Seth Temple
sdtemple@uw.edu

Project

On "Probability functions on complex pedigrees"

Pedigrees encode how individuals in a sample are related to each other. Coupled with genotypic or phenotypic data, we begin to observe the evolutionary process of heredity. As a result, pedigrees can be useful in modeling inheritance patterns for a sample of related individuals and extrapolating to more general inheritance patterns. Throughout this quarter, we have calculated probabilities conditional on relatedness using Bayes theorem, the total law of probability, and techniques developed by Jacquard [3] and Thompson [8]. For example, we computed $P(Y|\mathcal{R})$ where Y encoded genotypic information and $\mathcal{R}$ encoded familial relationships. Cannings, Thompson, and Skolnick address such conditional probabilities in their classic 1978 paper "Probability functions on complex pedigree", but, whereas we resigned ourselves to pairs of related individuals (typically of 3rd degree or less), the aforementioned trio build the mathematics necessary to determine likelihoods for large and complicated pedigrees. In this review, we discuss the intuition behind Cannings et al.'s pioneering recursive calculations (referred to as the peeling procedure), and we place the procedure in its historical context.

This article is first and foremost a treatise on computing probabilities whereby a graph encapsulates conditional dependencies. To that effect, Cannings et al. formally define a probability model that describes phenotypes as the outward expression of an essential essence (referred to as ousiotypes by the authors and as genotypes to modern audiences). Of note, this framework treats the above Y as phenotypic information, as this article predates widespread sequencing. Additionally, the framework is flexible enough to allow for multiallelic ousiotypes and environmental influences on the phenotypic expression. Runtimes and storage do however limit the model to 10 genotypes ($n = 4$) [1].

$$\binom{n}{2} + \binom{n}{1} = \frac{n(n+1)}{2}$$

A graph that relates individuals in a sample to each other is essential to the peeling procedure. Cannings et al. introduce a graph theoretic paradigm by which a graph $G = (V, E)$ is composed of V vertices and E edges. They draw a distinction between individual vertices $\circ$ versus marriage vertices $\bullet$ and marriage edges $\rightarrow\rightarrow$ versus descent edges $\rightarrow$. These distinctions are crucial in differentiating between relatedness in a social context and biological relatedness [1]. The canonical example of a pedigree is a (family) tree that expands outwards and downwards. Whereas such examples may require less care, the purpose of this article is to assess likelihoods for pedigrees that may have complicated looping structures.

Loops occur in 3 ways: marriage rings, inbreeding loops, and exchange loops [1]. (1) Marriage rings arise due to sequences of marriages. The simplest example is a marriage square. We have males A and B and females C and D, and both males have at least 1 child with each female. (Two individuals are related via marriage edges and a marriage vertex if they bear children together.) (2) Inbreeding loops happen when biologically related individuals procreate. Examples include a parent marrying a child, siblings marrying, and first-cousins marrying. (3) Exchange loops result from multiple marriages relating ancestors in a specific way. For example, brothers A and B marry sisters C and D respectively. Here there is no explicit inbreeding present, but all children of the two marriages are connected to the same set of ancestors. Another common example is when brothers A and B both bear children with female C. Cannings et al. propose other types of loops that are combinations of or subcategories of these 3 loops [1]. The novel development in this paper is to handle these looping structures in computing pedigree likelihoods.

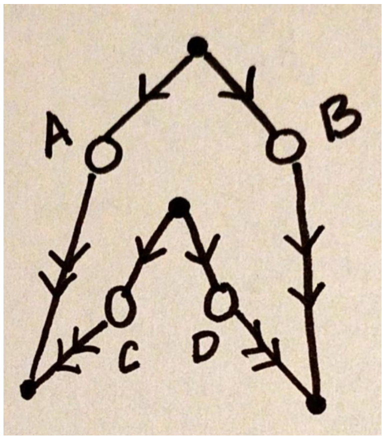

Figure 1: An example of an exchange loop

The peeling procedure is a sequence of recursive relations employed to simplify pedigree likelihood calculations. Sections 4 and 5 introduce more graph theory and define the peeling operations with mathematical rigor [1]. Here we provide intuition for peeling operations and in the appendix we give annotations for graph theoretic concepts where possible. Peeling can be thought of as cleaving a vertex and passing the information it contains to other vertices. This general idea of passing information (in the form of probability calculations) mirrors that of Rabiner's later forward-backward algorithm for hidden Markov models [6]. By successively peeling vertices from the graph, we eventually end up with a single vertex that contains all of the information in the pedigree. For tree-like connected components of a pedigree graph, each vertex contains information R^* and R^+, where R^* is (essentially) the probability of the phenotypes of descendants given the genotypes of a set of individuals and R^+ is (essentially) the probability of the phenotypes of ancestors given the genotypes of a set of individuals. For these tree-like connected components, we have the following peeling operations:

(i) peeling upwards; passing information from a child to its parents

(ii) peeling sideways; passing information from a spouse to a spouse

(iii) combining ancestral information of a marriage unit

(iv) peeling downwards; passing information from parents to their child

Cannings et al. work through an example of a possible peeling of a pedigree [1]. There are possibly many acceptable peeling sequences for a given pedigree, in which a peeling sequence is acceptable if the R-functions required at each recursive step are available. Moreover, with appropriate care, marriage rings (a loop) can be handled using the above peeling operations for trees [1].

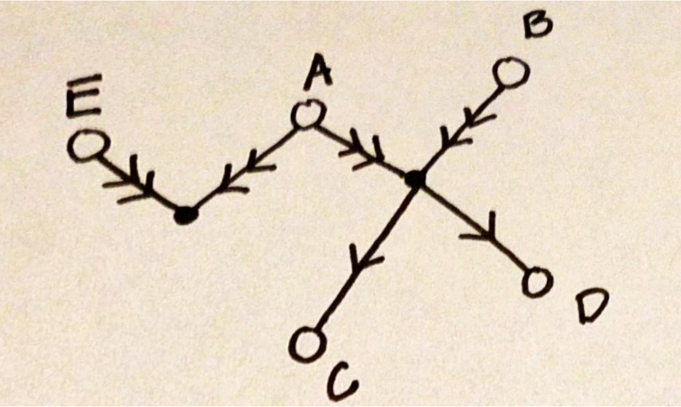

Figure 2: A zero-looped pedigree. One possible peeling is to peel E sideways, combine ancestral info from A and B, peel C upwards, and finally peel A and B downwards onto D. Another possible peeling is to instead peel D upwards and peel downards onto C.

For other types of loops, more graph theoretic definitions, more general peeling operations, and another R-function are derived because the bifurcation into upper and lower sections of the pedigree graph loses meaning when loops are present. Theorems 1 and 2 and some corollaries detail the specifics of these peeling operations for complex pedigrees [1]. Essentially, Theorem 1 gives a formula for how to pass information when there is a meaningful sequential ordering to the peeling and Theorem 2 gives a formula for how to pass information when we want to combine disjoint sections. As before, the R function is updated and passes information, and, besides the R function, the formulas include penetrance probabilities and (allele) transmission probabilities. All of this is brought together via summations and products. These peeling operations and the R function are generalizations of the tree peeling operations and the R^* and R^+ functions. In practice, to compute the likelihood for a complex pedigree we start by peeling the zero-looped components of the graph and defer to complex peeling operations to address loops.

Cannings et al. close with a peeling example for a pedigree with 96 vertices and loops present. This computation is an impressive feat for a 1970s computer, but there are some issues to make note of. First, the user must input a peeling sequence to the program. In the paper, the peeling sequence to solve the aforementioned complex pedigree encompasses nearly 2 pages. This could be error-prone and prohibitively requires that the user have expert knowledge of the peeling operations. Second, finding an optimal peeling sequence is difficult. Third, finding a peeling sequence for a complex pedigree is not yet fully solved due to computational complexities [1].

Other researchers in the 1970s competed with Cannings et al. to develop optimal techniques to calculate complex pedigree likelihoods. Most notably, Elston and Stewart developed a similar recursive procedure to calculate zero-looped pedigree likelihoods [2], and Lange and Elston extended this technique to handle complex pedigrees [4]. The approach of Lange and Elston is to split up loops in a complex pedigree by introducing phenotypically and genotypically identical copies of an individual. The split results in trees that we can calculate likelihoods on using a recursive procedure for zero-looped pedigrees. Cannings et al. scrutinize this approach, remarking that the algorithm has a polynomial runtime complexity in the number of individuals split up. However, they acknowledge that their peeling procedure often requires more computer memory [1].

Computing pedigree likelihoods in the 1970s was an active research area because of its many immediate applications. One application is to predict phenotypes or genotypes for individuals not yet born into the pedigree. For example, two carriers for cystic fibrosis, a trait characterized by Mendelian inheritance, may be encouraged by a genetic counselor to not mate. Another application is to describe the pedigree structure for individuals missing from the pedigree. Thompson demostrates this in “Ancestral inference. II. The founders of Tristan da Cunha”. Due to several bottleneck events, the inhabitants of the small island of Tristan da Cunha are all related to an identifiable subset of original founders. She uses the peeling procedure to infer how the contemporary inhabitants of Tristan da Cunha are related to each other and to the founders [7]. The most important application is to estimate parameters of a parametric (likelihood-based) model. Ott illustrates this by using the Elston and Stewart model to estimate the recombination fraction from a sample family afflicted with myotonic dystrophy [5]. Cannings et al. mention work on implementing recombination fractions in their model as well [1]. Recombination fractions, allele frequencies, penetrance rates, and sporadic rates are parameters common to likelihood-based models that we have discussed in class and are fundamental to understanding patterns of inheritance.

Estimating parameters of a probability model is a ubiquitous approach by which statisticians learn about the processes they model. This quarter we discussed multiple likelihood-based models and the estimation of their parameters. Specifically, we focused on parametric linkage analysis in weeks 8 and 9 and performed an analysis in lab using the MERLIN software. Parametric linkage analysis gained much popularity in the 1990s and 2000s as a method to map the genome and identify causal variants for genetically-related diseases. Such analyses require *a priori* the ability to compute likelihoods. The achievements of Cannings et al., Elston and Stewart, and others in prior decades paved the way for these developments. While “Probability functions on complex pedigrees” is a well-cited paper, the Elston and Stewart paper and the Lander-Green algorithm are more frequently referenced. These approaches involve simpler Markov models and are better suited for analysis of a more traditional family tree. We learn in GENOM 540 that simplicity and efficiency are often important in genetics due to the size and diversity of the data.

Nevertheless, “Probability functions on complex pedigrees” is a neat paper, and reading it is great training for a statistical geneticist. The very nature of inheritance involves dependencies. Being able to visualize these dependencies via a graph and take advantage of that structure to aid in computations is an important skill. Markov properties through which we describe these dependencies are present in many statistical genetics models, namely any hidden Markov model. As mentioned previously for the forward-backward algorithm, due to time and memory limitations, recursive approaches are often essential to make the computation feasible. I am glad to have selected and read this paper because it expanded my understanding of how we compute likelihoods for pedigrees. Since there are so many conditional probabilities to account for, we cannot take a naïve product as in the case of i.i.d. samples in STAT 512-3. On the other hand, these computations are challenging, so even in classes like this one we limit ourselves to 2-3 related individuals. This paper takes a very general look at problems in statistical genetics and constructs language and paradigms to think about them. In this respect, it shows us how to approach future problems where many conditional dependencies exist.

Appendix

- components of a graph with respect to a set of individuals P are the connected subgraphs resulting from the removal of vertices P and all edges going into or out of them
- sections are unions of connected components
- the spirit of the upper section of P is to capture the ancestors of P; likewise, the spirit of the lower section of P is to capture the descendants of P
- R-functions with a subscript ω are introduced to simplify calculations for individuals with multiple marriages by treating each marriage one at a time
- a division divides a subset of the vertices such that one subset of the subset can only connect to the rest of the graph via another subset of the subset, respectively referred to as the division, the section, and the split set
- a simple peeling sequence is a sequence of successive divisions such that the division and the section are monotonically increasing
- divisions are disjoint if the intersection of their sections is the empty set and the sections are disjoint from the split sets
- simple peeling sequences are disjoint if their final divisions are disjoint
- a complex peeling sequence is built by combining simple peeling sequences (see paper)
- an entry set of a set of vertices w.r.t. another set of vertices is the set of vertices that have edges into it from the other set of vertices
- $R(\cdot)$ with inputs a division and a vector of genotypes is the probability of phenotypes for a section and genotypes for the entry set of the split set w.r.t. to the section conditional on the genotypes for the entry set of the split set w.r.t. all other vertices

Figure 3: IBM 370 computer

References

[1] C Cannings, EA Thompson, and MH Skolnick. "Probability functions on complex pedigrees". In: *Advances in Applied Probability* 10.1 (1978), pp. 26–61.

[2] Robert C Elston and John Stewart. "A general model for the genetic analysis of pedigree data". In: *Human heredity* 21.6 (1971), pp. 523–542.

[3] Albert Jacquard. "Genetic information given by a relative". In: *Biometrics* (1972), pp. 1101–1114.

[4] K Lange and RC Elston. "Extensions to pedigree analysis". In: *Human heredity* 25.2 (1975), pp. 95–105.

[5] Jurg Ott. "Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies." In: *American journal of human genetics* 26.5 (1974), p. 588.

[6] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[7] EA Thompson. "Ancestral inference. II. The founders of Tristan da Cunha." In: *Annals of Human Genetics* 42.2 (1978), pp. 239–253.

[8] Elizabeth A Thompson. "Statistical inference from genetic data on pedigrees". In: *NSF-CBMS regional conference series in probability and statistics*. JSTOR. 2000, pp. i–169.