Chapter 1

Introduction

1.1 Context and scope

This thesis deals with spatial statistics, or more precisely, with the statistical analysis of spatial data. The spatial data we consider are random configurations of points, line segments or discs in the two dimensional plane. Such random 'processes' could serve, and have served, to model a wide variety of real life phenomena. The points may represent the locations of trees in a forest, the line segments could be fractures in a rock surface and the discs could be cells under a microscope.

Usually, we observe these spatial processes through a bounded window while we imagine they extend throughout the entire plane. As a result, we often have to deal with 'edge-effects'. For instance, observing a collection of line segments through a bounded window some of the segments will be clipped at the window's boundaries. Such *censoring* has been studied extensively in the one dimensional case. The line segments then represent lifetimes which are sometimes not completely observed. Estimation of the lifetime distribution from incomplete observations is the main subject of the field of survival analysis. Laslett (1982a, b) was the first to point out the analogy of censoring in survival analysis and edge effects in spatial statistics.

Our goal in this thesis is always to estimate the probability distribution of the (partly) observed random spatial process. Sometimes this distribution is determined by a *finite* dimensional parameter, such as the expected number of points in a given set. At other times this will not be the case and then we speak of a semi- or nonparametric problem or of a problem with a 'large parameter space'. In recent years estimation of infinite dimensional parameters has received much attention and has made great headway. Bickel et al. (1993) is a standard reference for modern semi-parametric theory. In-

complete observation, such as due to edge effects, is well incorporated in the theory.

One of the main themes of this thesis is the application to spatial data of semi-parametric methods. Another, closely related theme is maximum likelihood estimation from spatial data. Maximum likelihood estimation is not common practice in spatial statistics as the method of moments and minimum contrast methods are more often applied (Stoyan et al. 1987). One might view our effort in the light of a development which has been described—somewhat provocatively—by Geyer (1999):

"Exotic areas of statistics, such as spatial statistics in general and spatial point processes in particular often recapitulate the history of statistics. The first formal inference is nonparametric, using method of moments estimators. At this stage there is no modeling. Ordinary statistics was in this phase a century ago with Pearson families of curves fit by method of moments. Time series was in this phase in the 1950s with spectral analysis. Spatial statistics was in this phase in the 1970s. Ripley's K-functions are an example. So is spatial autocorrelation analysis of lattice processes.

The next phase involves the introduction of parametric statistical models, efficient estimation methods, and hypothesis tests, which happened in ordinary statistics in the 1920s and 1930s, in time series in the 1960s, and is only now happening in spatial statistics. "

The thesis is organized around three problems: Laslett's line segment problem; estimation of the chord length distribution; and estimation for the bombing model. We describe these problems in section 1.2 of this introduction and return to each of them in the next three chapters. In the remainder of this chapter we gather some theory and tools for later reference. In particular, we briefly discuss the issue of independence in spatial processes and introduce the Poisson process and the related Boolean model. Next, we discuss some topics in semi-parametrics and finally consider Markov Chain Monte Carlo methods and perfect sampling. We want to draw the reader's special attention to two topics to which we contribute something new. The first is a general discussion of consistency of maximum likelihood estimators in section 1.4.2. The second is a modification of a stochastic version of the well-known EM algorithm (Dempster et al. 1977, Celeux and Diebolt 1986) in section 1.5.3.

1.2 Three problems

In this thesis we analyze three spatial missing data problems. In all of these we are unable to write down the likelihood of the data, let alone maximize it. This seems typical of spatial statistical estimation problems and explains why maximum likelihood is usually not the preferred approach.

In each of the first two of our problems we manage to find 'analogous' problems, which are simpler yet similar. For these analogous problems we *can* perform maximum likelihood estimation. Thus we obtain estimators that are MLEs—only not for the actual problems at hand! Still, if the simplifications are minor we may expect that the nice properties that an MLE typically possesses will be preserved, at least to a certain degree. Efficiency will generally fail, but consistency and asymptotic normality may not. We shall study consistency only, but we feel confident that it would be also possible to tackle asymptotic normality.

Although the first two problems both concern line segments and our approach to them is similar, we should point out that they really are quite different. In the first problem we have to deal with spatial dependence among the data. The difficulty with the second problem is that the data are in the form of random functions on random sets.

In the third problem, concerning the Boolean model of randomly placed discs, our difficulties are of a different nature still. The likelihood is easy enough, except for an intractable normalizing constant. No simplification will help and so the emphasis shifts from analytic methods to a Monte Carlo approach. In particular, we present a new algorithm to sample from the distribution of a wide variety of point processes.

We now briefly state our three problems and point out their particular difficulties. We do not as yet strive for mathematical rigor, but merely wish to indicate where we are going—and hopefully whet the reader's appetite.

Laslett's line segments

Figure 1.1 is a map of an area of about 160 by 160 meters of the Canadian Shield near Manitoba. The black parts are the rock surface, the white parts are soil, vegetation or water. The white lines through the black regions indicate fractures and the problem is to estimate the distribution of their lengths. This is a missing data problem because the fractures are unobservable in the white regions and outside the 160 by 160 meters area of the map. The particular difficulty here lies in the fact that the (black) observation region is not convex. This means that of a single fracture we might observe several fragments. Because the resolution of the map is rather poor, it is impossible

to know if two fragments belong to the same fracture.

The fractures are conveniently modeled by a location and an associated line segment. By 'location' we mean, for instance, the location of a preferred point on the segment, such as the mid point or the left-hand end-point.

In earlier contributions, the case of a convex observation window was considered. Assuming the locations and line segments to be a Poisson marked point process one finds that the fractures are independent. Standard methods can then be applied to compute the (non-parametric) maximum likelihood estimator (Gill 1994, Wijers 1995a and b). For us there is no gain in assuming that the locations of the fractures are Poisson, because even if the *fractures* are independent some of the *fragments* are certainly dependent.

Chord Length Estimation

Consider a random closed set $\mathcal{W} \subset \mathbb{R}^2$. With each point in $\mathcal{W}$ we can associate a *chord* in a given direction which is the longest line segment through that point which lies wholly within $\mathcal{W}$. Also, with each point in $\mathcal{W}^c = \mathbb{R}^2 \setminus \mathcal{W}$ we can associate a *contact segment* (our terminology) in a given direction which is the longest line segment starting at that point which is completely contained in $\mathcal{W}^c$. Of course these two concepts are closely related.

The problem is estimation of the distribution function of a typical chord and a typical contact segment, when we observe $\mathcal{W}$ through a compact set B. This is a missing data problem because some chords and contact segments are clipped at the boundaries of B. The particular difficulty here is that for *every* point in $\mathcal{W}$, respectively $\mathcal{W}^c$ we have a chord or a contact segment. Such continuous data is awkward and we are unable to apply standard methods for missing data problems.

The Bombing Model

Consider the so-called Boolean model in $\mathbb{R}^2$ where at each point of a Poisson point process of intensity λ, a disc of random radius is placed. This model is also known as the bombing model. We observe only the union of the discs through a compact window W. In this example data is missing as some discs are not observed because they are covered by others. Also there is the usual censoring at the observation window's boundaries. For our present purpose, estimation of the intensity parameter λ, the latter censoring is not the main difficulty. The difficulty is that the conditional distribution of the complete data, given the observed data involves a normalizing constant which cannot possibly be evaluated. Again, standard methods for missing data problems fail.

Figure 1.1: Fractures in 160 m. by 160 m. granitic rock of the Lac du Bonnet batholith in Manitoba. From Stone, Kamineni and Brown (1984). Digitized and post-processed by professor A.J. Baddeley of the University of Western Australia. The irregular black region is the rock's surface. The white parts represent areas where the rock cannot be observed due to soil, vegetation or water. The white lines through the black regions indicate fractures.

1.3 Independence, Poisson processes and Boolean models

This section borrows its title from section 1.5 of Baddeley (1999a). It contains some basic facts concerning the Poisson process in more general spaces than $\mathbb{R}^+$, and introduces the Boolean model. The Poisson process and the Boolean model—which is derived from it—are perhaps the simplest of all spatial models. They have been studied in detail, for instance in Daley and Vere-Jones (1988) and Stoyan et al. (1987). The content of this section can be said to be part of the folklore of stochastic geometry.

Classical statistics deals typically with independent, identically distributed random variables. An essential element of spatial statistics is a lack of independence. Suppose that X is a random spatial process of, say, randomly placed points. If there is any spatial structure or regularity, then typically we fail to have independence of $X \cap A$ and $X \cap B$, even when A and B are disjoint. The prime example of a *completely random* point process is the Poisson process. We now give a rather general definition.

Let $(S, \mathcal{B}, \Lambda)$ denote a measure space with a finite diffuse Radon measure Λ. A typical example is the case where S is a compact subset of $\mathbb{R}^2$ with the Borel sigma-algebra and $\Lambda(\mathrm{d}s) = \lambda \mathrm{d}s$.

Consider the 'Carter–Prenter exponential space' Ω of all finite subsets x of elements of S with a sigma-algebra $\mathcal{F}$ generated by sets $\{x \in \Omega : n(x \cap B) = n\}$ $(B \in \mathcal{B})$, where $n(x)$ denotes the number of elements in x.

The *Poisson* process Φ of *intensity* Λ can be seen as a random (counting) measure on $(S, \mathcal{B})$ or as a measure on $(\Omega, \mathcal{F})$. By $\Phi(B)$ we denote the (random) number of points of Φ that fall in a $\mathcal{B}$-measurable set B. The Poisson process is determined by the following two properties

1. $P(\Phi(B) = n) = \Lambda(B)^n e^{-\Lambda(B)}/n!, \qquad$ for $n = 0, 1, 2, \ldots,$ for all $B \in \mathcal{B}$
2. $\Phi(B_1), \Phi(B_2), \ldots, \Phi(B_k)$ are independent for disjoint $B_1, B_2, \ldots, B_k$.

Note that $E\Phi(B) = \Lambda(B) = \int_B \Lambda(\mathrm{d}s)$.

The expected number of points in a set B has the Poisson distribution with parameter $\Lambda(B)$. Conditionally on $\Phi(B) = n$, the points of $\Phi \cap B$ are distributed as a sample of size n from the 'normalized intensity' $\Lambda(S)^{-1}\Lambda(\mathrm{d}s)$.

Poisson processes have a nice 'superposition' property: The superposition (union) of independent Poisson processes is again a Poisson process, whose intensity is the sum of the intensities of the individual processes.

It is often useful to associate with each point of a point process a characteristic or 'mark'. For instance, when the points indicate locations (of the

midpoint) of fractures in a rock surface, an extra mark could indicate each fracture's length or direction. When the points represent (the center of) trees in a forest the mark could be their stem's diameter.

Now suppose we have a point process $\Phi = \{S_i\}$ on a space S. Suppose with each point S_i is associated a random mark M_i in some 'mark space' M, equipped with a probability measure P. Then the 'Marking Theorem' states that the following are equivalent

- The point process Φ is Poisson with intensity Λ and conditional on Φ, the marks are i.i.d. according to P.
- The marked point process $\{(S_i, M_i)\}$ is a Poisson process on $S \times M$ with intensity measure $\Lambda \times P$.

As a consequence of the Marking Theorem we have the 'thinning' property of Poisson processes: Suppose we have a Poisson process of intensity Λ. Now if we delete each point, independently of the others, with probability $1 - p$, the result will be a Poisson process with intensity $p\Lambda$.

Above we gave two examples of marked point processes; one representing a random collection of line segments (fractures), the other of circles (trees with specified stems). When the underlying point process is Poisson, such collections of randomly placed objects are called *Boolean models* (Matheron 1975, Stoyan et al. 1987). More precisely, suppose that $\Phi = \{S_i\}$ is a Poisson process on $\mathbb{R}^d$ and suppose that $\Xi_1, \Xi_2, \ldots$ are a sequence of independent, identically distributed random closed sets in $\mathbb{R}^d$. Then the Boolean model is defined as $\cup_i (S_i \oplus \Xi_i)$. The symbol $\oplus$ means Minkowski addition defined as $A \oplus B = \{a + b,\ a \in A,\ b \in B\}$.

1.4 Semi-parametrics

In this section we introduce the maximum likelihood estimator (MLE) $\hat{P}_n$ of a probability measure P_0 among a class $\mathcal{P}$ which we allow to be infinite dimensional. In this situation, the MLE is often called the nonparametric MLE (NPMLE). Also, we discuss differentiable, parametric sub-models through $\mathcal{P}$. Each such submodel will have a score function. With these scores, we can consider solutions of score equations. We discuss consistency of the NPMLE and of solutions of score equations. Other desirable properties we might look for in estimators, such as efficiency and asymptotic normality will not be touched upon in this thesis and are therefore omitted in the present discussion. Finally, we also consider so-called missing data problems, which go also by the name 'mixture models'.

A standard reference on semi-parametrics is the monograph by Bickel et al. (1993). Our exposition here owes much to the very readable chapter 25 on semi-parametrics of a recent book by van der Vaart (1998).

1.4.1 Nonparametric maximum likelihood estimation

In this section we define the so-called nonparametric maximum likelihood estimator (NPMLE). Consider random variables $X_1, X_2, \ldots, X_n$ which are independent and identically distributed according to some distribution P_0 in a class $\mathcal{P}$ of probability measures.

When $\mathcal{P}$ is not dominated by a single measure, the ordinary definition of the maximum likelihood estimator fails because there really is no likelihood. Kiefer and Wolfowitz (1956) proposed a way out of this difficulty by considering only pairwise comparisons. Now let $\mathbb{P}_n$ denote the empirical measure.

Definition 1.1. *A maximum likelihood estimator $\hat{P}_n$ of $P_0 \in \mathcal{P}$ is any element of $\mathcal{P}$ such that*

$$\int \log \frac{\mathrm{d}\hat{P}_n}{\mathrm{d}(\hat{P}_n + P)} \mathrm{d}\mathbb{P}_n \geq \int \log \frac{\mathrm{d}P}{\mathrm{d}(\hat{P}_n + P)} \mathrm{d}\mathbb{P}_n \quad \forall P \in \mathcal{P}, \tag{1.1}$$

while for every $P \in \mathcal{P}$, the left hand side is greater than minus infinity.

The latter requirement is a technicality. Its meaning will become clear in Lemma 1.1 below.

The definition says that $\hat{P}_n$ should 'beat' any other candidate on the data in pairwise comparison. Note that as we only compare two measures at a time, $\mathcal{P}$ need not be dominated. Note also that if $\mathcal{P}$ is dominated the definition reduces to the ordinary definition of maximum likelihood.

The definition depends on which versions of the Radon-Nikodym derivatives are used. These versions are only determined up to $(P_1 + P_2)$ null sets (when comparing P_1 and P_2 in $\mathcal{P}$), which are not necessarily $\mathbb{P}_n$ null sets. This problem has nothing to do with the fact that $\mathcal{P}$ may be undominated or infinite dimensional. It arises in parametric models just the same. It is generally felt that there should not be a problem—certainly not asymptotically—if we assume "what are known in some quarters as the British regularity conditions", to quote Cox (2000) (on a different matter). For a more thorough discussion see Scholz (1980).

For definiteness, we imagine that we have some mechanism to choose a version of each Radon–Nikodym derivative $\mathrm{d}P/\mathrm{d}(P+Q)$ for all $P, Q \in \mathcal{P}$ in advance, thus making sure that our choices do not depend on the data.

Moreover, we do this in such a way that $0 \leq \mathrm{d}P/\mathrm{d}(P+Q) \leq 1$ and $\mathrm{d}P/\mathrm{d}(P+Q) + \mathrm{d}Q/\mathrm{d}(P+Q) = 1$.

We find the definition of the NPMLE in its present form (1.1) slightly awkward and prefer to have a definition in terms of 'log likelihood ratios', see (1.2) and (1.3) below. Let us now define the likelihood ratio and log likelihood ratio.

Consider two probability measures P_1 and P_2 in $\mathcal{P}$, where P_2 need not dominate P_1. Define the *likelihood ratio* to be

$$\frac{\mathrm{d}P_1}{\mathrm{d}P_2}(x) = \frac{(\mathrm{d}P_1/\mathrm{d}(P_1+P_2))(x)}{(\mathrm{d}P_2/\mathrm{d}(P_1+P_2))(x)}.$$

This likelihood ratio should of course not be confused with the Radon–Nikodym derivative, which conventionally is defined only when $P_1 \ll P_2$. Because we can choose versions of the numerator and denominator which add up to 1, there are no problems with dividing zero by zero. If the numerator is positive while the denominator is zero, the ratio is defined to be infinite.

Lemma 1.1. *The defining property (1.1) of the NPMLE $\hat{P}_n$ is equivalent to*

$$\int \log \frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P} \mathrm{d}\mathbb{P}_n \geq 0 \quad \forall P \in \mathcal{P}, \tag{1.2}$$

and also to

$$\int \log \frac{\mathrm{d}P}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n \leq 0 \quad \forall P \in \mathcal{P}. \tag{1.3}$$

Proof This lemma is immediate from our definition of likelihood ratio. The assumption that the left hand side of (1.1) is greater than minus infinity is equivalent to (1.2) and (1.3) being well-defined. □

We should mention that an NPMLE need not exist and if it exists it need not be unique. A modification called the method of *sieves* due to Grenander (1981) is sometimes useful. We choose submodels $\mathcal{P}_n \subseteq \mathcal{P}$ such that $\cup_n \mathcal{P}_n$ will almost surely eventually include the true P_0. Then we define a 'sieved' NPMLE $\hat{P}_n^s$ to be such that

$$\int \log \frac{\mathrm{d}\hat{P}_n^s}{\mathrm{d}(\hat{P}_n^s + P)} \mathrm{d}\mathbb{P}_n \geq \int \log \frac{\mathrm{d}P}{\mathrm{d}(\hat{P}_n^s + P)} \mathrm{d}\mathbb{P}_n \quad \forall P \in \mathcal{P}_n, \tag{1.4}$$

while for every $P \in \mathcal{P}_n$, the left hand side is greater than minus infinity.

The $\mathcal{P}_n$ should, if possible, be chosen in such a way that $\hat{P}_n^s$ exists and is easily computed.

It is often useful to choose $\mathcal{P}_n$ in a way that is dependent on the data. For instance, we can take $\mathcal{P}_n = \{P \in \mathcal{P} : P \ll \mathbb{P}_n\}$. The resulting sieved NPMLE we should perhaps call *data* sieved NPMLE. However, we shall generally omit *both* modifiers 'data' and 'sieved'.

1.4.2 Consistency of the (NP)MLE

In this section we argue that the 'natural' approach to proving consistency of an MLE leads to *Hellinger* consistency. This usually overshoots the mark. Hellinger consistency—being equivalent to consistency in total variation—is too strong for most purposes. In many situations, especially when the model $\mathcal{P}$ is infinite dimensional, it just does not hold. For instance, when the NPMLE $\hat{P}_n$ is discrete while P_0 is absolutely continuous (with respect to Lebesgue measure) the Hellinger distance between $\hat{P}_n$ and P_0 is never less than 1.

One way to escape the rigidity of Hellinger consistency is to try to find a sequence P_n such that P_n tends to P_0 in some reasonable sense (say weak convergence), while at the same time the Hellinger distance between P_n and $\hat{P}_n$ tends to zero. Such a comparative sequence has successfully been used by Murphy (1994), Wijers (1995b) and Gill (1994) and it will also be used several times in this thesis.

Now we define two concepts of 'distance' between two probability measures. The first, Hellinger distance, is a true distance in the mathematical sense; the second, Kullback–Leibler information is not. Kullback–Leibler information and the Hellinger metric play an important role in proving consistency of maximum likelihood estimators as they fit in well with the MLE definition as we will see later on.

Definition 1.2. *The* Hellinger distance $H(P, Q)$ *between two probability measures P and Q is defined to be the square root of*

$$\begin{aligned} H^2(P,Q) &= \int \left| \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{1/2} - \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1/2} \right|^2 \mathrm{d}\mu \\ &= 2 - 2\int \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{1/2} \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1/2} \mathrm{d}\mu, \end{aligned}$$

for μ dominating both P and Q

This definition does not depend on the choice of μ. It can be shown that the topologies induced by Hellinger metric and the—perhaps more familiar—total variation metric are the same. In fact, we have

$$\frac{1}{2}H^2(P,Q) \leq \|P-Q\|_{TV} \leq H(P,Q),$$

where $\|.\|_{TV}$ denotes the total variation norm. The variational metric is always between 0 and 1 and the Hellinger metric is between 0 and $\sqrt{2}$.

Definition 1.3. *The* Kullback–Leibler information *of a probability measure P with respect to Q is defined to be*

$$K(P:Q) = \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P. \tag{1.5}$$

Conventionally, $\mathrm{d}P/\mathrm{d}Q$ is only defined when Q dominates P and the Kulback–Leibler information is defined to be infinite when that is not the case. Since we defined $\mathrm{d}P/\mathrm{d}Q$ even when Q does not dominate P, (1.5) completely specifies the Kullback–Leibler information. One can check that, also according to our definition (1.5), $K(P:Q) = \infty$ when Q does not dominate P. Note that $K(P:Q)$ can well be infinite even when Q dominates P.

Even though the Kullback–Leibler information is often used to quantify the difference between two measures, it is not a distance in the mathematical sense. However, as a simple consequence of Jensen's inequality we do have that $K(P:Q) \geq 0$ with equality if and only if $P = Q$.

The following lemma relates Kullback–Leibler information to Hellinger distance.

Lemma 1.2. *For any two probability measures P and Q we have*

$$K(P:Q) \geq H^2(P,Q).$$

Proof For all $x > 0$ we have $\log x \leq x - 1$. Hence,

$$\begin{aligned}
K(P:Q) &= \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P \\
&= -2\int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^{1/2}\mathrm{d}P \\
&\geq 2\int 1 - \left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^{1/2}\mathrm{d}P \\
&= 2 - 2\int \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1/2}\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{1/2}\mathrm{d}\mu \\
&= H^2(P,Q).
\end{aligned}$$

where μ is some measure dominating both P and Q. □

We now turn to consistency. Suppose for now that

$$K(P_0 : \hat{P}_n) = \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}P_0 < \infty.$$

We can then rewrite (1.3) as

$$\begin{aligned} 0 &\geq \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n \\ &= \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_0) + \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}P_0 \\ &= \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_0) + K(P_0 : \hat{P}_n), \end{aligned}$$

so if

$$\int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_0) \to 0, \qquad P_0\text{-almost surely} \tag{1.6}$$

then $K(P_0 : \hat{P}_n) \to 0$, which by the above lemma implies $H(P_0 : \hat{P}_n) \to 0$. In other words, (1.6) implies the almost sure Hellinger consistency of the NPMLE $\hat{P}_n$.

Condition (1.6) is an obvious consequence of a uniform law of large numbers over a class of functions. Such uniform LLN's are the object of much study in modern empirical process theory. In 'empirical process parlance', (1.6) follows directly from saying that the random function $\log(\mathrm{d}P_0/\mathrm{d}\hat{P}_n)$ is, for n large enough, in a so-called Glivenko–Cantelli class.

The convergence of (1.6) establishes 'Kullback–Leibler' consistency, which by lemma 1.2 implies Hellinger consistency. We can also get at Hellinger consistency directly. The equality we will now prove is due to van de Geer (1993), though she considered only the case where $\mathcal{P}$ is dominated by a single sigma-finite measure. We make no assumptions on $\mathcal{P}$ here. Taking $P = P_0$

in (1.2) and then using the fact that $\log x \leq x - 1$ $(x > 0)$ have

$$\begin{aligned}
0 &\leq \int \log \frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \mathrm{d}\mathbb{P}_n \\
&= 2 \int \log \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}\mathbb{P}_n \\
&\leq 2 \int \left[\left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} - 1 \right] \mathrm{d}\mathbb{P}_n \\
&= 2 \int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_0) + 2 \int \left[\left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} - 1 \right] \mathrm{d}P_0 \\
&= 2 \int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_0) - 2H^2(P_0 : \hat{P}_n).
\end{aligned}$$

Because $H^2(P_0; \hat{P}_n)$ is always between 0 and 2 it is not a problem to add and delete it as we did in the one-but-last line above. Now if

$$\int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_0) \to 0 \tag{1.7}$$

we can conclude that $H(P_0, \hat{P}_n) \to 0$, which is to say that $\hat{P}_n$ is Hellinger consistent. Again is suffices to verify a uniform law of large numbers. Van de Geer (1993) proceeds to use empirical process theory to investigate the rate of convergence of $\hat{P}_n$ to P_0 (when indeed they converge).

In case $\mathcal{P}$ is convex, she suggests comparing $\hat{P}_n$ not to P_0 but to a convex combination of $\hat{P}_n$ and P_0, say P_n. The idea is that it may be easier to establish $H(\hat{P}_n, P_n) \to 0$ than $H(\hat{P}_n, P_0) \to 0$ when the two are actually equivalent.

The two methods we presented for proving consistency of the NPMLE yield Hellinger consistency, which—being equivalent to consistency in total variation norm—is a *very* strong form of consistency. It may be too much to ask for. It may also be more than we want as we are often only interested in weak consistency (i.e. consistency in the Prohorov metric). In many non– and semi-parametric models the NPMLE is discrete whereas the true P_0 could be absolutely continuous. As the total variation distance between singular measures is 1, the Hellinger distance between $\hat{P}_n$ and P_0 cannot be less than 1.

We propose the use of a sequence $\{P_n\}$ which tends to P_0 in some desired sense, say weakly, while $H(\hat{P}_n, P_n)$ tends to zero. Such a sequence was successfully applied by Murphy (1994) in a case where Hellinger consistency indeed did not hold.

We cast this idea in the form of a theorem. Suppose that d is a metric on $\mathcal{P}$ which weaker than (or equivalent to) Hellinger metric.

Theorem 1.1. *Let $\hat{P}_n$ denote the NPMLE based on independent observations $X_1, X_2, \ldots, X_n$ which are all distributed according to $P_0 \in \mathcal{P}$ and let $\mathbb{P}_n$ denote the empirical measure. If there exists a sequence $\{P_n\} \in \mathcal{P}$ such that $d(P_n, P_0) \to 0$ (P_0-a.s.) while either*

$$K(P_n : \hat{P}_n) < \infty \quad \textit{and} \quad \int \log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_n) \to 0, \qquad P_0\textit{-almost surely}$$

or

$$\int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_n} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_n) \to 0, \qquad P_0\textit{-almost surely}$$

then $d(\hat{P}_n, P_0) \to 0$ (P_0-almost surely)

Proof Replacing P_0 with P_n in the preceding discussion, we see that either condition on P_n allows us to conclude that $H(\hat{P}_n, P_n) \to 0$. This theorem's claim now follows from the assumption that $d(P_n, P_0) \to 0$. □

1.4.3 Score equations

In the parametric case the 'score function', i.e. the derivative of the log likelihood, is a familiar object. Often the maximum likelihood estimator solves the score equations: sum of scores (is less than or) equals zero. This sometimes enables us to calculate the MLE and it also helps us to prove efficiency and asymptotic normality. As it turns out, the notion of a score can be extended very neatly to the infinite dimensional case. Hence we can also write down 'non-parametric' score equations. In the infinite dimensional case the score equations usually do not characterize the NPMLE. However, solutions of the score equations are often relatively easy to obtain, when the real NPMLE is much more elusive. Besides, we tend to prove favorable properties of maximum likelihood estimators by considering them as solutions of score equations. Why then not simply consider such solutions from the outset? We shall at times be sloppy and even call any solution of the score equations an NPMLE.

Scores are defined through what are called 'differentiable submodels'. Following van der Vaart (1998), we consider maps $t \mapsto P_t$ from a neighborhood of 0 in $[0, \infty)$ to $\mathcal{P}$ such that there exists a measurable function $g : \mathcal{X} \to \mathbb{R}$ for which

$$\int \left[\frac{p_t^{1/2} - p^{1/2}}{t} - \frac{1}{2} g p^{1/2} \right]^2 \mathrm{d}\mu \to 0. \tag{1.8}$$

where p_t and p are Radon–Nikodym derivatives of P_t and P with respect to a sigma-finite measure $\mu(P, P_t)$, which dominates both P_t and P.

We call $t \mapsto P_t$ a differentiable submodel through P. If the integrand converges pointwise then $g = \frac{\mathrm{d}}{\mathrm{d}t}|_{t=0} \log \mathrm{d}P_t$ which is just the familiar score function. The collection of scores associated with all differential submodels is called the *tangent set* of the model $\mathcal{P}$ at P and denoted $\dot{\mathcal{P}}(P)$. Lemma 25.14 in van der Vaart (1998) asserts that from (1.8) it follows that $\int g \mathrm{d}P = 0$ and $\int g^2 \mathrm{d}P < \infty$. Hence, $\dot{\mathcal{P}}(P)$ can be identified (up to equivalence) with a subset of $L_2^0(P)$.

Here a clear distinction can be made between semi- and nonparametric models: a nonparametric model has a tangent set equal to the whole of $L_2^0(P)$, while the tangent set of a semi-parametric model is a true subset of $L_2^0(P)$. Moreover, the tangent set of a semi-parametric model is not contained in a finite dimensional subspace of $L_2^0(P)$. In other words, it is infinite dimensional but not everything. Indeed, if the model consists of all probability measures (on a given measurable space) then $\dot{\mathcal{P}}(P) = L_2^0(P)$.

Now consider the 'score inequalities'

$$\int g \mathrm{d}\mathbb{P}_n = \frac{1}{n} \sum_i g(X_i) \leq 0 \qquad \forall g \in \dot{\mathcal{P}}(\hat{P}_n) \tag{1.9}$$

An NPMLE as defined in the previous section is a solution of these inequalities. The converse does not hold. Still, we will call a solution of these equations an NPMLE.

The most familiar NPMLE, the empirical measure, solves a the score inequalities when the model $\mathcal{P}$ is non-parametric so that $\dot{\mathcal{P}}(\hat{P}_n) = L_2^0(\hat{P}_n)$. Taking scores $g(x) = \mathbf{1}_F(x) - \hat{P}_n(F)$, for all $F \in \mathcal{F}$, we see that the empirical $\hat{P}_n = \mathbb{P}_n$ indeed arises as a solution of (1.9).

We now discuss how the methods of the previous section can be used to prove consistency of solutions of the score equations. To this end we must choose particular submodels. Let $\hat{P}_n$ be a solution of the score equations and consider a sequence $\{P_n\} \in \mathcal{P}$ such that $P_n \ll \hat{P}_n$.

Now suppose that $\log(\mathrm{d}P_n / \mathrm{d}\hat{P}_n) \in L_2(\hat{P}_n)$ and that $\hat{P}_n^{(1-t)} P_n^t \in \mathcal{P}$ for t in an open neighborhood of 0 in $[0, \infty)$. Then $t \mapsto \hat{P}_n^{(1-t)} P_n^t$ is a differentiable

submodel with score

$$\log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \in \dot{\mathcal{P}}(\hat{P}_n).$$

Because $\hat{P}_n$ solves all the score equations it follows that

$$\int \log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n \leq 0,$$

which coincides with (1.3). Now we can apply the reasoning of the previous section to try to prove that a particular $\hat{P}_n$ is consistent.

Unfortunately, it is not usually true that $\hat{P}_n^{(1-t)} P_n^t \in \mathcal{P}$. However, it *is* often true that $\mathcal{P}$ contains the straight line $(1-t)P + tQ$ for all P and Q in $\mathcal{P}$ (i.e. $\mathcal{P}$ is convex). If (and only if) $P_n \ll \hat{P}_n$ and $\mathrm{d}P_n/\mathrm{d}\hat{P}_n \in L_2(\hat{P}_n)$ then $t \mapsto (1-t)\hat{P}_n + tP_n$ is a differentiable submodel with score

$$\frac{\mathrm{d}P_n - \mathrm{d}\hat{P}_n}{\mathrm{d}\hat{P}_n} = \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \in \dot{\mathcal{P}}(\hat{P}_n)$$

and since $\hat{P}_n$ solves all score equations, it follows in particular that

$$\int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}\mathbb{P}_n \leq 0. \tag{1.10}$$

We are assuming that $\mathrm{d}P_n/\mathrm{d}\hat{P}_n \in L_2(\hat{P}_n)$, so in other words

$$\int \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}P_n < \infty.$$

Hence, we can rewrite (1.10) as

$$\begin{aligned} 0 &\geq \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}\mathbb{P}_n \\ &= \int \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_n) + \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}P_n. \end{aligned}$$

Now since $x - 1 > \log x$ for all $x > 0$ we have

$$\int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}P_n > \int \log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}P_n = K(\hat{P}_n, P_n) \geq 0.$$

Hence if

$$\int \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_n) \to 0 \qquad P_0 - \text{almost surely} \tag{1.11}$$

then $\int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1\right) \mathrm{d}P_n$ tends to zero, which implies that $K(\hat{P}_n, P_n)$ tends to zero which in turn implies that $H(\hat{P}_n, P_n)$ tends to zero. Thus we have a theorem which is very similar to Theorem 1.1.

Let d be a metric on $\mathcal{P}$ which is weaker than (or equivalent to) Hellinger metric.

Theorem 1.2. *Let $\hat{P}_n$ denote a solution of the score equations (1.9) based on independent observations $X_1, X_2, \ldots, X_n$ which are all distributed according to $P_0 \in \mathcal{P}$ and let $\mathbb{P}_n$ denote the empirical measure. Suppose also that $\mathcal{P}$ is convex. If there exists a sequence $\{P_n\} \in \mathcal{P}$ such that*

1. *$d(P_n, P_0) \to 0$ (P_0-a.s.)*
2. *The submodel $t \mapsto (1-t)\hat{P}_n + tP_n$ is Hellinger differentiable at $\hat{P}_n$*
3. *(1.11) holds*

then $d(\hat{P}_n, P_0) \to 0$ (P_0-a.s.)

1.4.4 Missing data problems and the EM algorithm

In this section we study missing data models which are also known as information loss models or mixture models. The issue is that for some reason we do not (completely) observe the random variables whose common distribution $Q_0 \in \mathcal{Q}$ is of interest.

Let X be a generic random variable on some (measurable) space $\mathcal{X}$ with distribution $Q_0 \in \mathcal{Q}$ and let C be another random variable whose conditional distribution given X is known to us. Let $Y = \Phi(X, C)$, where Φ is a known many-to-one mapping. Y takes values in a measurable space $\mathcal{Y}$. Now suppose we observe independent $Y_1, \ldots, Y_n$ which are distributed as Y and from these we want to estimate Q_0. We call the X_i the complete data and the Y_i the incomplete or observed data. Estimation of $Q_0 \in \mathcal{Q}$ (or functionals thereof) from the observations Y_i is called a missing data problem.

The setup we have described is so general that one could maintain that *every* estimation problem is a missing data problem. Still, it will prove a convenient way to think about all sorts of censoring mechanisms.

We now give another way to describe the situation. Consider a *Markov kernel* $K(\mathrm{d}y; x)$. This means that $K(\mathrm{d}y; x)$ is a probability measure for every $x \in \mathcal{X}$ and $K(A; x)$ is a measurable function of x when $A \subseteq \mathcal{Y}$ is a measurable set. Now for all $Q \in \mathcal{Q}$ define the 'mixture' $P = KQ$ by

$$P(\mathrm{d}y) = KQ(\mathrm{d}y) = \int_{\mathcal{X}} K(\mathrm{d}y; x)\mathrm{d}Q(x). \tag{1.12}$$

Clearly P is a probability measure on the space $\mathcal{Y}$. Q is called the 'mixing distribution'. We suppose K is known and we observe an i.i.d. sample $Y_1, \ldots, Y_n$ with common distribution $P_0 = KQ_0$, where Q_0 is known to lie in a class $\mathcal{Q}$. The model for the distribution of a generic observation Y is of course $\mathcal{P} = K\mathcal{Q} = \{KQ : Q \in \mathcal{Q}\}$. The objective, again, is to estimate Q_0.

We have explained what we mean by a 'missing data model' and by 'a mixture model'. In fact, these two descriptions are the same. A missing data model can be described as a mixture model by defining $K(\mathrm{d}y; x)$ to be the conditional distribution of $Y = \Phi(X, C)$ given $X = x$. Conversely, a mixture model can be described as a missing data model as follows. First, define a random variable C which, conditional on $X = x$, has distribution $K(\mathrm{d}y; x)$. Next, define a random variable $Y = \Phi(X, C) = C$.

We shall now give two useful properties of $\mathcal{P} = K\mathcal{Q}$ which follow directly from properties of $\mathcal{Q}$.

Lemma 1.3. *If $\mathcal{Q}$ is convex then so is $\mathcal{P} = K\mathcal{Q}$.*

Proof Take two arbitrary elements $P, P' \in \mathcal{P}$. Then there exist $Q, Q' \in \mathcal{Q}$ such that $P = KQ$ and $P' = KQ'$. Now for $0 \leq t \leq 1$

$$\begin{aligned} tP + (1-t)P' &= t \int K(\mathrm{d}y; x)\mathrm{d}Q(x) + (1-t) \int K(\mathrm{d}y; x)\mathrm{d}Q'(x) \\ &= \int K(\mathrm{d}y; x)\mathrm{d}(tQ(x) + (1-t)Q'(x)) \in \mathcal{P} \end{aligned}$$

because $tQ + (1-t)Q' \in \mathcal{Q}$. □

Even more importantly,

Lemma 1.4. *If $t \mapsto Q_t$ is a differentiable submodel through $Q \in \mathcal{Q}$ with score function g , then $t \mapsto P_t = KQ_t$ is a differentiable submodel through $P = KQ \in \mathcal{P}$ with score $E_Q(g(X)|Y)$.*

Proof For a proof refer to Bickel et al. (1993) proposition A.5.5 or Le Cam and Yang (1988) proposition 4. □

We shall concern ourselves exclusively with completely *nonparametric* missing data / mixture problems, which means that $\dot{\mathcal{Q}}(Q) = L_2^0(Q)$ for all $Q \in \mathcal{Q}$. In other words, every square integrable function, which integrates to zero is a score function belonging to a differentiable submodel. The fact that scores are so plentiful is useful when we want to estimate Q_0 by solving score equations. The nonparametric situation typically arises when we have

no prior information about Q_0 so that we must take our model $\mathcal{Q}$ to consist of all probability measures on $\mathcal{X}$. In passing, we note that if $\mathcal{Q}$ consists of all probability measures on a given space, then $\mathcal{Q}$ is convex. By the first of the above lemmas, convexity of $\mathcal{Q}$ implies convexity of $\mathcal{P} = K\mathcal{Q}$. This will be very useful as this puts us in the realm of Theorem 1.2.

The fact that we only observe Y and not X may cause a lack of *identifiability*. We may well have that $KQ = KQ'$ while $Q \neq Q'$. This is unfortunate as it means that from the data we can not tell Q and Q' apart. We can of course reduce $\mathcal{Q}$ in such a way that we do have identifiability, but then our missing data problem will most likely fail to be completely non-parametric.

In the problems we study in this thesis we can construct a *different* missing data problem which *is* identifiable, but also completely non-parametric. It features a different model $\tilde{\mathcal{Q}}$ of probability measures on a different space $\tilde{\mathcal{X}}$ and also a different Markov kernel $\tilde{K}$. Of course we make sure that $\tilde{K}\tilde{\mathcal{Q}} = K\mathcal{Q}$ so that the model for the data is the same. For the remainder of this section we simply assume identifiability.

Turning to maximum likelihood estimation, the NPMLE $\hat{Q}_n$ of Q_0 is defined to satisfy

$$\int \log \frac{\mathrm{d}P}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n := \int \log \frac{\mathrm{d}KQ}{\mathrm{d}K\hat{Q}_n} \mathrm{d}\mathbb{P}_n \leq 0, \qquad \forall Q \in \mathcal{Q},$$

where $\mathbb{P}_n$ is the empirical distribution of the *observed* data $Y_1, \ldots, Y_n$.

Instead of trying to find the NPMLE, we might be already be content with a solution $\hat{Q}_n$ of the score equations

$$\int E_{\hat{Q}_n}(g(X)|y)\mathrm{d}\mathbb{P}_n(y) = \frac{1}{n}\sum_i E_{\hat{Q}_n}(g(X_i)|Y_i) = 0, \quad \forall g \in \dot{\mathcal{Q}}(\hat{Q}_n). \tag{1.13}$$

If we have a completely nonparametric missing data problem every function g in $L_2^0(\hat{P}_n)$ is a score. Thus we can take $g(x) = \mathbf{1}_A(x) - \hat{Q}_n(A)$ for any measurable set A to obtain the so-called *self consistency equations*

$$\hat{Q}_n(A) = \int E_{\hat{Q}_n}(\mathbf{1}_A(X)|y)\mathrm{d}\mathbb{P}_n(y)) = \frac{1}{n}\sum \hat{Q}_n(X_i \in A|Y_i). \tag{1.14}$$

It is usually assumed that for all x the kernels $K(\mathrm{d}y; x)$ have densities with respect to a single sigma-finite measure. This may not hold, and in particular it does not hold in the problems that are considered in this thesis. Fortunately, to solve the score equations we only need to compute integrals with respect to the conditional distribution under $\hat{Q}_n$ of X given Y. Hence,

it suffices if the $K(\mathrm{d}y; x)$ are dominated for $\hat{Q}_n$-almost all x. Thus suppose that a sigma-finite measure μ_n dominates $K(\mathrm{d}y; x)$ for $\hat{Q}_n$-almost all x. The measure μ_n is allowed to depend on the data $Y_1, \ldots, Y_n$. We denote $K(\mathrm{d}y; x)/\mathrm{d}\mu_n(y) = k_n(y; x)$. Then the conditional distribution of X given $Y = y$ under $X \sim \hat{Q}_n$ is

$$\frac{k_n(y; x)\mathrm{d}\hat{Q}_n(x)}{\int_{x'} k_n(y; x')\mathrm{d}\hat{Q}_n(x')} \tag{1.15}$$

and the score equations (1.13) become

$$\sum_i \int \frac{g(x)k_n(Y_i; x)\mathrm{d}\hat{Q}_n(x)}{\int_{x'} k_n(Y_i; x')\mathrm{d}\hat{Q}_n(x')} = 0, \quad \forall g \in \dot{\mathcal{Q}}(\hat{Q}_n). \tag{1.16}$$

Again, if a missing data problem is completely nonparametric then we can take $g(x) = \mathbf{1}_A(x) - \hat{Q}_n(A)$ for any measurable set A and (1.14) becomes

$$\hat{Q}_n(A) = 1/n \sum_i \int_A \frac{k_n(Y_i; x)\mathrm{d}\hat{Q}_n(x)}{\int_{x'} k_n(Y_i; x')\mathrm{d}\hat{Q}_n(x')}. \tag{1.17}$$

These equations are the well known self-consistency equations. Solving them iteratively is an instance of the EM algorithm. When we start the iterations with an initial guess $Q_n^{(0)}$ then all subsequent iterates will be dominated by it. Hence we can never do better than compute a 'sieved' NPMLE over the model $\{Q \in \mathcal{Q} : Q \ll Q_n^{(0)}\}$.

There has been much discussion of the merits and flaws of the EM algorithm (Dempster et al. 1977). Vardi and Lee (1993) point out the wide applicability of EM, but—obviously—blind application of the algorithm to every problem one meets is unwise. However, it does seem to perform rather well in 'nice' cases, i.e., parametric– and $\sqrt{n}$-nonparametric problems. EM is wholly unsuitable to determine the support of a distribution.

The rate at which EM converges is painfully slow, and has even been described as 'lethargic'. On the other hand, only infantile computer skills are required to implement it. A clever mathematician who is good with computers can generally do much better (cf. Meilijson, 1989).

1.5 Perfect Simulation

The distribution of a spatial process is often very complicated. Typically, if there is any spatial dependence we have a normalizing constant that is utterly intractable. To appreciate this fact, consider a Poisson point process

of intensity $\lambda > 0$ on the unit square. Suppose we are interested in the conditional distribution of the process, given that no two points are closer than at distance $r > 0$. This is called the hard core model. The normalizing constant here is an integral over all finite configurations of points that are all further apart than r. Performing this integration analytically appears impossible. Stochastic integration would be feasible if we can efficiently obtain realizations of the hard core model. Rejection sampling, i.e. generating Poisson processes and discarding those that do not meet the requirement, may not be efficient depending on λ and r.

Coupling-from-the-past (CFTP) (Propp and Wilson, 1996) is a clever method to obtain samples from complicated distributions which arise as the stationary distribution of a Markov chain—as quite often happens in spatial models. For instance, the hard core model arises as the stationary distribution of a so-called spatial birth and death process, where births that violate the condition are rejected.

In chapter 4 we consider estimation of the intensity of the Boolean model of discs (also known as the 'bombing model'). As it turns out we need to sample from a conditional Poisson point process. We have implemented a CFTP algorithm which is extremely quick.

This Boolean discs problem is a missing data problem, and we can use a stochastic version of the EM algorithm (StEM) to estimate the intensity. This algorithm produces an ergodic Markov chain on the parameter space. The StEM estimator is a sample (or an average of samples) from the stationary distribution of this chain. We can again use CFTP to actually obtain such samples. This we call the *perfect* StEM algorithm.

This section is based on van Zwet (2000). It is organized as follows. Below we first describe coupling-from-the-past, then the stochastic EM and finally the perfect stochastic EM algorithm.

1.5.1 Coupling From The Past

Consider an ergodic (i.e., irreducible and aperiodic) Markov chain X_t on a state space S and suppose we want to simulate its equilibrium distribution. Starting the chain from some arbitrary initial state and then running it for a very long, but finite time will generally not ensure that samples are from the stationary distribution. Recently, Propp and Wilson (1996) devised a method, called *coupling from the past* (CFTP), to produce perfect or exact samples. We closely follow Kendall and Thönnes (1998) to explain how it works.

For now, let us assume that the state space S is finite. A Markov chain X_t on S can be described by means of i.i.d. 'random transition maps' H_t :

$S \to S$. Such a random transition map H_t determines for each state $i \in S$ that the chain X might be in at time $t-1$ where the chain would jump to next. If p_{ij} are the transition probabilities of the Markov chain to move from state i to j, then the common distribution of the H_t should be such that $P(H_t(i) = j) = p_{ij}$. Then a realization of the Markov chain X is generated by fixing $X(0)$ at some X_0 and setting

$$X(t) = H_t(X(t-1)).$$

Coupling from the past now works as follows. We select a time $-T < 0$ in the past and run chains starting from each state of S from time $-T$ to time 0. The chains are coupled by using the *same* realizations of the transition maps for all the chains. Hence, it follows that if two chains started at different initial states coalesce, they will forever after stay together. Now we check if all chains have coalesced at time 0. If so, then the state at time zero must be a sample from the stationary distribution. This is understood as follows. Imagine that at some time long before $-T$ we also started a chain from an initial state selected according to the stationary distribution. This chain will remain in equilibrium, so in particular its state at time zero is distributed according to the stationary distribution. However, we have arranged it so that *all* chains, no matter which state they were in at time $-T$, are at time zero in the same state.

If not all chains have met, then we run chains from time $-2T$ to time 0, making sure that we use the same realizations of $H_{-T+1}, H_{-T+2}, \ldots, H_0$. If the paths still have not coalesced, we run chains from time $-4T$ and so on.

If the state space S has more than just a few elements it will not be feasible to run chains starting from all possible states.

Let us now consider finite, countable or even uncountable S. Suppose that S admits a partial ordering $\preceq$, and that there are a minimal and a maximal element, $\underline{s}$ and $\overline{s}$ such that

$$\underline{s} \preceq s \preceq \overline{s}, \qquad \text{for all } s \in S.$$

Also suppose that the chain is 'monotone' in that it respects the ordering.

$$H_t(s) \preceq H_t(s'), \qquad \text{for all } s \preceq s' \text{ a.s.}$$

Now we can imagine that we run coupled chains starting at all possible states, while in practice we only run chains from states $\underline{s}$ and $\overline{s}$. Because all paths are ordered, it follows that if the latter two have met then all paths starting from intermediate states would have met as well. Of course it remains to verify in each application that the algorithm will almost surely terminate in finite time.

1.5.2 Stochastic EM

The stochastic EM (StEM) algorithm (Celeux and Diebolt (1986), Wei and Tanner (1990)), which we describe here, is an iterative procedure to estimate a parameter in a missing data problem. It produces an ergodic Markov Chain on the parameter space. The equilibrium distribution of this chain has certain desirable properties (Nielsen (1997a and b). In the next section we will investigate if and when the ideas of the previous section can be applied to actually obtain a sample from this equilibrium distribution.

Suppose that X is distributed according to a probability measure P_{θ_0}. Suppose we can observe only the result of a many-to-one mapping $Y(C, X)$, where the conditional distribution of C given X is known. The goal is to estimate θ_0, where θ_0 is assumed to be in some general set Θ, from observing $Y = y$. This is sometimes called a missing data problem (see section 1.4.4). Often the EM algorithm (Dempster, Laird and Rubin (1977)) provides a method to find the maximum likelihood estimator of θ_0. There are two drawbacks. The first is that it is not known how many iteration steps are needed to bring one close enough to convergence. The other is that sometimes the E-step, computation of the conditional expectation of the likelihood given the data, is impossible.

In this latter case, the stochastic version of the EM algorithm (StEM) may be a viable alternative. For a review and large sample results see Nielsen (1997a and b). The algorithm works as follows. Suppose the complete data maximum likelihood estimator is readily computable: $\hat{\theta}_{MLE} = M(X)$.

1. Fix a $\theta(0)$ in Θ;
2. Sample $X(1)$ from $P_{\theta(0)}(.|Y = y)$;
3. Set $\theta(1) = M(X(1))$.

By iterating this procedure, we obtain a sequence $\theta(0), \theta(1), \theta(2), \ldots$. If step 2 is carried out using independent Unif(0,1) variables for each t, the sequence $\theta(t)$ is a time homogeneous Markov chain. Under certain conditions it is also ergodic. If so, the algorithm converges in that the $\theta(t)$ converge in distribution to a random variable, say $\hat{\theta}$, which is distributed according to the stationary distribution of the Markov chain. Then $\hat{\theta}$ is the StEM estimate. In other words, a StEM estimate is a sample from the stationary distribution of the 'StEM Markov chain'.

It is not clear for how long we should run the StEM chain to allow it to reach equilibrium. This lack of a termination rule seems to be something StEM and EM have in common. However, we shall find that in some cases

we can apply coupling-from-the-past to obtain a sample that is guaranteed to come from the stationary distribution of the StEM chain.

1.5.3 Perfect Stochastic EM

In this section we combine the ideas from the previous two sections. We use the set-up and notation of the preceding section. Let us suppose that the parameter space Θ admits a partial ordering $\preceq$. Fix a time $-T < 0$ in the past. Suppose that for $t = 0, -1, -2, \ldots$ we can construct independent collections of random variables $\{X_\theta(t),\ \theta \in \Theta\}$ such that

1. $X_\theta(t) \ \sim \ P_\theta(.|Y = y)$
2. $\theta \preceq \theta' \Rightarrow M(X_\theta(t)) \preceq M(X_{\theta'}(t))$ almost surely.

Recall that $M(X)$ is the complete data maximum likelihood estimator of θ.

To simulate a Markov chain $\theta_{-T}(-T), \theta_{-T}(-T+1), \ldots, \theta_{-T}(0)$ we define random transition maps

$$H_t(\theta(t-1)) = M(X_{\theta(t-1)}(t)), \qquad t = -T+1, -T+2, \ldots, 0.$$

Now suppose that there are 'minimal' and 'maximal' elements $\underline{\theta}$ and $\overline{\theta}$ such that $\underline{\theta} \preceq \theta \preceq \overline{\theta}$ for all $\theta \in \Theta$. Consider two coupled paths $\theta^L_{-T}(.)$ and $\theta^U_{-T}(.)$ starting at $\theta^L_{-T}(-T) = \underline{\theta}$ and $\theta^U_{-T}(-T) = \overline{\theta}$. It now suffices to check if these paths have coalesced at time zero, i.e. if $\theta^L_{-T}(0) = \theta^U_{-T}(0)$. If so, then we have a perfect StEM estimate $\hat{\theta} = \theta^L_{-T}(0)$. If not we have to go back further in time.

We demonstrate the Perfect StEM algorithm in a simple example. We also apply it to a more involved problem in section 4.4.2.

An Example

Suppose $X = (X_1, X_2, \ldots, X_n)$ is an i.i.d. sample from $P_{\theta_0} = \text{Exp}(\theta_0)$; the exponential distribution with intensity (reciprocal of the mean) θ_0. We wish to estimate the intensity θ_0. Suppose that θ_0 is known to be bigger than some $\underline{\theta} > 0$, that is $\theta_0 \in \Theta = [\underline{\theta}, \infty)$ Of course the maximum likelihood estimator of θ_0 is $\hat{\theta}_{MLE} = M(X) = (n/\sum X_i) \vee \underline{\theta}$. Suppose we only observe

$$\tilde{X}_i = X_i \wedge C \qquad \Delta_i = \mathbf{1}_{\{X_i > C\}},$$

for some fixed positive constant C. Write $Y_i = (\tilde{X}_i, \Delta_i)$ for the observed data.

The maximum likelihood estimator of θ_0 based on the observed data is known to be the 'occurrence' divided by the 'exposure'.

$$\frac{n - \sum \Delta_i}{\sum \tilde{X}_i}.$$

So there is really no need to apply the StEM algorithm here. The purpose of this example is strictly illustrative.

We now describe how the StEM algorithm works here. When multiplying vectors we mean coordinate wise multiplication (mapping two vectors to one vector).

1. Fix $\theta(0) > 0$
2. Generate $X(1) \sim P_{\theta(0)}(X|Y = y)$:

$$X(1) = \tilde{X} + \Delta E, \quad \text{where } E = (E_1, \dots, E_n),\ E_i \sim \text{Exp}(\theta(0))$$

3. $\theta(1) = M(X(1)) = \frac{n}{\sum X_i(1)} \vee \underline{\theta}$.

Repeating this procedure, we obtain an ergodic Markov chain $\theta(0), \theta(1), \dots$.

Now we demonstrate how we can apply perfect sampling to this chain. Recall the usual ordering on $\mathbb{R}^n$: $x \leq y$ if $x_1 \leq y_1$ and $x_2 \leq y_2 \dots$ and $x_n \leq y_n$. Note that if $x \geq y$ then $M(x) \leq M(y)$. Set $\overline{\theta} = M(\tilde{X})$. Since, for all t, $X(t) \geq \tilde{X}$ it follows that $\theta(t) = M(X(t))$ will always be less than $\overline{\theta}$. Hence, we may reduce our parameter-space Θ to $[\underline{\theta}, \overline{\theta}]$. Of course $\underline{\theta}$ and $\overline{\theta}$ are minimal and maximal elements of Θ with the usual ordering.

We now construct collections $\{X_\theta(t),\ \theta \in \Theta\}$ as required at the beginning of this section. For $t = -T+1, -T+2, \dots 0$ and $i = 1, 2, \dots, n$ generate independent

$$E_{\underline{\theta},i}(t) \sim \text{Exp}(\underline{\theta}) \quad \text{and} \quad E_i(t) \sim \text{Exp}(\overline{\theta} - \underline{\theta})$$

and define

$$E_{\theta,i}(t) = E_{\underline{\theta},i}(t) \wedge \frac{\overline{\theta} - \underline{\theta}}{\theta - \underline{\theta}} E_i(t).$$

Then, evidently, $E_{\theta,i}(t) \sim \text{Exp}(\theta)$, and $\theta \leq \theta'$ implies $E_{\theta,i}(t) \geq E_{\theta',i}(t)$. For $t = 0, -1, -2, \dots$ and all $\theta \in \Theta$ we define

$$X_{\theta,i}(t) = \widetilde{X}_i + \Delta_i E_{\theta,i}(t),$$

Thus, for $t = 0, -1, -2, \dots$ we have collections of random variables $\{X_\theta(t),\ \theta \in \Theta\}$ such that

1. $X_\theta(t) \ \sim \ P_\theta(.|Y = y)$

2. $(\theta \leq \theta' \Rightarrow M(X_\theta(t)) \leq M(X_{\theta'}(t)))$ a.s.

as required. We can now run a 'lower' chain $\theta^L_{-T}(-T) = \underline{\theta}, \theta^L_{-T}(-T+1), \ldots, \theta^L_{-T}(0)$ and an 'upper' chain $\theta^U_{-T}(-T) = \overline{\theta}, \theta^U_{-T}(-T+1), \ldots, \theta^U_{-T}(0)$ and check if $\theta^L_{-T}(0) = \theta^U_{-T}(0)$. We do need to make sure that the algorithm will terminate in finite time. It is enough to check that for some fixed $-T$ the event $\theta^L_{-T}(0) = \theta^U_{-T}(0)$ has positive probability. Well,

$$\begin{aligned} \Pr(\theta^L_{-1}(0) = \theta^U_{-1}(0)) &\geq \Pr(X_{\underline{\theta},i}(0) = X_{\overline{\theta},i}(0), \quad \forall i) \\ &\geq \Pr(\mathrm{Exp}(\underline{\theta}) \leq \mathrm{Exp}(\overline{\theta} - \underline{\theta}))^n > 0. \end{aligned}$$

This concludes the example.

The difficulty in making the StEM algorithm perfect lies in the construction of collections of random variables satisfying points 1 and 2 at the beginning of this section. There probably is no general recipe to achieve this. Here we ask ourselves when such a construction is at all possible. It turns out that the concept of *realizable monotonicity* (Fill and Machida (1998)) is essentially what we need.

Realizable monotonicity is very close to stochastic monotonicity, which is a more familiar concept and which is also easier to check. Suppose that the complete data are random elements in some space E, and that the complete data maximum likelihood estimator of a parameter $\theta \in \Theta$ is given by a function M on E. We assume that the parameter space Θ admits a partial order $\preceq_1$.

Suppose that E admits a partial ordering $\preceq_2$ such that, for any $x_1, x_2 \in E$

$$x_1 \preceq_2 x_2 \;\Rightarrow\; M(x_1) \preceq_1 M(x_2).$$

Then we need to be able to construct a collection $\{X_\theta,\ \theta \in \Theta\}$ such that

1. $X_\theta \sim P_\theta$,

2. $\theta \preceq_1 \theta' \Rightarrow X_\theta \preceq_2 X_{\theta'}$, almost surely.

We shall have a look at two notions of monotonicity for a collection $\{P_\theta\ ,\ \theta \in \Theta\}$ of probability measures: realizable and stochastic monotonicity.

Definition 1.4. *The collection* $\{P_\theta\ ,\ \theta \in \Theta\}$ *is called* realizably monotone *if there exists a collection of random variables* $\{X_\theta\ ,\ \theta \in \Theta\}$ *such that*

1. $X_\theta \sim P_\theta$,

2. $\theta \preceq_1 \theta' \Rightarrow X_\theta \preceq_2 X_{\theta'}$, *almost surely.*

A subset U of E is said to be an *up-set* in $(E, \preceq_2)$ if $y \in U$ whenever $x \in U$ and $x \preceq_2 y$. If P_1 and P_2 are probability measures on E then P_1 is *stochastically smaller* than P_2 if $P_1(U) \leq P_2(U)$ for all up-sets U in $(E, \preceq_2)$. we then write $P_1 \preceq_2^D P_2$.

Definition 1.5. *The collection* $\{P_\theta\ ,\ \theta \in \Theta\}$ *is called* stochastically monotone *if*

$$\theta \preceq_1 \theta' \Rightarrow P_\theta \preceq_2^D P_{\theta'},\ a.s.$$

Realizable monotonicity is clearly what we need, while stochastic monotonicity is usually easier to check. It is easily seen that realizable monotonicity implies stochastic monotonicity. That the converse is not always true is demonstrated by an example in Ross (1993). However, for various finite classes of $(E, \preceq_2)$, Fill and Machida (1998) give conditions on finite index sets $(\Theta, \preceq_1)$ such that stochastic monotonicity *does* imply realizable monotonicity. For instance, realizable and stochastic monotonicity are equivalent when $(E, \preceq_2)$ or $(\Theta, \preceq_1)$ is a finite linearly ordered set. Recall that a set is linearly ordered if each pair of elements is comparable.

Fill and Machida (1998) coined the term 'realizable monotonicity' and noted its relevance for perfect sampling. Their results for finite sets are all the more useful because of the following unpublished result by Ross (1993).

Theorem 1.3. *Suppose that* $(\Theta, \preceq_1)$ *is a partially ordered set and* $(E, \preceq_2)$ *is a complete separable metric space with closed partial order. Then* $\{P_\theta,\ \theta \in \Theta\}$ *is realizably monotone if and only if for every finite* $\Psi \subseteq \Theta$ $\{P_\theta,\ \theta \in \Psi\}$ *is realizably monotone.*

Now if for some separable set with a closed partial order we can use the results of Fill and Machida (1998) to check realizable monotonicity for all its finite subsets then Ross's theorem allows us to conclude realizable monotonicity for the entire infinite set.