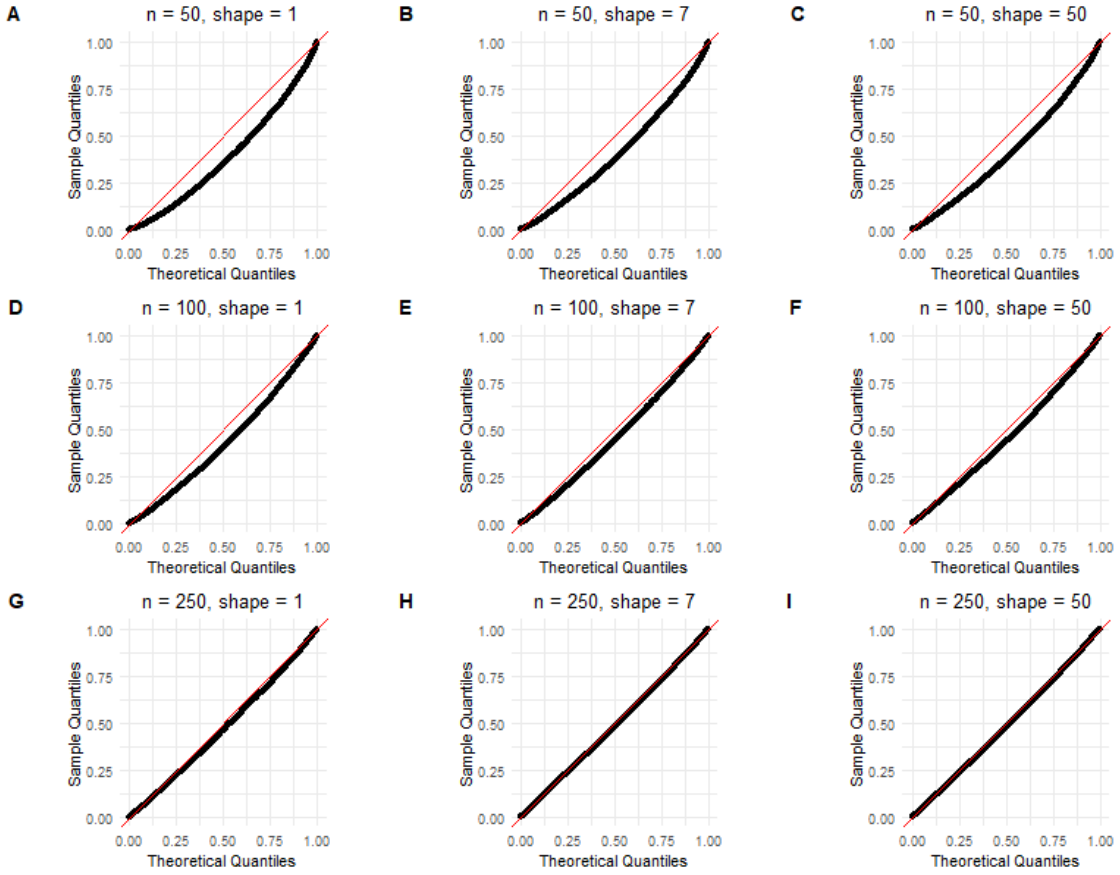


Figure 4.30: Generalized linear model, simulation 5, k=5. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

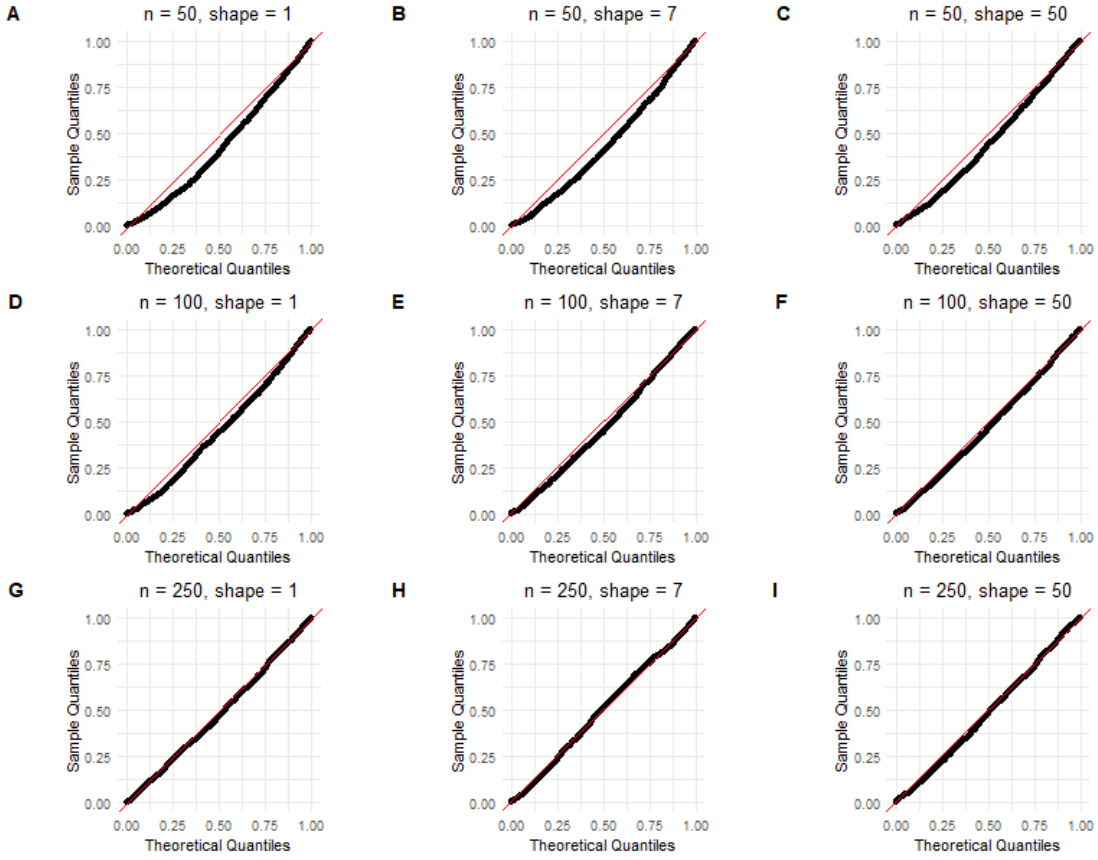


Table 4.24: Generalized linear model, simulation 5, k = 5. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0372	0.0278	0.0265
n = 100	0.0263	0.0179	0.0163
n = 250	0.0133	0.0130	0.0140

Table 4.25: Generalized linear model, simulation 5, k = 5. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.1192	0.1001	0.0938
n = 100	0.0876	0.0749	0.0695
n = 250	0.0633	0.0528	0.0576

Number of explanatory variables = 10

Figure 4.31 shows the Q-Q plot of all P-values from goodness-of-fit test in each simulation setting. Figure 4.32 shows the Q-Q plot of P-values less than or equal to 0.10. The x and y-axes in all panels are the theoretical quantiles of Uniform distribution and quantiles of the sample, respectively. The figure is arranged in the same way as before, where each row belongs to a different sample size and each column represents a parameter setting. The results of each shape parameter of $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$ are arranged in first, second, and third columns, respectively.

In Figure 4.31, a noticeable disagreement between theoretical and sample quantiles is observed, particularly for sample sizes of $n = 50$ and $n = 100$, regardless of the chosen value for the shape parameter. This disagreement persists for $n = 250$ and a shape parameter of one. However, for $n = 250$ and larger shape values, there is a slight improvement. For P-values less than 0.10 in Figure 4.32, we also observe a discrepancy between sample and theoretical quantiles for values of $n = 50$ and $n = 100$. But increasing the sample size to $n = 250$ resolves the issue. Table 4.26 and Table 4.27 present the estimated type-one error rates at the significance levels of $\alpha = 0.01$ and $\alpha = 0.05$, respectively. It is evident that the type-one error rate is not adequately controlled in this simulation at either level. Increasing the sample size from $n = 50$ to $n = 100$ does not appear to improve the control of the type-one error rate. Only for a large shape value, such as $\alpha = 50$, is the type-one error rate slightly inflated when the sample size is $n = 250$. Overall, in the case of ten explanatory variables, we observe poor performance.

Figure 4.31: Generalized linear model, simulation 5, k=10. Theoretical quantiles vs sample quantiles of obtained P-values from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

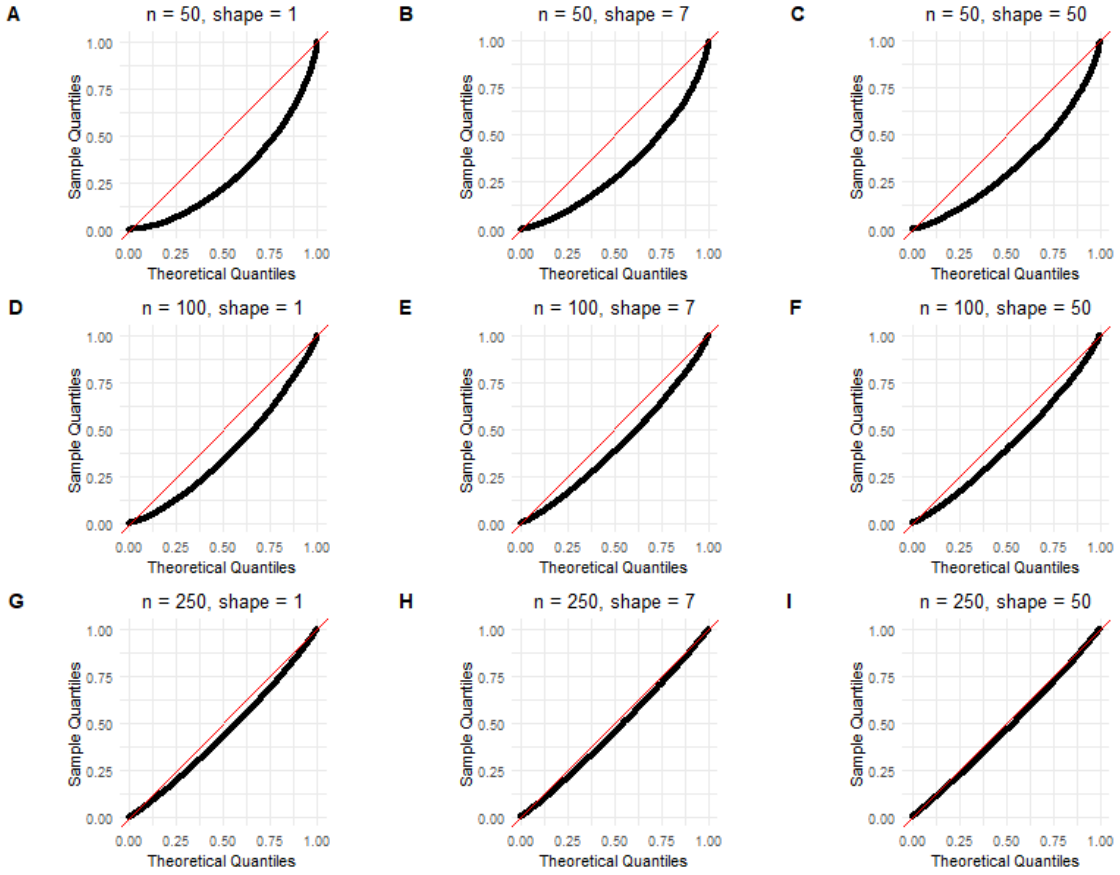


Figure 4.32: Generalized linear model, simulation 5, k=10. Theoretical quantiles vs sample quantiles of obtained P-values less than or equal to 0.1 from goodness-of-fit test for different parameter settings in each panel: panels (A), (B), and (C) are for sample size $n = 50$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (D), (E), and (F) are for sample size $n = 100$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively. Panels (G), (H), and (I) are for sample size $n = 250$, and $\alpha = 1$, $\alpha = 7$, and $\alpha = 50$, respectively.

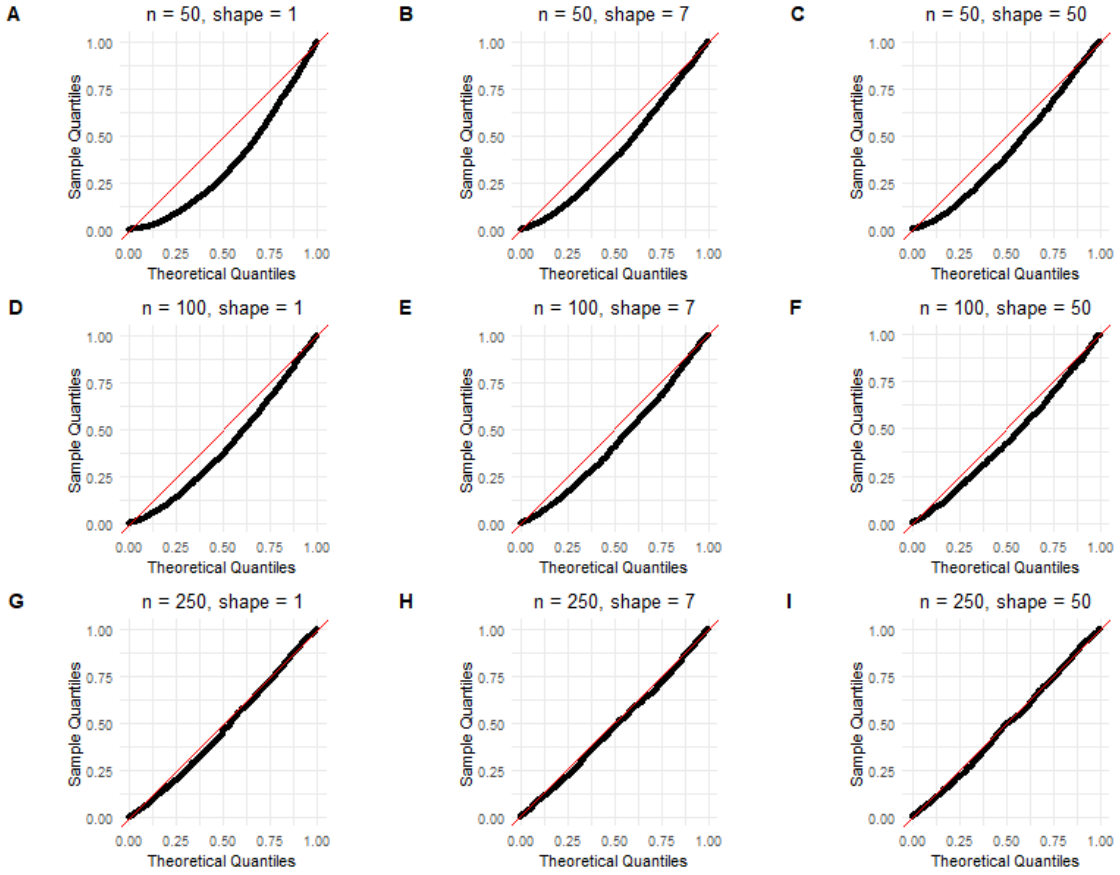


Table 4.26: Generalized linear model, simulation 5, k = 10. The estimated type one error rate at level 0.01. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.0899	0.0528	0.0490
n = 100	0.0444	0.0295	0.0272
n = 250	0.0196	0.0153	0.0152

Table 4.27: Generalized linear model, simulation 5, k = 10. The estimated type one error rate at level 0.05. Rows are sample size and columns are shape parameters.

	$\alpha = 1$	$\alpha = 7$	$\alpha = 50$
n = 50	0.2225	0.1594	0.1443
n = 100	0.1318	0.1008	0.0986
n = 250	0.0770	0.0679	0.0619

4.6 Real data example

In this section, we demonstrate the application of the proposed goodness-of-fit test by using publicly available empirical data as an illustrative example. We have chosen the third-party motor insurance claims from Sweden in 1977 as our sample data set; this data is conveniently available in the *faraway* package in R. The dataset comprises 1797 observations across 8 distinct variables as follows [56]:

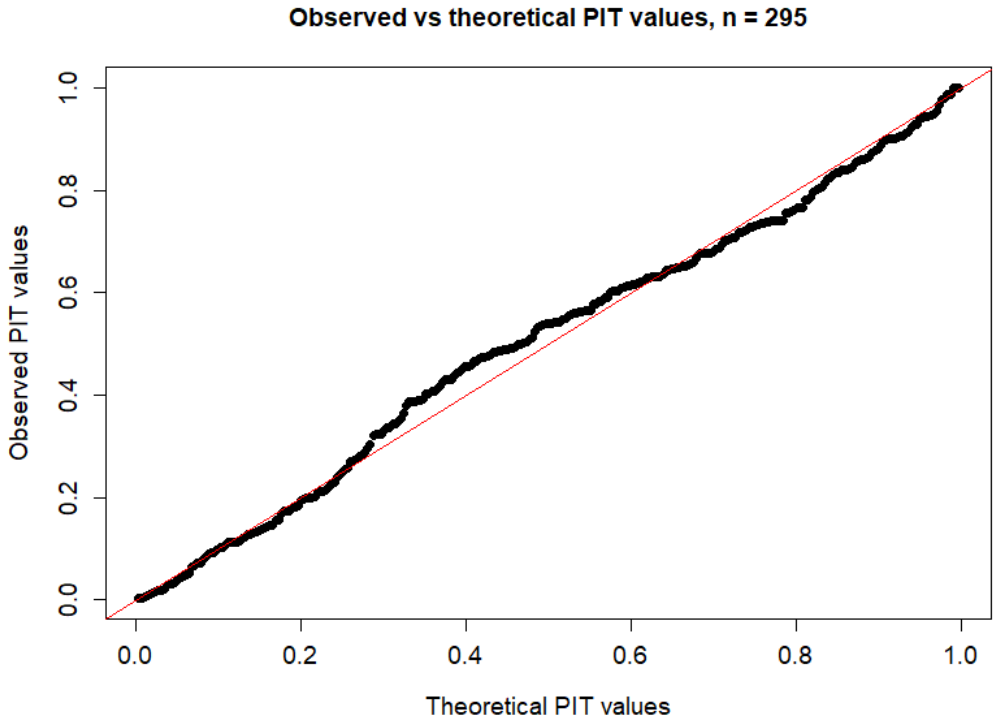
- `Kilometres` is a categorical variable that quantifies the annual mileage of a vehicle (in km) and has five discrete levels. Level 1 represents vehicles with less than 1000 km, level 2 covers the range of 1000-15,000 km, level 3 corresponds to 15,000-20,000 km, level 4 pertains to the range of 20,000-25,000 km, and level 5 comprises vehicles with a mileage exceeding 25,000 kilometers.
- `Zone` is a categorical variable that represents the geographical area of the vehicle. There are seven distinct levels including level 1: Stockholm, Goteborg, Malmo with surroundings area; 2: Other large cities with surroundings; 3: Smaller cities with surroundings in southern Sweden; 4: Rural areas in southern Sweden; 5: Smaller cities with surroundings in northern Sweden; 6: Rural areas in northern Sweden; 7: Gotland.
- `Bonus` represents the number of years that have elapsed since the last claim, plus one.
- `Make` categorizes car models into nine different levels, with levels 1-8 representing common cars and level 9 representing any other model.
- `Insured` is the number of insured in policy-years.
- `Claims` shows the number of insurance claims.
- `Payment` represents the total value of payments in Swedish Kronor.
- `perd` is the payment per claim which is `Payment` divided by `Claims`.

The insurance companies were interested in modeling `Payment` as the response variable and investigating its relationship with other explanatory variables in the dataset. This analysis can help them better understand the factors that influence `Payment` and identify any patterns or trends that can improve their decision-making processes. The dataset has also been studied in different statistical and actuarial science literature and textbooks. For example, Edward W. Frees in [57] studied this data and fitted a generalized linear model with a Gamma assumption for the response variable and a log link function. The systematic component of the model includes `Zone`, `Make`, and an offset term for the log number of claims (`Claims`). In addition, the online tutorial of the *faraway* package fits a Gamma regression model with a logarithmic link function to model the response variable. The systematic component of the model includes `make`, `bonus`, `kilometres` and an offset term for logarithmic of `Insured`. The author of the tutorial applied this model to a dataset consisting of records from the cities of Stockholm, Goteborg, Malmo, and their surrounding areas (zone = 1), which contains 295 observations.

We use the data and fitted model presented in the *faraway* tutorial to illustrate the application of goodness-of-fit tests for validating the assumptions made about the response variable. We test the null hypothesis that the response variable of this generalized linear model follows a Gamma distribution using the proposed method. The Cramer-von-Mises statistic for this test is $W_n^2 = 0.2051$ with a corresponding approximate $pvalue = 0.0052$. The result indicates that we reject the null hypothesis and thus suggests that the response variable model does not follow a Gamma distribution. The maximum likelihood estimate of the shape parameter, computed using the *gamma.shape* function from the *MASS* package in R, is 2.053 (with a standard error of 0.157).

If the null hypothesis about the distribution of the response is correct, we expect the probability integral transformed (PIT) values to be evenly distributed across the interval [0,1]. We further assess this assumption visually by plotting the PIT values of the sample versus the expected PIT values under the null hypothesis as shown in figure 4.33 below.

Figure 4.33: Real data. Probability integral transformed values of response variable vs expected probability integral transformation values.



Upon visual inspection of the probability integral transformed values, it is apparent that there is a curvature in the middle section of the plot. The null hypothesis that the response variable follows a Gamma distribution is rejected with a significant p-value at level $\alpha = 0.01$ based on the results of the goodness-of-fit test. Both theoretical and visual analysis indicate that the response variable do not conform to a Gamma distribution assumption. Therefore, it seems necessary to investigate the model assumptions further.

4.7 Conclusion and future research

In this Chapter, we reviewed results from several simulations and a real data analysis followed by some practical suggestions for goodness-of-fit tests based on empirical distribution functions. We discuss the strengths and limitations of the proposed method, offer recommendations for the application in real data analysis, and outline a future research direction.

4.7.1 Conclusion

We performed four simulations in the case of an i.i.d sample from normal distribution and four simulations in the case of an i.i.d sample from Gamma distribution. In conclusion, all simulations demonstrated an excellent performance and successfully controlled the type one

error rate at both the 0.01 and 0.05 nominal levels. We observed a strong agreement between the theoretical quantiles and sample quantiles of computed P-values, in particular for ones that are less than or equal to 0.10. Moreover, we observed that using probability integral transformed values for estimating the eigenvalues of the covariance function has a clear advantage over choosing n data points equally spaced over the [0,1] interval. As a result, we recommend the use of probability integral transformed values; this choice is implemented in our package discussed in the next chapter. We estimated the Fisher information matrix by either the variance of the score function or the negative Hessian evaluated at the MLEs. However, it is important to note that estimating the Fisher information matrix by the negative Hessian evaluated at the MLEs resulted in a highly conservative test, with P-values consistently smaller than expected based on theory. We think this happens because the off-diagonal elements of the Hessian matrix in both normal and gamma examples are zero and are not the same with their corresponding values from matrix obtained by the variance of score. As the sample size increased from $n = 50$ to $n = 250$, the test was no longer conservative. Therefore, one should expect a lower chance of rejecting the null hypothesis when the sample size is small and the Fisher information matrix is computed using the Hessian matrix.

We also examined the case of a linear model and a generalized linear model in which the distribution of the response variable may depend on some covariate. For the linear model case, as demonstrated in Section 3.9, the covariance function of the $\hat{W}_n(u)$ process does not depend on β, the vector of coefficients. Therefore, we considered only one set of values for the coefficient vector in our simulations. In this simulation, the type one error rate is well controlled at both the 0.01 and 0.05 nominal levels, and the proposed method shows excellent performance. It is worth noting that the type one error rate at a nominal level of 5 percent, when $n = 50$, is slightly inflated, but it does not change the final conclusion. Additionally, the sample quantiles of computed P-values match very well with their theoretical counterparts.

In the final simulation, we examined a generalized linear model (GLM) with inverse link function, using various numbers of explanatory variables. When using a GLM with a single explanatory variable, the sample quantiles closely align with the theoretical quantiles. The type one error rate is effectively controlled at the 0.01 level, although it is slightly elevated at the 0.05 level. This tendency is particularly evident for a sample size of $n = 50$ and $\alpha = 1$, but improves for larger sample sizes of $n = 100$ and $n = 250$. This conclusion holds true for all coefficients considered in the model.

Furthermore, we investigated the impact of including additional explanatory variables into the model. When considering $k = 2$ explanatory variables, a sample size of $n = 50$, and a shape parameter of $\alpha = 1$, the type one error rate exhibits a slight inflation above 0.01. But increasing the sample size to $n = 100$ and $n = 250$ effectively reduced the type one error rate, maintaining it at the desired level. In all different scenarios in this simulation,

the type one error rate is well controlled, with no notable disparity observed between the theoretical and sample quantiles of the computed p-values.

After adding $k = 5$ explanatory variables in the model, there is some deviation between the theoretical and sample quantiles of the computed p-values, particularly for small shape and sample sizes. Consequently, the type one error rate is slightly inflated at both nominal levels. However, increasing the sample size to $n = 250$ seems to help in this case. Finally, for a model with $k = 10$ explanatory variables, a significant discrepancy emerges between the theoretical and sample quantiles of the computed p-values. In this case, the type one error rate is inflated at both levels when the sample sizes are $n = 50$ and $n = 100$. However, increasing the sample size to $n = 250$ effectively controls the type one error rate and maintains it at the desired level.

We think the difficulty arises as a result of poor maximum likelihood estimation of coefficients in the generalized linear model. In our simulations, we see a significant difference between the true value that we used for the β vector and the estimated values that we obtained from the *glm* package in R. We used the *glm2* package in R since some convergence issues with *glm* package were reported in the literature [58]. This resolved the convergence issue but did not help with type one error rate. It is worth mentioning that we also considered $k = 2, 5, 10$ explanatory variables in the linear model simulations (results are not shown in the thesis). The conclusion remains the same as the case of a GLM simulation with $k = 2, 5, 10$ explanatory variables.

We applied our proposed goodness-of-fit test, which is based on the empirical distribution function, to a real data set. We fit a generalized linear model with a log link function and a Gamma response variable. The results of the test indicated that the assumptions about the distribution of the response variable do not seem to be true with $p = 0.0052$, which is significant at both the 0.01 and 0.05 levels. There were 11 parameters estimated in total in this model, the sample size was $n = 295$, and the maximum likelihood estimate of the shape parameter was $\hat{\alpha} = 2.053$. Comparing this with the results of our simulations in the GLM section, this would be close to simulation 5 with ten explanatory variables. The results from this simulation indicate that the sample quantiles of p-values match very well with their theoretical counterparts. The estimated type one error rate is well controlled at the 0.01 level and is somewhere between 0.07 and 0.06 at the 0.05 level. Considering the very small p-value of the test, visual investigation of probability transformed values, and comparing the results with the simulation section, we are confident that the assumption about the distribution of the response variable do not seem to be correct and further investigation is required.

For all the simulations presented in this thesis, we used $w_j = \frac{1}{m}$ as the quadrature weight for computing the eigenvalues (as discussed in Section 3.4). We decided to enhance the results of our simulations for the GLM with five or ten explanatory variables. For this purpose, we tried quadrature weights of the form $w_j = (U_{(j+1)} - U_{(j-1)})/2$, where

$U_{(j)} = F(Y_{(j)}; \theta)$, and we repeated the simulations under the same settings. The results (not shown here) clearly demonstrate an improvement. Therefore we have included these weights into our R package (see Chapter 5) for estimating the eigenvalues.

In conclusion, the proposed goodness-of-fit test based on empirical distribution function has an excellent performance for i.i.d samples from a Normal and Gamma distribution. This method can also be used for checking the response variable assumptions in a linear model or generalized linear model. Based on the conclusion from the simulations, it is important to use a sufficient sample size relative to the number of parameters in the case of GLM. The performance is affected if the sample size is small and there are too many parameters in the model. In addition, the accuracy of the maximum likelihood estimates of the linear model or generalized linear model impacts the performance and accuracy of this method. In the next section, we introduce two possible idea for the future research.

4.7.2 Future research

We considered an estimation method for the covariance function of the stochastic process in the context of goodness-of-fit tests based on empirical distribution functions. Our simulation included two cases: one involving i.i.d. samples, and the other where the response variable depends on some covariates. Our simulation specifically focused on continuous data. Furthermore, in our simulations of generalized linear models, we tested the assumptions about the distribution of the response variable and assumed the correct link function. Based on the results of our simulations, there are two areas of interest for future investigation.

First, it would be interesting to apply the method of estimating the covariance function to a sample with discrete data. The problem of the goodness-of-fit test for discrete data has been reviewed in the literature, and authors have suggested different methods. For example, the classic Pearson chi-square test and Kolmogorov-Smirnov statistic are well-known examples. Choulakian et al. [59] defined the Cramér-von-Mises and Anderson-Darling statistics for discrete data and studied the asymptotic theory, where the distribution under the null hypothesis is fully specified. Lockhart et al. [60] extended this work and included the asymptotic theory where the distribution under the null hypothesis is not fully specified, requiring estimation of some or all parameters. Additionally, Spinelli et al. [61] provided tests to examine the assumptions about the response variable in a Poisson regression model.

It should be noted that the methods just described do not use the probability integral transform because that transform does not produce uniformly distributed values. Instead, in [60], the integral defining the Cramér-von Mises statistic is replaced by a weighted sum over the possible values of the PIT.

We have not tried our proposal of estimating a covariance function in this context and the examination of the relevant limiting distribution have not been considered yet. It is worth noting that the discrete nature of the problem makes it challenging to investigate the limiting distribution of the covariance function.

As a second direction, we hope to apply the introduced goodness-of-fit method to check if the assumption about the link function in any GLM is indeed correct. For example, let $Y_1, Y_2, Y_3, \ldots, Y_n$ be a random sample from a population dependent on some covariates. One might be interested in testing whether the true expected value of Y_i, given the covariates X_i, follows a known function of β and X_i, i.e., $m(\beta^T x_i)$. Therefore, the following null hypothesis is of interest:

$$H_0 : E[Y_i|X_i] = m(\beta^T X_i)$$

The idea starts with the definition of:

$$H(u, \beta) = E[(Y - m(\beta^T X))I(\beta^T X) \leq u]$$

and estimating it with:

$$\hat{H}(u, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \Big((Y_i - m(\hat{\beta}^T X_i))I(\hat{\beta}^T X_i \leq u) \Big)$$

Now the covariance function of the following stochastic process is of interest:

$$W_n(u) = \sqrt{n} \Big(\hat{H}(u, \hat{\beta}) - H(u, \beta) \Big)$$

which could be computed in a similar way by writing the Taylor expansion as described in Section 3.4 of Chapter 3. The eigenvalues of this covariance function might be estimated to approximate the P-value.

Chapter 5

Package `gofedf` in R

5.1 Overview

We have developed an R package called `gofedf` to facilitate goodness-of-fit tests based on empirical distribution functions. This package includes functions and routines for computing p-values for two goodness-of-fit tests: Cramér-con Mises and Anderson-Darling. As mentioned earlier, the computation of p-values relies on estimating the covariance function of a stochastic process. Our implementation utilizes the sample-based covariance estimation method described in Section 3.5. The package `gofedf` offers a small number of tests for specific distributional assumptions and a suite of functions to handle a wide range of other models. In particular, there are functions to test independent and identically distributed samples for the Normal or Gamma distributions. Additionally, `gofedf` provides tools for checking the normality assumptions of residuals in linear models and for evaluating whether the response variable in a generalized linear model follows a Gamma distribution with any valid link function. The most significant feature of the package is to provide routines for formal model evaluation in general likelihood models. Section 3.5 describes this feature in more detail. The package is now published on the CRAN repository and is available for download [62].

5.2 Main functions in the package

In this section, we briefly review the main functions included in the package. The first and second section review the functions for testing the hypothesis that a sample is drawn from a Normal or a Gamma distribution. The next two sections consider linear models with a Gaussian error assumption and generalized linear models with a Gamma response variable. In these cases, the assumption is that the expected value of response depends on some covariates. A more detailed review is available in the vignette page of the package [62].

5.2.1 Normal distribution

To test an i.i.d. sample against the normal distribution, one can use the *testNormal* function in the package. The function requires a vector of numeric values as input. By default, the probability integral-transformed (PIT) values of the sample are used as a grid to compute the covariance function. Alternatively, one can define the number of equally spaced points to be used for computing the covariance function. By default, the variance of the score function is used to estimate the Fisher information matrix. However, it is also possible to use the Hessian matrix computed directly from the second partial derivatives. The function computes the maximum likelihood estimates (MLE) of the mean and standard deviation, as well as the score function. Finally, it returns the requested statistic, i.e., Cramer-von Mises, Anderson-Darling, or both statistics, along with the corresponding p-value. The following lines of code show an example of testing a sample against the Normal distribution by `gofedf` package:

```
> library(gofedf)

> set.seed(123)
> simdata = rnorm(n = 50)

> testNormal(x = simdata, method = 'both')

$Statistics
Cramer-von-Mises Statistic  Anderson-Darling Statistic
                 0.03781322                  0.21797039

$pvalue
pvalue for Cramer-von-Mises test          Anderson-Darling test
                        0.6766974                      0.9426823
>
```

5.2.2 Gamma distribution

The package can also be used to test an i.i.d sample against Gamma distribution by calling the function *testGamma*. This function uses the function *gamma.shape* from the *MASS* package to compute the maximum likelihood estimate of the shape parameter for the Gamma distribution [63]. The input requirements for the function are the same as before and users have control over the estimation of covariance function and Fisher information matrix. The probability integral-transformed (PIT) values of the sample can be used as a grid to compute the covariance function. This is the default setting. It is also possible to define the number of equally spaced points to be used for computing the covariance func-

tion. In addition, Fisher information can be estimated by the variance of the score (default value) or Hessian matrix. The following code gives an example of testing a sample against the Gamma distribution by *gofedf* package:

```
> library(gofedf)

> set.seed(123)
> simdata = rgamma(n = 50, shape = 1)

> testGamma(x = simdata, method = 'both')
$Statistics
Cramer-von-Mises Statistic Anderson-Darling Statistic
                 0.03313641                 0.20812957

$pvalue
pvalue for Cramer-von-Mises test          Anderson-Darling test
                        0.7367917                      0.6732966
>
```

5.2.3 Linear models

The *gofedf* package can also be used when the expected value of each observation depends on covariates. In the first example, we consider a linear model with constant variance and normal error assumptions. The main function for this case is *testLMNormal* which can be used to test the normality assumption. You can provide a matrix of covariates, where rows represent observations and columns represent explanatory variables. Alternatively, you can pass an object of class "linear model" for convenience, typically returned by the function *lm* in R. If you choose to use the latter option, make sure to set the x and y arguments to *TRUE* in the *lm* function to return the design matrix and response variable. The other arguments to the function remain the same as previous examples.

In the following example, we begin by randomly generating a set of coefficients, a matrix containing explanatory variables, and some error terms from a standard Normal distribution. The response variable is computed accordingly, and then we apply the *testLMNormal* function.

```
> library(gofedf)
> set.seed(123)
> n = 50
> p = 5
> b = runif(p)
> X = matrix( runif(n*p), nrow = n, ncol = p)
```

```
> e = rnorm(n)
> y = X %*% b + e
>
> testLMNormal(x = X, y, method = 'cvm')

$Statistic
[1] 0.02285164

$pvalue
[1] 0.9089065
>
> lm.fit = lm(y ~ X, x = TRUE, y = TRUE)
> testLMNormal(fit = lm.fit, method = 'cvm')
$Statistic
[1] 0.02285164

$pvalue
[1] 0.9089065
>
```

5.2.4 Generalized linear models

The second example involves a generalized linear model with a Gamma response variable. The package is designed to assess assumptions related to the response variable. Currently, it supports only the Gamma distribution, but we have plans to include more common distributions in future versions.

In this case, the main function is *testGLMGamma()*, which takes arguments we now describe. Similar to the linear model example, you can provide either a matrix containing explanatory variables and a response vector, or an object of class generalized linear model returned from the *glm* or *glm2* function. If you choose the latter option, the requirement to return the design matrix and response variable remains the same.

The "l" argument is a character vector that indicates the link function to be used. For the Gamma distribution, valid choices are 'log' and 'inverse' link functions. The 'start.value' parameter serves as the starting point for the *glm* or *glm2* functions. This value is crucial for the iteratively reweighted least squares (IRLS) algorithm, which is used to compute the maximum likelihood coefficients. We also need to estimate shape parameter. In the code, the MLE of shape parameter is estimated by *gamma.shape* function from *MASS* package. The other arguments for the function are the same as in previous examples.

```
> library(gofedf)
```

```
> library(glm2)
> set.seed(123)
> n = 50
> p = 5
> X = matrix(rnorm(n*p, mean = 10, sd = 0.1), nrow = n, ncol = p)
> b = runif(p)
> e = rgamma(n, shape = 3)
> y = exp(X %*% b) * e
> testGLMGamma(x=X, y, l = 'log', method = 'cvm')
$Statistic
[1] 0.0870493

$pvalue
[1] 0.1896532

$converged
[1] TRUE
>
> glm.fit <- glm2(y ~ X, family=Gamma(link = 'log'),
x=TRUE, y=TRUE)
> testGLMGamma(fit = glm.fit, l = 'log')
$Statistic
[1] 0.0870493

$pvalue
[1] 0.1896532

$converged
[1] TRUE
```

During our simulation study in Chapter 4 and while fitting the model using the *glm* function from the R *stats* package, we encountered convergence difficulties with the iteratively reweighted least squares (IRLS) algorithm in some Monte Carlo samples. Specifically, our examination showed that the problem arose during the optimization step when step-halving should have been invoked but was not. Consequently, the algorithm produced negative values for the linear predictor, leading to a failure in the convergence. To address this issue, we considered using the *glm2* function from the *glm2* package [58] for computing the maximum likelihood estimation of model coefficients. The estimation process in *glm2* is similar to that of the *glm* function in the R *stats* package but includes modifications to ensure greater stability in convergence. This includes employing a more rigorous step-halving than

that found in the *glm* function to ensure that the deviance decreases during each iteration. Further details about this algorithm can be found in [58]. While we cannot give strong advice about the choice of starting values with real data we were able to give good starting values in our simulation by use of the true parameter values.

5.3 Example: Inverse Gaussian Distribution

In addition to the functions mentioned in the previous sections, the package can be applied to conduct goodness-of-fit tests based on empirical distribution function statistics for any general likelihood model. The models considered will have independent but not necessarily identically distributed variables $Y_1, \ldots, Y_n$ and a p-dimensional parameter vector. The only requirements for users to supply are as follows: (i) a function to compute the maximum likelihood estimate of the parameters (ii) a function to compute the probability integral transformed (PIT) values of the responses (iii) a function that calculates the matrix S whose ith row contains the component of the score due to observation Y_i.

We provide an example of an inverse Gaussian distribution with constant known weights to illustrate the concept. Let's consider a sample of size n drawn from an inverse Gaussian distribution with constant known weights, denoted as w_i, and characterized by the following probability distribution function:

$$f(Y_i; \mu, \lambda) = \sqrt{\frac{\lambda w_i}{2\pi Y_i^3}} \exp\left(-\frac{\lambda w_i (Y_i - \mu)^2}{2\mu^2 Y_i}\right)$$

where w_i are constant and known weights. Therefore we can write the likelihood function as:

$$\begin{aligned} L(\mu, \lambda) &= \prod_{i=1}^{n} f(Y_i; \mu, \lambda) \\ &= \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \left(\prod_{i=1}^{n} \frac{w_i}{Y_i^3}\right)^{\frac{1}{2}} \exp\left(\frac{\lambda}{\mu}\sum_{i=1}^{n} w_i - \frac{\lambda}{2\mu^2} - \frac{\lambda}{2\mu^2}\sum_{i=1}^{n} w_i Y_i - \frac{\lambda}{2}\sum_{i=1}^{n}\frac{w_i}{Y_i}\right) \end{aligned}$$

It is easy to verify that the maximum likelihood estimates of μ and λ are:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} \qquad \hat{\lambda} = \frac{n}{\sum_{i=1}^{n}(\frac{w_i}{Y_i} - \frac{w_i}{\hat{\mu}})}$$

The score is a matrix with n rows and two columns with the following elements:

$$\left[\frac{-\lambda}{\mu^2} w_i + \frac{\lambda}{\mu^3} w_i Y_i \qquad \frac{1}{2\lambda} + \frac{w_i}{\mu} - \frac{w_i Y_i}{2\mu^2} - \frac{w_i}{2Y_i}\right]$$

For illustrative purposes, the package includes functions to compute the maximum likelihood estimates of parameters (i.e. $\hat{\mu}$ and $\hat{\lambda}$), score functions, and PIT values. These functions are named *inversegaussianMLE*, *inversegaussianScore*, and *inversegaussianPIT* in the package.

The *testYourModel* function from the package can be used to apply a goodness-of-fit test based on the empirical distribution function for any general likelihood model. For instance, in the case of the Inverse Gaussian model we've just described, you can apply the test using the following lines of code. To simulate the data, we randomly generate weights from a uniform distribution over the interval [5,6], which are then scaled. We also generate a sample of size $n = 50$ from an Inverse Gaussian distribution using the *statmod* package in R. The mean of the distribution, μ, is set to two. The shape parameter in this case depends on the weight of each observation. We calculate the Maximum Likelihood Estimates (MLE) of the model parameters, the score matrix, and Probability Integral Transform (PIT) values by calling their respective functions. Finally, we invoke the *testYourModel* function to compute the test statistic and P-value. It's important to note that the only requirements for this process are a vector of observations, a score function, and PIT values.

```
> library(gofedf)
> set.seed(123)
> n = 50
> weights = runif(n, min = 5, max = 6)
> weights = weights / sum(weights)
> mio = 2
> lambda = 2
> y = statmod::rinvgauss(n, mean=mio, shape=lambda*weights)
> thetahat = inversegaussianMLE(obs=y, w=weights)
> score.matrix = inversegaussianScore(obs=y, w=weights, mle=thetahat)
> pit.values = inversegaussianPIT(obs=y,w=weights,mle=thetahat)
> testYourModel(x = y, pit = pit.values, score = score.matrix)
$Statistic
Cramer-von-Mises Statistic
                 0.03292151

$pvalue
[1] 0.8436222
>
```

Bibliography

[1] E.A. Thompson. Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

[2] L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, 62(5):1198–1211, 5 1998.

[3] C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet*, 54(3):535–543, 3 1994.

[4] M. Boehnke. Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Am J Hum Genet*, 55(2):379–390, 8 1994.

[5] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, 9 2007.

[6] Sharon R. Browning and Elizabeth A. Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, 2012.

[7] Jurg Ott, Jing Wang, and Suzanne M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5):275–284, 2015.

[8] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, 12(5):1–22, 2016.

[9] Dominic Nelson, Jerome Kelleher, Aaron P. Ragsdale, Claudia Moreau, Gil McVean, and Simon Gravel. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genetics*, 16(5):1–12, 2020.

[10] Albert Tenesa, Pau Navarro, Ben J. Hayes, David L. Duffy, Geraldine M. Clarke, Mike E. Goddard, and Peter M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520–526, 2007.

[11] Catarina D. Campbell, Jessica X. Chong, Maika Malig, Arthur Ko, Beth L. Dumont, Lide Han, Laura Vives, Brian J. O'Roak, Peter H. Sudmant, Jay Shendure, Mark Abney, Carole Ober, and Evan E. Eichler. Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, 44(11):1277–1281, 2012.

[12] Henry R. Johnston and David J. Cutler. Population demographic history can cause the appearance of recombination hotspots. *American Journal of Human Genetics*, 90(5):774–783, 2012.

[13] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S. A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 7 2012.

[14] Seunggeun Lee, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani, Mark M. Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2):224–237, 2012.

[15] Seunggeung Lee, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, 95(1):5–23, 2014.

[16] Bo Eskerod Madsen and Sharon R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), 2009.

[17] Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, 2011.

[18] Charith B. Karunarathna and Jinko Graham. PerfectphyloR: An R package for reconstructing perfect phylogenies. *BMC Bioinformatics*, 20(1):1–9, 2019.

[19] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.

[20] Kelly M. Burkett, Brad McNeney, Jinko Graham, and Celia M.T. Greenwood. Using gene genealogies to detect rare variants associated with complex traits. *Human Heredity*, 78(3-4):117–130, 2014.

[21] L. Beckmann, D. C. Thomas, C. Fischer, and J. Chang-Claude. Haplotype sharing analysis using mantel statistics. *Human Heredity*, 59(2):67–78, 2005.

[22] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

[23] Nathan Mantel. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(2):209–220, 1967.

[24] Julie Josse and Susan Holmes. Measuring multivariate association and beyond. *Statistics Surveys*, 10(0):132–167, 2016.

[25] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[26] Gilles Guillot and François Rousset. Dismantling the Mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344, 2013.

[27] Jerome Kelleher, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

[28] Charith B. Karunarathna and Jinko Graham. Using Gene Genealogies to Localize Rare Variants Associated with Complex Traits in Diploid Populations. *Human Heredity*, 83(1):30–39, 2018.

[29] A. Derkach, J. F. Lawless, and L. Sun. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol*, 37(1):110–121, 1 2013.

[30] S. P. Dickson, K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein. Rare variants create synthetic genome-wide associations. *PLoS Biol*, 8(1):e1000294, 1 2010.

[31] J. E. Bailey-Wilson and A. F. Wilson. Linkage analysis in the next-generation sequencing era. *Hum Hered*, 72(4):228–236, 2011.

[32] C. Francks, F. Tozzi, A. Farmer, J. B. Vincent, D. Rujescu, D. St Clair, and P. Muglia. Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13. *Mol Psychiatry*, 15(3):319–325, 3 2010.

[33] Peter E. Smouse, Jeffrey C. Long, and Robert R. Sokal. Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. *Systematic Biology*, 35(4):627–632, 1986.

[34] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–246, 2012.

[35] M. Bouaziz, J. Mullaert, B. Bigio, Y. Seeleuthner, J. L. Casanova, A. Alcais, L. Abel, and A. Cobat. Controlling for human population stratification in rare variant association studies. *Sci Rep*, 11(1):19015, 09 2021.

[36] J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets. *Nat Genet*, 51(9):1330–1338, 09 2019.

[37] L. Speidel, L. Cassidy, R. W. Davies, G. Hellenthal, P. Skoglund, and S. R. Myers. Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Mol Biol Evol*, 38(9):3497–3511, 08 2021.

[38] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[39] D. A. Darling. The Kolmogorov-Smirnov, Cramér-von Mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838, 1957.

[40] D'Agostino Ralph B. and Michael Stephens. *Goodness-of-fit-techniques*. Routledge, 1986.

[41] H. Cramér. *Mathematical methods of statistics*. Princeton University Press, 1946.

[42] N. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou*, 2(2):3–14, 1939.

[43] A.N. Kolmogorov. Grundbegriffe der wahrscheinlichkeitsrechnung. *Julius Springer*, 4:83–91, 1933.

[44] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[45] R. A. Lockhart and M. A. Stephens. Tests of fit for the von Mises distribution. *Biometrika*, 72(3):647–652, 1985.

[46] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[47] Frank J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[48] G.R. Shorack and J.A. Wellner. *Empirical Processes with Applications to Statistics*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2009.

[49] M. A. Stephens. Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters. *The Annals of Statistics*, 4(2):357 – 369, 1976.

[50] R. Lockhart and T. Swartz. Computing asymptotic p-values for EDF tests. *Statistics and Computing*, 2(3):137–141, 1992.

[51] J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426, 1961.

[52] R. W. Farebrother. Algorithm AS 204: The distribution of a positive linear combination of chi-squared random variables. *Journal of the Royal Statistical Society*, 33(3):332–339, 1984.

[53] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

[54] Gemai Chen and Richard A. Lockhart. Weak convergence of the empirical process of residuals in linear models with many parameters. *The Annals of Statistics*, 29(3):748 – 762, 2001.

[55] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[56] Marc Hallin and Jean-François Ingenbleek. The Swedish automobile portfolio in 1977: a statistical study. Technical report, ULB – Universite Libre de Bruxelles, 1983.

[57] E.W. Frees. *Regression Modeling with Actuarial and Financial Applications*. International Series on Actuarial Science. Cambridge University Press, 2009.

[58] Ian C. Marschner. glm2: Fitting generalized linear models with convergence problems. *R J.*, 3:12, 2011.

[59] V. Choulakian, R. A. Lockhart, and M. A. Stephens. Cramér-von Mises statistics for discrete distributions. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 22(1):125–137, 1994.

[60] Richard A Lockhart, John J Spinelli, and Michael A Stephens. Cramér–von Mises statistics for discrete distributions with unknown parameters. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 125–133, 2007.

[61] John J. Spinelli, Richard A. Lockhart, and Michael A. Stephens. Tests for the response distribution in a Poisson regression model. *Journal of Statistical Planning and Inference*, 108(1):137–154, 2002.

[62] Richard Lockhart and Payman Nickchi. *gofedf: Goodness of Fit Tests Based on Empirical Distribution Functions*, 2023. R package version 0.1.0.

[63] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

Appendix A

Hybrid simulation

We used the *msprime* package to simulate sequences and their ancestry. This tool is efficient for simulating data under various population genetic models and is widely employed in population-genetic studies. We will now provide a brief description of two fundamental models that *msprime* can be used to simulate data from: 1) Wright-Fisher model, and 2) the standard coalescent theory.

The Wright-Fisher model describes the genetic dynamics in a population with constant size and non-overlapping generations and random mating. In each generation, individuals contribute offspring to the next generation through random sampling. For example, the genes in the next generation choose their parents with a random draw among a pool of $2N$ genes from the current generation ($2N$ is used for a haploid population). Therefore, the probability of two genes having the same parent (coalesce to the same gene) in the previous generation is $\frac{1}{2N}$. In general, the probability that two genes coalesce at generation t in this model is:

$$P(T = t) = (1 - \frac{1}{2N})^{t-1} \frac{1}{2N} \qquad t = 1, 2, 3, \ldots$$

The standard coalescent theory is a backward-in-time model used in population genetics to study the common ancestry of gene copies in a sample of individuals. It provides a theoretical framework for understanding how gene lineages coalesce backward in time until they share a common ancestor. This model is widely used to investigate the genetic history and diversity within populations. In standard coalescent theory, time is measured in a continuous scale instead. The probability that two genes coalesce after generation t in this case is:

$$P(T > t) = \left(1 - \frac{1}{2N}\right)^{t} \qquad t \geq 0$$

For large values of N we can approximate this probability using an exponential distribution as follows:

$$P(T > t) = e^{-\frac{t}{2N}}$$

Note that the Wright-Fisher model describes the changes in allele frequencies over generations in a population with random mating, discrete generations, and non-overlapping generations. On the other hand, the coalescent is a backward-in-time model that focuses on the common ancestry of a sample of genes within a population. It provides a way to study how gene lineages coalesce or merge backward in time until they share a common ancestor.

The approximations of coalescent theory break down when the number of simulated sequences (sample size) is large relative to the effective population size. In this case, the large-scale coalescent simulations produce unrealistic relatedness among simulated sequences [9]. For a coalescent with recombination, this produces an overabundance of simulated ancestors for each sequence, while this is not the case for Wright-Fishser model. To correct this and avoid unreasonable relatedness among sequences, there is a hybrid strategy included in *msprime* package to use backward discrete Wright-Fisher model to simulate sequences up to a certain generation back in the time. Then the simulation continues from there by the standard coalescent with recombination. More details about backward discrete Wright-Fisher model (refereed to as hybrid simulation) are available at *msprime* documentation page and also described in [9].

Appendix B

Allele frequency spectrum(AFS)

Figure B.1 shows the variant allele frequency spectrum of example dataset.

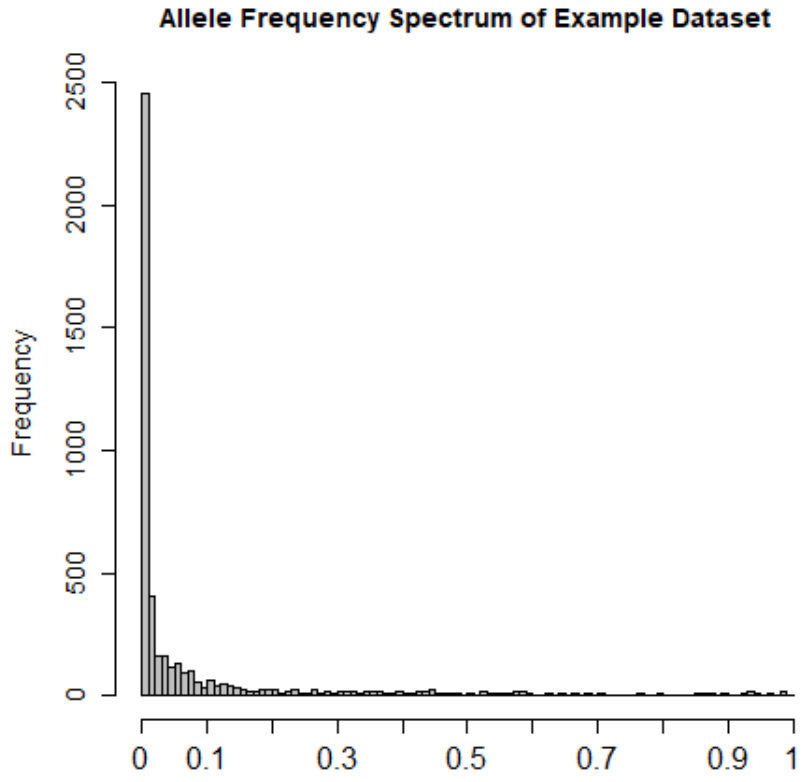


Figure B.1: Variant allele frequency spectrum of the example dataset.

Appendix C

Causal variant selection

The disease-trait model specifies that $logit(P(D=1|G)) = \beta_0 + \beta_1 \sum_{j=1}^{m} G_j$ so that

$$P(D=1|G) = \begin{cases} \frac{exp(\beta_0)}{1+exp(\beta_0)}, & \sum_{j=1}^{m} G_j = 0 \\ \frac{exp(\beta_0+\beta 1)}{1+exp(\beta_0+\beta 1)}, & \sum_{j=1}^{m} G_j = 1. \end{cases}$$

As the cSNVs are rare, we make the simplifying assumption that each individual carries at most one copy of a cSNV. Then the population prevalence of the disease is

$$P(D) \approx P(D|\sum_{j=1}^{m} G_j = 0)P(\sum_{j=1}^{m} G_j = 0) + P(D|\sum_{j=1}^{m} G_j = 1)P(\sum_{j=1}^{m} G_j = 1).$$

Setting the prevalence to 0.05, we obtain

$$0.05 \approx \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \times \frac{N_0}{3100} + \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \times \frac{N_1}{3100},$$

where N_0 and N_1 are respectively the number of individuals in the population that carry zero and one copy of a cSNV.

Setting $\beta_0 = -10$, $\beta_1 = 16$ and $N_1 = 3100 - N_0$, we obtain $N_0 \approx 2945$ and $N_1 \approx 155$. We select 15 variants of roughly equal frequency in the population such that their total number of copies is around 155. Thus each cSNV has a frequency of about $155/15 = 10.33$ in the population of 6200 sequences. We also have $N_0 + N_1 = 3100$ so we may solve for N_0 and N_1 to get $N_1 \approx 155$, and $N_0 \approx 2945$. We select 15 variants of roughly equal frequency in the population such that their total number of copies is around 155.

Appendix D

Sequence distances on partition

Figure D.1 shows an example partition with 4 sequences labeled as 1 to 4 and distances assigned from the *rdistMatrix()* function in the *perfectphyloR* package. As illustrated in the figure, the distance between a sequence and its ancestral node is one and the distance between two neighboring nodes that descend from the same most-recent common ancestral node is two.

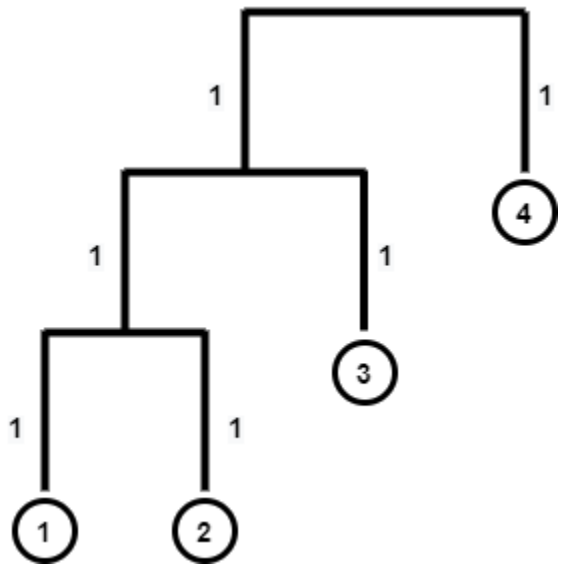


Figure D.1: Distance between sequences assigned from the *rdistMatrix()* function in the *perfectphyloR* package. 1, 2, 3, and 4 are sequences.

Appendix E

A worked example for calculation of GNN

To illustrate the GNN calculation, we consider a worked example of four sequences of length 10 kbp, as shown in Figure E.1. In the figure, the subregion from 0 kbp to 6 kbp is spanned by partition A, and the rest of the region by partition B. In other words, when we reconstruct the partition at each SNV position within the first 6 kbp, we have only one partition structure (partition A), and the rest of the region has the structure of partition B. The GNN proportion for a given target sequence can be computed as follows. For example, suppose we choose sequence 1 as our target sequence. Starting from sequence 1, we go upward in the reconstructed partition until we find the first internal node. We call this internal node a. All the sequences that descend from a, excluding the target sequence 1, are the genealogical nearest neighbors of sequence 1. The GNN proportion for the target sequence is the proportion of these neighbours that are case sequences within the clade below a. We repeat this calculation for all the target sequences and arrange the proportions in a vector indexed by target sequence. These proportions for the genomic region labeled as A comprise a vector, G_A. For example, in partition A, the GNN proportion for the four target sequences are: $1/1 = 1$ for sequence 1, $1/1 = 1$ for sequence 2, $2/2 = 1$ for sequence 3, and $2/3 = 0.67$ for sequence 4. Thus $G_A = [1,\ 1,\ 1,\ 0.67]$. Similarly in partition B we obtain $G_B = (1,\ 1,\ 0.67,\ 1)$.

Once we compute G_A and G_B, these vectors are weighted by the proportion of genomic region spanned by their respective partitions. Since partition A spans 60% of the total region of 10 kbp, the corresponding weight, $W_A = 0.6$. Similarly $W_B = 0.4$. We are now ready to compute the average GNN proportion by taking the weighted average of all these proportions in both partitions. By taking the weighted average, we assign more weight to the partitions corresponding to long physical lengths of sequence than partitions corresponding to short physical lengths of sequence. This weighted average summarizes the proportion of nearest neighbours to the target sequence that are case sequences. In our example, the average GNN proportion can be computed as:

$$\frac{G_A W_A + G_B W_B}{W_A + W_B} = [1,\ 1,\ 0.868,\ 0.802]$$

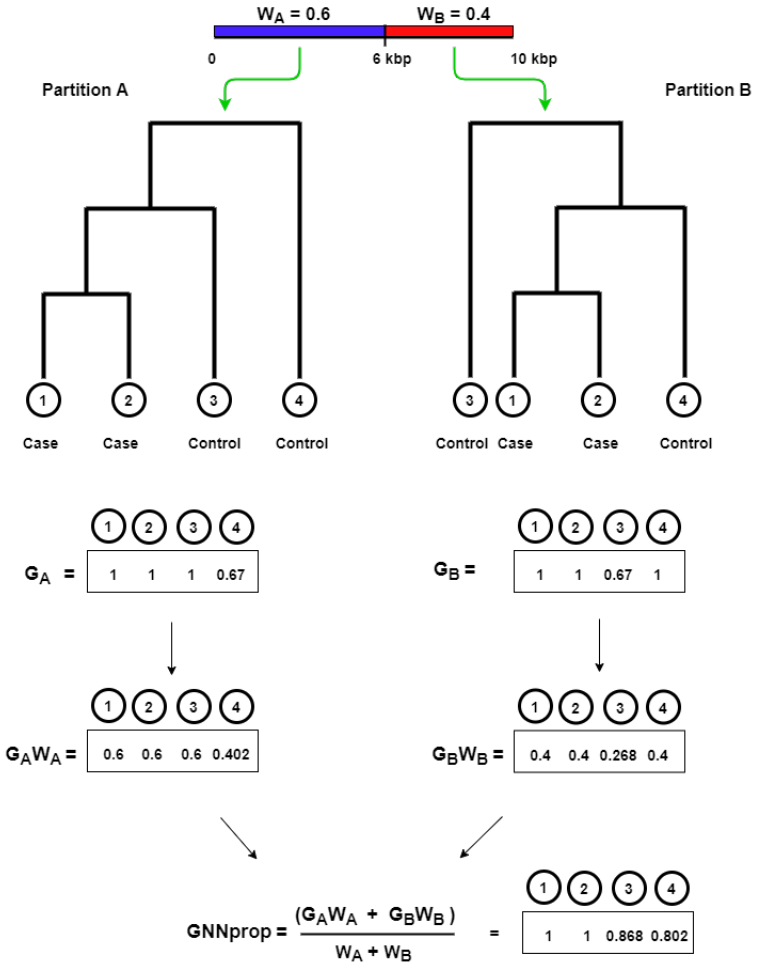


Figure E.1: A worked example to illustrate the calculation of average GNN proportion. Four sequences are considered, labeled with circles 1 to 4, over a 10 kbp region.

Appendix F

The estimated type-I error rate

Table F.1: The estimated type-I error rate or proportion of 500 null datasets that incorrectly reject the null hypothesis ($\hat{p}$) and associated approximate 95% confidence interval. Four methods are compared: 1) Fisher's exact test (FET), 2) SKAT-O, 3) distance correlation (dCor), and 4) Mantel.

Method	$\hat{\mathbf{p}}$	**Approximate 95% CI**	
		Lower bound	**Upper bound**
FET	0.040	0.022	0.057
SKAT-O	0.042	0.024	0.060
dCor	0.038	0.021	0.055
Mantel	0.048	0.029	0.067