

*NSF-CBMS Regional Conference Series
in Probability and Statistics
Volume 6*

**STATISTICAL
INFERENCE
FROM
GENETIC
DATA
ON
PEDIGREES**

Elizabeth A. Thompson

University of Washington

Institute of Mathematical Statistics
Beachwood, Ohio

American Statistical Association
Alexandria, Virginia

Conference Board of the Mathematical Sciences

*Regional Conference Series
in Probability and Statistics*

Supported by the
National Science Foundation

The production of the *NSF-CBMS Regional Conference Series in Probability and Statistics* is managed by the Institute of Mathematical Statistics: Barry Arnold, IMS Managing Editor, Statistics; Patrick Kelly, IMS Production Editor; Julia Norton, IMS Treasurer; and Elyse Gustafson, IMS Executive Director.

Library of Congress Control Number: 00-134575

International Standard Book Number 0-940600-49-8

Copyright © 2000 Institute of Mathematical Statistics

All rights reserved

Printed in the United States of America

Contents

Preface	xi
Table of Notation	xiii
1 Genes, Pedigrees and Genetic Models	1
1.1 DNA, alleles, loci, genotypes, and phenotypes	1
1.2 Mendel's laws and meiosis indicators	3
1.3 Pedigrees: the conditional independence structure	4
1.4 Models, parameters, and inferences	7
2 Likelihood, Estimation and Testing	11
2.1 Likelihood and log-likelihood.	11
2.2 Estimation, information, and testing	13
2.3 Population allele frequencies	16
2.4 The EM algorithm; general formulation	20
2.5 Gene counting and the ABO blood types	22
2.6 EM estimation for quantitative trait data	25
3 Gene Identity by Descent	29
3.1 Kinship and inbreeding coefficients	29
3.2 Methods of computation	30
3.3 Data on inbred individuals	32
3.4 Multi-gamete kinship and gene <i>ibd</i>	34
3.5 Patterns of gene <i>ibd</i> in pairs of individuals	36
3.6 Observations on related individuals	39
3.7 Monte Carlo estimation of expectations	44
3.8 Reduction of Monte Carlo variance	46
4 Genetic Linkage	49
4.1 Linkage and recombination: genetic distance	49
4.2 Haplotypes, linkage, and association	51
4.3 Lod scores for two-locus linkage analysis	53
4.4 Power, information and <i>Elods</i>	55
4.5 Two-locus kinship and gene identity	59

4.6	Homozygosity mapping with a single marker	61
4.7	Meiosis at multiple linked loci	64
4.8	Multi-locus kinship and gene identity	65
5	Models for Meiosis	69
5.1	The meiosis process	69
5.2	From chromatids to crossovers	71
5.3	From chiasmata to recombination patterns	72
5.4	The chiasmata avoidance process	73
5.5	Chromatid interference	75
5.6	Count-location models for chiasmata	76
5.7	Renewal process models of chiasma formation	77
6	Likelihoods on Pedigrees	81
6.1	The Baum algorithm and “Peeling”	81
6.2	Exact likelihoods for multiple markers	83
6.3	Computations on large but simple pedigrees	84
6.4	Example of peeling a zero-loop pedigree	86
6.5	Computations on complex pedigrees	90
6.6	Models with Gaussian random effects	91
7	Monte Carlo Estimates on Pedigrees	93
7.1	Baum algorithm for conditional probabilities	93
7.2	An EM algorithm for map estimation	95
7.3	Importance sampling for likelihoods	96
7.4	Risk probabilities and reverse peeling	97
7.5	Elods and SIMLINK	99
7.6	Sequential imputation	100
8	Markov chain Monte Carlo on Pedigrees	103
8.1	Simulation conditional on data: MCMC	103
8.2	Single-site updating methods	107
8.3	Combining exact computation and Monte Carlo	109
8.4	Tightly-linked loci: the M-sampler	111
9	Likelihood Ratios for Genetic Analysis	115
9.1	Monte Carlo likelihood ratio estimation	115
9.2	Monte Carlo relative likelihood surfaces	116
9.3	Monte Carlo EM for the mixed model	118
9.4	Likelihood estimators for complex models	120
9.5	Likelihood estimation of gene locations	123
9.6	Marker <i>ibd</i> and complete-data log-likelihoods	125

10 Case studies using the M- and LM-samplers	129
10.1 Background to a study	129
10.2 Conditional gene <i>ibd</i> probabilities	131
10.3 Likelihoods and log-likelihoods	133
10.4 Gene <i>ibd</i> in a smaller example	135
10.5 MCMC lod score estimation	137
10.6 Better MCMC lod scores	140
11 Other Monte Carlo Likelihoods in Genetics	147
11.1 Improving pedigree samplers	147
11.2 Interference by Metropolis-Hastings	149
11.3 Inference of typing or pedigree error	154
11.4 Other Monte-Carlo procedures for linkage analysis	156
11.5 Monte-Carlo likelihoods in population genetics	156

List of Tables

2.1	Conditional and joint probabilities of feasible mother-child genotype combinations	17
2.2	Data and estimated frequencies for Bernstein's analysis of <i>ABO</i> blood type determination	18
2.3	Sequence of EM iterates for the example of estimation of the frequency of a recessive allele	23
2.4	EM iterates for the estimation of <i>ABO</i> allele frequencies. The iterates of allele frequencies, and the resulting conditional probabilities of genotype <i>AO</i> and <i>BO</i> , given phenotypes <i>A</i> and <i>B</i> , respectively, are shown in the upper left panel. Then are shown the resulting expected genotype frequencies, given the observed phenotype frequencies and current allele frequency estimates (E-step). Finally, in the lower right are shown the new iterates of the allele frequencies (M-step)	24
3.1	States of gene <i>ibd</i> among the four genes of two individuals	37
3.2	Values of κ , and kinship coefficient ψ , for some standard relationships between two non-inbred individuals	37
3.3	Gene <i>ibd</i> state probabilities at a single locus for a pair of sisters with an aunt, niece, or half-sib. The states are given in the reduced genotypic state-class form, in which the paternal and maternal genes of the three individuals are not distinguished	43
4.1	Critical values for a test size $\alpha = 0.025$ and base-10 lod scores for binomial samples	56
4.2	The groups of offspring genotypes in an intercross design. Note the A_1A_1, B_1B_2 type includes both double-heterozygote two-locus genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1 . The third group includes the four types heterozygous at one of the two loci: $A_1A_1, B_1B_2, A_1A_2, B_1B_1, A_2A_2, B_1B_2$ and A_1A_2, B_2B_2	56
4.3	Probabilities of data observations in an intercross design. Given are the total probabilities of each group of types shown in Table 4.2, under the three alternative hypotheses	57
4.4	Comparison of the information in linkage designs per offspring individual sampled: Kullback Leibler information for testing $\rho = 1/2$ as a function of the true value of ρ	58

4.5	Distinguishing relationships among three individuals who are putatively a pair of sisters with an aunt, niece, or half-sib	61
4.6	Prior autozygosity probabilities over three linked loci for the final individual of the pedigree of Figure 3.1	66
10.1	True gene identity by descent simulated on the modified Icelandic pedigree	131
10.2	Conditional probabilities of gene identity by descent given the marker data simulated on the modified Icelandic pedigree. Shown are probabilities $\times 1000$. For details of the cases (1)–(4), see text	132
10.3	Conditional probabilities ($\times 1000$) of gene <i>ibd</i> among the four <i>C</i> alleles on the pedigree of Figure 10.3, with five equally spaced marker loci, M1 to M5, and for a recessive trait unlinked to the markers . .	136
10.4	Conditional probabilities ($\times 1000$) of gene <i>ibd</i> among the four <i>C</i> alleles on the pedigree of Figure 10.3, with five equally spaced marker loci, M1 to M5. The trait is now in the map, midway between M2 and M3	136
10.5	Summary of LM-sampler runs on the example of section 10.5. The penultimate run, designated (*), is the run also used for the results of Figures 10.9 and 10.10. The first column shows the M-sampler run discussed in section 10.5. The runs were done on a DEC alpha workstation 400-233, with 192 MB memory	141
11.1	Single-site and joint updating schemes on a pedigree	148
11.2	Probabilities of recombination (<i>r</i>) and non-recombination (<i>n</i>) in four equal marker intervals, under interference models I and II and under the Haldane model of no interference (model 0)	151
11.3	Gene <i>ibd</i> probabilities ($\times 1000$) for single loci, and under no interference (Haldane model)	152
11.4	Gene <i>ibd</i> probabilities ($\times 1000$) under the recombination pattern probabilities given for interference models (I) and (II) in Table 11.2. Each run consisted of 10,000,000 whole-meiosis Gibbs/Metropolis updates, and took about 1 hour CPU on a DEC Alpha 400-233 work-station with 256MB memory	153

List of Figures

1.1	An example pedigree from Goddard et al. (1996)	5
1.2	Meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, at any given locus j . The indicators $S_{i,j}$ are shown under the offspring individual, while the resulting labeled founder genes are shown within each individual	6
1.3	The conditional independence neighborhood structure on a pedigree: (a) the individual neighborhood, and (b) the haplotype neighborhood. The reference individual (a) or haplotype (b) is dark shaded. The individuals [haplotypes] defining the local dependence structure for the reference individual [haplotype] are light shaded	7
3.1	An example pedigree. The structure is the same as that of Figure 1.1 of section 1.3. The four individuals shaded grey are bilateral ancestors of the final individual	30
3.2	The relationship triangle for non-inbred relatives	38
3.3	The relationship of quadruple-half-first-cousins	39
3.4	Meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, and patterns of gene identity by descent, at any given locus j : see Figure 1.2	40
3.5	Determination of probabilities $\Pr(Y_{\bullet,j} S_{\bullet,j})$. The gene descent pattern is assumed to be that of Figure 1.2, and the pairs of genes are shown, rather than the individuals. Five individuals, shown as dashed circles, are assumed to be observed, with marker genotypes as indicated: see text for details. (a) Only genes present in observed individuals are constrained in type. (b) Two genes in a single observed individual are jointly constrained	42
4.1	Example of recombination in a three-generation family	53
4.2	Examples of (a) phase-known and (b) phase-unknown backcross linkage designs	54
4.3	Multi-locus genetic marker data are available on a pair of sibs, and on a third related individual, who may be an aunt, niece, or half-sister of the pair	60

5.1	The processes of mitosis and meiosis, shown for a single pair of homologous chromosomes in the nucleus of a cell of a diploid organism. See text for details	70
5.2	The formation of chiasmata, and the crossovers resulting in the chromosomes of the four offspring gametes. The crossovers occurring are the same as in Figure 5.1(e)	71
6.1	The conditional independence structure of data, in the absence of genetic interference	82
6.2	Pedigree without loops. Shaded individuals are those for whom phenotypic data are assumed to be available	86
8.1	The conditional independence structure for MCMC sampling	110
9.1	Model parameters for estimation of a location likelihood curve . . .	123
10.1	The modified Icelandic pedigree. The four individuals marked "Aff" are affected. Those shaded black have marker data available at the majority of the 17 marker loci. The affected half-shaded individual is typed at only two of the marker loci	130
10.2	Expected complete-data log-likelihood components for the simulated data on the modified Icelandic pedigree. Shown are $E_{\gamma_0}(\log_e \Pr(\mathbf{Y} \mathbf{S}) \mathbf{Y})$ (upper curve), and $E_{\gamma_0}(\log_e P_{\gamma}(\mathbf{S}) \mathbf{Y})$ for $\gamma = \gamma_0$ (\bullet , lower curve), and for γ to the left (Δ) and right ($+$) of γ_0 . The location U denotes unlinked. For additional details see text	134
10.3	Hypothetical phenotypic data assumed at each marker locus on the pedigree of Figure 1.1. The four potentially distinct C alleles are labeled C_1 to C_4	135
10.4	Marker ($M1$ to $M5$) and trait (Tr) locations for the example of Figure 10.3. The trait locus is at the midpoint of the ($M2, M3$) interval, so $d_0 = 12.77cM$ and $\rho_0 = 0.1187$	136
10.5	Exact base-10 location lod scores computed using GENEHUNTER 2. The solid lines correspond to having marker data on five pedigree members, and the broken lines to having marker data on only the final affected inbred individual. In each pair, the upper curve corresponds to a trait allele frequency $q = 0.001$, and the lower to $q = 0.05$	138
10.6	Expected complete-data log-likelihoods with the hypothetical data of Figure 10.3 assumed at each of five equally spaced linked marker loci. The notation is as in Figure 10.2	139
10.7	Estimated Monte Carlo location base-10 lod score curve for the hypothetical data of Figure 10.3	139
10.8	Base-10 location score curves for the example of section 10.5 re-estimated, shown also with the exact value	141

10.9 Expected complete-data log-likelihoods for the example of section 10.5, shown for the penultimate run of Table 10.5. The notation is as in Figure 10.2. As in that figure, the contribution from penetrance terms is shown separately from that for segregation terms 143

10.10 Estimated conditional probabilities of recombination in the five map intervals for the example of section 10.5, shown for the penultimate run of Table 10.5. For details, see text 144

11.1 A multiplex meiosis consisting of an ancestral chain of four meioses. These meioses may be jointly updated. For additional details, see text 148

Preface

This monograph is based primarily on material presented at the CBMS Summer Course on **Inferences from genetic data on pedigrees** given at Michigan Technical University, Houghton, Michigan, in July 1999. This monograph is not a textbook; it contains no exercises, and is insufficiently detailed for that purpose. However, it could be used as a textbook, either in conjunction with the excellent texts of Weir (1996), Lange (1997) and Ott (1999), or by advanced students who will consult the cited literature for details.

The notes used at the Summer Course have been augmented by material from two lecture classes given at the University of Washington. A Special Topics class was given in January-March, 1999, and additional background on Markov chain Monte Carlo and Monte Carlo EM are included from that class. Some details were also first presented at a SEMSTAT workshop in Eindhoven in March 1999 (Thompson, 2000*b*). Although material has been added, the examples in Chapter 10 and on identity by descent under interference (section 11.2) were first presented at a Royal Statistical Society Meeting in London, in March 1999 (Thompson, 2000*a*). Versions of Figures 9.1, 10.1, 10.2, 10.6, and 10.7 first appeared in Thompson (2000*a*). However, the 11-chapter monograph follows closely the ten sessions of the Summer Course presentations, with chapter 2 being the only addition, providing statistical background with genetic examples. The order of Chapters 8 and 9 has been reversed from the Summer Course; a case can be made for either ordering.

A more basic Statistical Genetics class was given in Fall 1999, at University of Washington, and led to extensive revision of Chapters 1-4. It is hoped that the monograph can thus serve two purposes. For example, a more introductory course could cover of Chapters 1-4, with final material taken from sections 6.1, 6.2, 7.1, and 7.2. More advanced students could skip Chapters 1-2, skim Chapters 3-5, and study the later chapters more thoroughly.

I would like to thank Dr.Anant Godbole and Dr.JianPing Dong, for their excellent organization of the CBMS Regional Research Conference at Michigan Technical University. I am also grateful to the many students who attended this course, and to students attending the two University of Washington courses, for their helpful comments and criticisms. In particular, I would like to thank Eric Anderson, Nicky Chapman and Dr.Ellen Wijsman for help with LaTeX, BibTeX, Xfig, and GENEHUNTER, and for many discussions. I am grateful to Amy Anderson for her thorough and critical reading of Chapters 1 to 5, and to Eric Anderson, Dr.Erin Conlon, Dr.Mary Kuhner, Anne-Louise Leutenegger, and Jessica

Maia, who all read and commented on other chapters.

Some of the MCMC work was undertaken in collaboration with Dr. Simon Heath. In particular, the implementation of the algorithm described in section 3.6 and the initial incorporation of the L-sampler of Heath (1997) into our M-sampler software to create the LM-sampler (section 10.6) are both due to Dr. Heath. Figures 1.1, 1.2, 3.4, 3.5, and 10.3, first appeared in Thompson and Heath (1999), and are also due to Dr. Heath. I am grateful to Dr. Heath for our continuing collaboration.

The CBMS Regional Research Conference was funded by NSF grant number 98-13767 to Dr. Jianping Dong and Dr. Anant Godbole of The Mathematical Sciences Department of Michigan Technical University, Houghton, MI.

Table of Notation

Since there are an insufficient number of user-friendly letters and symbols, some must be used for more than one purpose. However, for convenience, we summarize the principal usages here

Notation	Usage
Parameters	
θ	the general (set of) parameters of a model
ρ	a recombination frequency parameter
γ	a (trait) locus location
Γ_M	a marker map; set of marker locations
β	a trait model penetrance parameter
r	number of multinomial outcomes (or phenotypes)
p_1, \dots, p_r	probabilities of multinomial outcomes
k	number of alleles at a locus
q_1, \dots, q_k	population allele frequencies at this locus
q	an allele frequency, often for a recessive allele
ψ	a kinship coefficient
ϕ	chiasmata avoidance function
$\kappa_i, i = 0, 1, 2$	gene-identity probabilities
Indices and labels	
i	an index used primarily for individuals or meioses
j	an index used primarily for alleles or loci
k, k_i	a label for a gene
L	a number of loci ordered on a chromosome
m	a count, often of the number of meioses
v	miscellaneous other counts, of genes for example
n	sample size
F	father, or paternal, often as subscript
M	mother, or maternal, often as subscript also marker, as in marker data \mathbf{Y}_M
N	Monte Carlo sample size also (Chapter 5) the random number of chiasmata)
τ	an index of Monte Carlo or MCMC realizations
\mathcal{T}	a set of indices of latent variables
\mathcal{D}	a set of indices of data observations
Variables	
A_1, \dots, A_k	the alleles at a locus
\mathbf{Y} , value \mathbf{y}	the data random variables (usually phenotypes)
\mathbf{Y}_M	phenotypes at marker loci, in linkage mapping
\mathbf{Y}_T	trait phenotypes; $\mathbf{Y} = (\mathbf{Y}_T, \mathbf{Y}_M)$
\mathbf{X} , value \mathbf{x}	latent variables
\mathbf{X}^\dagger	a proposed value of \mathbf{X} in Monte Carlo sampling
\mathbf{X}^*	a sampled or resampled value of \mathbf{X} in Monte Carlo
$\mathbf{G} = \{G_i\}$	the set of genotypes of individuals i
g	a genotype — a possible value of G_i

Notation	Usage
Variables continued	
$\mathbf{S} = \{S_{i,j}\}$	set of meiosis indicators for meioses i and loci j
$S_{\bullet,j}$	the vector of $S_{i,j}$ at given locus j
$S_{i,\bullet}$	the vector of $S_{i,j}$ at given meiosis i
$G_{\bullet,j}, G_{i,\bullet}$	similarly for genotypes, locus j , individual i
$Y_{\bullet,j}, Y_{i,\bullet}$	similarly for phenotypes, locus j , individual i
$Y^{(j)}$	the data $\{Y_{\bullet,1}, Y_{\bullet,2}, \dots, Y_{\bullet,j}\}$; $\mathbf{Y} = Y^{(L)}$
$\mathbf{J} = \mathbf{J}(\mathbf{S})$	a gene <i>ibd</i> pattern, a function of \mathbf{S}
I_1, \dots, I_{L-1}	the intervals between L ordered loci
$\mathbf{R} = (R_j; j = 1, \dots, L-1)$	the recombination indicators in intervals I_j
\mathbf{r}	a vector of recombination indices; value of \mathbf{R}
$\mathbf{C} = (C_j; j = 1, \dots, L-1)$	the chiasmata presence/absence indicators in intervals I_j
\mathbf{c}	a vector of chiasma indices; value of \mathbf{C}
T , value t	a count (often binomial)
t_j	a multinomial count, e.g. of latent genotypes; also (Chapter 5) a set of binary indicators
n_{jl}, n_j	multinomial data counts, of observable phenotypes or genotypes
m_j	multinomial counts, often of alleles
Functions and probabilities	
Pr	probability, when not indexed by a parameter
$\text{Pr}(E; \theta)$	probability of event E under model θ
$P_\theta(\cdot)$	a probability distribution, indexed by θ
$P^*(\cdot)$	a probability distribution, used for the sampling or resampling distribution in Monte Carlo methods
$E_\theta(\cdot)$	Expectation, under a model indexed by θ
$\Phi(\cdot)$	the standard Normal (Gaussian) cumulative distribution function
$I(\cdot)$	the indicator function of an event
$L(\theta)$ or $L_{\mathbf{y}}(\theta)$	the likelihood for parameter θ given data \mathbf{y}
$L(\theta; \mathbf{Y})$	the likelihood function, considered also as a function of data random variables \mathbf{Y}
ℓ or $\ell(\theta)$	the log-likelihood function for parameter θ
$K_n(\theta; \theta_0)$	Kullback-Leibler information in a sample size n
$K_{\mathbf{y}}(\theta; \theta_0)$	K-L information in latent \mathbf{X} given data \mathbf{y}
$H_{\mathbf{y}}(\theta; \theta_0)$	expected complete-data log-likelihood given $\mathbf{Y} = \mathbf{y}$: $E_{\theta_0}(\log P_\theta(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$
$R(\cdot)$ and $R^*(\cdot)$	cumulative probabilities of data used in computing probabilities on graphs or pedigrees
$Q(\cdot), Q^*(\cdot), Q^\dagger(\cdot)$	cumulative conditional probabilities of latent variables given data on graphs or pedigrees
$h(\mathbf{X}^\dagger; \mathbf{X})$	Hastings ratio for proposed \mathbf{X}^\dagger when at state \mathbf{X}
$q(\mathbf{X}^\dagger; \mathbf{X})$	proposal probability for \mathbf{X}^\dagger when at state \mathbf{X}
α	the Metropolis-Hastings acceptance probability