RESEARCH PLANS

DAVID STEINSALTZ

Over the course of my career I have braided together my interests in probability and analysis with an interest in population genetics and biomedical applications. Most of my research in recent years has been connected, in some cases rather indirectly, with questions of life-history theory, either the mathematics of evolutionary theory of life histories or statistical issues connected with genomics or the empirical study of population-based ecology and medicine. The mathematical methods, while disparate, have generally been from probability theory in the broad sense. They include one-dimensional diffusions, measure-valued dynamical systems, and random-matrix theory. The statistical work has been concentrated in survival analysis, time series, and more recently Bayesian statistics and Markov-chain Monte Carlo.

1. Genomics and social data

I have started a very new collaboration with Melinda Mills and her team in the Oxford University department of sociology. They have been working for some time to analyse the heritability and particular genome-linked factors influencing social-demographic behaviours such as age at first birth, total lifetime fertility, and educational attainment.

The first piece of work to be produced as part of this project is a paper [11] that analyses the statistical behaviour of random-effects models that have become widely used in the social sciences for genome-wide association studies. Questions have been raised [5] about the reliability of this method, which is the basis of much of the work of the Mills group. Drawing on concepts from random-matrix theory, we show in the paper that the method is reliable in principle if its assumptions are satisfied, and we also provide estimates of the errors that are to be expected under certain reasonable scenarios of model misspecification (that is, violation of model assumptions). We develop new tools for addressing questions of bias due to "untagged variation" (gene polymorphisms that are causative but not included in the data) and the possibilities of "negative heritability".

We are now working on extending these methods to phenotypes that are longitudinal or survival times, blending classical survival analysis methods with the high-dimensional random-effects models. We are also looking at the interaction between evolutionary population models and the spectra of genetic-data matrices, and trying to understand how these may affect the estimability of heritability of particular traits.

In joint work with Professor Chris Holmes (University of Oxford department of statistics) and our doctoral student Ryan Christ I have been developing an efficient algorithm for detecting single nucleotide polymorphisms (SNPs) that have been under recent selection, and are hence buried in large regions of generally linked traits. The method pulls together a wide variety of techniques: hidden Markov models on genealogical trees, Gibbs measures, concentration inequalities and efficient numerical linear algebra, to scan rapidly through an entire genome and detect deviations from independence. The method also shows promise for testing interactions within rapidly changing social networks. A software package to make the method available to the wider research community is currently under development.

2. Hidden Markov models for longitudinal behavioural data

My paper [4], joint with Martin Kolb, provided an almost complete classification of the conditions under which a one-dimensional diffusion converges to a quasistationary distribution. The original motivation was to provide a basis for the exploration of certain theoretical models of ageing, where vitality states are given by a Markov process that determines the mortality rate. I have recently taken up this work again, in collaboration with Prof. G Roberts (U Warwick) and our joint doctoral student Andi Wang, because of the application of quasistationary distributions in Prof Roberts's *ScaLE* algorithm for sampling from complicated distributions in multidimensional

Euclidean space. The abstract functional analysis developed by Prof Kolb and myself turns out to be the most direct and general way to demonstrate the necessary convergence properties for the algorithm. We are currently working on a paper based on this work.

While Markov mortality models are widely applied, very little attention has been paid to verifying the relevance of such models, or fitting them to data. It is clear that mortality curves do not provide enough information to distinguish between models. What we need are longitudinal data throughout the life course. These are now becoming available from experiments being performed by Jim Carey at U.C. Davis, whereby an automatic sensing system tracks the behaviour of fruitflies second by second through their whole lives. By fitting these data to a number of different Markov mortality models we hope, first, to demonstrate (or dismiss) the validity of the oft-proposed Markov mortality models for ageing, such as those whose analysis is discussed above; and second, to demonstrate (or dismiss) the existence of a coherent "ageing process".

Our early approaches — which were developed in collaboration by my Queen's University PhD student Andrey Pavlov and myself — represent the vitality as a hidden Brownian motion with drift V_t, which drives the transition probabilities between behaviours according to a logistic formula

$$P_{ij} := \frac{e^{a_{ij}+b_{ij}V_t}}{\sum_{k=1}^{B} e^{a_{ik}+b_{ik}V_t}},$$

where a_{ij} and b_{ij} (as well as the drift and diffusion constant of V_t) are parameters to be estimated. We developed an approximate MLE procedure, based on Gaussian approximation and Kalman filter. As it happens, very similar models have become an object of serious interest in just the last couple of years, so we have refrained from publishing a basic exposition of the model and our preliminary results. Instead, we are working to develop the project in the following directions:

- Model testing: It is difficult to make realistic comparisons of models in situations such as this one, where all the models are large (in numbers of parameters) and false (in the sense of not being taken to be a literally accurate representation of the driving process). We are working on measures of model fit based on Conditional Expected Remaining Lifetime, proceeding from the intuition that predicting future lifetime
- Dynamic versus fixed effects: Recent studies have found that individuals with high variability in blood pressure are more susceptible to strokes. Is this because of the variability itself, or simply because of the amount of time that is spent with relatively high blood pressure? We will be using our Gaussian models — based on available data about individual variability in blood pressure — to suggest monitoring schemes that be most helpful in distinguishing between these possibilities.

Together with my student Gurjinder Mohan I am working to improve the statistical model by applying sequential Monte Carlo methods in place of the Kalman filter. This should make the estimation both more efficient and more flexible (that is, allowing for more general underlying processes.

This line of work is currently in the process of merging with my interest in social data and life histories. I am currently working with postdoctoral research assistant Maria Christodoulou on an analysis of reproductive ageing in the 1958 UK Birth Cohort. This is an extremely rich data set, but also quite challenging to come to grips with, because of the haphazard re-interviewing process and data missing very much not at random. This is exactly the sort of case where Bayesian methods excel, and we hope that estimating filtered ageing processes may provide a new tool for coming to grips with longitudinal data sets.

3. Bayesian survival analysis

Drawing on my work on the theory of ageing, I have been working (together with my former doctoral student DW Bester) to apply Bayesian methods to survival analysis problems with missing

data or covariates with errors. We used these methods in a paper on the influence of nutritional supplements on child behaviour [15]. Further collaboration to analyse results of a series of nutrition studies in UK prisons is in progress.

We were also able to find, using publicly available longitudinal survey data, a powerful effect of blood-pressure variability on individual mortality rates extending over a period of 15 years. This paper [10] is currently being written up in collaboration with Prof D Rehkopf and his student B Seligman at Stanford University medical school.

Longer-term we are looking to develop more systematic tools for model-checking in Bayesian survival models, and to write an `R` package.

I have also become interested in applying Bayesian concepts to meta-analysis. With my graduate student Karla Fox in Canada I worked on a general framework for adapting conventional meta-analysis concepts to survey data. My recent work [9] with G Spence and T Fanshawe (Nuffield Department of Primary Care and Health Sciences) created a Bayesian analogue of standard decision procedures for sequential meta-analysis, in order to determine more efficiently when a sequence of clinical trials may be considered decisive. I hope to find a graduate student interested in pursuing the many open questions left by this work, in particular the optimal choice of stopping parameters and the robustness of the method with respect to uncertainty in the prior distribution.

4. Mutation-selection models

Steven Evans, Kenneth Wachter, and I have been working on the foundations of the theory of mutation-selection balance. A leading theory of ageing states that senescence results from the inability of natural selection quickly to purge deleterious mutations from the genome, if the nocive effects are experienced only late in life. Senescence then appears as a consequence of large numbers of mutant alleles, each of small effect. The overlapping effects of large numbers of nearly neutral mutations falls well outside the scope of standard models of evolution.

In an earlier work we applied the Feynman–Kac formalism to develop a framework which allows for arbitrary sets of mutations, with varying rates and effects, which allows for arbitrary gene interactions ("epistasis"). Our research monograph in the *Memoirs of the American Mathematical Society* series [2] modifies the model to add recombination, which seems to make qualitatively different predictions. This paper proves that moderate rates of recombination, in the context of weak selection of mutation, should be enough to make the whole evolving distribution hew to a dynamical system of a fairly simple sort, defined on Poisson random measures. A series of papers [18, 16, 17] has further developed the implications of this model for evolutionary theory in general, and for the theory of ageing more specifically.

Plans for the near future are to further develop the mathematical theory of this model. While we have focused so far on the infinite-dimensional setting, the recombination version has a natural finite-dimensional version (where individuals accumulate multiple copies of a finite number of mutation types). These may be analysed with the standard methods of low-dimensional dynamical systems. In addition, our methods offer one approach to shoring up some theoretical weaknesses in Ronald Lee's groundbreaking model [6] of social-support effects in the evolution of ageing, which has been criticised for its assumption of an equilibrium without any consideration of dynamics, and unwarranted assumptions of linearity.

There are also questions one could raise about the scaling regime in which the earlier results were set. Since selection and mutation rates are assumed to scale equally (as the inverse of the time scale), each individual ends up with a handful of mutations, which are implicitly assumed to have small phenotypic effect. This makes it hard to interpret as an actual model for the evolution of ageing. More flexible scaling would force us to give up the elegance of a universal limit model, but an intermediate solution depending on the principles of quasi-linkage equilibrium seems feasible.

5. Iterated function systems, Lyapunov exponents, and ecology

Early in my career I published several papers — in particular "Locally contractive iterated function systems" (*Annals of Probability*, 1999) — which study the convergence properties of Markov chains defined by composing a succession of randomly chosen functions. These methods have found some application to problems of statistical estimation, and of theoretical ecology. Ecologists have increasingly come to recognise the importance of environmental stochasticity for determining population growth rates, and hence, ultimately, driving evolution. More recently, concerted efforts have been made to measure environmental stochasticity, and growth rates in different environments, and so estimate stochastic growth rates.

A standard model for an age-structured (or stage-structured) population is to view the current population structure as a vector X_t in $\mathbb{R}^n$ (when there are n possible ages), with the transitions given by $X_{t+1} = M_{t+1}X_t$, where M_{t+1} is an $n \times n$ matrix whose entries include the appropriate mortality and fertility rates in the population in the given time period. It is common to see these matrices as representing changing environmental circumstances, and to model them with a stationary Markov chain.

I have been working with biologists Shripad Tuljapurkar (Stanford) and Carol Horvitz (Miami) on estimating the elasticities of stochastic growth rates — also known as the top Lyapunov exponents. Lyapunov exponents are usually impossible to compute analytically, but may be estimated from simulations. This makes computation of the elasticities — the derivative of the growth rate with respect to changes in the parameters — most challenging. These elasticities are important directly, for understanding how the population may respond to environmental change, both in the short term and over evolutionary timescales; and indirectly, as ingredients in confidence estimates for the Lyapunov exponents themselves. Our first paper [13] on this topic applied Markov coupling techniques to estimate the elasticities with precise error bounds. A second paper, begun several years ago but not completed, implements this theory for two plant systems: one an understory shrub in which climate-change is expected to increase the probability of high intensity hurricanes; the other the African Mahogany tree in which the probability of high intensity harvest of bark and foliage is expected to be impacted by changes in local economic conditions coupled with indigenous perceptions of the state of the resource. We expect the results to help planners understand how changes in the frequency and sequence of environmental states will impact population dynamics. I have recently written extensive C++ computer code to carry out the necessary calculations connected with the algorithm described in the paper, and hope that this will enable the work to be completed in the not-too-distant future.

I am also working on generalising the mathematical and statistical theory introduced in that paper. One useful development, although straightforward, is to recognise that the techniques applied here to discrete-state Markov chains driving linear maps on Euclidean spaces could be applied with relatively little change to more general stationary processes driving non-linear iterated function systems. It will take some work, though, to determine exactly how far the generalisation can practically be taken, and exactly how to translate the terms from the specific setting to the more general. We are also thinking about how to apply these techniques to evolutionary ecological models in continuous time, turning the matrix models into stochastic partial differential equations. While tractable versions of such models may be written down, in principle, it has proven challenging to find sensible scaling regimes in which the different forms of stochasticity — random mating, survival, and environmental stochasticity — interact in a consistent way.

A new collaboration with S Tuljapurkar [12] — posted last year as a preprint, and now being revised — applied this theory to the problem of "diapause", random delays in development that some species have. In a deterministic environment, delays will always be disfavoured, because they slow population growth. Where environmental conditions vary randomly from year to year, though, a species with a fixed lifecourse will receive improvements to population growth from occasional

random delays. Using Banach-space probability techniques we show that a small perturbation to the life history of magnitude ϵ produces a relatively large increase to the population growth rate of magnitude $-1/\log\epsilon$. The same theory may be applied as well to species living on isolated patches that introduce some migration. We have produced an approximate solution for this setting as well, with the increase being on the order of ϵ^p where p is a power depending on the relative favourability of the patches. An exact computation of the power p remains an open question.

Together with my graduate student Fan Wang I have been working to generalise an efficient algorithm for calculating top Lyapunov exponents introduced a few years ago by M Pollicott [8]. Pollicott's algorithm was introduced for sequences of independent 2×2 positive matrices. Our work [14] (in preparation) generalises this to Markovian sequences of matrices of arbitrary dimension. We are currently working to weaken the assumption of positivity, and so make it applicable as well to computing all Lyapunov exponents. More generally, by associating the method with the problem of computing traces on von Neumann algebras (of transfer operators) we are working to explain more clearly why the method works, and the obstructions that can cause it to fail.

6. Understanding the human sex ratio

A major development in the evolutionary theory of aging has been the creation of economic-evolutionary models that incorporate resource transfers into the forces shaping the selective cost of mortality at different ages. Among the first great steps was [3], and soon after Ronald Lee [6] produced what has been touted as a great synthesis, combining the economic-anthropological modelling of Kaplan and Robson with classic stable-population theory and basic selective mechanisms. Unfortunately, serious flaws in Lee's mathematical framework have become apparent, precisely in the elements that seemed so original. The theory cannot be applied literally as it stands. The analysis depends on application of Fisherian stable-population theory, together with a shifting population structure. It requires a separation of time scales which cannot be assumed in general. In one preliminary study, I formulated a simplified version of Lee's model in a mathematically coherent dynamical-systems framework (using only two age classes and stochastic maturation). Once the dynamics had been made explicit, it was clear that the population structure might never settle down, as Lee's model presupposes. My new model has not yet been formalised in a draft paper.

Recognising the link between lifespan and reciprocal investment between generations, and between siblings, leads us to investigate the most conspicuous shaper of family structure and parental investment, the determination of offspring sex. Despite a century or more of research on this question, there is still no clear understanding of the sex ratio in human populations. For this reason, I worked for over a decade with a team of biologists and medical experts to assemble new data, and to integrate them with current evolutionary theory. We have made some intriguing discoveries: A thorough analysis of embryos from preimplantation genetic diagnosis (3 to 5 days), chorionic villus sampling (9 to 12 weeks), and amniocentesis (14 to 17 weeks), and first-trimester elective abortions, shows a nearly balanced sex ratio early in pregnancy, and some evidence of an initial female bias. This would contradict the longstanding belief that the 1.05:1 male:female sex ratio at birth arises from steadily higher male mortality acting throughout pregnancy on an even greater male bias earlier in pregnancy. I collaborated on the study design and on the analysis of the large number of different data sets, which needed to be combined into a unified statstical framework. Since its appearance in PNAS in March 2015 the paper [7] has been downloaded more than 10,000 times.

I have started collaborating with R. Catalano of the UC Berkeley School of Public Health to explore the causes of environmentally-determined fluctuations in sex ratio. Our paper [1] showed that these likely result from a modified tendency to miscarry less fit fetuses, rather than from a modified fitness of the fetuses. With his student A Gemill I have discussed ideas for a meta-analysis

of miscarriage data, which could, if correctly handled, fill in some of the holes in early gestation left by the earlier data sets. The data are extremely disparate, though, so will require careful consideration. I hope we will be able to take this up in the coming year.

This empirical result points up the need for new theoretical consideration of the basis of sex-ratio evolution. The classic Fisherian approach says there should be equal "investment" in males and females, but ignores the interaction of sex with age, and the differing possibilities for facultative sex biasing. This will be a project for the coming years, to update sex ratio theory to account for shifting sex ratios by age.

References

[1] RA Catalano, RJ Currier, and David Steinsaltz. Hormonal evidence of selection in utero revisited. *American Journal of Human Biology*, 27(3):426–431, 2015.

[2] Steven Neil Evans, David Steinsaltz, and Kenneth W Wachter. *A mutation-selection model with recombination for general genotypes*, volume 222 of *Memoirs of the American Mathematical Society*. AMS, 2013.

[3] Hillard S. Kaplan and Arthur J. Robson. The emergence of humans: The coevolution of intelligence and longevity with intergenerational transfers. *Proceedings of the National Academy of Sciences, USA*, 99(15):10221–6, 2002.

[4] Martin Kolb and David Steinsaltz. Quasilimiting behavior for one-dimensional diffusions with killing. *The Annals of Probability*, 40(1):162–212, 2012.

[5] Siddharth Krishna Kumar, Marcus W Feldman, David H Rehkopf, and Shripad Tuljapurkar. Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences*, 113(1):E61–E70, 2016.

[6] Ron Lee. Rethinking the evolutionary theory of aging: Transfers, not births, shape senescence in social species. *Proceedings of the National Academy of Sciences, USA*, 2003.

[7] Steven Hecht Orzack, J William Stubblefield, Viatcheslav R Akmaev, Pere Colls, Santiago Munné, Thomas Scholl, David Steinsaltz, and James E Zuckerman. The human sex ratio from conception to birth. *Proceedings of the National Academy of Sciences*, 112(16):E2102–E2111, 2015.

[8] Mark Pollicott. Maximal lyapunov exponents for random matrix products. *Inventiones mathematicae*, 181(1):209–226, 2010.

[9] Graeme T Spence, David Steinsaltz, and Thomas R Fanshawe. A bayesian approach to sequential meta-analysis. *Statistics in Medicine*, 2016.

[10] David Steinsaltz, DW Bester, Ben Seligman, and David H. Rehkopf. Very short term blood pressure variability and long-term mortality: evidence from the third national health and nutrition examination study. Manuscript in preparation, 2016.

[11] David Steinsaltz, Andrew Dahl, and Kenneth W Wachter. Statistical properties of simple random-effects models for genetic heritability. Under revision for Electronic Journal of Statistics. bioRxiv: 087304, November 2016.

[12] David Steinsaltz and Shripad Tuljapurkar. Stochastic growth rates for life histories with rare migration or diapause. arXiv:1505.00116, 2016.

[13] David Steinsaltz, Shripad Tuljapurkar, and Carol Horvitz. Derivatives of the stochastic growth rate. *Theoretical Population Biology*, 80(1):1–15, 2011.

[14] David Steinsaltz and Fan Wang. Transfer operators of markov matrix products. In preparation, Dec. 2016.

[15] Jonathan D Tammam, David Steinsaltz, DW Bester, Turid Semb-Andenaes, and John F Stein. A randomised double-blind placebo-controlled trial investigating the behavioural effects of vitamin, mineral and n-3 fatty acid supplementation in typically developing adolescent schoolchildren. *British Journal of Nutrition*, 115(02):361–373, 2016.

[16] Kenneth W Wachter, Steven N Evans, and David Steinsaltz. The age-specific force of natural selection and biodemographic walls of death. *Proceedings of the National Academy of Sciences*, 110(25):10141–10146, 2013.

[17] Kenneth W. Wachter, David Steinsaltz, and Steven N. Evans. Vital rates from the action of mutation accumulation. *Journal of Population Ageing*, 2(1–2):5–22, 2009.

[18] Kenneth W Wachter, David Steinsaltz, and Steven N Evans. Evolutionary shaping of demographic schedules. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10846–10853, 2014.