Bayesian learning ecosystem dynamics with delayed dependencies from incomplete multiple source data : an application to plant epidemiology

Stéphane DUPAS^{1,2}

1 Laboratoire Evolution, Génome, Comportement et Écologie, Univ Paris-Sud, CNRS, IRD 247, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France

2 Departamento de Ciencias Biologicas, Universidad de Los Andes, Bogota, Colombia

* E-mail: Stephane.Dupas@ird.fr

Abstract

Ecosystem dynamics forecasting is central to major problems in ecology, society, and economy. The existing models serve as decision tools but their parameters validity are usually not confronted to real data in a formalized approach. Dynamics bayesian network inference is promissing but limited when dealing with incomplete multiple source time series with delayed time dependencies. We propose here a temporal bayesian network with time delay and aproximate inference algorithm, to learn altogether cryptic ecosystem variables, missing data, and model parameters. The novelty in the approach is that it combines simulation-based and likelihood-based aproximate bayesian inference. The advantage of simulation based is that it allows to sample hidden processes. The advantage of likelihood based is that it provides a summary statistics that is really representing the model we are interested in. The ecosystem variables and the missing data are simulated from indicator variables using the probabilistic indicator-ecosystem model. The likelihood is estimated by averaging the probability of observed-simulated data over simulations, the parameter space is sampled with Metropolis Hasting algorithm. Another innovative proposition is to parametrize the network structure in order to learn model structure within a space provided by prior distribution. We apply to plant epidemiology.

24 Introduction

25 The systems we live in, be they natural or human constructions, are ecosystems. Understanding and
 26 predicting their dynamics is central to many of our stakes on earth. We can model their emerging
 27 properties from the knowledge of the interactions, but this pose problems of extrapolation and it is more
 28 difficult to analyse interactions separately than to study the system as a whole and infer interactions
 29 from node dynamics. Yet, this requires methods for statistical inference. Many studies attempts to infer
 30 interaction from statistical relationship in static data, but the causal inference can only be attained from
 31 temporal relationships among system nodes.

32 For deterministic linear models such as lotka voltera, systems parameters solutions can be obtained
 33 by solving systems of equations through matrix algebra [1].

34 When dealing with statistical inference of probabilistic ecosystem dynamics statistical inference, the
 35 most well polished techniques are Kalman filters for continuous state linear dynamic systems (LDS) and
 36 Hidden Markov Models (HMMs) for discrete state sequences. But real world ecosystems are determined
 37 by the interaction between variables of different nature (information, energy and material) that can be
 38 discrete or continuous. In such hybrid bayesian network, the main solution is to discretize continuous
 39 variable and use discrete model inference technique. Truncated exponential has been proposed to better
 40 approximate continuous variable discretization [2]. Note that inference of DBN structure from discrete
 41 or discretized data is promising but has to be interpreted with caution [3]. Another front of research
 42 is nonlinear relationships between variables. It has been approximated as a succession of linear models
 43 indexed by switching variable (Switching linear dynamic systems, SDLS) [4]. Particle filter monte carlo
 44 methods accompanied with resampling have been proposed to treat, in principle any type of distribution,
 45 non linearity and non stationarity [5, 6]. When possible, Rao blackwellized marginalization can treat
 46 problems of sampling in high dimensional space [7]. These particle filters and sequential Monte Carlo
 47 methods can allow to infer posterior bayesian distribution

48 Another issue is the time delay dependance. In DBN, each time slice depend only on the previous

time slice variables (X_t depends only on X_{t-1}). However time delays are comonly observed in ecosystem dynamics and stability properties are highly dependent on time lag and turnover rates properties of the system [8]. Time lags are essential to predict important properties of ecosystems, in particular collapsing risk. A framework for learning temporal bayesian network with deeper time dependencies would be useful.

Finally most of the data set for ecosystem dynamics inference contain missing data time series data is usually incomplete. And the probability that some of the measures in the time series is lacking increases exponentially with the number of data. Missing data can include a whole variable. Variables interacting ecosystems and constituting the ecosystem can be unmeasured but estimated though indicators or samples.

When considering a new ecosystem, a specific mathematical and software solution may or may not be found. A generic inference tool allowing to encompass the different ecosystem types would however be of interest.

Approximate Bayesian Computation is now comonly used for model inference in ecology and evolution [9]. Parameters are sampled from prior distribution, data are simulated and summarized with statistics and the parameters that gave the summary statistics closest to the data are conserved in the posterior [10, 11]. It has been suggested of use for modelling dynamic bayesian networks. In many ecosystem dynamics, simulating time series based on model parameters would bring a lot of noise in summary statistics and reduce method power due to historical contingency. Like for kahlman filter it would be interesting to update historical simulations with data in real time. Another issue with ABC is the choice of summary statisttics. In this work we propose to combine likelihood and Monte Carlo data simulation to get the advantage of each. Ecosystem variable needed for likelihood calculation are simulated from observation and model altogether. Then posterior model parameters can be sampled using likelihood averages times prior probability.

Example of plant epidemiological models. Use prior information of known epidemiological model [12]

Probabilistic model

Let us consider the data X_i^t representing several incomplete times series where X_i is the variable, and t the time when it was sampled. X_i^t are directly related to underlying interacting ecosystem variables Y_i^t . X represent the measured variables and Y represent the interacting variables in the ecosystem. For instance, X_i^t is an observation of a species at time t and Y_i^t is the density of this species at time t , or, X_i^t is an event of pesticide application in a field, and Y_i^t is the concentration of pesticide residues in the leaf at time t . We assume we can define a recursive form of the ecosystem model which parameter are time independent. Note that to model the ecosystem dynamics among Y_i we may need additional hidden variables Z_j . For instance the fungus targetted by the pesticide can have a cryptic density Z_j^t at latent development stage within leaves cells where it cannot be observed directly [12]. In addition, for the observed variables X_j^t there can be periods with missing data. The difficulty in learning such dynamics models is to learn hidden variables Z and unobserved variables at all time. In order to get rid of these latent variables we propose a recursive sheme with deeper time dependencies than from t to $t + 1$ so that we focus on modeling dependencies among variables directly linked to observations. The rationale of our approach is to be able to adapt machine learning to any kind of data and ecosystem.

The solution I propose in this paper to this problem of inference of non linear probabilistic ecosystem dynamic model with hidden variable and deep time dependence from incomplete data is to combine data simulation with likelihood. First the ecosystem variables are simulated from indicator data using indicator and ecosystem models, then joints probability of observed and simulated data is calculated and aeraged over simulations for each prior sample. The algorithm is not going to learn hidden variables but to simulate them. Data completed by simulations will allow to calculate an estimate of the likelihood of the model. If the residual variation of the simulated data is not too important averaging over an acceptable number simulations can provide estimates of model likelihood and permit bayesian inference without to make use of non heuristic summary statistics as in Aproximate Bayesian Computation. I name this approach, MCL (Monte Carlo likelihood), or bayesian computation with monte carlo likelihood.

Figure 1. Generic ecosystem - indicator model

Let us consider the ecosystem dynamics model presented in figure 1. It is constituted with a time series of indicators $[X_i^t]$ collected at time t for variable X_i related to ecosystem variables $[Y_i^t]$. $p(X/Y, \theta_{X,prior})$ is the sampling distribution model of the indicator variables, and $p(Y/\Gamma(Y), \theta_{Y,prior})$ is ecosystem dynamics model of underlying ecosystem variables described in figure 1 for its unrolled version, where $\Gamma(Y)$ are the parents of Y . Note that $\Gamma(Y)$ can be given as prior or parametrized through θ_{prior} . The model $\theta, \Gamma(Y)$ is time independent but not markovian since Y_t does not depend only on Y_{t-1} . Time dependence can also be delayed with $\Gamma(Y)$ given by θ value. In the case of potato late blight disease the fungus development includes a latent period where development is cryptic in the leaves and it's survival is not susceptible to environmental conditions (Figure 2). Sporangium produced and infesting leaves the day before the sampling will not be observable as mycelium on the leaves. Since the time dependance can be delayed, these cryptic variables can be analytically removed from the model with time delayed responses. In the case of potato late blight disease, the cryptic variable is the latent stage of the fungus, that give birth to exposed micelium outside the leave and originated from infestation by sporangium a few days before (the development time depends on the temperature). We can propose a prior function relating the number of newly emerged mycelium from past infesting sporangium as a result of the number of sporangium at

113 a period before which length depend on development rate of the latent stage, and its infestation rate at
114 that stage, that depends on climatic conditions. We recomend removing analytically any criptic variable
115 from the model.

116 We assume we can define a prior model $p(Y/\theta_Y, \Gamma(Y))$ for ecosystem variable dynamical interactions
117 with prior probability distributions for θ_Y . We assume we can define a prior model for measured variable
118 sampling $p(X_i^t/Y_i^t, \theta_X)$ where θ_X are sampling parameters. For instance for temperature they determine
119 the precision of the thermometer or the difference between the temperature at the plant leaves its value
120 in the meteorological station. The model can also include deterministic fixed parameters.

121 The likelihood for a sample prior will be calculated by averaging probability of simulating ecosystem
122 data from indicator and ecological sample prior as detailed in Dupas et al unpublished.

123 Condider x indicator data. If we simulate y from $p(y/x, \theta_X, \theta_Y)$, then model likelihood is (Dupas et
124 al unpublished)

$$\mathcal{L} = \sum_{k=1}^{n_{sim}} p(x, y/\theta'_X, \theta'_Y) \quad (1)$$

125 *Algorithm*

- 126 1. Discretize spatio-temporally the data
- 127 2. Define indicator prior sampling model $p(X/Y, \theta_X)$ and ecological prior sampling model structure
128 $\Gamma(Y)$ and probability distribution $p(Y/\Gamma(Y), \theta_Y)$
- 129 3. Define parameter proposal rules of metropolis hasting sampler
- 130 4. Sample $\theta^{(0)} = [\theta_X^{(0)}, \theta_Y^{(0)}]$ from their prior
- 131 5. Repeat for $j \in [1..n_{generations}]$:
 - 132 (a) define θ' from $\theta^{(j)}$ using proposal rules
 - 133 (b) if $j > 0$ set $p = p'$

134 (c) Repeat for $s \in [1..n_{simulations}]$:

135 i. For $t \in [t_{min}, t_{max}]$ and for $i \in [1..n_{variables}]$, complete time series data by simulating

136 $y_i^t \hookrightarrow p(Y_i^t/X_i^t) \propto p(X_i^t/Y_i^t) \times p(Y_i^t) \times p(Y_i^t/\Gamma(y_i^t))$, where $\Gamma(y_i^t)$ represent the parents of

137 y_i^t , already simulated.

138 ii. Calculate (x, y) probability $p(x, y/\theta'_X, \theta'_Y)$ over the completed time series

139 (d) estimate the likelihood as the average of (x, y) probability among simulations

140 $\mathcal{L} = \sum_{s=1}^{n_{simulations}} p(x, y/\theta'_X, \theta'_Y)$

141 (e) Calculate posterior $p' = p(\theta'_X, \theta'_Y/x, y) \propto \mathcal{L} \times p(\theta'_X, \theta'_Y)$

142 (f) if $j > 0$ Calculate $\alpha = \frac{p'}{p}$ and

143 i. if $\alpha \geq 1$ set $\theta^{(i)} = \theta'$

144 ii. if $\alpha < 1$ set $\theta^{(i)} = \theta'$ with probability α

145 iii. if $thinning \mid j$ then save θ in θ_{sample}

146 6. return θ_{sample}

147 Plant epidemiology example : the potato late blight

148 I will propose non linear probabilistic models for the different components of the likelihood in the case

149 of potato late blight disease (figure 2). Plant epidemiology depends on historical and environmental

150 constraints which are very specific to each locality. We need to be able to learn from these specificities

151 in real time.

152 In this model we consider the prior information of *Phytophthora infestans* life cycle relationships

153 with environmental variables proposed by [12] where the epidemics is determined by

- 154 1. $rLGR$, the relative lesion growth rate, function of T , the mean daily air temperature
- 155 2. rSR , the sporulation rate function of TH , the mean daily air temperature when relative humidity
- 156 is above 90%,

Figure 2. Potato late blight ecosystem model. We distinguish between indicator variables in yellow and ecosystem variables in green. Indicator variable and ecosystem variable are linked by sampling or indicator probabilistic model through bayesian rules (yellow arrows). Ecosystem variable are linked together by ecosystem temporal bayesian network model with deep time dependence (black arrows). Nodes in grey are undelying missing data

157 3. LP , the latent period, or the period of intracellular development during which the disease is unde-
158 tectable in the leaf tissues, function of T .

159 In [12] the relationship between these parameters and climatic variables are polynomial of second or third
160 order with 3 or 4 parameters each.

161 Acknowledgments

162 This work has been supported by IRD EGCE research unit project and BIO_INCA international mixed
163 laboratory

164

165 References

- 166 1. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rtsch G, Pamer EG, et al. Ecological Modeling
167 from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. PLoS
168 Computational Biology. 2013;doi:10.1371/journal.pcbi.1003388.
- 169 2. Cobb BR, Shenoy PP. Inference in hybrid Bayesian networks with mixtures of truncated exponen-
170 tials. 2005;doi:10.1016/j.ijar.2005.06.002.
- 171 3. Sander EL, Wootton JT, Allesina S. Ecological Network Inference from Long-Term Presence-
172 Absence Data. Scientific Reports. 2017;doi:10.1038/s41598-017-07009-x.
- 173 4. Pavlovic V, Rehg J, Tat-Jen C. A Dynamic Bayesian Network Approach to Tracking Using Learned
174 Switching Dynamic Models. In: Lynch N, Krogh BH, editors. Hybrid Systems: Computation and
175 Control. HSCC 2000. Lecture Notes in Computer Science, vol 1790. Berlin, Heidelberg: Springer;
176 2000. p. 366–380.

- 177 5. Doucet A, de Freitas JFG, Gordon NJ. Sequential Monte Carlo Methods in Practice. Springer-
178 Verlag; 2000.
- 179 6. Guo D, Wang X, Chen R. New sequential Monte Carlo methods for nonlinear dynamic systems.
180 Statistics and Computing. 2005;doi:10.1007/s11222-005-6846-5.
- 181 7. Doucet A. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In: Uncertainty
182 In Artificial Intelligence Proceedings; 2000. p. 176–183.
- 183 8. May RM. Time-Delay Versus Stability in Population Models with Two and Three Trophic Levels.
184 Ecology. 1973;54(2):315–325. doi:10.2307/1934339.
- 185 9. Beaumont Ma. Approximate Bayesian Computation in Evolution and Ecology. Annual Review
186 of Ecology, Evolution, and Systematics. 2010;41(1):379–406. doi:10.1146/annurev-ecolsys-102209-
187 144621.
- 188 10. Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in
189 practice. Trends in Ecology and Evolution. 2010;25:410–418. doi:10.1016/j.tree.2010.04.001.
- 190 11. Robert CP, Marin Jm, Pillai NS. Why approximate Bayesian computational (ABC) methods
191 cannot handle model choice problems. Systems Biology. 2011;arXiv:1101:1–22.
- 192 12. Andrade-Piedra JL, Hijmans RJ, Forbes GA, Fry WE, Nelson RJ. Simulation of Potato Late Blight
193 in the Andes. I: Modification and Parameterization of the LATEBLIGHT Model. Phytopathology.
194 2005;95(10):1191–1199. doi:10.1094/PHYTO-95-1191.