A pedigree-transmission likelihood for multiplex families

by

Tianyu Yang

B.Sc., Queen's University, 2021

Project Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in the

Department of Statistics and Actuarial Science

Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2023

Declaration of Committee

Name: **Tianyu Yang**

Degree: **Master of Science**

Thesis title: **A pedigree-transmission likelihood for multiplex families**

Committee:

Chair: Liangliang Wang
Associate Professor, Statistics and Actuarial Science

Jinko Graham
Supervisor
Professor, Statistics and Actuarial Science

Rhonda Rosychuk
Committee Member
Adjunct Professor, Statistics and Actuarial Science

Brad McNeney
Examiner
Associate Professor, Statistics and Actuarial Science

Abstract

Family-based studies of a genetically inherited disease ascertain so-called *multiplex* pedigrees with several disease-affected members. A parameter of interest in these studies is the probability that a parent transmits a given genetic variant to their child. I describe a likelihood method for estimating the transmission probability of a rare genetic variant based on DNA sequencing data in affected relatives. These pedigree-transmission likelihoods are based on data for the presence or absence of the variant in the affected relatives. I describe how to implement the likelihoods using software for Bayesian Networks in R. The ideas are illustrated with several example pedigrees of various shapes and sizes.

Keywords: Multiplex Pedigrees; Likelihood Curves; Transmission Probability; Bayesian Networks

Table of Contents

List of Figures

Chapter 1

Introduction and background

Medical genetic studies typically collect *pedigrees* or extended families. Pedigrees are sampled to contain multiple affected relatives to increase the chances that the disease has a genetic cause. *Multiplex pedigrees* contain several disease cases, presumably because a disease-influencing genetic variant is *segregating* or being transmitted from parent to offspring. Multiplex pedigrees are often complex and contain more than two generations of relatives. Family members can be categorized as affected or not affected by the disease and as founders or non-founders. Founders are the individuals for whom parental information is unavailable. By contrast, non-founders have information available about parents[1]. For example, the multiplex pedigree of Figure 1.1 has three affected (IDs 7,8,9) and six unaffected (IDs 1-6) members, four founders (IDs 1,2,4,6), and five non-founders (IDs 3,5,7,8,9).

Many medical genetic studies of multiplex pedigrees have revealed genetic variants associated with disease susceptibility. Understanding the biological mechanisms underpinning these disease associations is potentially useful for developing disease therapies. In this project, we focus on rare variants (RVs) because our work is motivated by sequencing studies of familial cancer. Rare DNA variants are those with frequencies of less than 1 percent in the population[2]. Sequencing allows investigators to interrogate rare genetic variations that cannot be accessed by traditional chip-based genotyping methods. To economize on sequencing costs, the living affected members of a pedigree are frequently the only individuals approached for DNA collection. The idea is to focus the investigation on the affected relatives and look for shared genetic variants among them.

This project develops transmission-probability likelihoods to evaluate the association between a RV and heritable disease in multiplex pedigrees. Unfortunately, in complex pedigree structures, calculating the likelihood of a transmission probability,

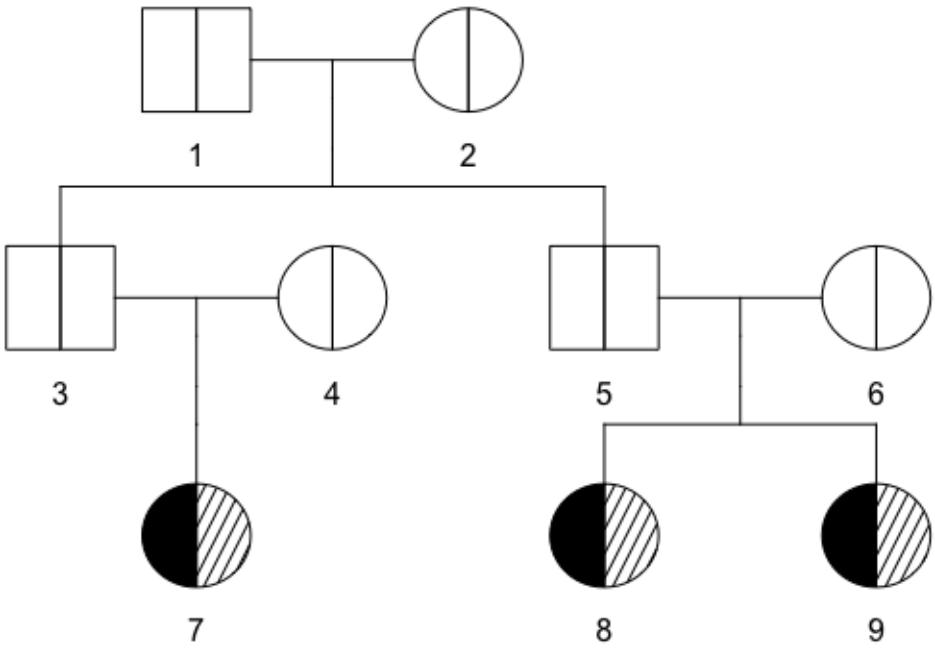

Figure 1.1: Male individuals are indicated by squares and females by circles. Co-parents are connected by a horizontal line. Disease-affected individuals are indicated by solid coloring on the left and individuals with DNA available (i.e. genetic data) are indicated by hashing on the right.

τ, from parent to offspring is not straightforward. The likelihood of τ arises from the joint probability of an RV configuration in the disease-affected relatives with sequencing data. We describe how to calculate this likelihood by casting the pedigree as a Bayesian network. We find Bayesian networks to be a useful and convenient tool for calculating joint probabilities of RV configurations and therefore likelihoods.

As families have been ascertained for multiple disease cases, we assume there is no protective genetic variant, so that the transmission probability of a genetic variant must be $\tau \geq 1/2$. When $\tau = 1/2$, the genetic variant is randomly transmitted to disease-affected family members according to Mendel's laws and therefore unassociated with the disease[3]. When $\tau > 1/2$, the genetic variant is preferentially trans-

mitted to disease-affected family members and therefore associated with the disease. Since the genetic variant is rare, we assume it comes into the pedigree through at most a single founder.

To calculate likelihoods, we rely on Bayesian networks which are a special case of *graphical independence networks* or probabilistic graphical models. Graphical independence networks model the joint distribution of a set of variables in terms of their conditional dependencies in a graph and are widely used in many domains[4]. For example, any physical and biological processes can be naturally modeled as a network of causal influences[5].A graphical independence network consists of nodes (variables) and components known as *conditional probability tables* (CPTs). CPTs specify the local conditional dependencies between variables in the network.

Bayesian networks are a type of probabilistic graphical model comprised of nodes and directed edges and are used to compute joint probabilities with Bayes' rule. A Bayesian network corresponds to a *Directed Acyclic Graph* (DAG) where no self-connection or loop is allowed[6]. Each node in a Bayesian network represents a unique random variable, and each directed edge represents the conditional probability of the random variable the edge enters given information on the random variable from which the edge exits. Bayesian networks models can be used to learn from data, then used for inference to estimate the probabilities of causal or subsequent events.

The natural structure of a pedigree, and in particular the conditional independencies of genetic data, permits it to be cast as a Bayesian network[7]. We take the RV status of individuals in a pedigree as the variables or nodes in a network so that the pedigree can be seen as a directed graph with children's RV status depending on parents'. Computational methods for building graphical models may then be applied to construct the probabilistic network and calculate the joint probability of the observed genetic data more efficiently. We use this probabilistic network to compute conditional probabilities and infer the variant transmission probability from the pedigree.

Chapter 2

Likelihood

2.1 Formulas

This section describes how to calculate the likelihood of a RV transmission probability in a multiplex pedigree, given data on the RV status of the affected members with available DNA and assuming that a single founder introduced the RV into the pedigree. To obtain the likelihood, we view the pedigree as a directed graph in which the RV status of family members represents the random variables.

Being rare, the RV is assumed to enter the pedigree through at most a single founder. Therefore, in a pedigree with no inbreeding, the possible values for RV status are 0 (no copies) and 1 (a single copy). We take the prior probability that any founder is the one that introduced the RV to be uniform over all the pedigree founders. The likelihood is obtained from the conditional probability of the RV configuration, C, for affected pedigree members with available DNA, given that exactly one of the founders introduced a single copy of the RV into the pedigree.

The event in which exactly one founder introduced the RV is a union of the mutually exclusive events, F_i, that each of the founders, i, introduced the RV. We define these events to be mutually exclusive because the chance of more than one founder carrying a RV is essentially zero. The probability of the RV configuration C in the affected pedigree members with available DNA is thus

$$\begin{aligned} P(C|\cup_i F_i) &= \frac{P(C,\cup_i F_i)}{P(\cup_i F_i)} \\ &= \frac{\sum_i P(C,F_i)}{\sum_i P(F_i)} \\ &= \sum_i P(C|F_i)\frac{P(F_i)}{\sum_j P(F_j)}. \end{aligned}$$

Founders are assumed to have the same prior probability of introducing the RV; i.e., $P(F_1) = P(F_2) = P(F_3) = \cdots$. Thus $P(F_i)/\sum_j P(F_j)$ on the right-hand side of the above equation is equal to one over the number of founders. The conditional probability, $P(C|F_i)$, of the RV configuration given that founder i introduced the RV is obtained by *path counting* the number of independent transmissions connecting founder i to the affected members.

For example, in the example pedigree of Figure 1.1, the affected members with available DNA are IDs 7, 8 and 9. Hence, the configuration vector C has a first element corresponding to the RV status of the individual with ID 7, a second element corresponding to the RV status of the individual with ID 8 and a third element corresponding to the RV status of the individual with ID 9. In this pedigree, an RV from the founder with ID 1 can reach the affected pedigree member with ID 7 through two independent transmissions from the pedigree member with ID 1 to the pedigree member with ID 3 and then from the pedigree member with ID 3 to the pedigree member with ID 7. Similarly, an RV from the founder with ID 1 can reach the pedigree members with IDs 8 and 9 through one independent transmission to the pedigree member with ID 5 and then two independent transmissions from the pedigree member with ID 5 to the pedigree member with IDs 8 and 9. Therefore, $P(C = (1,1,1)|F_1) = \tau^{2+1+2}$ and $P(C = (1,1,0)|F_1) = \tau^{2+1+1}(1-\tau)$, for example. In the calculation for $P(C = (1,1,0)|F_1)$, we count two transmissions of the RV from the individual with ID 1 to the individual with ID 7, one from the individual with ID 1 to the individual with ID 5, one from the individual with ID 5 to the individual with ID 8, and one transmission of the other variant (i.e. the non-RV) in the individual with ID 5 to the individual with ID 9. Similarly, the likelihood for $\tau = 3/4$ when the data are $C = (1,1,1)$ is:

$$\begin{aligned} P(C = (1,1,1)|\cup_i F_i) &= \frac{1}{4}\left\{[\tau^2 \times \tau(\tau^2)] + [\tau^2 \times \tau(\tau^2)] + [\tau \times 0] + [\tau^2 \times 0]\right\} \\ &= \frac{1}{4}\left\{[\tau^5] + [\tau^5] + [0] + [0]\right\} \\ &= \frac{1}{4}\left\{2 \times \left(\frac{3}{4}\right)^5\right\} \\ &= .1187 \end{aligned}$$

2.2 Implementation

In this section, we present two new R functions to create a Bayesian network for a pedigree and to evaluate the likelihood of a transmission probability in that pedigree. The likelihood function is then applied to a second example pedigree depicting a nuclear family with three children.

2.2.1 Bayesian network function

The Bayesian network function, `BNcreate`, takes two arguments: a pedigree data-frame, `pedfile`, in a format defined by the `kinship2` R package[8] and a RV transmission probability, τ. Given a pedigree data-frame, `BNcreate` defines the conditional probability of RV status for each child given their parents using a RV transmission probability, τ. The conditional probability of the child is in the first entry of a triplet, given its parents in the second and third entries. Then `BNcreate` constructs the *conditional probability tables* (CPTs) for both founders and non-founders in the pedigree and creates the Bayesian network using the `grain` function [9]. The R code for `BNcreate()` is:

```
BNcreate = function(pedfile, tau){
  ## conditional prob of child in the first entry of the triplet
  ## given parents in the second and third entries of triplet with transmission
  ## probability tau
  geno_prob = c(
    # 000 100 200 010 110 210 020 120 220
    1, 0, 0, 1-tau, tau, 0, 0, 1, 0,
    # 001 101 201 011 111 211 021 121 221
    1-tau, tau, 0, (1-tau)^2, 2*tau*(1-tau), tau^2, 0, 1-tau, tau,
    # 002 102 202 012 112 212 022 122 222
    0, 1, 0, 0, 1-tau, tau, 0, 0, 1
  )


   # Construct the CPTs for founders
   # Founders: columns of father or mother are labelled as 0.

   founders = which(pedfile[, "father"] == 0) # the row numbers of founders
   founders_vec = rep(0, length(founders))
   # store the id number of founders (for user configuration)
   founders_cpt = list()
```

```
  # store the CPTs of founders in a list, which is denoted as [[i]]
  for (i in 1:length(founders)) {
    id = pedfile[founders[i], "id"]
    founders_vec[i] = id
    node = cptable(c(id), values = rep(1/3, 3), levels = copy)
    # CPTs of founders
    founders_cpt[i] = list(node)
  }

  # Construct the CPTs for non-founders

  pedfile_c = pedfile[-c(founders), ]
  # Pedigree data after removing founders
  nonfounders_cpt = list()
  # Store the CPTs for non-founders
  for (i in 1:nrow(pedfile_c)) {
    c = pedfile_c[i, 'id'] # id numbers of child
    f = pedfile_c[i, "father"] # id numbers of father
    m = pedfile_c[i, "mother"] # id numbers of mother
    node_nf = cptable(c(c, f, m), values = geno_prob, levels = copy)
    # CPTs for non-founders
    nonfounders_cpt[i] = list(node_nf)
  }

  plist = compileCPT(founders_cpt, nonfounders_cpt)
  gin = grain(plist)
  # Create the Bayesian network from the CPTs of both founders and non-founders
  return(gin)
}
```

The CPTs specify the local dependency of a child's RV status given the RV status of its parents. The local dependency structure is straightforward, with children depending only on their parents. The Bayesian network for a pedigree specifies that an individual may have 0, 1, or 2 copies of an RV. Each combination of a child and its two parents, therefore, has $3^3 = 27$ possible RV configurations. The `cptable()` function in the R package `gRain` [9] is called for each child-parent combination to specify the probability of the 27 possible RV configurations in the CPT. First, CPTs are created for founders and for non-founders of the pedigree. Next all the CPTs are combined into a list for

the Bayesian network using the function `compileCPT`. Each individual's CPT can be extracted from the combined CPTs in the list returned by `compileCPT`. For example, suppose an individual with ID 3 is the child of parents with IDs 1 and 2. If we extract the CPT for the individual with ID 3, we get a 3×3×3 table, each cell is a conditional probability associated with the child given different cases of RV status of the parents, ID 1 and ID 2. Each table gives the conditional probability that ID 3 carries 0, 1, or 2 copies of the RV given the RV status of its parents, ID 1 and ID 2. Once the CPTs are specified, the Bayesian network is created from the list of CPTs using the `grain` function.

2.2.2 Likelihood function

The likelihood function, `likehd()`, takes three arguments: `pedfile`, `tau`, and `config`. The first argument,`pedfile`, is a `kinship2` pedigree object specifying the pedigree structure. The second argument, `tau`, is the transmission probability of the RV. The third argument, `config`, is the numeric vector C giving the RV status of affected individuals with DNA available. Each element of the `config` vector C corresponds to an affected pedigree member with DNA. The elements are ordered by their ID numbering. For example, if the affected individuals with available DNA have ID numbers 7, 8 and 9 and $C = (1, 0, 1)$, IDs 7 and 9 carry one copy of the RV and ID 8 carries no copies. The function applies `BNcreate()` to a pedigree data-frame to create a Bayesian network for a given value of the transmission probability. Then the function loops through all the pedigree founders and, for each founder i, calculates $P(C|$ founder i introduced the RV$)$, the conditional probability of the RV configuration, C, given that the RV enters the pedigree on founder i. The loop sums over the product of this conditional probability and the probability that founder i is the founder who introduced the RV to get the desired likelihood from $P(C|$ exactly one founder introduced the RV$)$. The R code for the likelihood function is:

```
likehd = function(pedfile, tau, config){

  gin = BNcreate(pedfile, tau)
  # Bayesian networks of the specified pedigree with transmission prob tau
  likelihood = 0

  founders = which(pedfile[, "father"] == 0)
  # find the row numbers of founders
  founders = as.character(pedfile$id[founders])
```

```
  # find the ID numbers of founders
  affs = which(pedfile[, "affected"] == 1&pedfile[,"avail"]==1)
  # find the row numbers of affected individuals with DNA available
  affs=as.character(pedfile$id[affs])
  # find the ID numbers of affected individuals with DNA available

  # Assume the RV enters pedigree on exactly one founder and
  # loop through all the founders
  for (i in 1:length(founders)) {
    state = rep(0, length(founders))
    # vector state specifies which founder introduces the RV
    state[i] = "1" #1 means carrying the RV.
    bn = setEvidence(gin, nodes = founders, states = state)
    # Set founder i to introduce the rare variant

    # calculate P(config | founder i introduced the rv)
    prob=1
    for(n in affs){
      if(prob > 0) # prevents conditioning on zero prob events
      {
        # calculate probability for this node, n
        p=unlist(querygrain(bn,nodes=n,exclude=FALSE)) #get marginal prob
        names(p)=c("0","1","2")
        prob=prob*p[config[n]]
      # condition on node n's state
        bn = setEvidence(bn, nodes=n, states=config[n])
      }
    }
    likelihood = prob*1/length(founders)+likelihood
  }

  return(likelihood)
}
```

2.2.3 Application

We apply the likelihood function to a second example pedigree depicting a nuclear family. This nuclear family is shown below in Figure 2.1 and has only two gener-

ations. The founders have IDs 1 and 2, and they have three affected children, IDs 3, 4 and 5, with DNA available. We next apply the function `likehd()` to the data-frames for pedigrees in Figure 1.1 and Figure 2.1 and check the results against manual calculations from the *path-counting* approach described in the previous section 2.1.

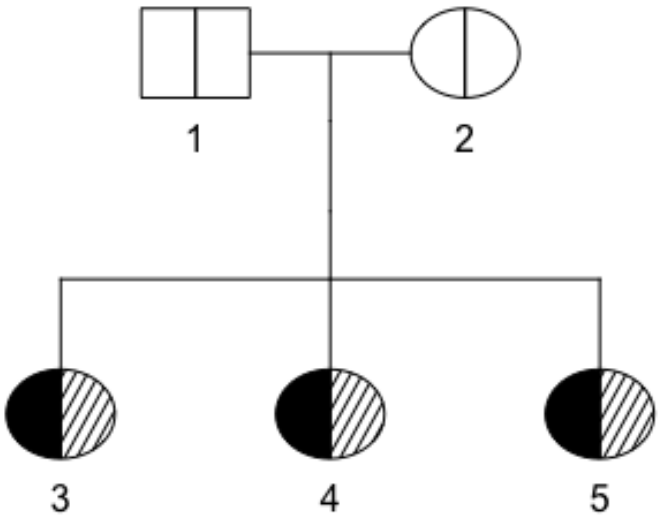

Figure 2.1: Example pedigree depicting a nuclear family

For the nuclear family in the pedigree of Figure 2.1, we evaluate the likelihood of $\tau = 3/4$ when the RV configuration for IDs 3, 4, and 5 is $C = (1, 1, 1)$. The output of `likehd()` is 0.421875, which agrees with our hand calculation of $\frac{1}{2}(\tau^3+\tau^3) = 0.421875$ from the path-counting approach. When the RV configuration for IDs 3, 4, and 5 is $C = (1, 0, 1)$, which mean IDs 3 and 5 carry 1 copy of the RV and ID 4 carries 0 copies of the RV (as might occur if the disease in ID 4 is environmentally caused). The output of `likehd()` is 0.140625, which again agrees with our hand calculation of $\frac{1}{2}(\tau^2(1-\tau) + \tau^2(1-\tau)) = 0.140625$.

For the extended pedigree in Figure 1.1, when the RV configuration is $C = (1, 1, 1)$, `likehd()` evaluates the likelihood of $\tau = 3/4$ to be about 0.1187. This likelihood

value complies with our hand calculation from the path-counting approach. Next, we evaluate the likelihood of $\tau = 3/4$ when the RV configuration for IDs 7, 8, and 9 is $C = (1, 1, 0)$. In this configuration, IDs 7 and 8 both carry the RV but ID 9 does not. The output of `likehd()` is .0396. This likelihood value agrees with our hand calculation of $\frac{1}{2}\tau^4(1-\tau) = \frac{1}{2}\left(\frac{3}{4}\right)^4\left(1-\frac{3}{4}\right) = 0.0396$ from the path-counting approach.

When the RV configuration is $C = (1, 1, 1)$, all three of the affected relatives with DNA share the RV. In this case, the transmission probability of $\tau = 3/4$ has a higher likelihood value in the nuclear pedigree of Figure 2.1 than in the extended pedigree of Figure 1.1 (compare 0.421875 to 0.1187). The three affected siblings in the nuclear pedigree of Figure 2.1 are separated by fewer transmission events than the two affected siblings and one affected cousin in the extended pedigree of Figure 1.1. The three affected siblings in the nuclear pedigree therefore have a higher probability of sharing the RV than the two affected siblings and one affected cousin in the more extended pedigree.

2.3 Likelihood Curves

This section presents plots of likelihood values versus a sequence of transmission probabilities, τ, for various RV configurations C in the example pedigrees of Figures 1.1 and 2.1. Values of $\tau > 1/2$ suggest that the RV is positively associated with the disease, values of $\tau < 1/2$ suggest that the RV is negatively-associated with the disease and values of $\tau = 1/2$ suggest that the RV is unassociated with the disease.

2.3.1 Extended Pedigree

We start by considering the extended pedigree in Figure 1.1 when the RV configuration for IDs 7, 8, and 9 is $C = (1, 1, 1)$. The corresponding likelihood curve is shown in Figure 2.2 below. The likelihood values increase rapidly once the transmission probability is greater than 0.6, up to a likelihood value of 0.5. The maximum-likelihood estimate of the transmission probability is $\hat{\tau} = 1$.

When the RV configuration for the extended pedigree is changed to $C = (1, 1, 0)$, the resulting likelihood curve is shown in Figure 2.3 below. The likelihood function reaches a peak at a transmission probability of approximately 0.8. Likelihood values increase with the transmission probability when the transmission probability is less

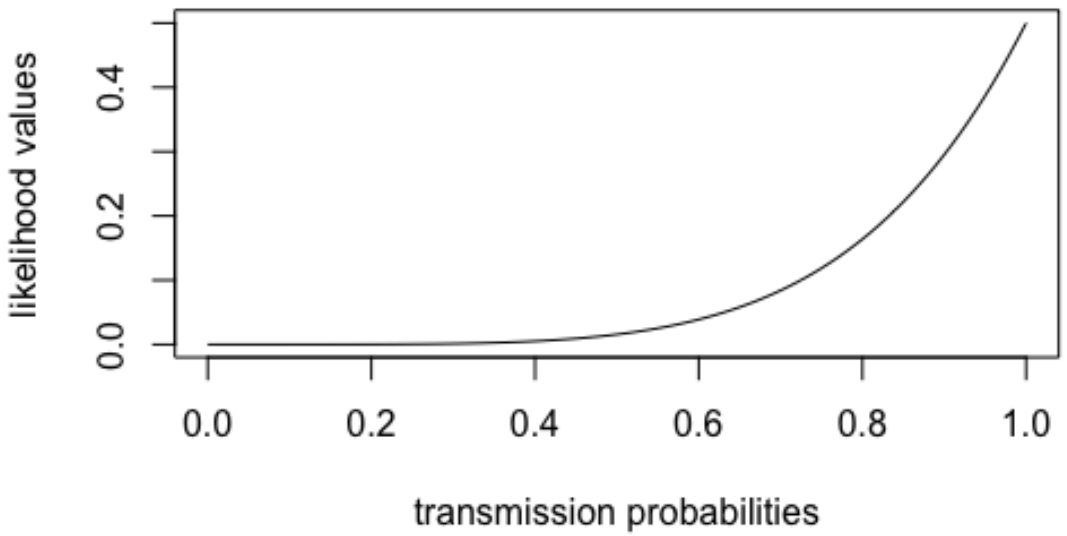

Figure 2.2: Likelihood curve for the extended pedigree in Figure 1.1 when $C = (1, 1, 1)$

than 0.8, and decrease when the transmission probability is greater than 0.8. The maximum-likelihood estimate of the transmission probability is thus $\hat{\tau} \approx 0.8$.

2.3.2 Nuclear Pedigree

Next, we consider likelihood curves for the nuclear family in Figure 2.1. Figure 2.4 below shows the likelihood curve for $C = (1, 1, 1)$. The likelihood values increase rapidly once the transmission probability is greater than 0.4 up to a likelihood value of 1. The maximum-likelihood estimate of the transmission probability is $\hat{\tau} = 1$.

When $C = (1, 1, 1)$, the likelihood curve for the nuclear family in Figure 2.4 rises more gently (i.e. is flatter) than its counterpart for the extended pedigree in Figure 2.2. The likelihood curve for the nuclear family is flatter than the one for the extended pedigree because it has fewer transmission events separating the three affected relatives. As a result, likelihood values at transmission probabilities $0.2 < \tau < 0.6$ are higher in the nuclear family than in the extended pedigree. Transmission probabilities $0.2 < \tau < 0.6$ are therefore more likely in the nuclear pedigree than in the extended pedigree.

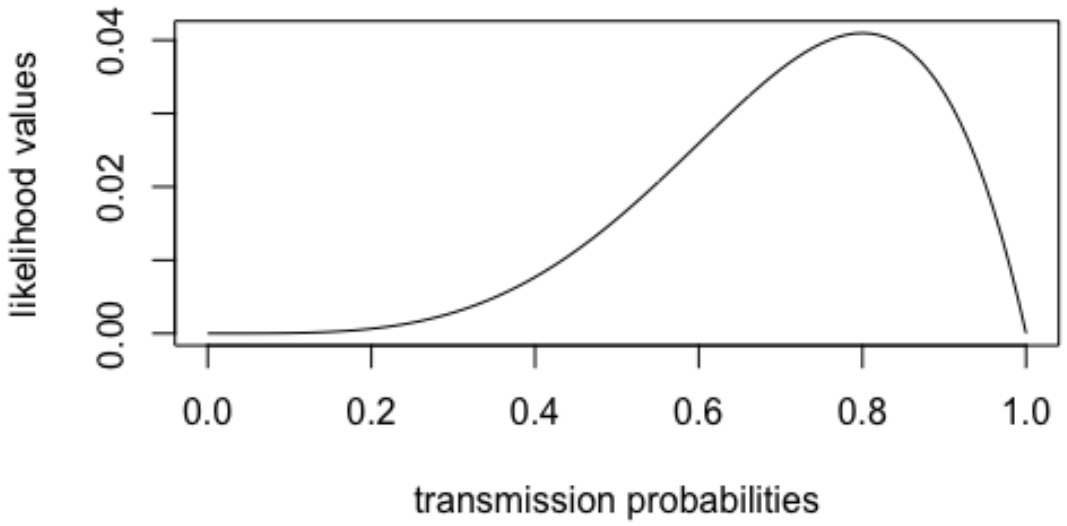

Figure 2.3: Likelihood curve for the extended pedigree in Figure 1.1 when $C = (1, 1, 0)$

When the RV configuration for the nuclear pedigree of Figure 2.1 is changed to $C = (1, 1, 0)$, the likelihood curve is shown in Figure 2.5 below. The likelihood curve reaches a peak at likelihood value 0.15 with a transmission probability of approximately 0.7. Likelihood values increase with the transmission probability when the transmission probability is less than 0.7, and decrease when the transmission probability is greater than 0.7. The maximum-likelihood estimate of the transmission probability is thus $\hat{\tau} \approx 0.7$. When $C = (1, 1, 0)$, the peak of the likelihood curve for the nuclear family is shifted to the left of the peak of the likelihood curve for the extended pedigree in Figure 2.4. The MLE for the nuclear family is therefore smaller than for the extended pedigree. Moreover, the likelihood curve for the nuclear family has slightly less curvature around its MLE (i.e. is flatter) than the likelihood curve for the extended pedigree in Figure 2.4. The likelihood curve for the nuclear family is flatter than the one for the extended pedigree because it has fewer transmission events separating the three affected relatives.

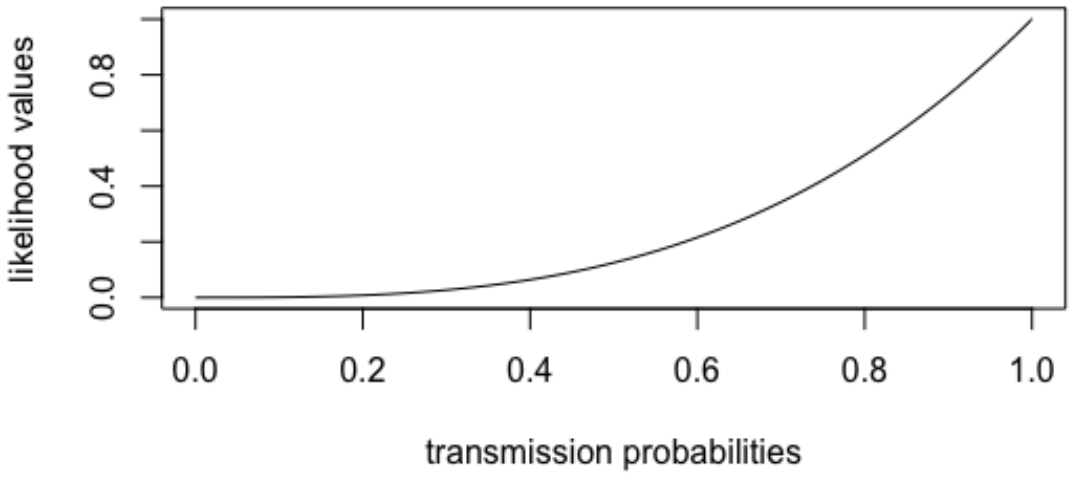

Figure 2.4: Likelihood curve for the nuclear pedigree of Figure 2.1 when $C = (1, 1, 1)$

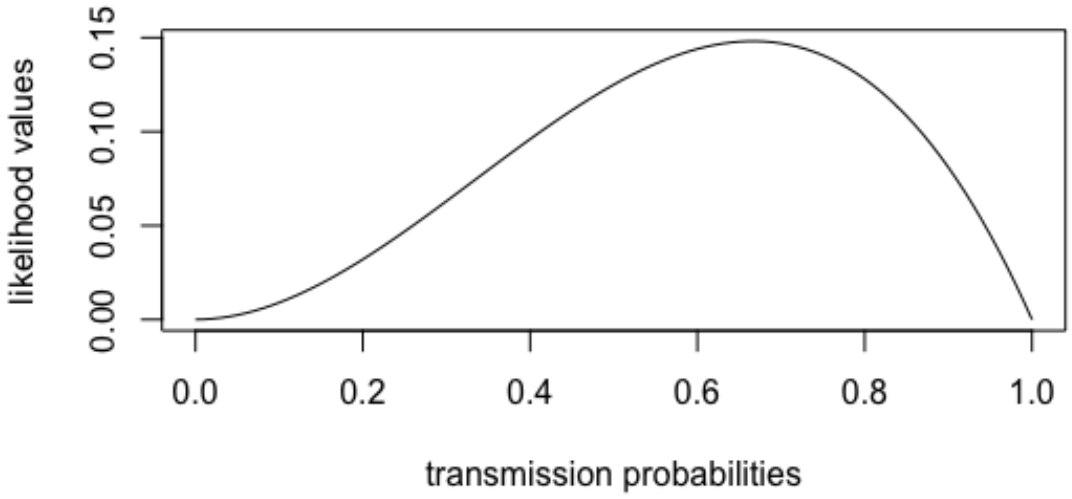

Figure 2.5: Likelihood curve for the nuclear pedigree of Figure 2.1 when $C = (1, 1, 0)$

Chapter 3

Examples

This chapter considers examples involving more complex pedigrees obtained by modifying the second of the two sample pedigrees in the `kinship2` R package [8].

3.1 First pedigree

Our first example of a more complex pedigree is shown in Figure 3.1 below. From the figure, we see that ID 205 has unknown affection status, as marked by a "?". This individual is not sequenced and so is not relevant to our likelihood. The affected family members with DNA available are relevant and have IDs 206, 207, 211 and 214. The founders have IDs 201, 202, 203, and 209. Let's evaluate the likelihood of $\tau = 1/2$ for various configurations of the RV status in these individuals. First, suppose that all four of them share a copy of the RV; i.e. $C = (1, 1, 1, 1)$. For this RV configuration, the `likehd()` function returns a value of 0.0078125 at $\tau = 1/2$. This likelihood value complies with our hand calculation from the path-counting approach:

$$\begin{aligned} P(C = (1,1,1,1) | \cup_i F_i) &= \frac{1}{4}\left\{[\tau^2 \times \tau \times \tau \times \tau^2] + [\tau^2 \times \tau \times \tau \times \tau^2] \right. \\ &\quad \left. + [1 \times \tau \times 0] + [1 \times 0]\right. \\ &= \frac{1}{4}\left\{[\tau^6] + [\tau^6] + [0] + [0]\right\} \\ &= \frac{1}{4}\left\{2 \times \left(\frac{1}{2}\right)^6\right\} \\ &= .0078125 \end{aligned}$$

Next, we evaluate the likelihood of $\tau = \frac{1}{2}$ when the RV configuration for IDs 206, 207, 211, 214 is $C = (1, 0, 0, 0)$, which indicates that only ID 206 carries the RV.

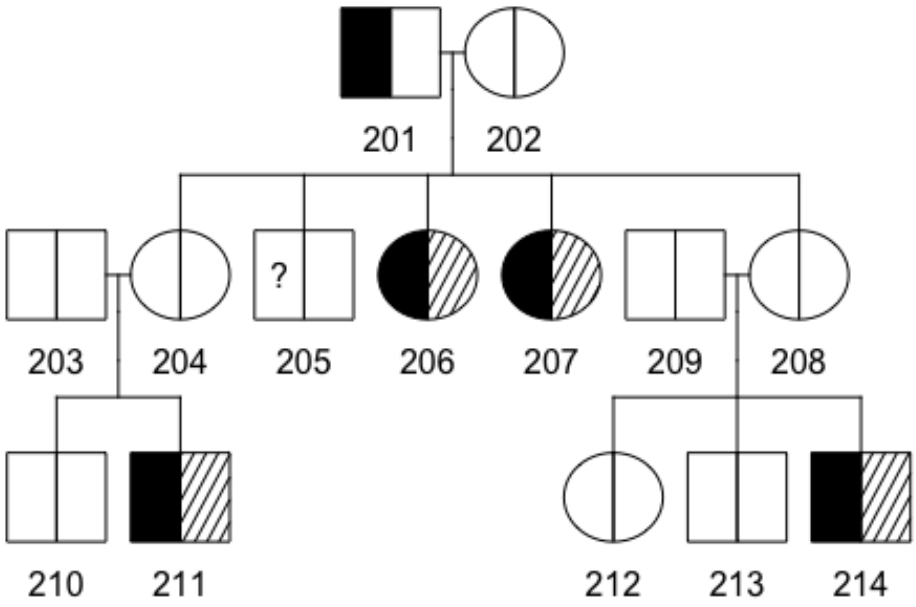

Figure 3.1: Example of complex pedigree

When the RV configuration is C = (1,0,0,0), the `likehd()` function returns a value of about 0.0703125 at $\tau = 1/2$. This likelihood value also complies with our hand calculation from the path-counting approach:

$$\begin{aligned} P(C=(1,0,0,0)|\cup_i F_i) &= \frac{1}{4}\Big\{[\tau\times(1-\tau)\times((1-\tau)+\tau\times(1-\tau))^2] \quad + \\ &\qquad [\tau\times(1-\tau)\times((1-\tau)+\tau\times(1-\tau))^2] \;+\; [0] \;+\; [0]\Big\} \\ &= \frac{1}{4}\left\{\frac{1}{2}\times\frac{1}{2}\times\left(\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\right)^2 \;+\; \frac{1}{2}\times\frac{1}{2}\times\left(\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\right)^2\right\} \\ &= .0703125 \end{aligned}$$

3.2 Second pedigree

Consider an updated version of the pedigree in Figure 3.1, in which the founder with ID 201 has DNA available. The updated pedigree is shown below in Figure 3.2. The hash marks on the right side of the symbol for ID 201 indicate that this disease-affected founder has DNA available.

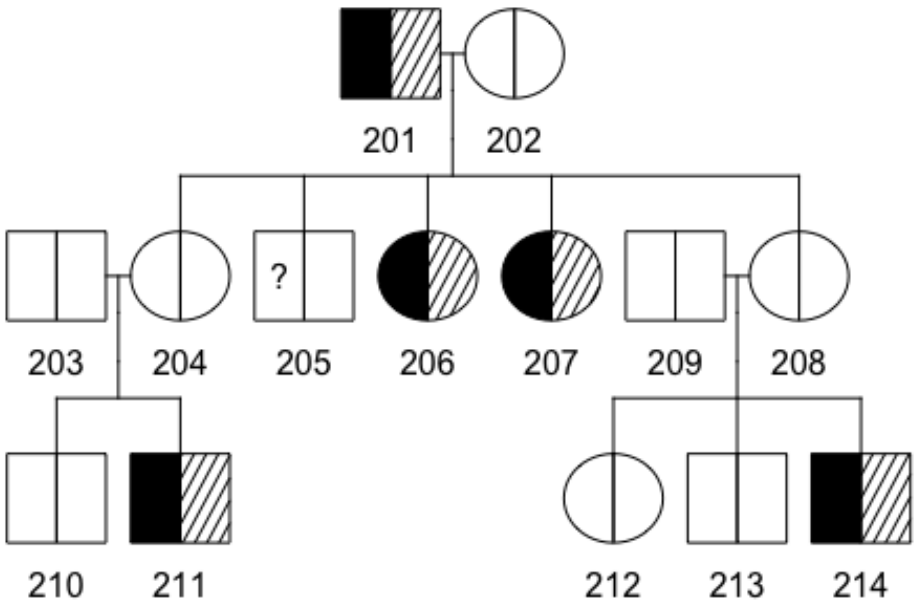

Figure 3.2: Updated complex pedigree in which ID 201 has DNA available

Let's work through the likelihood calculation when all the affected pedigree members with DNA in Figure 3.2 carry the RV. In this case, the RV configuration is $C = (1, 1, 1, 1, 1)$ for IDs 201, 206, 207, 211 and 214. This pedigree has four founders with IDs 201, 202, 203 and 209. If ID 201 is the founder introducing the RV, the conditional probability of the configuration is

$$P(C = (1, 1, 1, 1, 1) \,|F_{201}) = 1 \times \tau \times \tau \times \tau^2 \times \tau^2 = \tau^6.$$

If ID 202 is the founder introducing the RV, the conditional probability of the configuration is

$$P(C = (1,1,1,1,1)\,|F_{202}) = 0 \times \tau \times \tau \times \tau^2 \times \tau^2 = 0.$$

Similarly, if IDs 203 or 209 are the founders introducing the RV, the conditional probability of the configuration is zero. Therefore, the likelihood function for the transmission probability is $P(C = (1,1,1,1,1)\,|\cup_i F_i) = \frac{1}{4}\tau^6$. The configuration probability is 0.00390625 at $\tau = 1/2$, as confirmed by the `likehd()` function.

Now suppose the RV configuration is $C = (0,1,1,1,1)$. If ID 201 is the founder introducing the RV, the conditional probability of the configuration is $P(C = (0,1,1,1,1)\,|F_{201}) = 0$ because ID 201 not carrying the RV conflicts with the conditioning event that ID 201 introduced the RV into the pedigree. If ID 202 is the founder introducing the RV, the conditional probability of the configuration is $P(C = (0,1,1,1,1)\,|F_{202}) = \tau \times \tau \times \tau^2 \times \tau^2 = \tau^6$. If IDs 203 or 209 introduce the RV, the conditional probability of the configuration is $P(C = (0,1,1,1,1)\,|F_{203}) = P(C = (0,1,1,1,1)\,|F_{209}) = 0$. The likelihood function for τ is therefore, again, $P(C = (0,1,1,1,1)\,|\cup_i F_i) = \frac{1}{4}\tau^6$. The configuration probability is 0.00390625 at $\tau = 1/2$, as confirmed by the `likehd()` function.

3.3 Likelihood Curves

This section presents likelihood curves for the first complex pedigree in Figure 3.1. We start by considering the configuration $C = (1,1,1,1)$ when all four of the affected pedigree members with available DNA carry a single copy of the RV.

The likelihood curve looks similar to the curve for the extended pedigree of Figure 1.1 when all its affected pedigree members with DNA carry a single copy of the RV. From the above plot, we see that the likelihood values increase rapidly when the transmission probability is greater than 0.6. The maximum likelihood estimate of the transmission probability in this case is $\hat{\tau} = 1$.

Next suppose that we observe the configuration $C = (1,1,0,1)$ in the complex pedigree of Figure 3.1. Figure 3.4 below shows the resulting likelihood curve. The likelihood curve reaches its peak value when the transmission probability is approximately 0.8. Likelihood values increase with the transmission probability when the transmission probability is less than 0.8, and decrease when the transmission prob-

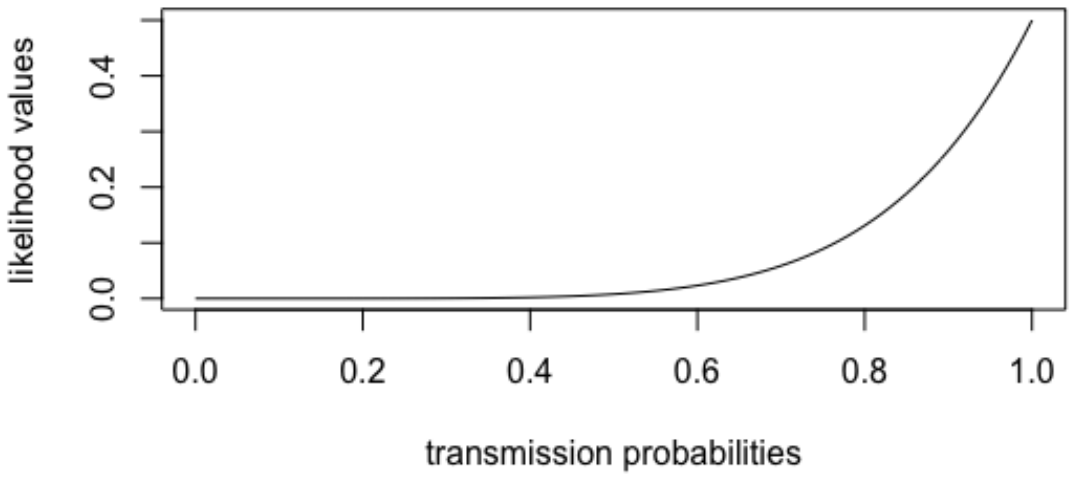

Figure 3.3: Likelihood curve of complex pedigree in Figure 3.1 when $C = (1, 1, 1, 1)$

ability is greater than 0.8. The maximum likelihood estimate of the transmission probability is, therefore, $\hat{\tau} \approx 0.8$.

Finally, let us conclude by considering the configuration $C = (1, 0, 0, 0)$; that is ID 206 carries a single copy of the RV and IDs 207, 211 and 214 carry no copies. The resulting likelihood curve is shown in Figure 3.5 below. The likelihood curve reaches a peak value when the transmission probability is approximately 0.4. The maximum likelihood estimate of the transmission probability in this case is $\hat{\tau} \approx 0.4$.

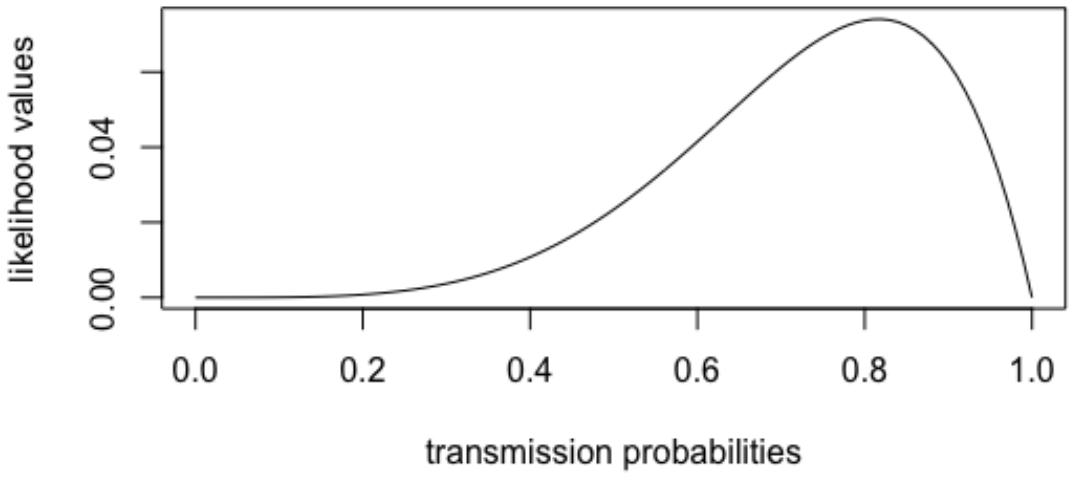

Figure 3.4: Likelihood curve of complex pedigree in Figure 3.1 when $C = (1, 1, 0, 1)$

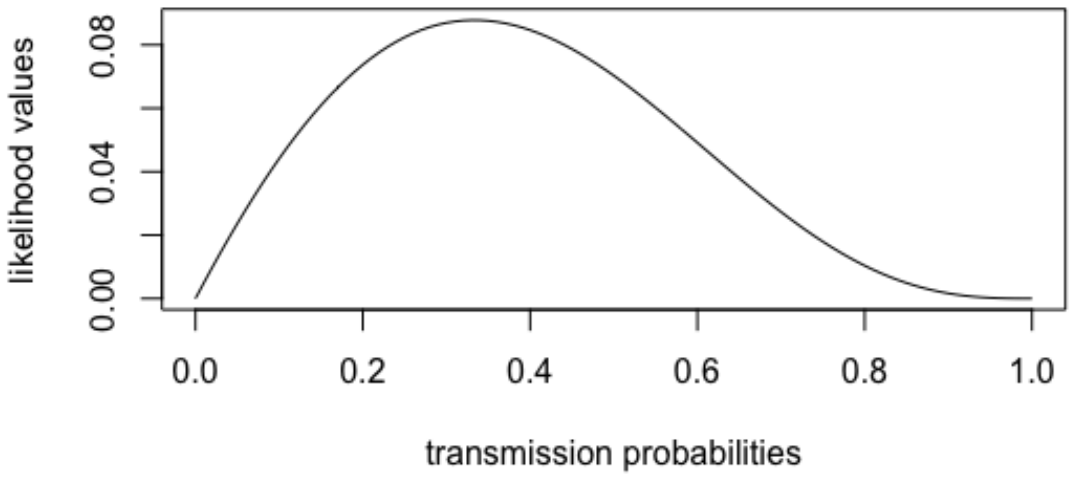

Figure 3.5: Likelihood curve of complex pedigree in Figure 3.1 when $C = (1, 0, 0, 0)$

Chapter 4

Conclusions

In this project, we are interested in the likelihoods of the probability that a rare variant (RV) is transmitted to the disease-affected members of an extended family, or pedigree. These likelihoods are obtained from the joint probability of an RV configuration in the diseased individuals with DNA in the pedigree. We use the method of path-counting to develop mathematical formulas for the likelihood of a transmission probability, provided data on the RV status of the affected members with available DNA, and assume that a single founder introduces the RV into the pedigree. We explain how a pedigree can be viewed as a Bayesian network; i.e, a directed graphical model of the RV status of its members, with the RV status of a child depending causally on the RV status of its parents. Since Bayesian networks are a special case of graphical independence networks, we show how to implement the likelihood using the R package 'gRain' for graphical independence networks.

This project introduces examples of a simply extended pedigree containing three generations (Figure 1.1), a nuclear pedigree containing two generations (Figure 3.1), and more complex pedigrees involving three generations (Figures 3.1 and 3.2). We show how to calculate the likelihoods manually using path-counting methods. We also show how to implement the likelihoods in R. The manual calculations are then used to validate the results of the R implementation. Additionally, we present likelihood curves for the transmission probabilities arising from different configurations of the RV status and compare these curves between different pedigrees.

In principal, the functions we have developed should be applicable to large extended pedigrees containing thousands of individuals such as those from extended kindreds of Hutterites in Alberta or of the Saguenay-Lac-Saint-John region of Quebec. It would be of interest to test the functions on these massive pedigrees to see if they break down computationally; we leave this for future work. Another direction

for future research is to extend our functions to determine the most likely founder to introduce the RV into the pedigree given a fixed transmission probability. The most likely founder would have the largest likelihood contribution for the transmission probability, and could easily be returned with the other function output.

Bibliography

[1] R.L. Bennett, K.S. French, R.G. Resta, and et al. Standardized human pedigree nomenclature: Update and assessment of the recommendations of the national society of genetic counselors. *J Genet Counsel*, 17:424–433, 2008.

[2] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, 2011.

[3] Joanna Spahis. Human genetics: constructing a family pedigree. *The American Journal of Nursing*, 102(7):44–50, 2002.

[4] Todd Andrew Stephenson. An introduction to Bayesian network theory and usage. RR 00-03, IDIAP, 2000.

[5] J. Puga, M. Krzywinski, and N. Altman. Bayesian networks. *Nat Methods*, 12:799–800, 2015.

[6] Xin-She Yang. *Introduction to Algorithms for Data Mining and Machine Learning*. Academic Press, 2019.

[7] Timo Koski and John Noble. *Bayesian Networks: An Introduction*. John Wiley & Sons, 2011.

[8] JP Sinnwell, TM Therneau, and DJ Schaid. The 'kinship2' R package for pedigree data. *Hum Hered*, 78(2):91–93, 2014.

[9] S Højsgaard. Graphical independence networks with the 'grain' package for R. *Journal of Statistical Software*, 46(10):1–26, 2012.

Appendix A

Code

An RMarkdown tutorial explaining the R functions described in this thesis is available on GitHub at https://github.com/SFUStatgen/TJ2022.