Computational Statistics and Data Analysis 78 (2014) 82–99

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Parameter estimation via stochastic variants of the ECM algorithm with applications to plant growth modeling

S. Trevezas [a,*], S. Malefaki [b], P.-H. Cournède [a]

[a] Laboratory of Mathematics Applied to Systems, École Centrale Paris, Grande Voie des Vignes, 92290 Châtenay-Malabry, France

[b] Department of Mechanical Engineering & Aeronautics, University of Patras, GR 26500 Rio Patras, Greece

ARTICLE INFO

Article history:
Received 20 February 2013
Received in revised form 7 April 2014
Accepted 9 April 2014
Available online 18 April 2014

Keywords:
Plant growth model
Hidden Markov model
Monte Carlo ECM-type algorithm
Metropolis-within-Gibbs
Automated Monte Carlo EM algorithm
Sequential importance sampling with resampling

ABSTRACT

Mathematical modeling of plant growth has gained increasing interest in recent years due to its potential applications. A general family of models, known as functional–structural plant models (FSPMs) and formalized as dynamic systems, serves as the basis for the current study. Modeling, parameterization and estimation are very challenging problems due to the complicated mechanisms involved in plant evolution. A specific type of a non-homogeneous hidden Markov model has been proposed as an extension of the GreenLab FSPM to study a certain class of plants with known organogenesis. In such a model, the maximum likelihood estimator cannot be derived explicitly. Thus, a stochastic version of an expectation conditional maximization (ECM) algorithm was adopted, where the E-step was approximated by sequential importance sampling with resampling (SISR). The complexity of the E-step creates the need for the design and the comparison of different simulation methods for its approximation. In this direction, three variants of SISR and a Markov Chain Monte Carlo (MCMC) approach are compared for their efficiency in parameter estimation on simulated and real sugar beet data, where observations are taken by censoring plant's evolution (destructive measurements). The MCMC approach seems to be more efficient for this particular application context and also for a large variety of crop plants. Moreover, a data-driven automated MCMC–ECM algorithm for finding an appropriate sample size in each ECM step and also an appropriate number of ECM steps is proposed. Based on the available real dataset, some competing models are compared via model selection techniques.

1. Introduction

Mathematical modeling of plant development and growth has gained increasing interest in the last twenty years, with potential applications in agricultural sciences, plant genetics or ecology. Functional–structural plant models (FSPMs, Sievänen et al., 2000) combine the description of plant architectural development and ecophysiological functioning, and offer the most promising perspectives for a better understanding of plant growth (Vos et al., 2010). However, the parameterization of FSPMs is generally impeded by several difficulties: the complex and interacting mechanisms which guide plant's evolution generally are translated into strongly nonlinear models involving a large number of equations and parameters; experimental protocols to collect detailed data are heavy and often inaccurate; finally, plant models are generally developed without an appropriate statistical framework. As a consequence, plant growth models often remain

* Corresponding author. Tel.: +33 0141131798.
E-mail addresses: samis.trevezas@ecp.fr (S. Trevezas), smalefaki@upatras.gr (S. Malefaki), paul-henry.cournede@ecp.fr (P.-H. Cournède).

http://dx.doi.org/10.1016/j.csda.2014.04.004
0167-9473/

descriptive without a real predictive capacity. Efforts have thus been undertaken in the recent years to develop methods for parameter estimation and uncertainty assessment adapted to complex models of plant growth (Ford and Kennedy, 2011; Cournède et al., 2011; Trevezas and Cournède, 2013). In this paper, a certain class of plants with known organogenesis (in plants, organogenesis is the process of creation of new organs) is studied, whose growth is modeled by GreenLab FSPM (de Reffye and Hu, 2003). A lot of agronomic plants can be modeled in this way, like maize (Guo et al., 2006), rapeseed (Jullien et al., 2011), grapevine (Pallas et al., 2011) or even trees (Mathieu et al., 2009). The parameters of the model are related to plant functioning. The vector of observations consists of organ masses, measured only once by censoring plant's evolution at a given observation time (destructive measurements).

In Cournède et al. (2011), a first approach for parameter estimation was introduced but based on the rather restrictive assumption of an underlying deterministic model of biomass production and uncorrelated errors in the mass measurements of different organs in the plant structure. In Trevezas and Cournède (2013), the authors proposed a more general framework for statistical analysis which can potentially be applied to a large variety of plant species by taking into account process and measurement errors. They provided a frequentist-based statistical methodology for parameter estimation in plants with deterministic organogenesis rules. This framework can also serve as the basis for statistical analysis in plant models with stochastic organogenesis (see Kang et al., 2008 for the description of GreenLab with stochastic organogenesis). The basic idea consists in describing data (organ masses) measurements as resulting from the evolution of a non-homogeneous hidden Markov model (HMM), where the hidden states of the model correspond to the sequence of unknown biomasses (masses measured for living organisms) produced during successive growth cycles. In such a complex model, the maximum likelihood estimator (MLE) cannot be derived explicitly and for this reason a Monte Carlo ECM-type (Expectation Conditional Maximization) algorithm (Dempster et al., 1977; Meng and Rubin, 1993; Jank, 2005b; McLachlan and Krishnan, 2008) was adopted to compensate for the non-explicit E-step and also the non-explicit M-step. The authors used sequential importance sampling with resampling (SISR) to simulate from the hidden states given the observed data. The M-step is performed with a conditional maximization approach (see, ECM in Meng and Rubin, 1993), in which the parameters of the model are separated into two groups, one for which explicit updates can be derived by fixing the parameters of the other group, and one for which updates are derived via numerical maximization.

Due to the typically large number of equations and time steps to consider in plant growth models, the computational load is an important factor to take into account. Consequently, the efficiency of the estimation algorithms is a key issue to consider, especially when the final objective is decision-aid in agriculture. Likewise, as one of the objectives of FSPM is to be able to differentiate between genotypes (two different genotypes should be characterized by two different vectors in the parameter space Letort, 2008 and Yin and Struik, 2010), the accuracy of the estimation algorithms has to be assessed. In this context, it is very important to profit from advanced simulation techniques in order to reduce the Monte Carlo error associated with a given estimation algorithm. For this reason, we focus on the comparison of different simulation techniques which are performed in the E-step. The resulting approximation of the so-called Q-function (computed in the E-step) is crucial to the quality of parameter estimation. The most efficient algorithm can subsequently be used to calibrate agronomic plants with the method of MLE and then make model comparison and selection. An example of this type is presented in the current paper based on a dataset from the sugar beet plant. Moreover, in order to enhance computational efficiency, the design of automated and data driven algorithms should help by making an efficient use of Monte Carlo resources, for the benefit of the users. The above arguments motivate the current study.

In this paper, we compare different simulation techniques to perform the E-step. In particular, three variants of sequential importance sampling with resampling (SISR) and a Markov Chain Monte Carlo algorithm (MCMC). The three variants concern: (i) the SISR presented in Trevezas and Cournède (2013), where the resampling step is multinomial (Gordon et al., 1993), (ii) a modification of the previous algorithm by performing the resampling step with a combination of residual and stratified resampling (see, eg., Cappé et al., 2005 and references therein) and (iii) a sequential importance sampling algorithm with partial rejection control (see, eg., Liu et al., 1998, 2001). The variant of the MCMC algorithm that we developed is a hybrid Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990), where simulations from the full conditional distributions of the hidden states were replaced by Metropolis–Hastings (MH) steps (Metropolis et al., 1953; Hastings, 1970). Having as target to further optimize the MCMC approach and to simplify a routine use of the method, a data-driven automated algorithm is proposed. Moreover, a data-driven automated algorithm is proposed in order to simplify a routine use of the proposed method. The main benefits of such an algorithm in the EM context concern the automatic determination of the Monte Carlo sample size in each EM step and of the total number of EM steps. One of the most commonly used in practice algorithms of this type which is efficient and computationally cheap is the one presented in Caffo et al. (2005). We adapted this algorithm to our context to find an appropriate sample size in each ECM step and also an appropriate number of ECM steps. The details can be found in Section 4. Simulation studies from a synthetic example and also a real dataset from the sugar-beet plant were used to illustrate the performance of the competing algorithms.

The MCMC–ECM algorithm proved to be the most efficient in parameter estimation for the plant growth model that we study in this paper. Additional to the significant reduction of the Monte Carlo error, the MCMC algorithm revealed another advantage compared to SISR in the specific context of this study. Since plant organs have generally large expansion durations, then by censoring plant's evolution at a given time, a whole batch of organs will have not completed their expansion (immature organs). As will be explained in Section 3, the MCMC algorithm can better handle this type of asymmetry. The automated version of MCMC–ECM was thus selected to make a further statistical inference. In particular, two different types of hidden Markov models were described and tested on a real dataset for their fitting quality.