On perfect simulation and EM estimation

Kajsa Larson
Abstract

Perfect simulation and the EM algorithm are the main topics in this thesis.

In paper I, we present coupling from the past (CFTP) algorithms that generate perfectly distributed samples from the multi-type Widom–Rowlinson (W–R) model and some generalizations of it. The classical W–R model is a point process in the plane or the space consisting of points of several different types. Points of different types are not allowed to be closer than some specified distance, whereas points of the same type can be arbitrary close. A stick-model and soft-core generalizations are also considered. Further, we generate samples without edge effects, and give a bound on sufficiently small intensities (of the points) for the algorithm to terminate.

In paper II, we consider the forestry problem on how to estimate seedling dispersal distributions and effective plant fecundities from spatially data of adult trees and seedlings, when the origin of the seedlings are unknown. Traditional models for fecundities build on allometric assumptions, where the fecundity is related to some characteristic of the adult tree (e.g. diameter). However, the allometric assumptions are generally too restrictive and lead to nonrealistic estimates. Therefore we present a new model, the unrestricted fecundity (UF) model, which uses no allometric assumptions. We propose an EM algorithm to estimate the unknown parameters. Evaluations on real and simulated data indicates better performance for the UF model.

In paper III, we propose EM algorithms to estimate the passage time distribution on a graph. Data is obtained by observing a flow only at the nodes – what happens on the edges is unknown. Therefore the sample of passage times, i.e. the times it takes for the flow to stream between two neighbors, consists of right censored and uncensored observations where it sometimes is unknown which is which. For discrete passage time distributions, we show that the maximum likelihood (ML) estimate is strongly consistent under certain weak conditions. We also show that our proposed EM algorithm converges to the ML estimate if the sample size is sufficiently large and the starting value is sufficiently close to the true parameter. In a special case we show that it always converges. In the continuous case, we propose an EM algorithm for fitting phase-type distributions to data.

Keywords: Perfect simulation; coupling from the past; Markov chain Monte Carlo; point process; Widom–Rowlinson model; EM algorithm; dispersal distribution; fecundity; first-passage percolation
Acknowledgement

I would like to thank people who have helped and inspired me during my doctoral study.

I especially want to thank my supervisor Professor Sara Sjöstedt de Luna, this thesis would not have been possible without her encouragement, support, guidance, and feedback on my work. It has really been a true pleasure to discuss research (and life) with her.

I would like to thank my co-supervisor Professor Olle Häggström for suggesting two very nice and interesting problems and for his support. I also would like to thank my co-supervisor Professor Lennart Bondesson for careful reading of this manuscript and insightful, invaluable and constructive comments.

I was delighted to be involved in Nikos Nanos tree project in Madrid. It was very exciting to work with real data, and our discussions have been a great inspiration for my research.

I would like to thank Jessica Fahlén for entertaining, inspirational and informative discussions about the design of this thesis. Thanks to Mia Han at Print & Media for her help with the cover. I want to express my gratitude to Anders Muszta for enlightening comments. Thanks to Anders Lundquist for Latex templates. Thanks to my other, former colleges at the Statistics department, in particular Ingrid Svensson for her warmth and wisdom.

My bridge friends have given me many joyful and stimulating moments. Especially thanks to my team ”Karamellen”, my robbert and dinner buddies, and Peder Linder for his humor and long stream of new bridge conventions. Thanks also to my friends in my ”book circle”. I always feel happy after meeting you.

Thanks to my friends from the West coast, especially my precious Mona Grähs, Ingerun Syrén-Sjösvärd, Maria Franzén, and Cecilia and Elisabet Thelander.

Finally, I want to thank my parents and my family. I feel very happy and fortunate for my two wonderful children Ellis and Dina, and thank you Mats for all your support during these years.

Umeå, May 2010

Kajsa Larson
List of papers

This thesis consists of the following papers:


Contents

1 Introduction 11

2 Perfect simulation 11
  2.1 Markov chain Monte Carlo . . . . . . . . . . . . . . . . . . . . 12
  2.2 Coupling from the past . . . . . . . . . . . . . . . . . . . . . 14

3 The EM algorithm 17

4 Summary of papers 19
  4.1 Paper I: Perfect simulation of some spatial point processes 19
  4.2 Paper II: Inverse modeling for effective dispersal: do we need
tree size to estimate fecundity? . . . . . . . . . . . . . . . . . . 22
  4.3 Paper III: Estimation of the passage time distribution on a
graph via the EM algorithm . . . . . . . . . . . . . . . . . . . 23

Papers I-III 29
1 Introduction

Due to the rapid growth of computer capacity during the last decades, computational methods play an increasingly important role in science, particularly in statistics. Many practical problems that were previously difficult, and even impossible to approach by analytical techniques, can now be addressed using computational methods in combination with mathematical modeling. This thesis focuses on applications of two computational techniques: perfect simulation and the EM algorithm.

The starting point for the field of perfect simulation came in 1996 when Propp and Wilson [36] presented the seminal algorithm called Coupling From The Past (CFTP). Perfect simulation build on Markov chain Monte Carlo (MCMC) methods, and is an ingenious procedure to provide samples that are distributed exactly according to the target distribution, in contrast to ordinary MCMC, where only approximate samples are obtained.

The EM algorithm has a long history, but the name EM algorithm and main breakthrough came in 1977 with Dempster, Laird and Rubin [8]. The EM algorithm is a method for finding maximum likelihood (ML) estimates in situations when the observed sample have missing data or is incomplete in some way, in which case the likelihood function often is complicated and difficult to maximize directly. The EM algorithm is an iterative procedure in two steps, an E-step (short for expectation) and an M-step (short for maximization), and is based on the so called complete sample, which is the sample that one would have if there was no missing information.

Perfect simulation and the EM algorithm are further described in Section 2 and 3, respectively.

2 Perfect simulation

The CFTP algorithms introduced by Propp et al. [36] generate exact samples from a range of discrete distributions, in particular the Ising model\(^1\). After this pioneering work, CFTP has been extended into many different areas, for instance spatial point processes in Häggström, van Lieshout, and Møller et al. [18], Kendall [20], Kendall and Møller [22], Thöennes [43], van Lieshout and Stoica, [47], and Larson in Paper I; Markov random fields in Häggström and Nelander [15, 16], Häggström and Steif [17], and van

---

\(^{1}\)The Ising model consists of discrete variables called spins that can be in one of two states. The spins are arranged in a lattice or graph, and each spin interacts only with its nearest neighbors. An applet of the Ising model can be found at web site [34].
Lieshout and Schreiber [46]; stochastic geometry in Kendall and Thønnes [23]; and Bayesian inference in Murdoch and Green [33]. An overview on statistical inference for spatial point processes can be found in the book by Möller and Waagepetersen [32].

The phrase ”perfect simulation in space” refers to algorithms that generates samples without edge-effects (for lattice and spatial interaction models), see van den Berg and Steif [45], Hågström and Steif [17], Brix and Kendall [3], Möller and Rasmussen [30], and Larson in Paper I. The idea behind the algorithms is to not only simulate backwards in time, but also outwards in space. Apart from the advantage of not suffering from edge effects it can also be useful in quantifying the edge effects suffered from ordinary perfect simulation.

Another perfect simulation algorithm is Fill’s algorithm, see Fill [10] and Möller and Schladitz [31]. Further references in perfect simulation, as well as illustrating web animations and programs, see Wilson’s online web site [50].

Since MCMC is the basis of perfect simulation, we now give a brief presentation of MCMC methods, and in particular the Gibbs sampler, which is used in Larson, Paper I.

2.1 Markov chain Monte Carlo

MCMC methods are a collection of stochastic simulation algorithms where the goal is to generate a sample from a certain probability distribution. The main idea in these algorithms is to construct a Markov chain with equilibrium distribution equal to the target distribution, run the chain for a long time, and give the final state as a sample.

The first MCMC algorithms was introduced in 1953 by Metropolis et al. [28], where they considered Gibbsian point processes with a fixed and finite number of points. The next big step came in 1970 when Hastings [19] introduced general accept-reject methods, known as Metropolis-Hastings algorithms. The well-known Gibbs sampler is a special case of the Metropolis-Hastings algorithm, and was developed by Geman and Geman [13] in the context of image-processing models. Gelfand and Smith [12] generated new interest in the algorithm in 1990, when they showed that it could be applied to a variety of inference problems, in particular in Bayesian statistics (by using MCMC to numerically calculate multi-dimensional integrals\(^2\) in

\(^2\)Multi-dimensional integrals also often arise in computational physics, computational biology and computational linguistics.
posterior distributions). The advancement of faster computers has made MCMC methods increasingly popular during the last two decades. Excellent treatments of MCMC methods can be found in Tanner [42], Gilks, Richardson, and Spiegelhalter [14], and Robert and Casella [37].

MCMC methods apply to a wide range of distributions for which direct simulation may be difficult or even impossible. For instance, for spatial point processes, where the joint density typically is unknown, since the normalizing constant is practically impossible to compute.

A problem with MCMC methods is to know how long the Markov chain should be run to be sufficiently close to equilibrium. Although there are many different empirical techniques and theoretical bounds for determining how many steps that are needed, the so called burn-in period, see Frigessi [11], Brooks, Dellaportas, and Roberts [4], Cowles and Carlin [7], and Diaconis [9], in many practical situations it is more or less impossible to guarantee when a Markov chain is in or close to equilibrium. Another important issue beside the burn-in period is whether the Markov chain mixes rapidly, i.e. if the chain reaches the equilibrium distribution quickly, starting from an arbitrary state, see Levin, Peres, and Wilmer [24].

To construct a Markov chain with state space $S$ such that the stationary distribution is $\pi$, one has to find transition probabilities $P(x, \cdot)$ satisfying the equation,

$$\pi(\cdot) = \int \pi(dx) P(x, \cdot). \quad (1)$$

Hastings shows in [19], that if a Markov chain is reversible with respect to $\pi$, that is,

$$\pi(dx) P(x, dy) = \pi(dy) P(y, dx) \quad \forall x, y \in S,$$ \quad (2)

then equation (1) holds. If in addition the Markov chain is irreducible and aperiodic, then $\pi$ is the unique stationary distribution, and the convergence to $\pi$ is guaranteed. Thus, instead of trying to find transition probabilities satisfying (1), it is only needed to consider the simpler equation (2).

The basic principle for the Gibbs sampler’s is that given a target joint multivariate distribution, it sequentially samples from its conditional distributions. Let $f(x_1, \ldots, x_k)$ denote the target joint density for the random variables $X_1, \ldots, X_k$, and update the Markov chain by, for $i = 0, 1, 2, \ldots,$
draw samples,
\[ \begin{align*}
x_{i+1}^1 & \text{ from } f(x^1|x_i^2, x_i^3, \ldots, x_i^k), \\
x_{i+1}^2 & \text{ from } f(x^2|x_{i+1}^1, x_i^3, \ldots, x_i^k), \\
& \quad \vdots \\
x_{i+1}^k & \text{ from } f(x^k|x_{i+1}^1, x_{i+1}^2, \ldots, x_{i+1}^{k-1}).
\end{align*} \]

The iteration is continued until the Markov chain is sufficiently close to equilibrium. For more details we refer to Casella and George [5]. Since the Gibbs sampler only requires knowledge of how to sample from conditional distributions, it can be useful for spatial point processes, where the joint density typically is unknown. In Larson, Paper I, the presented CFTP algorithms are based on the Gibbs sampler.

### 2.2 Coupling from the past

CFTP is a brilliant algorithm that knows by itself during run time when to stop, and moreover, the output is distributed exactly according to the target distribution, see Propp et al. [36]. The idea behind CFTP is that if a Markov chain has been running for all time (from \(-\infty\)) then the state of the Markov chain at a fixed time, for instance 0, is distributed exactly according to the target distribution. CFTP figures out what state the Markov chain is in at time 0, without simulating the whole chain from \(-\infty\) (which of course is impossible!). The basic idea is to run several coupled\(^3\) Markov chains, starting from every possible state, from some (finite) time in the past to time 0. Since the chains are coupled they use the same pseudo-random numbers, and hence, chains that coalesce (i.e. are in the same state) will thereafter always follow the same path. Thus, if all the chains coalesce, then the state that the chains are in at time 0 is distributed exactly according to the target distribution. Otherwise, start new Markov chains further back in the past and repeat the procedure (using new, and the earlier used, pseudo-random numbers).

It is common to double the starting times of CFTP (i.e. \(-2, -4, -8, \ldots\)), which is binary search and yields a running time less than 4 times the optimal choice, see Propp et al. [36].

\(^3\)Coupling and stochastic domination are fundamental concepts in perfect simulation. For further details we refer to Lindvall [25] (revised in [26]), Thorisson [44], and Larson in Paper I.
Now we give three illustrative examples of CFTP in the discrete case.

**Example 1.** Consider a Markov chain with state space \( \{1, 2, 3\} \) and equilibrium distribution \( \pi \). We run three coupled chains, starting in the states 1, 2, and 3, from time \(-2\) to time 0, see Figure 1. The chains have not coalesced by time 0. Chain one (lined path) and chain three (circled path) are in state 2 at time 0, and chain two (dotted path) is in state 1. Therefore, we must start further back in the past, and we choose time \(-4\), see Figure 2. Now the chains have coalesced by time 0, all chains are in state 2, and thus, 2 is an unbiased sample from \( \pi \).

![Figure 1: Realizations of three coupled Markov chains, starting in the states 1, 2, and 3, from time -2 to time 0.](image1)

![Figure 2: Realizations of three coupled Markov chains, starting in the states 1, 2, and 3, from time -4 to time 0.](image2)
If a Markov chain is monotone or antimonotone\(^4\) (with respect to a partial order of the state space), then the CFTP algorithm can be reduced, from simulating as many chains as there are states, to just simulating two chains, one starting in the maximal state and one in the minimal state. This is of course a major advantage when the state space is large, which it usually is.

**Example 2.** Consider a monotone Markov chain with state space \(\{1, 2, 3\}\) and equilibrium distribution \(\pi\). We run three coupled chains, starting in the states 1, 2 and 3, from time \(-4\) to time 0, see Figure 3. The chain that starts in state 2 is sandwiched between the maximal chain (which started in state 3) and the minimal chain (which started in state 1). The chains have coalesced by time 0, and 2 is an unbiased sample from \(\pi\). \(\square\)

**Example 3.** Consider an antimonotone Markov chain with state space \(\{1, 2, 3\}\) and equilibrium distribution \(\pi\). We run three coupled chains, starting in the states 1, 2 and 3, from time \(-4\) to time 0, see Figure 4. The chain that starts in state 2 is sandwiched between the other two chains, which are exchanging roles as the maximal and minimal chain while running. The

\(^4\)A Markov chain is monotone (antimonotone) with respect to a partial order of the state space, if its transition matrix preserves (reverses) the partial order for all states in the state space.
chains have coalesced by time 0, and 1 is an unbiased sample from $\pi$. \hfill \Box

For more details of these examples, we refer to Larson in Paper I.

Point processes have a continuous state space, which rules out the possibility to start from every possible state. A solution to this difficulty can be in line with the sandwiching idea used in the finite, monotone case. However, CFTP is not limited to monotone situations, there are several generalizations of CFTP, an overview can be found in Kendall [21]. In Larson, Paper I, CFTP algorithms are presented for multi-type W–R models, in which case there is no obvious partial order of the state space.

3 The EM algorithm

The EM algorithm is a powerful technique that provides an iterative procedure for finding maximum likelihood estimates in cases when there are incomplete or missing data, and the likelihood function is difficult to maximize directly. The idea behind the algorithm is to instead work with a sequence of easier maximizations whose limit is the answer to the original problem. The replacing sequence is based on the so called complete sample, which is the sample that one would have had if there were no missing information. The EM algorithm is an iterative procedure in two steps, an E-step (short for expectation) and an M-step (short for maximization).
In the E-step the expectation of the loglikelihood of the complete sample, given the observed sample and the current estimate for the parameter (which is chosen by the user the first turn), is computed. In the M-step the new conditional loglikelihood is maximized and a new updated estimate of the parameter is obtained. The procedure is continued until convergence is obtained (or 'close' enough).

The EM algorithm is particularly useful when the complete (unobserved) sample is distributed according to the exponential family. In the E-step the expectation of the sufficient statistic, given the observed sample and the current estimate for the parameter, is computed. In the M-step, the new conditional loglikelihood is a linear function in the parameters, and thus it is usually possible to derive closed form updates, see Sundberg [38, 39]. These simpler expressions are used in both Nanos, Larson, Millerón, Sjöstedt de Luna in Paper II and in Larson, Paper III.

The EM algorithm is guaranteed to increase the likelihood at each iteration, see Dempster et al. [8]. However, this climbing property does not automatically imply convergence of the corresponding parameter estimates, since it can be a local maximum or a saddle point. Under some rather weak continuity conditions, which holds in most practical situations, in particular for the exponential family, the likelihood sequence converges monotonically to some finite value, see Wu [51]. The regularity conditions that lead to convergence of the corresponding parameter estimates, see Wu [51] and Boyles [2], are unfortunately often difficult to show analytically, even if they are fulfilled in fairly general situations. In Larson, Paper III, we show that our proposed EM algorithm, for the discrete case, converges to the strongly consistent ML estimate if the sample is sufficiently large and the starting value is sufficiently close to the true value.

However, even if convergence is difficult to show analytically, in practice we can in many cases almost guarantee convergence, since we can start from several different starting values and that way ensure that the global maximum is reached. This procedure is used in Nanos et al. in Paper II. The convergence properties of the EM algorithm are discussed in detail by McLachlan and Krishnan [27].

One major criticism of the EM algorithm is that the convergence can be quite slow. There are some methods for speeding up the convergence, see for instance Minami [29], Varadhan and Roland [48], and McLachlan et al. [27]. The disadvantages of speeding methods in general is that the stability and simplicity often decreases.
Example 4. Here we give an illustration of the convergence of an EM algorithm presented by Nanos et al. in Paper II. The dataset includes the spatial coordinates of 15 adult trees and 674 seedlings (European beech, Fagus sylvatica L., stand of northern Madrid) found in a square plot of 2500 m$^2$. The fecundities (i.e. the number of successfully established offspring of the adult trees), $S_1, \ldots, S_{15}$, and the two dispersal parameters $\mu$ and $\sigma$ in a lognormal distribution, were estimated using the EM algorithm. Depending on the initializing values in the algorithm, the iteration paths may vary extensively, and also have very different convergence rates, see two iteration sequences in Figure 5.

There are many extensions of the EM algorithm, for instance stochastic versions. Stochastic EM is mainly used in situations when the E-step is analytically intractable, typically involving a high-dimensional integral without closed-form solution. The integral is instead approximated via simulation, which is the common idea in the stochastic EM algorithm (SEM) introduced by Celeux and Diebolt [6] and the Monte Carlo EM algorithm (MCEM) introduced by Wei and Tanner [49]. An application of MCEM in the context of mixture models can be found in Svensson, Sjöstedt de Luna, and Bondesson [41], with corresponding consistency properties in Svensson and Sjöstedt de Luna [40]). For further details about EM algorithms and extensions, see the book by McLachlan et al. [27], and the references therein.

4 Summary of papers

4.1 Paper I: Perfect simulation of some spatial point processes

In this paper, being my licentiate thesis, we present CFTP algorithms that generate perfectly distributed samples from the multi-type Widow–Rowlinson (W–R) model and some generalizations of it. The classical W–R model is a point process in the plane or the space consisting of points of several different types. Points of different types are not allowed to be closer than some specified distance, whereas points of the same type can be arbitrary close. An application can be to describe certain gases consisting of several types of particles.

Initial inspiration to this work comes from Häggström et al. [18], where they present a CFTP algorithm for the W–R model with exactly two types
Figure 5: Two iteration sequences of the EM algorithm, estimating $S_1, \ldots, S_{15}, \mu$ and $\sigma$. Both sequences eventually converge to the same values. (Note that the scales in the two figures differs.)
of points. They exploit monotonicity and the sandwiching principle by letting the minimal state be represented by a configuration full of points of type one, and the maximal state be represented by a configuration full of points of type two. However, when there are three types or more, there is no useful partial order of the state space. We have found a partial order on an extended space, that makes it possible to use monotonicity and sandwiching. By introducing so called ghost points and real points we make our CFTP algorithm intuitive.

We first present a CFTP algorithm, based on the Gibbs sampler, for the classical (hard-core) multi-type W–R model. We also consider a soft-core W–R model, where points of different types are not completely forbidden to be close to each other, just inhibited in various degrees. Then we extend results to allow the hindrance between two points of different types to be explained by more than the Euclidean distance between them. In particular we consider a stick-model where the hindrance is defined by imaginary sticks, with centers at the associated points, and where sticks are not allowed to cross each other. The different directions of the sticks (a finite number), represent the different types of the points. A CFTP algorithm is also given for a soft-core version of the stick-model.

The runtime of the CFTP algorithm for the multi-type W–R model, is determined by the number of different types and their intensities (i.e. the expected numbers of points on a 'unit') and the hard-core (soft-core) distance. Simulation studies indicate that in the symmetric case (i.e. all types have the same intensity), the runtime first grows exponentially with the intensity, but then suddenly, when the intensity becomes larger seems to be superexponential. This change in growth may be explained by a phase transition, which loosely speaking is when points of a certain type(s) ”win” and tend to be in a majority.

We also present a CFTP algorithm that yields samples from the multi-type W–R model without edge effects, inspired by van den Berg et al. [45]. The underlying idea is to not only simulate backwards in time, but also outwards in space. This algorithm does not always terminate for large intensities, in particular if there is a phase transition. A bound on sufficiently small intensities for the algorithm to terminate is given.

Finally, we consider a generalization of the stick-model where we let all directions of the sticks to be possible. We propose a CFTP algorithm based on a spatial birth-and-death process, using results in Kendall [20] and Kendall et al. [22].
4.2 Paper II: Inverse modeling for effective dispersal: do we need tree size to estimate fecundity?

A great challenge in plant ecology is to estimate seed and seedling dispersal distributions, as well as effective plant fecundity (the number of successfully established offspring produced by an adult tree). Data which estimation is based on often consists of spatial coordinates of adult trees and counts of seeds/seedlings in small quadrats, where the origin of the seeds/seedlings are unknown. Current models used to simultaneously estimate fecundities and dispersal distribution build on allometric assumptions, where the fecundity is related to some (easily) measured characteristic of the adult tree, for instance the diameter, the age, or the number of cones. However, the allometric assumptions are generally too restrictive, which may lead to not only nonrealistic fecundity estimates, but also severely biased estimates of the dispersal distribution.

Here we present a new model, the unrestricted fecundity (UF) model, which avoids allometric assumptions on the fecundities, and instead allow them to vary freely\(^5\), and even to be zero. The UF model is flexible and can easily adjust to datasets with different properties, and moreover, it is the only model that can detect adult trees without any progeny at all. Parentage analysis in forest populations show that the female reproductive success in the seedling stage is largely unequal across adults, suggesting that classical allometric models may not be adequate enough to capture its variability. The number of parameters to estimate when using the UF model is the number of adult trees plus two dispersal parameters, as compared to the classical allometric models having only a few parameters to describe fecundities. To handle the larger number of parameters and the missing information in the data, we propose an EM algorithm for parameter estimation.

We have fitted the UF model, via the EM algorithm, to a real dataset from an expanding European beech population of central Spain, consisting of 15 adult trees and 674 seedlings, as well as to simulated data. In comparisons with a classical allometric model, where the fecundity is assumed to be proportional to the diameter of the adult tree at breast height, the UF model fitted the data better and the parameter estimates were less biased.

The reliability of the dispersal and fecundity estimates using the UF model depends on several factors. The pattern of the adult trees influences

\(^5\)The fecundities of the adult trees are assumed to follow Poisson distributions with unknown (nonequal) parameters.
primarily the accuracy of the fecundity estimates, the more isolated the adult tree – the more trustworthy. For adult trees that stands in dense clusters, it is difficult to judge from which of the trees a seedling originates from. Individual fecundity estimates for adult trees within dense clusters should therefore be interpreted with care, and maybe instead be summed to a cluster estimate instead. However, a simulation study indicates that the parameter estimates of the dispersal distribution seem to be rather stable in general, even when clusters are present.

We provide formulas for standard errors for (non-null) estimates, via the information matrix (computed after omitting trees with estimated null fecundity). A simulation study reveals that the standard error estimates for the UF model seem reliable for the parameters of the dispersal distribution and for the fecundity estimates of isolated trees, but (somewhat) overestimated for the estimated individual fecundities of clustered trees. We believe that more reliable standard error estimates, and parameter estimates, on cluster level can be achieved by doing fusions\(^6\) of dense clusters before estimation, and thus reducing the number of parameters and null fecundity estimates.

### 4.3 Paper III: Estimation of the passage time distribution on a graph via the EM algorithm

A certain flow is spreading on a graph\(^7\), along the edges, from neighbor to neighbor, as time goes by. It is only possible to observe the spreading at the nodes – what happens on the edges is unknown. The passage times, i.e. the times it takes for the flow to stream between two neighbors, are assumed to be independent and identically distributed on all edges. An application of the model can for instance be the spread of an infection in a population, by letting the nodes represent the individuals and the edges represent the possible infection-paths.

In this paper we estimate the passage time distribution, first via a general discrete distribution and then via phase-type distributions. A phase type distribution is shortly the distribution of the time until absorption in a finite-state Markov process with a single absorbing state. Phase type distributions are important since they can approximate any distribution with

\(^6\)A fusion of a cluster of adult trees is a replacement of the trees in the cluster to a virtual tree located in the "middle" of the former cluster (for instance in the center of gravity).

\(^7\)A graph is a set of nodes together with a set of edges that connect pairs of nodes, and two nodes with an edge between them are called neighbors.
positive support on $[0, \infty]$ to arbitrary accuracy.

The main consequence of only being able to observe the spreading via the nodes is that when a node, with more than one wet neighbor (i.e. a neighbor which is already reached by the flow), is reached by the flow, it is unknown from which neighbor the flow came from. Thus, in a sample of passage times, some will be right censored and some uncensored, and sometimes it is unknown which is which. We handle this missing information by using the EM algorithm.

When the passage time distribution is discrete, we show that the ML estimate is strongly consistent under certain weak conditions. We also show that our proposed EM algorithm converges to the ML estimate if the sample size is sufficiently large and the starting value of the EM algorithm is sufficiently close to the true parameter. Furthermore, we show that in a special case (when the maximum degree of the graph is equal to two and there are three possible passage times), the EM algorithm always converges to the ML estimate irrespective of the starting value.

In a simulation study, the performance of the EM algorithm for a discrete passage times distribution, with ten possible passage times, all equally likely, on the 2-dimensional lattice is studied. In this example the EM algorithm always converges irrespective of the starting value. The convergence rates of the EM algorithm are in general faster for the probabilities of the lower outcomes. Furthermore, the resulting ML estimates for the probabilities of the discrete outcomes have low bias, and are in general smaller for the lower outcomes.

For continuous passage times we propose an EM algorithm for fitting phase type distributions. The algorithm build on our proposed EM algorithm for the discrete case, combined with results from Asmussen, Nerman and Olsson [1] and Olsson [35], where they fit phase type distributions to samples with uncensored and censored observations via the EM algorithm. Phase type distributions typically do not have unique representations, and therefore one cannot refer to consistency of the estimates. We present a small simulation study, which indicate good performance for Weibull distributed passage times.

References


[34] Bernd Nottelmann. Web site for perfectly random sampling with markov chains, that was developed as part of the STP project. http://webphysics.davidson.edu/applets/ising/default.html.


