STATISTICS AND NONLINEAR DYNAMICS IN BIOLOGY AND MEDICINE

Giles Hooker (Cornell University),
Jiguo Cao (Simon Fraser University),
David Earn (McMaster University),
Edward Ionides (University of Michigan),
Darren Wilkinson (University of Newcastle)

July 28, 2014 - August 1, 2014

1 Overview of the Field and Recent Developments

This workshop brought together researchers from several different disciplines to work on the problem of performing statistical inference with mechanistic models of real-world systems in biology and medicine. This involved

- Statisticians, whose work is traditionally associated with estimating parameters from data, assessing uncertainty in those estimates and developing methods to perform inference about data and the processes that generate it.
- Applied Mathematicians and Probabilists whose work is focussed on understanding mechanistic models and describing their behavior. This is particularly the case for stochastic models that evolve probabilistically over time.
- Subject area experts who wish to use mechanistic models and statistics to answer substantive questions about their fields of interest. In this meeting we had representatives from fields in Systems Biology, Immunology, Epidemiology and Ecology.

Historically, there has been little communication between these fields: statisticians generally have not worked with the type of mechanistic modeling that applied mathematicians tend to employ and applied mathematicians have rarely been interested in data or in formal statistical questions beyond "the proposed model exhibits the type of behavior we appear to see in the real world".

This has resulted in little attention being given to the problems of conducting statistical inference in these models – estimating parameters, establishing methods to test for goodness of fit, inferential tests about relevant parameter values or properties of the system. Moreover, development of methods to address these problems has frequently been carried out within application disciplines and there has been little cross-talk between the communities carrying out this research.

This workshop was designed to address these problems. The last decade has seen the development of a small community of statisticians, applied mathematicians and probabilists with research interests specifically

addressed towards these issues and who also have established collaborations with researchers in relevant application areas. This workshop provided a venue for all these communities to come together, share ideas, perspectives and establish new collaborative links. Informal feedback from workshop participants has been very positive with many new ideas emerging and the common refrain that the participants had all learned a great deal.

Mathematically, the field is concerned with the problem of fitting models of systems dynamics to data. That is, we posit a Markovian model for the evolution of the state vector x of a system over time:

$$P(x(t+\delta)|x(t),\theta) = g(x(t),\theta)$$

in which x is a vector of values describing the state of the system and θ a vector of unknown parameters and the model describes the (probabilistic) evolution of $x(t)$ from from times t to $t+\delta$. This framework encompasses a very general class of models including finite population models, based on the Gillespie algorithm (Gillespie, 1976), diffusion processes:

$$dx = f(x,\theta)dt + \Sigma(x,\theta)dW$$

and should be taken to be broad enough to include ordinary differential equations (ODEs)

$$\dot{x} = f(x,\theta).$$

Ordinary differential equation models make the evolution of a system's state explicitly deterministic. However, these models have been extensively studied with the applied mathematics literature and produce important challenges for statistical inference. A significant portion of the existing statistical methodological development focusses on these models and they are therefore also included here.

In addition to a model of system evolution, a model is also proposed for the process that generates observations, these can be as simple as the addition of Gaussian noise to the value of the state at a given time

$$y_i = x(t_i) + \epsilon_i, \epsilon_i \sim N(0,\Sigma)$$

but may be more complex. Frequently, not all components of the state vector x can be directly observed, or observed at all, and often only some transformation of x is observable. In general we write

$$P(y_i|x(t_i),\theta) = h(x(t_i),\theta)$$

and assume that observations between observation times are independent. This provides the framework of a *partially observed Markov process* (pomp) under which framework much of the existing methodology is based.

Given this set-up of mechanistically-inspired mathematical models of the evolution of a process that is itself imperfectly observed, key statistical challenges include:

1. Estimating θ from the data $y_1,\ldots,y_n$ and quantifying uncertainty about that estimate, including whether it is identifiable from data at all.
2. Testing hypotheses about θ directly, or about the long-term behavior of the system.
3. Designing systems that improve the estimate of parameters.
4. Assessing the fit of these systems and whether the data contains evidence that the model is insufficient to characterize system properties.

Each of these tasks is computationally and methodologically challenging. While the first of these has received most attention – including at this workshop – none of these can be said to be well understood.

The indirect nature of observations, in this case, makes statistical inference relatively difficult and several approaches have been developed to render it tractable. Broadly, these can be classified into two approaches:

Direct Methods explicitly optimize fit to the data. In the case of ordinary differential equations, this translates to minimizing the squared error between the data and a solution of the ODE. This task is possibly the best studied of problems addressed in this workshop, but represents a particularly challenging optimization problem because the squared error surface that results from nonlinear dynamic models can exhibit many local modes and ridges; a problem that can be exacerbated by numerical solution methods for ODE models. For this reason, many purpose-driven optimization methods have been produced since the late 1970's; for instance in Varah (1982); Bellman and Roth (1971); Bock (1983); Biegler et al. (1986); Baake et al. (1992); Graepel (2003); Li et al. (2005); Chen and Wu (2008); Ramsay et al. (2007) as well as stochastic search methods such as simulated annealing (Jaeger et al., 2004; Colijn and Mackey, 2005). In this context, Bayesian MCMC methods can also be employed, Huang et al. (2006) or using specialized annealing methods Campbell and Steele (2012) or proposals that make use of local geometry (Calderhead and Girolami, 2011; Transtrum et al., 2011).

For models of stochastic dynamics, parameters are obtained by maximizing a log likelihood

$$l(\theta) = \log P(y_1, \ldots, y_n|\theta).$$

This likelihood has no closed-form solutions, making estimation challenging. See, for example, expressions in Aït-Sahalia (2008); Greenwood and Wefelmeyer (1998). To counter this, most estimation methods employ particle filters (Doucet et al., 2000; Carvalho et al., 2010) have generally been employed either to maximize the likelihood via stochastic optimization – an important version of this is in the iterated filtering of Ionides et al. (2006); similar developments in Bayesian analysis can be found in Golightly and Wilkinson (2011); Picchini and Ditlevsen (2011). Alternatively, approximations can be sometimes used (Koyama et al., 2010; Overgaard et al., 2005). Applications can be found in Earn (2009); Cauchemez and Ferguson (2007); He et al. (2010); Goldstein et al. (2009); Hooker et al. (2011); He et al. (2011); Earn et al. (2012); He et al. (2013).

Indirect Methods seek to transform the data so as to produce a simpler numerical problem, although this may entail a reduction in statistical precision. There are a large number of way in which this has been proposed, particularly:

- Fitting numerically estimated derivatives in ODE models. In particular, the following two-stage estimate has received substantial attention in the statistics literature:
  1. Produce a non-parametric estimates $\hat{x}(t)$ and $\hat{\dot{x}}(t)$ of the state vector and its derivative as it evolves over time.
  2. Choose parameters to minimize the deviation between $D\hat{x}(t)$ and $f(\hat{x}(t), \theta)$.

  This method, called either "two-stage least squares" or "gradient matching" has the advantage of avoiding the need to numerically solve the ODE at multiple values of θ, and implicitly avoids the estimation of initial conditions, but requires that enough data be recorded on each state variable to produce reliable non-parametric estimators in the first step. See for example Nevers (1966); Bellman and Roth (1971); Varah (1982); Ramsay (1996); Ellner et al. (2002); Poyton et al. (2006); Lu et al. (2011); Wu et al. (2012). An alternative form is provided by matching the integrals of the data; see Himmelblau et al. (1967); Tang (1971); Yermakova et al. (1982); Vajda et al. (1986); Font and Fabregat (1997), an approach which Itai Dattner updated at this workshop. While a substantial literature on its asymptotic properties exists under the assumption of data generated from an ODE model, little literature exists on its performance in stochastic models.
- More general approaches focus on fitting a set of summary statistics of the data. These are inspired as approximations to sufficient statistics, but can be very general in nature. They may represent qualitative features such as the duration and amplitude of cycles (Tien and Guckenheimer, 2008), if these are observed, or may be more indirect and given in terms of estimated parameters from time-series models or power spectra as in (Reuman et al., 2006; Benton, 2006).

  Once calculated, the model (including the data-generating model) is then simulated many times at any given parameter value and the statistics are recalculated for every statistic. Parameters are then selected based on the correspondence of their simulated statistics with those calculated

from the observed data. In Approximate Bayesian Computation (ABC), a posterior is built up of parameters generated from the prior for which simulations were close to those observed; see Ratmann et al. (2007, 2009); Toni et al. (2009) for examples. Synthetic likelihood assumes an approximately normal distribution for the statistics and optimizes a Gaussian log likelihood (Wood, 2010).

Many of these schemes are computationally intensive methods requiring multiple simulations of stochastic systems. Thus, a great deal of research focusses on improving the computational efficiency of these methods to allow both higher dimensional systems. This entails methods to approximate stochastic models of system behavior by simplified models which can be more efficiently simulated. These approximations take the form of reductions in the dimension of the state vector as well as simplifications of dynamic processes, and generic means to achieve these reductions remains an important open problem.

A final aspect of these set of problems is the roadblocks that exist in the practical implementation of these methods by practitioners. Mechanistic models require more substantial input from the user in specifying the model than the linear regression models more commonly used in Statistics. Moreover, the form in which the model must be put often changes depending on the estimation/inference methodology. This means that developing useful software interfaces to methods is a vitally important task and considerable effort has been made by a number of groups to make these methods accessible; vis NIMBLE Development Team (2014); Hooker et al. (2014); King et al. (2014); Raue et al. (2013); CBIM (2014).

2 Presentation Highlights

2.1 Monday, July 28

The starting session of the workshop featured three plenary talks from Hulin Wu, Simon Wood, and Suzanne Ditlevsen. These researchers have had long involvement in statistics for Ordinary Differential Equations, Stochastic Models especially in Ecology and in the mathematical analysis of these models respectively. They were all exceptionally well-placed to provide an overview of these areas and create a context for the workshop.

Hulin Wu discussed the extension of parameter estimation in Ordinary Differential Equation models to large scale systems biology data. Modern biological techniques allow high-throughput data to be generated on many thousands of genes as well as proteins, RNA and multiple additional components of biological system. The scale of these data pose both modeling and computational challenges. In particular, with thousands or tens of thousands of measured elements of a system acting at several different scales, it is not possible to derive equations for system dynamics from first principles. We must thus employ some form of model selection to choose the structure of these models automatically. Hulin presented methods based on combining the gradient matching ideas described above with LASSO type penalties to both circumvent the computational cost associated with solving ODEs many times and to allow a natural integration of modern model selection methods. These allow the problem of estimating parameters to scale up to 1000's of protein species. These ideas were demonstrated on data describing immune responses to the influenza virus.

Simon Wood provided a description of indirect methods for stochastic systems and their advantages. In particular, he described extensions to the synthetic likelihood methods developed in Wood (2010). These methods create summary statistics of the data and choose parameters so that the distribution of simulations of these summary statistics fit the observed summaries as well as possible. The purpose of this data reduction is to improve the conditioning of the resulting optimization problem. Especially when the observation error is low, directly maximizing a likelihood can be high numerically challenging in chaotic systems where likelihood surfaces exhibit multiple local optima. Further, a judicious choice of summary statistics will produce parameter estimates with close to the same precision as those obtained by maximum likelihood. A key challenge in this is that a normal sampling distribution is assumed for a high dimensional set of summary statistics; some solutions to this were later described by Matteo Fasiolo.

Suzanne Ditlevsen presented an analysis of the modeling choices that statisticians must make in choosing the level of detail that they include in a system under study. In many cases, a simplified description of a system will be more effective at eliciting scientifically-relevant information from data than a more complete, but less-well-determined model. As an example, she presented an example of determining firing times in a noisy neuron model. While it is possible to simulate systems such as a Hogkin-Huxley model of action potentials, these models contain much more detail than can be observed from data consisting of a sequence of inter-spike intervals. Rather, by reducing the model to a simplified oscillation with a suitably-characterized Brownian drift, it is still possible to estimate noise intensity and firing thresholds from such data. This is an important example of the role that applied mathematical analysis plays in the interface between statistics and nonlinear dynamics.

The Monday afternoon session was comprised of short, four-minute talks from all participants as a means of sparking informal discussion and presenting perspectives on these problems. Problems such as data-driven determination of the type of stochastic disturbances to a model, parameter identifiability and the scaling of inference were all touched on by various members and resulted in lively discussion over the succeeding days.

2.2 Tuesday, July 29

Tuesday Morning was devoted to statistical problems involving Ordinary Differential Equations. These models are mathematically relatively easy to work with, but pose numerical problems, both in approximating solutions to these equations and optimizing the choice of parameters for them. For this reason, most of the talks in the session focussed on indirect methods for estimating parameters, although Oksana Chkrebtii presented very novel work on viewing numerical error in estimating differential equations as a further source of uncertainty that can be accounted for under a Bayesian framework.

Tuesday afternoon was focussed on models arising in epidemiology and modeling infectious diseases and in this context talks were all based around statistical methods to uncover additional structure in complex disease models, incorporating climactic variability, spatial distribution and indirect observational processes.

Eberhard Voit discussed the identification of metabolic pathway models. He emphasized that the metabolic pathway systems have strong constraint on the parameters and some intrinsic features. He then discussed the difficulty of estimating suitable parameter values from time series data. He pointed out that there are many instances where the common criterion of the sum of squared residual errors may not be sufficient to evaluate the fit of the model. The even greater challenge is that the most popular functional forms are not known whether to be able to describe biological processes. So it is important to consider the structural uncertainty when modeling metabolic pathway systems.

Jens Timmer gave a wide-ranging talk on sources of uncertainty in parameter estimation in differential equation modeling. These include uncertainty about the performance of numerical optimizers as well as statistical uncertainty in both parameters and model predictions resulting that is inherited from noisy measurements as well as uncertainty about model structure. He advocated the use of local optimizers based on relaxation methods as well as profile likelihood as a tool for expressing uncertainty both about parameters and about predictions. Finally he presented a new observability criterion that allows the investigation of what measurements would most improve parameter identifiability and the precision of parameter estimates.

Itai Dattner focused the statistical inference for ordinary differential equations linear in the parameters. He proposed a new estimation method to address this problem. This method is based on matching the integral of the right hand side function to the observations at each time. The method requires a smooth, but avoids numerical integration and derivative estimation, and can be used in both fully or partially observed systems. He then showed his theoretical and numerical results of this method in the talk.

Giles Hooker discussed gradient matching for mis-specified ODE models. Two-stage methods are often motivated by the computational cost of repeatedly solving ordinary differential equations. This talk took a different view that these methods can also be motivated as providing robustness towards model mis-

specification. In particular, it proposed a model in which an ODE is forced by a smooth, stationary stochastic process. The autocovariance of this process can be estimated by the discrepancy between $\hat{\dot{x}}(t)$ and $f(\hat{x}(t), \hat{\theta})$ and this can then be used to correct confidence intervals for $\hat{\theta}$. Speculatively, it may be possible to employ similar techniques to produce confidence intervals for parameters resulting from other methods that do not insist on an exact solution of the ODE such as the profiling methods in Ramsay et al. (2007).

Oksana Chkrebtii considered the problem of exact Bayesian inference for the solution of intractable differential equation models. In the case where exact solutions are not available in closed form, many existing inferential tools rely on approximations based on time discretisation. Oksana showed that ignoring discretisation error can lead to biased parameter estimates, even for apparently simple ODE models. She went on to introduce a new formalism for the modelling and propagation of solution uncertainty via a Bayesian inferential framework, making extensive use of Gaussian process representations of uncertain functions, allowing exact inference and uncertainty quantification for discretised differential equation models. She illustrated the methods on some challenging chaotic ODE and PDE systems.

Aaron King demonstrated practical issues involved in fitting partially observed stochastic dynamic models to disease transmission data. He showed how time series data can inform key transmission parameters, and then moved on to the topic of how such models can and should be used for making forecasts. Studying cholera in Bangladesh, he showed how rainfall and global climate measures such as El Niño can be used to improve predictive properties of model-based forecasts. Fitting models using appropriate statistical criteria also enables quantification of the associated forecast uncertainty.

Vanja Dukic described a problem in epidemic modeling across spatially distributed locations. The model requires the inclusion of travel structure patterns between US states and incorporates data from Google's Flu Trends. This leads to computational problems in both developing a tractable expression of likelihood and in fitting a system that involves 204 state variables. Vanja presented a Bayesian proposal mechanism to allow tractable inference.

Jooh Ha Park presented a new statistical approach to inference for nonlinear dynamic systems of high dimension, which is a current computational challenge in applications such as the study of space-time systems. Joon Ha's motivating example was joint estimation of disease epidemics in multiple cities, for which full likelihood-based methods have previously been considered intractable. He proposed a variation of Sequential Monte Carlo (SMC) method for estimating latent states which can provide a computationally feasible solution to the joint estimation of dynamic trajectories of interacting systems. This method was shown to reduce the computational cost by a huge amount in a toy, linear-Gaussian example where the sub-systems are weakly interacting, while achieving the desired property that the sampled latent states form a proper sample from its true distribution according to the underlying model. The approach was then applied to the measles epidemic in the UK from 1950 to 1953. Preliminary results, fitting the five largest cities, showed that the proposed methods yield a reasonable estimate of epidemic history with relatively low computational cost. The conventional SMC method applied on the same set of data could not generate any result. Joon Ha also showed how to estimate key epidemic parameters using his space-time filter in conjunction with the Iterated Filtering method of Ionides et. al, 2011, recovering the known result that the transmission rate of measles is substantially different between during school term and during school holidays.

Jiguo Cao talked about selecting ordinary differential equation (ODE) models among competing candidates. He and his collaborators proposed a method for ODE model selection when the competing ODE models are special cases of a full model. Their model selection method has two steps: in the first step, the parameters in the full ODE model are estimated from noisy data; in the second step, the least squares approximation and the adaptive LASSO methods are combined to identify parameters in the full ODE model which are zero. He then talked about their theoretical and numerical results of this method.

2.3 Wednesday, July 30

The Wednesday half day was devoted to the presentation of mathematical analysis and approximation in support of statistical methodology. This takes the form both of supporting the development of computational tools via appropriate analysis and approximation methods, as presented by Matteo Fasiolo and Darren Wilkinson as well as tools for understanding the behavior of dynamic systems and hence aiding (or warning against) the reduction of them to simpler more-tractable models (Junling Ma and Lea Popovic). It was felt that this was a useful prelude to an afternoon left free for discussion and relaxation.

Junling Ma presented a mathematical analysis of epidemic models on structured graphs. In particular, when the contact network of the model is fixed, the basic reproduction numbers of the epidemics changes. Fixed network structure implies that a node cannot reinfect its neighbors before its neighbors recover. This means that the basic reproductive number differs between SIR and SIS models. In turn, this means that the basic reproduction number cannot be readily imputed by examining the the exponential growth rate of the epidemic. For large average degree networks this effect reduces but the result has important implications for forecasting epidemic models in small world networks.

Matteo Fasiolo discussed extensions of the pseudo-likelihood methods described by Simon Wood on Monday. In these methods, simulated data are generated from a stochastic process at a candidate set of parameters and summary statistics are generated from these data. The simulation is repeated enough times to estimate a mean and covariance of the summary statistics and this is used to construct a pseudo-likelihood for the observed test statistics based on a multivariate normal distribution. Parameters are then chosen to maximize this pseudo-likelihood. The methodology avoids many of the computational challenges associated with direct likelihood estimation of partially observed Markov processes and allows for measurements to be made at different scales than the models. For example, ecological models of forest growth simulate forests at the level of individual trees, but data are obtained from remote sensing measurements for which this level of resolution is not very relevant.

Matteo presented a relaxation of the normal assumptions employed for the summary statistics based around saddle point approximations. He demonstrated that the use of these approximations can substantially improve the performance of these estimators especially in cases where the simulation distribution of summaries can be expected to be skewed. Moreover, saddle point approximations can be obtained at much less computational cost than a fully non-parametric approach which would have to overcome the curse of dimensionality.

Lea Popovic considered biochemical mechanisms for which molecular-level stochasticity is critical to biological function. Cellular functions in biological organisms comprise of complex interactions of different proteins, DNA, mRNA molecules, and others. Sources of stochasticity in cells are multiple: some are due to inherent randomness of biochemical reactions between the species, while others are due to variations in cellular composition, cellular division mechanisms, etc. Biochemical reactions may be understood by writing down systems of differential equations or by writing down Markov chain models that explicitly represent these sources of stochasticity. Lea discussed examples where deterministic differential equations are insufficient to understand the biological processes, such as stochastic switching, diffusion moderated sensitivity, and sharpening of spatial patterns. Lea discussed and demonstrated mathematical tools which help to understand how these behaviors can be understood as appropriate limits of stochastic systems.

Darren Wilkinson gave an overview of computationally intensive Bayesian inference for partially observed Markov processes. Markov process models are often intractable in the sense that the discrete time transition density of the process cannot be directly evaluated, and so algorithms that are ”likelihood-free” (in the sense that they do not require evaluation of the likelihood of the Markov process) are particularly valuable. Fortunately, it turns out to be possible to develop likelihood-free inferential algorithms that target the exact posterior distribution, provided that one is able to simulate exact forward realisations from the model. The particle MCMC algorithm known as particle marginal Metropolis Hastings (PMMH) turns out to be especially effective in this context. Darren provided an explanation of the PMMH algorithm and its application to intractable Markov processes, and illustrated it in the context of inference for the rate constants

of stochastic kinetic biochemical network models using time course data. Such models are used extensively in system biology for mechanistic modeling of gene expression.

2.4 Thursday, July 31

Thursday morning sessions were devoted to tools specifically aimed at fitting stochastic dynamic models – ie, those in which the system evolves probabilistically. Presentations on particle filtering and Sequential Monte Carlo by Ed Ionides and Alexandre Bouchard-Coté covered new advances in sequential simulation methods, while Simon Preston examined a localized form of Approximate Bayesian Computation, based instead on simulating entire systems. Michael Dowd provided an introduction to larger scale problems in biological Oceanography in which PDE systems for fluid dynamics interract with stochastic models of Ocean Biology.

The afternoon session was devoted to ecological applications and examined models of animal movement, integral projection models for plant growth that move from individual-level rules to population observations, host-parasite dynamics and software.

Edward Ionides presented a new iterated filtering algorithm. Iterated filtering algorithms recursively combine parameter perturbations with latent variable reconstruction, providing stochastic optimization procedures for latent variable models. Previously, theoretical support for these algorithms was based on using conditional moments of the perturbed parameters to approximate derivatives of the log likelihood function. A new theoretical approach was presented based on the convergence of an iterated Bayes map. A new algorithm supported by this theory was shown to give substantial numerical improvements on a toy example and a computational challenge, inferring parameters of a partially observed Markov process model for cholera transmission.

Michael Dowd presented a survey of statistical methods in Oceanography. In particular, this task mus combine partial differential models of fluid dynamics with ecological, often stochastic, models for ocean biology. Data for these models are obtained by remote sensing as well as by tracts taken from marine gliders. Assimilating models with these data require complex computational tools involving ensembles of multiple models, approximate smoothing methods to impute states and Bayesian state space models to treat both model identification and sampling design.

Alexandre Bouchard-Coté described a Divide-and-Conquer Sequential Monte Carlo (SMC) method for statistical inference on a collection of auxiliary distributions organized into a tree. Compared with standard SMC method, their Divide-and-Conquer SMC exploits multiple populations of weighted particles while still being an exact approximate method. He then showed the application of this method for infer a phylogenetic tree.

Simon Preston described piecewise approximate Bayesian computation (PW-ABC), an approach to inference for discretely but perfectly observed Markov process models, based on dividing the dataset into subsets and using ABC within each subset. The approach is easy to parallelise, and naturally reduces the dataset used for ABC. This reduced dimension obviates the need for summary statistics and large tolerances, making the procedure simple to implement and accurate. Simon explained that the main challenge is the combination of ABC samples in order to form the full posterior density. He discussed two strategies — one involving Gaussian approximations and the other based on kernel density estimates. He discussed the behaviour of the two approaches in the large ABC sample limit, in addition to presenting some numerical results which illustrated the performance of the method in practice. Finally, Simon explored the use of such methods in the context of deterministic models, and the relationship to multiple shooting methods.

Greg Dwyer presented a field ecologists viewpoint on the topic of population dynamics models with a particular application in modeling Gypsy moth outbreaks. These insects undergo outbreaks in which population densities increase by orders of magintude and then crash because of epizootics of fatal, directly transmitted diseases known as baculoviruses. There is severe, density-dependent mortality caused by baculoviruses that suggests they help to drive the long-period, large-amplitude cycles observed in Gypsy moths. Greg pointed out that an important component of fitting these models is auxiliary experiments in which relevant

parameters can be estimated directly rather than resorting to methods in partially observed Markov processes. In particular, it is possible conduct field experiments in which transmission rates can be directly observed by isolating small populations. This then allows parameters to be estimated before being applied to larger-scale observation processes to be assessed for model validity. In the case of Gypsy moth outbreaks, individual host variability in infection risk was evident in small scale experiments and also provided a better fit tp long-term data.

Perry de Valpine focussed on the development of software tools to enable the use of complex dynamic models within statistical estimation routines. A particularly difficult aspect of developing software to implement generally-usable statistical methods with such models is the need to allow a very flexible set of model structures, along with the fact that different statistical methods are generally optimized for models expressed in different ways. Perry introduced a new software package called `NIMBLE` that is aimed at allowing a flexible interface between models and statistical methods. It instantiates models with a `BUGS`-like syntax that then allows an interface to methods in `R`. These models can be compiled into `C++` to allow for fast simulation. The framework is very recently developed and provides a promising interface for ecological modelers.

Steve Ellner focussed on challenges in developing statistical methods for estimating integral projection models – a recently developed set of tools in mathematical ecology (Easterling et al., 2000). These models provide a means of generating population-level observables from individual-level rules for a discrete-time Markov chain through integrodifference equations. Such models have the capacity to answer questions such as whether having a good environment very early on is important to reproductive success, or is total lifetime environment the most relevant quantity. Statistical problems for these models are largely unexamined, such as selecting covariates, indentification of key events, and combining stochastic and deterministic processes.

Scott McKinley presented an analysis of models of animal movement patterns. Many classical movement models assume a scale-free distribution of movement in animal displacement as a first approximation to modeling movement trajectories. Scott advocated for a multi-scale approach involving differing behaviors at local scales from long-range movement. A mathematical analysis of this multi-scale framework demonstrates important consequences for population processes in terms of encounter rates and the functional response of predators.

2.5 Friday, August 1

The final session of the morning was left aside for a discussion of modeling strategies. This was moderated by Edward Ionides and featured Priscilla Greenwood and James Ramsay as discussants: two of the most distinguished researchers associated with the field. A debate was suggested on the provocative proposition that "No one should fit an ODE to data who knows how to fit an SDE to data." A lively discussion ensued regarding the role of random effects in modeling and particularly the use of random processes a means of "hiding" or "accounting for" (depending on one's viewpoint) poor model fit. Both discussants presented provocative and thought-provoking arguments, Jim Ramsay particularly delighting the attendees with his presentation of the naive Albertan's view of statistical models.

The meeting concluded with the general agreement that it had been highly stimulating and effective in cross-fertilizing ideas between participants from different subject areas and a plan that further workshops would be useful and should be planned.

3 Scientific Progress Made

Much of the purpose of the meeting was to make connections between disparate disciplines and to establish cross-disciplinary links rather than to specifically develop new mathematics. However, some important new developments were announced as part of the program of talks. These all point towards making the practical use of methods for data from dynamical systems easier, more broadly applicable, and more computationally efficient.

Particularly important developments are

- A new iterative filtering method that significantly reduces the computational challenges of performing stochastic maximum likelihood with particle filters. Edward Ionides proposed a much simpler means of updating parameters in an older method (Ionides et al., 2006) that makes implementation of these methods significantly simpler and their description much easier.

- Methods to account for smooth stochastic disturbances to ordinary differential equation models when conducting statistical inference (Giles Hooker) present the potential to significantly broaden the scope of models to which two-stage methods can be applied. This helps to bridge the gap between deterministic and stochastic dynamical models – the breadth of models to which these methods can be validly applied remains under investigation.

- The introduction of saddle point approximations for synthetic likelihood (Matteo Fasiolo) serves to significantly broaden the range of models and summary statistics for which synthetic likelihood methods can be validly used. They also represent a modern revival of a rather classical set of tools that have new use in computationally-intensive statistics.

- The use of model selection methods within ordinary differential equations (Hulin Wu and Jiguo Cao) represent important new developments for high-dimensional state-space models. These are particularly valuable in the era of high throughput experiments, particularly in systems biology, when we have only a partial understanding of the processes involved. They allow us to discover a mechanistic model rather than to construct one from scratch.

- The new software platform `NIMBLE` (Perry de Valpine) represents an important new means of implementing statistical methodology for complex dynamical models. Developing general-use software that can be applied across a range of models and methods and that also does not incur unreasonable set-up costs for users is a significant challenge and this additional resource can be expected to make collaboration and methods development significantly easier.

4 Outcome of the Meeting

The goal of the meeting was to generate cross-disciplinary collaboration and understanding in the shared problem of statistical methodology with nonlinear dynamical systems models. In this the meeting was clearly successful; the most common sentiment being that "I learned a great deal." Indeed, there was a considerable amount of cross-disciplinary conversation. There was also general agreement that holding meetings on this topic on a regular basis will be very beneficial for the subject area.

In addition, some concrete projects emerged. The `NIMBLE` software project presented by Perry de Valpine is expected to grow and attract collaboration from attendees. Further, James Ramsay and Giles Hooker are currently writing a book, to be the first statistical treatment of the problem. The meeting furnished much by the way of material for this project, both in terms of updates to methodology and reminders about further subjects to be treated. A manuscript is expected by the end of 2015.

All participants heartily thanked BIRS and its staff for providing truly exceptional facilities and organization to support the meeting.

References

Aït-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics 36*(2), 906–937.

Baake, E., M. Baake, H. G. Bock, and K. M. Briggs (1992). Fitting ordinary differential equations to chaotic data. *Physical Review A 45*, 5524–5529.

Bellman, R. and R. S. Roth (1971). The use of splines with unknown end points in the identification of systems. *Journal of Mathematical Analysis and Applications 34*, 26–33.

Benton, T. G. (2006). Revealing the ghost in the machine : Using spectral analysis to understand the influence of noise on population dynamics. *Proceedings of the National Academies of Sciences 103*(49), 18387–18388.

Biegler, L., J. J. Damiano, and G. E. Blau (1986). Nonlinear parameter estimation: a case study comparison. *AIChE Journal 32*(1), 29–45.

Bock, H. G. (1983). Recent advances in parameter identification techniques for ODE. In P. Deuflhard and E. Harrier (Eds.), *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pp. 95–121. Basel: Birkhäuser.

Calderhead, B. and M. Girolami (2011). Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus*.

Campbell, D. and R. Steele (2012). Smooth functional tempering with application to nonlinear differential equation models. *Statistical Computing 22*, 429–443.

Carvalho, C. M., M. S. Johannes, H. F. Lopes, N. G. Polson, M. Johannes, and F. M. Associate (2010). Particle Learning and Smoothing. *Statistical Science 25*(1), 88–106.

Cauchemez, S. and N. M. Ferguson (2007). Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in london. *Journal of the Royal Society Interface 5*, 885–897.

CBIM (2014). *DEDiscover: Differential Equation Modeling Solution*.

Chen, J. and H. Wu (2008). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to hiv-1 dynamics. *Journal of the American Statistical Association 103*, 369–384.

Colijn, C. and M. C. Mackey (2005). A mathematical model of hematopoiesis I . Periodic chronic myelogenous leukemia. *Journal of Theoretical Biology 237*(2), 117–132.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequetial Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*, 197–208.

Earn, D. J. D. (2009). Mathematical epidemiology of infectious diseases. In M. A. Lewis, M. A. J. Chaplain, J. P. Keener, and P. K. Maini (Eds.), *Mathematical Biology*, Volume 14 of *IAS/ Park City Mathematics Series*, pp. 151–186. American Mathematical Society.

Earn, D. J. D., D. He, M. B. Loeb, K. Fonseca, B. E. Lee, and J. Dushoff (2012). Effects of school closure on incidence of pandemic influenza in Alberta, Canada. *Annals of Internal Medicine 156*(3), 173–181.

Easterling, M. R., S. P. Ellner, and P. M. Dixon (2000). Size-specific sensitivity: applying a new structured population model. *Ecology 81*(3), 694–708.

Ellner, S. P., Y. Seifu, and R. H. Smith (2002). Fitting Population Dynamic Models to Time-Series Data by Gradient Matching. *Ecology 83*(8), 2256–2270.

Font, J. and A. Fabregat (1997). Testing a predictor-corrector integral method for estimating parameters in complex kinetic systems described by ordinary differential equations. *Computers and Chemical Engineering 21*(7), 719–731.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics 22*, 403–434.

Goldstein, E., J. Dushoff, J. Ma, J. Plotkin, D. J. D. Earn, and M. Lipsitch (2009). Reconstructing influenza incidence by deconvolution of daily mortality times series. *PNAS 106*, 21825–21829.

Golightly, A. and D. J. Wilkinson (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus 1*(6), 1–14.

Graepel, T. (2003). Solving Noisy Linear Operator Equations by Gaussian Processes: Application to Ordinary and Partial Differential Equations. In *Twentieth International Conference on Machine Learning*.

Greenwood, P. and W. Wefelmeyer (1998). Cox's factoring of regression model likelihoods for continuous-time processes. *Bernoulli*.

He, D., J. Dushoff, R. Eftimie, and D. J. D. Earn (2013). Patterns of spread of influenza A in Canada. *Proceedings of the Royal Society B-Biological Sciences 280*(1770), 20131174.

He, D., E. L. Ionides, and A. a. King (2010). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface 7*(43), 271–283.

He, D. H., J. Dushoff, T. Day, J. L. Ma, and D. J. D. Earn (2011). Mechanistic modelling of the three waves of the 1918 influenza pandemic. *Theoretical Ecology 4*(2), 283–288.

He, D. H., J. Dushoff, T. Day, J. L. Ma, and D. J. D. Earn (2013). Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. *Proceedings of the Royal Society B-Biological Sciences 280*(1766).

Himmelblau, D., C. Jones, and K. Bischoff (1967). DETERMINATION OF RATE CONSTANTS FQR COMPLEX KINETICS MODELS. *IEC Fundamentals 6*(4), 539–543.

Hooker, G., S. P. Ellner, L. D. V. Roditi, and D. Earn (2011). Parameterizing state-space models for infectious disease dynamics by generalized profiling: Measles in ontario. *Journal of the Royal Society Interface 8*, 961–975.

Hooker, G., L. Xiao, and J. Ramsay (2014). *CollocInfer: Collocation Inference for Dynamic Systems*. R package version 1.0.0.

Huang, Y., D. Liu, and H. Wu (2006). Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. *Biometrics 62*(2), 413423.

Ionides, E. L., C. Bretó, and A. A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*.

Jaeger, J., M. Blagov, D. Kosman, K. Kolsov, Manu, E. Myasnikova, S. Surkova, C. Vanario-Alonso, M. Samsonova, D. Sharp, and J. Reinitz (2004). Dynamical analysis of regulatory interactions in the gap gene system of *drosophila melanogaster*. *Genetics 167*, 1721–1737.

King, A. A., E. L. Ionides, C. M. Bretó, S. P. Ellner, M. J. Ferrari, B. E. Kendall, M. Lavine, D. Nguyen, D. C. Reuman, H. Wearing, and S. N. Wood (2014). *pomp: Statistical inference for partially observed Markov processes (R package)*.

Koyama, S., L. C. Perez-Bolde, C. R. Shalizi, and R. E. Kass (2010). Approximate methods for state-space methods. *Journal of the American Statistical Association*, 170–180.

Li, Z., M. Osborne, and T. Prvan (2005). Parameter estimation in ordinary differential equations. *IMA Journal of Numerical Analysis 25*, 264–285.

Lu, T., H. Liang, H. Li, and H. Wu (2011). High Dimensional ODEs Coupled with Mixed-Effects Modeling Techniques for Dynamic Gene Regulatory Network Identification. *Journal of the American Statistical Association 106*(496), 1242–1258.

Nevers, N. D. E. (1966). Rate Data and Derivatives. *A.I.Ch.E. Jornal 12*(6), 1110–1115.

NIMBLE Development Team (2014). *NIMBLE Users Manual, Version 0.1*.

Overgaard, R. V., N. Jonsson, C. W. Tornø e, and H. Madsen (2005). Non-linear mixed-effects models with stochastic differential equations: implementation of an estimation algorithm. *Journal of Pharmacokinetics and Pharmacodynamics 32*(1), 85–107.

Picchini, U. and S. Ditlevsen (2011). Practical estimation of high dimensional stochastic differential mixed-effects models. *Computational Statistics & Data Analysis 2*(3).

Poyton, A. A., M. S. Varziri, K. B. McAuley, P. J. McLellan, and J. O. Ramsay (2006). Parameter estimation in continuous dynamic models using principal differential analysis. *Computational Chemical Engineering 30*, 698–708.

Ramsay, J. (1996). Principal Differential Analysis: Data Reduction by Differential Operators. *Journal of the Royal Statistical Society. Series B(Methodological) 58*(3), 495–508.

Ramsay, J. O., G. Hooker, D. Campbell, and J. Cao (2007). Parameter estimation in differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B 16*, 741–796.

Ratmann, O., C. Andrieu, C. Wiuf, and S. Richardson (2009). Model criticism based on likelihood-free inference , with an application to protein network evolution. *Proceedings of the National Academies of Sciences 106*(26), 10576–10581.

Ratmann, O., O. Jø rgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf (2007). Using Likelihood-Free Inference to Compare Evolutionary Dynamics of the Protein Networks of H . pylori and P . falciparum. *PloS Computational Biology 3*(11), 2266–2278.

Raue, A., M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmller, and J. Timmer (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE 8*(9), e74335.

Reuman, D. C., R. A. Desharnais, R. F. Costantino, O. S. Ahmad, and J. E. Cohen (2006). Power spectra reveal the influence of stochasticity on nonlinear population dynamics. *Proceedings of the National Academies of Sciences 103*(49), 18660–18665.

Tang, Y. (1971). On the Estimation of Rate Constants for Complex Kinetic Models. *Ind. Eng. Chem. Fundam. 10*(2), 321–322.

Tien, J. H. and J. Guckenheimer (2008). Parameter estimation for bursting neural models. *Journal of Computational Neuroscience 24*(3), 358–373.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface 6*, 187–202.

Transtrum, M. K., B. B. Machta, and J. P. Sethna (2011, Mar). Geometry of nonlinear least squares with applications to sloppy models and optimization. *Phys. Rev. E 83*, 036701.

Vajda, S., P. Valko, and A. Yermakova (1986). A Direct-indirect Proceedure for estimation of Kinetic Parameters. *Computers and Chemical Engineering 10*(1), 49–58.

Varah, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific Computing 3*, 28–46.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature 466*(August).

Wu, H., H. Xue, and A. Kumar (2012). Numerical discretization-based estimation methods for ordinary differential equation models via penalized spline smoothing with applications in biomedical research. *Biometrics 68*(2), 344–52.

Yermakova, A., S. Vajda, and P. Valko (1982). Direct integral method via spline-approximation for estimating rate constants. *Applied Catalysis 2*, 139–154.