Springer Theses
Recognizing Outstanding Ph.D. Research

George William Albert Constable

Fast Variables in Stochastic Population Dynamics

Springer

Springer Theses

Recognizing Outstanding Ph.D. Research

Aims and Scope

The series "Springer Theses" brings together a selection of the very best Ph.D. theses from around the world and across the physical sciences. Nominated and endorsed by two recognized specialists, each published volume has been selected for its scientific excellence and the high impact of its contents for the pertinent field of research. For greater accessibility to non-specialists, the published versions include an extended introduction, as well as a foreword by the student's supervisor explaining the special relevance of the work for the field. As a whole, the series will provide a valuable resource both for newcomers to the research fields described, and for other scientists seeking detailed background information on special questions. Finally, it provides an accredited documentation of the valuable contributions made by today's younger generation of scientists.

Theses are accepted into the series by invited nomination only and must fulfill all of the following criteria

- They must be written in good English.
- The topic should fall within the confines of Chemistry, Physics, Earth Sciences, Engineering and related interdisciplinary fields such as Materials, Nanoscience, Chemical Engineering, Complex Systems and Biophysics.
- The work reported in the thesis must represent a significant scientific advance.
- If the thesis includes previously published material, permission to reproduce this must be gained from the respective copyright holder.
- They must have been examined and passed during the 12 months prior to nomination.
- Each thesis should include a foreword by the supervisor outlining the significance of its content.
- The theses should have a clearly defined structure including an introduction accessible to scientists not expert in that particular field.

More information about this series at http://www.springer.com/series/8790

George William Albert Constable

Fast Variables in Stochastic Population Dynamics

Doctoral Thesis accepted by
the University of Manchester, UK

Springer

Author
Dr. George William Albert Constable
Department of Ecology
and Evolutionary Biology
Princeton University
Princeton, NJ
USA

Supervisor
Prof. Alan McKane
School of Physics and Astronomy
The University of Manchester
Manchester
UK

ISSN 2190-5053 ISSN 2190-5061 (electronic)
Springer Theses
ISBN 978-3-319-21217-3 ISBN 978-3-319-21218-0 (eBook)
DOI 10.1007/978-3-319-21218-0

Library of Congress Control Number: 2015943440

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Parts of this thesis have been published in the following journal articles:

G. W. A. Constable, A. J. McKane and T. Rogers
Stochastic dynamics on slow manifolds
J. Phys. A: Math. Theor. Vol 46, 295002 (2013)

G. W. A. Constable and A. J. McKane
Fast-mode elimination in stochastic metapopulation models
Phys. Rev. E Vol 89, 032141 (2014)

G. W. A. Constable and A. J. McKane
Population genetics on islands connected by an arbitrary network: An analytic approach
J. Theor. Biol. Vol 358, 149–165 (2014)

G. W. A. Constable and A. J. McKane
Models of genetic grift as limiting forms of the Lotka-Volterra competition model
Phys. Rev. Lett. Vol 114, 038101 (2015)

G. W. A. Constable and A. J. McKane
Stationary solutions for metapopulation Moran models with mutation and selection
Phys. Rev. E Vol 91, 032711 (2015)

Supervisor's Foreword

It is with the greatest of pleasure that I write the foreword to George Constable's Ph.D. thesis. At one level the thesis is a very clear description of a method of analysing models in population genetics, which I will outline in more detail below. But on a more general level it should also be very useful as an introduction to those wishing to understand the formalism of continuous-time stochastic processes. There are not so many textbooks on this subject, and those which do exist can be technically quite forbidding. This need not be so—the essential ideas and techniques are relatively straightforward to understand—and the opening chapters of this thesis provide an accessible introduction to them.

The main subject of the thesis is the stochastic time-evolution of populations when the individuals making up the population have a very simple genetic make-up: they are haploid and the focus is on a single gene which has only two alleles. The ideas can certainly be extended to more complex systems, but the aim was to develop techniques to allow models to be analysed, and testing them on the simplest situations to start with. Attention was also directed to aspects of the models which have perhaps received less attention than they deserve.

The first of these aspects is the careful specification of the model and its subsequent simplification through the use of a diffusion approximation. Theoretical physicists tend to take great care in distinguishing between microscopic, mesoscopic and macroscopic descriptions of the same system, and the specification of the approximations that are made to go from one level of description to another. There is less of a tradition of doing this in the context of biological systems. In this thesis there is a careful separation of the modelling and approximation processes, so that both the starting point and the nature of the approximations subsequently made are absolutely clear.

The second, and more substantial, contribution described in the thesis is in the development of an additional approximation which makes an intractable equation amenable to analysis. The standard method of specifying a model after the diffusion approximation has been made goes back to the work of Fisher, but was popularised by Kimura in the 1950s. It takes the form of a partial differential equation known as

a Fokker-Planck equation to physicists and as a Kolmogorov forward equation to those in many other disciplines. In simple situations it can be analysed but if, for instance, the population is subdivided into islands (or demes, as they are sometimes called), then this is a partial differential equation in many variables, and as such any in-depth analysis appears to be almost impossible.

As a consequence in some areas of population genetics, the general equations governing the dynamics have not been addressed directly, because of their intractability, and the focus has moved to simpler systems or to those where progress could be made. This is understandable, but what the work in this thesis shows is that this is not necessary; a simple approximation is available which reduces a many-island description to one which is effectively a one-island model, but with effective parameters. The resulting Fokker-Planck equation can now be used to calculate the probability that a particular allele becomes fixed, how long this will take on average or what the nature of the stationary probability distribution of alleles is. There had been a few previous attempts to do this, but the procedure outlined here is both more general and easier to understand than previous studies.

The procedure is based on a fast-mode elimination technique. The idea is very simple and long established in the theory of dynamical systems. Essentially, the variables in the model are decomposed into a set which decay at different rates, $|\lambda^{(i)}|^{-1}$, $i = 1, 2, \ldots$. It turns out that in many cases of interest there is a 'gap' between the smallest $|\lambda^{(i)}|$ (taken to be $|\lambda^{(1)}|$) and all the others. This means that after quite a short time, compared to the timescales of interest to us, only the mode characterised by $|\lambda^{(1)}|$ is left in the model describing the system. This is the 'slow' mode—all the other 'fast' modes have decayed away and dropped out of the theoretical description. What is left is a model with just one degree of freedom, which can then be analysed systematically.

Of course, although the idea just described seems simple enough, finding a concrete procedure which works, and which can be turned into a calculational tool, is not. However, remarkably, a method was found which is both rather straightforward to apply and which also gives results in excellent agreement with computer simulations of the original individual-based (that is, microscopic) model. It results in a 'reduced' model in which the parameters are given in terms of the island sizes, the scale of migration between the islands, or whatever parameters were present in the original (full) model. Although I have used the example of subdivided populations to illustrate the method, it should be more generally applicable to the reduction of complex population genetics models down to much simpler ones with just a few effective parameters which are explicitly given in terms of those of the full model.

I have already mentioned that those looking for an easy-to-understand introduction to the formalism of continuous-time stochastic processes would benefit from reading this thesis. But I would also hope that it would appeal to theoretical biologists seeking to extend the scope of problems it can be applied to, to mathematicians wanting to make the approach more rigorous, and to theoretical physicists looking for an application of the ideas and techniques of non-equilibrium

statistical mechanics. So my hope is that the publication of this thesis will allow a much wider range of people to appreciate the power of the methodology presented here, and also enable them to contribute to extending its range of applicability.

Manchester Alan McKane
April 2015

Abstract

In this thesis, I present two methods of fast variable elimination in stochastic systems. Their application to models of population dynamics from ecology, epidemiology and population genetics, is explored. In each application, care is taken to develop the models at the microscale, in terms of interactions between individuals. Such an approach leads to well-defined stochastic systems for finite population sizes. These systems are then approximated at the mesoscale, and expressed as stochastic differential equations. It is in this setting that elimination techniques are developed.

In each model a deterministically stable state is assumed to exist, about which the system is linearised. The eigenvalues of the system's Jacobian are used to identify the existence of a separation of timescales. The fast and slow directions are then given locally by the associated eigenvectors. These are used as approximations for the fast and slow directions in the full nonlinear system. The general aim is then to remove these fast degrees of freedom and thus arrive at an approximate, reduced-variable description of the dynamics on a slow subspace of the full system.

In the first of the methods introduced, the conditioning method, the noise of the system is constrained so that it cannot leave the slow subspace. The technique is applied to an ecological model and a susceptible-exposed-infectious-recovered epidemiological model, in both instances providing a reduced system which preserves the behaviour of the full model to high precision.

The second method is referred to as the projection matrix method. It isolates the components of the noise on the slow subspace to provide its reduced description. The method is applied to a generalised Moran model of population genetics on islands, between which there is migration. The model is successfully reduced from a system in as many variables as there are islands, to an effective description in a single variable. The same methodology is later applied to the Lotka-Volterra competition model, which is found under certain conditions to behave as a Moran model. In both cases the agreement between the reduced system and stochastic simulations of the full model is excellent.

It is emphasised that the ideas behind both the conditioning and projection matrix methods are simple, their application systematic, and the results in very good agreement with simulations for a range of parameter values. When the methods are compared, however, the projection matrix method is found in general to provide better results.

Acknowledgments

Writing this thesis has been a surprisingly enjoyable experience. It would not have been so without the help and support of those around me.

I would like to offer my most sincere gratitude to my supervisor, Alan McKane, whose advice on all matters has been invaluable, and whose company I have thoroughly enjoyed.

I thank Tim Rogers, who by virtue of infectious enthusiasm, tempered by bouts of cynicism, is a great guy to work with.

Over the past three years I have been lucky enough to be surrounded by people who are happy to take time out of their day to talk about ideas and clarify thoughts. In this regard I am grateful to Tobias, Diana, Louise and Luis, as well as the many who have shared Office 7.26—Alex, Joe, Tommaso, James, Toby, Peter and Cesar.

The funding for this Ph.D. has been provided by a University of Manchester Engineering and Physical Sciences scholarship, and so I thank the EPS for their generosity.

Finally, I am indebted to my family, which have been unfailingly supportive—David and Julie, Deborah and June.

Contents

Chapter 1
Introduction

Essentially, all models are wrong, but some are useful.
George Box [3]

In what follows I am going to explore how concepts and mathematical tools originally developed within physics can be applied to a variety of other fields. These can include, but are not limited to, population genetics, evolution, opinion dynamics, epidemiology and ecology. This thesis will focus primarily on models with an interpretation in population genetics, however models with an ecological and epidemiological flavour will also be explored. With this in mind, let us begin by discussing the questions, 'what do we mean by a model?' and 'what makes a good model?'. The answers to these questions are by no means unarguable, but rather serve to give the reader an impression of the philosophy to which I attempt to adhere.

A model, in its most general form, is a representation of the real world used to help better comprehend or predict its behaviour. In order to gain any tractability, these models feature some degree of abstraction from the real world; the most accurate model would have a one-to-one correspondence with its real world counterpart, but a replica is no more easy to understand than the original. The degree of abstraction is a modelling choice which, to some degree, depends both on the aim of the model and the degree of knowledge one has about the real system. A large degree of abstraction is preferable if knowledge of the real system is limited, or if one wishes the model to be very general. For these reasons, I will be concerned with the quantitative analysis of abstract models which aim to give a qualitative understanding of the behaviour of certain systems.

This thesis will deal exclusively with dynamic models which describe the evolution in time of a system which is characterised by a set of variables. If at some time t_0 the state of the system is described by a vector of variables $\boldsymbol{x}(t_0)$, we ask at some later time $t_0 + \Delta t$ how the properties of $\boldsymbol{x}(t)$ have changed. The model is said to be deterministic if the state of the system at $t_0 + \Delta t$ can be determined precisely. If the system is in state $\boldsymbol{x}(t_0)$ at time t_0, it will be in a calculable state $\boldsymbol{x}(t_0 + \Delta t)$ at time

G.W.A. Constable, *Fast Variables in Stochastic Population Dynamics*,
Springer Theses, DOI 10.1007/978-3-319-21218-0_1

$t_0 + \Delta t$. Such systems are most frequently described by a set of ordinary differential equations (ODEs) of the following form

$$\frac{d}{dt}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}), \tag{1.1}$$

where the vector-valued function $\boldsymbol{f}(\boldsymbol{x})$, along with initial conditions $\boldsymbol{x}_0 = \boldsymbol{x}(t_0)$, defines the model.

Within physics, the principle of Occam's razor is ubiquitous in theoretical work. This philosophy perhaps has its origins in the fact that the physical world has, historically, been found to obey laws which adhere so closely to their model counterparts as to make the line between a model and reality appear blurred. However there is also a more pragmatic factor in constructing simple models, that of gaining an intuitive grasp on how a system behaves. As eluded to earlier, though a model with more variables may provide a more accurate representation of its subject, it will also be harder to understand intuitively. The apotheosis of the interpretability of a mathematical problem comes in the form of an analytic solution. While a far greater range of mathematical problems can be tackled numerically or through simulation than analytically, such methods cannot rival the encompassing power of a single equation that can describe some behaviour of a system.

While physicists in general still hold to an aesthetic desire for simplicity, the last two hundred years has seen an undermining of the belief that simple models are sufficient. Among the issues at the core of this subversion have been the practical intractability and indeterminacy of many variable systems, chaos theory and the realisation that emergent phenomena can defy naïve interpretation. Together, these issues in some sense embody the class of problems at the centre of the discipline known as complex systems.

The first of these problems is encountered while trying to describe the thermodynamic behaviour of gas in terms of its constituent particles, and was tackled by an approach which came to be known as statistical mechanics [10]. If one wishes to understand the behaviour of a gas, one may consider the behaviour of each particle in the gas independently. However, even if we know the exact form of the interaction between the particles and identify the initial conditions of each particle in the system, we are still left with a formidably complicated system of equations to solve. The key to progress is in observing that we are not really interested in all the details that solving the full system would give us; even if we could solve the full set of equations, the velocity of the 13, 500, 303, 304th particle is superfluous information that gives us no understanding as to how the collection of particles behaves. If instead we assume each of the particles is indistinguishable from any other, we can begin to make statistical inferences about the system. We are then interested in the statistics of the mass of particles, about which we can make some analytic progress.

The second of the above problems is chaos. The central, powerful observation of chaos theory is that there exists simple systems (possibly of the form of Eq. (1.1) in three or more variables) whose behaviour is made essentially unpredictable by a limited knowledge of the initial conditions [9]. In a chaotic system, two trajectories

whose initial conditions lie very near one another will evolve in time in an entirely different manner. In order to accurately predict the trajectory a system would follow, one would need to know the initial conditions of the system to infinite precision. Clearly this nullifies the predictive power of the model. In this thesis I will not deal with this class of system. However, the indeterminacy that arises from this phenomena is important in motivating the work that follows, in that it highlights the failing of determinism and points instead towards models where predictability is inherently limited.

A stochastic model is one which incorporates inherent unpredictability. In contrast to a deterministic model, if the state of the system, $\boldsymbol{x}(t)$, is known at some time t_0, we do not know the state of the system at $\boldsymbol{x}(t_0 + \Delta t)$. The variable $\boldsymbol{x}(t)$ is a realisation of a stochastic variable X whose evolution in time cannot be predicted precisely. The exact way in which systems with stochastic components are modelled is once again a matter of choice. In this thesis however, I will concentrate on models which describe the time-evolution of the probability density function (PDF) of the stochastic variable. The PDF gives the probability that a system is in a state $\boldsymbol{x}(t)$. If the system is in state $\boldsymbol{x}(t_0)$ at t_0, rather than ask what state the system is in at time $t + \Delta t$ as in the deterministic description, given by Eq. (1.1), we ask what the probability of being in some state $\boldsymbol{x}$ is at time $t + \Delta t$. However, before we proceed to discuss in full the details of a dynamic-stochastic model, let us review some important results and intuitions from probability theory.

Say a fair, six sided die is thrown in the air. In principle, given enough information about the initial state of the die (its trajectory, weight, alignment etc.) one could calculate on exactly what side the die would fall. However, rather than model all of these parameters in a complicated model, it is common to embody all the dynamical processes in one probabilistic process. With each throw of the die, rather than determine a definite answer, we allow an equal probability of each of the six results occurring. What do these probabilities indicate? If we were to take a frequentist approach, they tells us that if we throw the die an infinite number of times, we would get each side of the die exactly 1/6th of the time [4]. Letting n_x denote the number of times an event x occurs, and N the total number of trials, the probability $p(x)$ of x occurring is then

$$p(x) = \lim_{N \to \infty} \frac{n_x}{N}. \tag{1.2}$$

Equivalently, we could say that if we threw an infinite number of dice, then 1/6th of them would show each side.

What happens if, quite reasonably, we do not have an infinite collection of dice? The expected probability of rolling a 3 is $1/6$. However, given a finite number of dice N, one may not (and possibly cannot) achieve this fraction of threes. Instead, one obtains a sample probability,

$$\bar{p}(x) = \frac{n_x}{N}, \tag{1.3}$$