Hierarchical Bayesian Methods in Ecology

Devin Goodsman (University of Alberta),
Christian P. Robert (Université Paris-Dauphine),
François Teste (University of Alberta)

September 10–September 12, 2010

1 Workshop Context

Ecosystems are dynamic in both space and time, hence involve multiple spatial and temporal scales, and are often heterogeneous in both of those dimensions, leading to spatial and temporal clustering. Accommodating this complexity in the context of scientific (statistical) hypothesis testing necessitates more advanced methods than those available within the classical null hypothesis testing paradigm.

Rather than ignoring ecological complexity, the modern approach in ecology is to incorporate this complexity into more realistic models. This leads to a more holistic portrayal and understanding of ecology. Once such models are constructed, they can be estimated based on the available data (and statistical principles). Environmental scientists can compare models of divergent ecological hypotheses by comparing their fit to data or predictive power. Thus an essential skill for modern ecologists is to be able to translate scientific hypotheses into ecologically relevant numerical constraints on natural processes, prior to more traditional statistical model comparison or model choice.

Ecologists use many types of statistical models to accommodate ecological complexity. These include random effect and multi-level models to incorporate clustering; ordinary and partial differential equations to model time-continuous changes through time and across space; and Markov models in discrete time that describe changes in ecosystems based on their previous state. There are many variations within each of these model types and many further types that are not covered here. Modern ecological models often have both stochastic and deterministic features, thereby accounting for the inevitable effects of measurement error, process error and natural variability on model performance. The Bayesian framework and paradigm [8] easily conforms to both temporal and spatial variability, as well as to both stochasticity and dynamical process models. However, from the ecologist user perspective, it requires a reasonably deep understanding of the tools of probability theory and probabilistic calculus, as well as statistical inference and stochastic approximation techniques.

Therefore, high-level "numerical literacy" has become an increasingly essential asset for modern ecologists. However, calculus, Monte Carlo approximation, and dynamical modelling courses are rarely part of graduate-level ecological training [1]. The resulting deficit in mathematical training leaves ecologists at a disadvantage and requires arduous self-teaching. This workshop was intended to introduce (mostly local) ecologists to advanced numerical and stochastic techniques and to modelling methodology for better accommodating ecological complexity.

2 Current Ecological Models

The heterogeneity of ecosystems means that they often require spatially explicit models: Moving from one domain to another corresponds to changes in ecosystem parameters and patterns. Incorporating heterogeneity in space and time is difficult using classical statistical methods while more tractable using Bayesian methods. As an illustration, ecologists have used a Bayesian model to reconstruct spatially correlated vegetation composition at a tree level based on fossilized pollen [2]. The complexity involved in reconstructing tree locations from pollen dispersal is evident. Wind impacts where pollen lands relative to the source plant and pollen disperses differently depending on the species.

Dynamical ecological models with unknown model parameters (to be estimated), incorporation of errors due to model inadequacy (termed "process error"), and observation errors are among the most sophisticated stochastic models used by ecologists [3]. These state-space models can be implemented in the frequentist statistical paradigm using the Kalman filter, with the limitation of the Gaussian requirement of this filter, or through Markov Chain Monte Carlo (MCMC) simulation methods in the Bayesian framework, which offers a much wider diversity in the modelling range. Although MCMC algorithms are now standard, their use in complex models remains limited by the current computing power. Due to their flexibility and novelty, Bayesian state space models are beginning to proliferate in the ecological literature despite the high learning cost due to their computational complexity. For instance, ecologists recently used such a model to predict tree growth based on sparse data from tree ring and diameter measurements [4]. State space models are appropriate for such data because they are able to fill in the gaps in the data with estimates, while acknowledging estimate uncertainty thereby enabling a more complete analysis.

Other types of dynamical ecological models involve changes in state variables over continuous time or space. The process models at the core of these dynamical models are often systems of ordinary differential equations. When many state variables are involved, and when there are many interactions in the model, these quickly become high-dimensional. At the current time, Bayesian methods provide the most feasible method of estimating parameters in high-dimensional systems while including stochastic processes to model uncertainty [5]. One example of such high-dimensional system is the ocean biogeochemical cycle involving nitrate, ammonium, dissolved organic nitrogen, phytoplankton, zooplankton and bacteria as state variables [6]. This biogeochemical cycle can be modeled with a system of six ordinary differential equations with parameters estimated using MCMC algorithms that take advantage of the availability of reliable prior knowledge [6]. The predictions of the model were true to observed temporal patterns demonstrating the utility of such methods [6].

3 Workshop Problems and Mathematical Approaches

Each of the participants brought models and datasets to the two-day workshop with the intention of exploring Bayesian parameter estimation techniques. The models that the participants brought to the workshop could be divided into three classes based on complexity and model type: The simplest models were random effect models which usually accounted for clustered experimental designs. The second class of problems were continuous time dynamical models, which typically had ordinary or partial differential equations modeling their processes. The third and most complex class of models were hidden Markov models which are to be fitted in discrete time. These include unmeasured latent variables, and allow for the estimation of observation and process error.

One ubiquitous model type in the whole of environmental science is the random effect model. Random effect models are a special case of both hierarchical and latent models [7]. One variant of random effect models is the random intercept model. A highlight of the workshop was an interactive and fruitful R programming session to estimate parameters for a random intercept Poisson regression of resin canals, a defensive tree trait, as predicted by two tree characteristics. The data were from lodgepole pine trees sampled in a clustered fashion in sites ranging from Grande Prairie to Sundre, Alberta. A Gibbs sampling algorithm was embedded in a Metropolis Hastings algorithm to build posterior probability densities for the parameters including the random intercepts. Posterior densities were built based on the data and uninformative priors. This elicitated a very strong response from the participants, because it showed how much a local decomposition by conditioning and the corresponding Gibbs sampler simplified the analysis of a globaly complex model.

Therefore, although far from being at the cutting edge of statistical science, this random effect model enabled participants to see Metropolis–Hastings and Gibbs sampling algorithms under a realistic perspective, applied to a hierarchical problem they knew. The simplicity of the model enabled participants to much better understand the underlying computational machinery, which can be difficult to apprehend at the conceptual level. Statistical approaches to the more complex model classes explored during the workshop involved application of the same computational machinery. The following models provided by participants are comparable in sophistication to those described in the Current Ecological Models section above.

Until recently, the dynamical models used by ecologists typically ignored observation uncertainty. However, dynamical models that incorporate observation uncertainty within the uncertainty around parameter estimates are increasingly common in the ecological literature. One such model explored during the workshop described the abundance and movement of planktonic larval lice from a salmon farm in the Broughton Archipelago. The process model was an diffusion-advection-decay equation:

$$\frac{\delta \mathrm{n}}{\delta \mathrm{t}} = D\frac{\delta^2 \mathrm{n}}{\delta x^2} - \gamma \frac{\delta \mathrm{n}}{\delta x} - (\mu_n + \theta_n)\mathrm{n}, \tag{1}$$

where n is the density of planktonic larval lice, the diffusion coefficient D represents the combined effect of tides and winds, and random movements of individuals, γ is the advection (flow) of larvae due to currents, and individuals die at a per capita rate of μn and transform to a post-larval life stage at a rate of θn.

The steady state solution of equation (1) gives the larval lice density along a 1-D corridor. One can obtain a likelihood function by assuming observed lice counts on fish are Poisson distributed, with an expected value equal to the model prediction. Using prior information for the parameters (D, γ, μ, θ), in combination with the calculated likelihood and the data for n, we can employ a Metropolis–Hastings algorithm to arrive at a posterior density function for each parameter that includes uncertainty due to variability in the data, and the uncertainty of prior information.

A hidden Markov model explored during the workshop differentiated between ice movement and polar bear movement based on data from global positioning collars that recorded bear movement, but did not distinguish between a bear's own movement and movement due to moving sea ice. The mathematical approach we discussed was to use a state-space model which includes two equations: The first was the observation equation which described the relationship between the movement of the ice and the movement recorded by the global positioning collars. The second equation was a process equation that described the movement strategies of polar bears.

Like with the simplest example, the parameters of this complicated model can be estimated using a combination of Gibbs sampling and Metropolis–Hastings algorithms. Gibbs sampling algorithms are especially useful for hierarchical structures such as those within state space models as they capitalize on conditional dependence relationships that result from the hierarchical structure [9]. In other words, they enjoy a local simplification that allows for a mostly straightforward implementation.

4 Outcome of the Meeting

The intention of this workshop was not to extend the boundaries of statistical science and to achieve new advances in statistical methodology *per se*. Rather, the objective was to enable ecologists, who are not directly involved in statistical research, to understand modern ecological models and to use modern statistical and Monte Carlo techniques in their on-going and future research. In order to accomplish this goal, it was essential that ecologists could freely communicate with statisticians at the forefront of statistical modeling. However, such an exchange requires some investment on the part of ecologists to raise their mathematical fluency and correspondingly a receptive ear from statisticians to understand which were the stumbling blocks towards a better understanding of those methods.

We are aware that many of the workshop participants found the pace of the workshop to be too fast and the material covered to be very challenging. We believe that future workshops dealing with complex ecological models require more than 2 days and more than a single interlocutor/statistical discourse. Two-day workshops—even such as the current running almost round-the-clock over the two days—provide enough time to understand basic applications without delving into mathematical complexity but obvious fall short of providing the "big picture" that would benefit the intended audience. Bearing this self-criticism in mind, we believe that the workshop nonetheless provided participants with the fundamental tools required to explore more complex ecological methods on their own, assuming they are willing to invest the time and effort. As evidence of this, an ecological modeling reading group was initiated at the workshop and continues to the present at the University of Alberta. Like the workshop, the objective of the reading group is to enable ecologists to understand and use sophisticated ecological models in their research. The reading group membership includes nine of the participants from the Hierarchical Bayesian Methods in Ecology workshop.

References

[1] A.M. Ellison and B. Dennis, Paths to statistical fluency for ecologists, *Frontiers in Ecology and the Environment* **8** (2010), 362–370.

[2] C.J. Paciorek and J.S. McLachlan, Mapping Ancient Forests: Bayesian Inference for Spatio-Temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record, *Journal of the American Statistical Association* **104** (2009), 608–622.

[3] B.M. Bolker, *Ecological Models and Data in R*, Princeton University Press, Princeton, New Jersey, 2008.

[4] J.S. Clark, M. Wolosin, M. Dietze, I. Ibanez, S. Ladeau, M. Welsh, B. Kleoppel, Tree growth inference and prediction from diameter censuses and ring widths, *Ecological Applications* **17** (2007), 1942–1953.

[5] J.S. Clark, *Models for Ecological Data*, Princeton University Press, Princeton, New Jersey, 2007.

[6] G.B. Arhonditsis GB, D. Papantou, W.T. Zhang, G. Perhar, E. Massos, M.L. Shi, Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management, *Journal of Marine Systems* **73** (2008), 8–30.

[7] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York, 2007.

[8] C.P. Robert *The Bayesian Choice*, Springer, New York, 2001.

[9] C.P. Robert and G. Casella, *Introducing Monte Carlo Methods with R*, Springer, New York, 2010.