Stochastic Geometry, Data Structures and Applications of Ancestral Selection Graphs

Nicoleen Cloete

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Mathematics at the University of Auckland, 2005

To Danie, Renier and Nicola

May all your dreams come true

Contents

Notation

n : Size of sample taken from a population of organisms, 19

N : Size of the population of organisms, 19

$\boldsymbol{\lambda_0}$ **:** Type 0 birth rate, 19

$\boldsymbol{\lambda_1}$ **:** Type 1 birth rate, 19

UA : Ultimate ancestor, 19, 22

MRCA : Most recent common ancestor, 3, 29, 75, 77

$\boldsymbol{t_{UA}}$ **:** Time to ultimate ancestor, 20, 22

$\boldsymbol{Z_0(t_0)}$ **:** Number of allele type 0 at time t, 20

$\boldsymbol{Z_1(t_0)}$ **:** Number of allele type 1 at time t, 20

$\boldsymbol{\theta}$ **:** Mutation parameter, 20

$\boldsymbol{\sigma}$ **:** Selection parameter, 20

z : Frequency of allele type 0, 20

$\boldsymbol{V_{AR}}$: Label of the vertex above the root, 56

$\boldsymbol{Tr_{(V,E)}}$: Genealogy of an ancestral selection graph with topology (V, E), 79

$\boldsymbol{\mathcal{E}}$: Set of events in a graph, 87

$(\boldsymbol{\beta_m}, \boldsymbol{\gamma_m})$: Label process describing the topology of the graph, 89

$\boldsymbol{\beta_m}$: The label of the edge branching if the m^{th} event is a branching event, 89

$\boldsymbol{\gamma_m}$: The labels of the two coalescent edges if the m^{th} event is a coalescent event, 89

$\boldsymbol{\Sigma}$: The set of selection parameter values used for experiments, 102

$\boldsymbol{w(g, B_Y, \sigma, \theta | B_I, Tr)}$, 102

$\boldsymbol{k(g, B_Y | \sigma \in \Sigma, B_I)}$, 102

LI : Ladder index, 107

$\boldsymbol{\zeta}$: Statistic of interest, 126

$\boldsymbol{\tau_\zeta}$: Integrated autocorrelation time, 126

$\boldsymbol{v_\zeta(s)}$: Auto-covariance at lag s, 126

$\boldsymbol{\hat{var}_\zeta}$: Estimated variance of statistic ζ, 126

$\boldsymbol{t_{FE}}$: Time to the first event, 127

D : Data at the leaf tips, 136

Abstract

The genealogy of a random sample of a population of organisms can be represented as a rooted binary tree. Population dynamics determine a distribution over sample genealogies. For large populations of constant size and in the absence of selection effects, the coalescent process of Kingman determines a suitable distribution. Neuhauser and Krone gave a stochastic model generalising the Kingman coalescent in a natural way to include the effects of selection.

The model of Neuhauser and Krone, determines a distribution over a class of graphs of randomly variable vertex number, known as ancestral selection graphs. Because vertices have associated scalar ages, realisations of the ancestral selection graph process have randomly variable dimensions.

A Markov chain Monte Carlo method is used to simulate the posterior distribution for population parameters of interest. The state of the Markov chain Monte Carlo is a random graph, with random dimension and equilibrium distribution equal to the posterior distribution.

The aim of the project is to determine if the data is informative of the selection parameter by fitting the model to synthetic data.

Acknowledgment

I want to acknowledge my supervisors Dr. Geoff Nicholls from the Department of Mathematics and Assoc. Prof. David Scott of the Department of Statistics for their help, suggestions and encouragement throughout the project. Our weekly discussions will be missed. Dr Nicholls helped me in formulating the project and spent hours helping me to find bugs in the code. Assoc. Prof Scott was very helpful especially at the end when Dr. Nicholls left the University of Auckland. His suggestions on the content of the thesis are greatly appreciated.

I wish to thank Prof. Allen Rodrigo of the Bio-informatics Institute of the University of Auckland for his involvement during the initial stages of this project.

As the recipient of a Bright Future Top Achiever Doctoral Scholarship, I want to thank the Foundation of Science and Industry of New Zealand for the financial support. Without the support this thesis would not have been possible.

I also want to thank all my friends - there are too many to name - but espe-

cially Prof. John Butcher for the moral support throughout my PhD.

The last but certainly not the least, thank you to the most important people in my life, my husband Danie and my two children Renier and Nicola. You guys are better than A1. Sometimes the going was tough, but still you supported me and were always there with smiles.