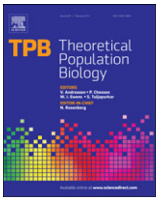
Theoretical Population Biology xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb



A parametric interpretation of Bayesian Nonparametric Inference from Gene Genealogies: Linking ecological, population genetics and evolutionary processes

José Miguel Ponciano

Biology Department, University of Florida, P.O. Box 118525 Gainesville, FL 32611-7320, United States

ARTICLE INFO

Article history:
Received 6 March 2017
Available online xxxx

Keywords:
Demographic stochasticity
Environmental stochasticity
Gaussian Process
Gene genealogies
Ornstein Uhlenbeck process
Population size trajectories

ABSTRACT

Using a nonparametric Bayesian approach Palacios and Minin (2013) dramatically improved the accuracy, precision of Bayesian inference of population size trajectories from gene genealogies. These authors proposed an extension of a Gaussian Process (GP) nonparametric inferential method for the intensity function of non-homogeneous Poisson processes. They found that not only the statistical properties of the estimators were improved with their method, but also, that key aspects of the demographic histories were recovered. The authors' work represents the first Bayesian nonparametric solution to this inferential problem because they specify a convenient prior belief without a particular functional form on the population trajectory. Their approach works so well and provides such a profound understanding of the biological process, that the question arises as to how truly "biology-free" their approach really is. Using well-known concepts of stochastic population dynamics, here I demonstrate that in fact, Palacios and Minin's GP model can be cast as a parametric population growth model with density dependence and environmental stochasticity. Making this link between population genetics and stochastic population dynamics modeling provides novel insights into eliciting biologically meaningful priors for the trajectory of the effective population size. The results presented here also bring novel understanding of GP as models for the evolution of a trait. Thus, the ecological principles foundation of Palacios and Minin (2013)'s prior adds to the conceptual and scientific value of these authors' inferential approach. I conclude this note by listing a series of insights brought about by this connection with Ecology.

1. Introduction

Statistical inference for stochastic processes in biology was central to the research in Paul Joyce's lab. I was humbled and challenged by the request to write a paper celebrating the memory of Paul Joyce's contributions to mathematical modeling and statistical inference in population genetics. Little by little, my fears became excitement when I envisioned a little note illustrating the type of interactions that would occur between the members of his lab, and anyone who approached him to talk about science. Those interactions often occurred very early in the morning, around seven AM, his favorite moment of the day to indulge in research (with coffee).

At the time I started to study under his guidance (summer 2002), professors Zaid Abdo and Vladimir Minin were my lab mates. I had the privilege to learn with and from them through day to day conversations, classes, homeworks, research problems and most importantly, from our successes and failures. The diversity of topics that we talked about and worked on was naturally, a reflection of Paul's innate fascination for any problem in biology having to do with mathematical statistics and stochastic processes. Indeed, he would often be the glue connecting the thinking and ideas among topics. Seeking to see beyond a particular area or application, and understand the connections between probabilistic results applied to one or another area in biology is perhaps, one of the most valuable lessons I got from him.

During one of my last visits to Idaho before his tragic accident, we reminisced about the times when "Vlad" (Minin) was a student. We naturally talked about his (then) latest work, a successful attempt to dramatically improve the accuracy and precision of Bayesian inference of population size trajectories from gene genealogies (Palacios and Minin, 2013). During the rest of our conversation, I proceeded to build a case to demonstrate why I thought that this novel methodology had a remarkable ability to recapitulate fundamental biological properties of the system: because unbeknownst to them, Palacios and Minin's contribution was strongly connected with theoretical concepts and results from statistical ecology. My argument met, of course, a skeptic listener but after my exposition and many interjections, Paul apparently

---

E-mail address: josemi@ufl.edu.

https://doi.org/10.1016/j.tpb.2017.10.007

Please cite this article in press as: Ponciano J.M., A parametric interpretation of Bayesian Nonparametric Inference from Gene Genealogies: Linking ecological, population genetics and evolutionary processes. Theoretical Population Biology (2017), https://doi.org/10.1016/j.tpb.2017.10.007.

conceded because he exclaimed: "Well I hope you're right, because if so, then this would be one of these cool instances in which mathematical population genetics learns from ecological thinking". The reasoning I presented to Paul, formally written, constitutes the contents of this note.

Palacios and Minin (2013) proposed a Bayesian nonparametric methodology to reconstruct past population dynamics using genomic data and the Coalescent process. This non homogeneous Markov death process specifies the relationship between ancestral genealogies of a random sample of genes and effective population size. Because changes in population size result in changes on the genetic pool in a population, at any point in time genomic data carries information regarding past demographic processes and population dynamics. Although estimating the effective population size amounts to estimating the total population size in an idealized Wright–Fisher model, studying changes in this parameter remains important because of its interpretation as a metric of relative genetic diversity.

Motivated by the lack of statistical methods to infer past population dynamics from a sample of genes that did not depend on strong parametric assumptions, Palacios and Minin (2013) proposed a transformed Gaussian Process (GP) as the prior for past population trajectories. These authors justify their choice because such process "does not adhere to a particular functional form, or hypothesis on past population dynamics" (Palacios and Minin, 2013). In this article, I borrow results from theoretical ecology, to show that Palacios and Minin (2013) prior choice, although justifiable under numerical and statistical grounds, can be interpreted as a class of stochastic population dynamics models, albeit one previously not studied and hence, one that brings novel insights into both population genetics and statistical ecology.

Engen et al. (1998) published what now is considered one of the standard references to understand the concepts of "demographic stochasticity" and "environmental variability (stochasticity)" in population dynamics modeling. These authors drew their ideas from the stochastic processes models of Keiding (1975) and Ludwig (1976) which incorporated two main sources of stochasticity: stochasticity due to random births and deaths, known as demographic stochasticity; and temporal stochasticity in any of the demographic rates (*e.g.* good years/bad years for survival, etc. . .). Traditional ecological concepts, such as density-dependence (the regulation of population growth rates according to the density of such population) were also explicitly incorporated in these models. Operationally, formulating a model with the so called 'demographic stochasticity' amounted to specify, for instance, a Branching Process (BP) model with a density dependent offspring distribution of individuals. To add temporal stochasticity into one of the demographic rates, or what came to be known as 'environmental stochasticity' (Lewontin and Cohen, 1969), a temporally uncorrelated random shock was added to the mean of the offspring distribution (often assumed to be Poisson). The result was a density-dependent, BP in Random Environments (BPRE) model (Tier and Hanson, 1981). At that time, various properties of simpler BPRE's had already been worked out by Athreya and Karlin (1971b, a).

Diffusion approximations of the BPRE models later opened the door to the study of animal abundance fluctuations as modeled by realistic, stochastic population dynamics models (Keiding, 1975, Ludwig, 1976, Tier and Hanson, 1981 see Appendix). Straightforward analytical expressions of the properties of the density-dependent BPRE models (such as stopping times and quasi-extinction probabilities) are often too unwieldy or difficult to obtain. Their approximation by means of diffusion processes however, have led to a remarkable improvement in the understanding of how stochasticity from demographic events (births, deaths, etc.) and hence persistence, are affected when the rates themselves are allowed to vary randomly over time. To date, research in this field has yielded a plethora of results that guide the decisions and questions of wildlife managers, population biologists and theoretical ecologists alike (Dennis et al., 1991; Dennis and Taper, 1994; Dennis et al., 1995; de Valpine and Hastings, 2002; Staples et al., 2005; Dennis et al., 2006; Sæther et al., 2007; Lele et al., 2007; Melbourne and Hastings, 2008; Nichols et al., 2009; Knape and de Valpine, 2012; Sæther et al., 2013; Lebreton and Gimenez, 2013; Dennis and Ponciano, 2014; Ferguson and Ponciano, 2014, 2015).

The diffusion approximation of ecological BP models are usually presented as a Stochastic Differential Equation (SDE) model (Tier and Hanson, 1981). The infinitesimal mean of these models usually corresponds to one of the well-known deterministic ODE models of population growth, such as the logistic equation. If only demographic stochasticity is considered (*i.e.*, if a BP model in constant environments is approximated with a diffusion), then the infinitesimal variance of the process scales proportionally to population size, whereas including both environmental and demographic stochasticities results in an infinitesimal variance with two terms, one proportional to population size and one that scales like the square of population size (see Dennis, 2002 and citations therein). Finally, a density-dependent (or density-independent) SDE model of population abundances where the infinitesimal variance scales only like the square of population size has been shown to correspond to a model that assumes no demographic stochasticity and only environmental stochasticity. In what follows, first I briefly summarize the approximation of BPRE's with diffusions. I then expose the relationship between Palacios and Minin (2013)'s prior for the effective population size and stochastic demography. I conclude by showing how, unbeknownst to Palacios and Minin (2013), their GP model brings about a novel parametric understanding of stochastic population dynamics.

2. Palacios and Minin's model and stochastic demography

At the core of these author's approximation is the usage of a transformation of a GP as a prior for the effective population size, $N_e(t)$. GP are stochastic processes such that any finite sample from the process has a joint multivariate normal distribution (Rasmussen and Williams, 2006). As I explain below, this defining property of GPs is crucial for Bayesian inference of a quantity that varies through time, like $N_e(t)$.

In the context of Bayesian statistics, the 'nonparametrics' labeling refers to placing priors to a potentially infinite number of parameters. This approach differs from the classic definition of nonparametric (*e.g.* distribution free) statistics. Palacios and Minin's inference is nonparametric in the sense that they do not adopt any particular functional form for past changes in effective population size (like exponential or logistic growth back from past to present). Their contribution is novel, because instead of choosing from a set of prior beliefs consisting of different functional forms of time for these changes, the authors chose to model the prior for the past trajectory of the effective population size as a collection of points all drawn at random from a general stochastic process. This stochastic process then becomes the prior for the parameter of interest: the entire trajectory of the effective population size $N_e(t)$. As Rasmussen and Williams (2006) put it, a function of time $f(t)$ can be loosely thought of as a very long vector where each entry in the vector specifies the function value $f(t)$ at a particular time t (Rasmussen and Williams, 2006). In Bayesian Inference, the difficulty imposed by having to specify an infinite dimensional object like a function of time as a prior is nicely overcome with GPs. Because finite samples from GPs are jointly multivariate normal, eliciting a prior for the function of interest at a finite number of points in time (here at a collection of points of $N_e(t)$) loosely amounts to sampling from a multivariate normal distribution at