# Models for point processes observed with noise

By JENS LUND

*Department of Mathematics and Physics, Royal Veterinary and Agricultural University, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark*

jlund@dina.kvl.dk

AND MATS RUDEMO

*Department of Mathematical Statistics, Chalmers University of Technology and Gothenburg University, 41296 Gothenburg, Sweden*

rudemo@math.chalmers.se

September 8, 1999

## Summary

Consider a pair of point processes, $X$ and $Y$, where $X$ is regarded as a 'true' point process and $Y$ is an imperfect observation of $X$. For the transformation from $X$ to $Y$, we consider a number of disturbance mechanisms covering random thinning, displacement, censoring of the displaced points, and, in addition, superposition of extra points. We present the conditional likelihood of $Y$ given $X$. When both point processes are observed the likelihood may be used for inference about the disturbance mechanisms: estimation of parameters, tests of model reductions et cetera. The likelihood is a sum, typically with extremely many terms, and we discuss an approximation with a small number of terms. The results are applied to an example, where $X$ denotes a set of 'true' positions of tree tops, while $Y$ denotes tree top positions estimated by template matching in a digital image obtained by high-resolution aerial photography. Using the conditional likelihood the parameters governing the various disturbance mechanisms are estimated.

*Some key words*: Censoring of point processes; Conditional likelihood; Image data; Incomplete observation; Random displacement; Superposition; Thinning; Matching of point sets.

## 1. Introduction

### 1·1 *Background*

Random displacement, thinning, and superposition are well-known operations on point processes, see Daley & Vere-Jones (1988) and Stoyan et al. (1995). An important problem is to decide when the distribution type, in particular, the Poisson process, is preserved under these operations. A related problem is to determine when iteration of these operations with suitable scaling gives a specific process, typically the Poisson process, as a limit.

In the present paper we will instead look at these mechanisms as disturbances of a point process, that is, as elements in models for incomplete observation of a point process. This type of modelling was used in Dralle & Rudemo (1997) where incomplete observations of tree positions from aerial photography data were studied. The combined effects of thinning, systematic and random displacement, censoring through displacement out of the study area, and superposition of additional points, called 'ghost points', were investigated. The statistical method used was iterative least squares estimation combined with a 'most probable' pairing of true and observed points. Here we will study the various disturbance mechanisms by use of maximum likelihood estimation. This involves derivation of the conditional likelihood function for the observed process given the true process, see Theorem 1 in § 2 with proof in the Appendix. A number of simplifications of the general disturbance model are discussed in § 3 such as pure displacement models, spatially homogeneous noise models and cluster processes.

The conditional likelihood for the observed point process contains a very large number of terms, each term corresponding to a specific mode of generation for the set of observed points. The model may be viewed as a missing data problem where each term specifies the origin of each observed point as either a displacement of one of the original points, or as an additional point. It is well-known that statistical inference for missing data may be complicated and the EM algorithm is one well known way to handle missing data, McLachlan & Krishnan (1997). However, the EM algorithm does not seem feasible in our case due to a complicated E-step, and in § 4 we study approximations of the likelihood considering a small number of comparatively large terms. We suggest an iterative procedure for finding the largest terms, including a starting procedure and search among terms similar to those currently considered according to specified 'neighbour' relations. Other computational considerations concern approximations close to the boundary of observations.

Likelihood estimation is applied to estimation of tree top positions from an aerial photo in § 5. Rather than using the kernel smoothing method of Dralle & Rudemo (1996, 1997) in the first step to find candidate tree positions, we employ the template matching method described in Larsen & Rudemo (1998) suitable also for aerial photographs obtained under off-nadir viewing angles. Point processes and marked point processes have a well established tradition in forestry, see Penttinen et al. (1992), although data have usually been acquired by ground-based field measurements.

### 1·2 *Basic assumptions and notation*

Let $Y$ be an imperfect observation of a point process $X$. We assume that $X$ and $Y$ are point processes on a subset $A$ of the $d$-dimensional Euclidean space $\mathbb{R}^d$ with a finite number of points, $X = \{X_i : i \in M\}$, $M = \{1, \ldots, m\}$, $Y = \{Y_j : j \in N\}$, $N = \{1, \ldots, n\}$. Assume further that $A$ is bounded with a positive $d$-dimensional volume $|A|_d$. Suppose that $Y$ is generated from the $X$-process by the following disturbance mechanisms:

(i) *Thinning.* Each point $X_i, i \in M$, is thinned with probability $1 - p(X_i)$ and retained with probability $p(X_i)$. If an $X$-point is thinned, then there will not be any corresponding $Y$-point. Thinnings are assumed to be independent for different points.

(ii) *Displacement.* For each remaining point $X_i$ a corresponding $Y$-point is generated by displacement to a position with probability density $k(\,\cdot\,|X_i)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. Given $X$, the displacements of different points are independent, mutually and of the thinnings.

(iii) *Censoring.* The displaced points are observed if they are within the observation region $A$; otherwise they are censored and not observed. Thus censoring of an unthinned point generated by $X_i$ occurs with probability $\int_{A^c} k(y|X_i)\,dy$. (Here $A^c$ denotes the complement $\mathbb{R}^d \setminus A$ of the set $A$.)

(iv) *Superposition of ghost points.* In addition to the points generated as described above we have superposition of extra 'ghost' points. These points are assumed to arise from a Poisson process on $A$ with intensity $g(\,\cdot\,|X)$, where $X$, as above, denotes the entire

$X$-process including thinned points. Given $X$, the ghost points are assumed to be independent of thinning, displacement and censoring.

The points generated from $X$ by the combination of thinning, displacement, censoring and superposition form the $Y$-process, which is thus restricted to the set $A$.

For a Borel set $B$, we let $Y(B)$ denote the number of $Y$-points in the set $B$, that is, we use the same symbol both for the point process and the associated counting measure. Further, $|S|$ will denote the number of elements in a finite set $S$. With this notation we thus have $Y(B) = |\{j \in N : Y_j \in B\}|$. Furthermore, we let $\mathcal{P}(M, M')$ denote the set of one-to-one mappings $\pi : M \to M'$ for two finite sets $M$ and $M'$ with the same number of elements.

## 2. The conditional likelihood

The conditional likelihood $L(Y|X)$ of a point process $Y$ observed on a bounded set $A \subseteq \mathbb{R}^d$, given another point process $X$, is the Radon-Nikodym derivative of the probability measure of $Y$ given $X$ with respect to a reference measure. We will let the reference measure be the probability measure for a Poisson process with a constant intensity $\lambda_0(y) = 1$ for $y \in A$. In the following theorem we assume for simplicity that the functions $g$ and $k$ are continuous, although this condition can be weakened.

Theorem 1. *Let $X$ and $Y$ be two finite point processes specified as in § 1·2 on a bounded set $A$. Suppose that $g(y|X)$ and $k(y|X_i)$, $i \in M$, are continuous functions of $y \in A$. Then the conditional likelihood of $Y$ given $X$ is*

$$L(Y|X) = \exp\left(|A|_d - \int_A g(y|X)\,dy\right) \sum_{\substack{M_1 \subseteq M \\ N_1 \subseteq N \\ |M_1| = |N_1|}} \sum_{\pi \in \mathcal{P}(M_1, N_1)} L_1 L_2 L_3, \tag{1}$$

*where*

$$L_1 = \prod_{i \in M_1} p(X_i) k(Y_{\pi(i)}|X_i),$$

$$L_2 = \prod_{i \in M \setminus M_1} \left(p(X_i) \int_{A^c} k(y|X_i)\,dy + 1 - p(X_i)\right),$$

$$L_3 = \prod_{j \in N \setminus N_1} g(Y_j|X),$$

*and the reference measure corresponds to the Poisson process on $A$ with intensity 1.*

*Proof*: See Appendix.

## 3. Submodels of the general model

### 3·1 *General considerations*

In most applications the model described in § 1·2 is too general to be really useful and we will describe some special cases of it. Firstly, the general model allows the thinning probability $1-p(x)$ to depend on the position $x$. Often it is reasonable to assume that the probability of thinning is constant on the set $A$, $p(x)=p$, and we will do so in the sequel.

Secondly, the displacement distribution with density $k(\,\cdot\,|X_i)$ is frequently chosen to be centred around $X_i+\mu$, where we may interpret $\mu$ as a systematic error made in the observations. A simple choice of the displacement distribution is a $d$-dimensional normal distribution.

### 3·2 *The pure displacement model*

Assume that $p=1$ and that $g(\,\cdot\,|X)=0$. This means that we do not have any extra ghost points, and that we observe the randomly displaced $X$-points, provided they are not censored.

### 3·3 *Homogeneous superposition noise*

An important special case for the intensity of the ghost points is the homogeneous noise, $g(\,\cdot\,|X)=\lambda$. Then the likelihood function (1) simplifies to

$$L(Y|X) = \sum_{\substack{M_1\subseteq M\\ N_1\subseteq N\\ |M_1|=|N_1|}} \sum_{\pi\in\mathcal{P}(M_1,N_1)} T(M_1,N_1,\pi) \tag{2}$$

with terms

$$\begin{aligned} T(M_1,N_1,\pi) &= \\ p^{|M_1|}\,\lambda^{|N\setminus N_1|}\ \exp\left((1-\lambda)|A|_d\right) &\left\{\prod_{i\in M_1} k(Y_{\pi(i)}|X_i)\right\} \prod_{i\in M\setminus M_1}\left(p\int_{A^c} k(y|X_i)\,dy+1-p\right). \end{aligned} \tag{3}$$

### 3·4 *A cluster Cox process*

When the intensity $g$ for the ghost points depend on the $X$-process, these points contain information on $X$. A specific example is when all the original $X$-points are thinned, $p = 0$, so we only observe the ghost points. Assume further that each $X_i$ gives rise to $N_i$ off-spring points, with $N_i$ Poisson distributed with mean $\nu(X_i)$. Assume also that, conditional upon $X$, all the $N_i$'s are independent, all off-spring points are independent, and the points generated by $X_i$ are distributed according to a probability density $h(\,\cdot\,|X_i)$. Conditional upon $X$ the $Y$-process is an inhomogeneous Poisson process with intensity $g(y|X) = \nu(X_1)h(y|X_1) + \ldots + \nu(X_m)h(y|X_m)$ and the $X_i$'s will be cluster centres. Conditional further on $|Y| = n$, the distribution of the $Y$-points may be regarded as a mixture, and estimation of the cluster centres in a mixture distribution is, for instance, considered in Titterington et al. (1985). Considering $X$ as random, we note that the $Y$-process is a Cox point process. In our forestry example a cluster may be generated for each tree as large branches may under certain conditions produce reflections similar to tree tops.

## 4. Approximate likelihood analysis

### 4·1 *Overview*

The likelihood function (2) is in theory simple to compute, but for even modest–sized datasets the number of terms in the sum becomes enormous and it is impossible to calculate all the terms within a reasonable time. Furthermore, the orders of magnitude of terms in the sum are very different, and we have to be careful with numerical computations where we add a large number of small terms.

We will now describe algorithms for finding approximations of the likelihood function by focussing on the values of a few large terms. In particular, we will consider algorithms for the choice of $r$ terms in (2) for a fixed small $r$. In the example in § 5 the likelihood function contains more than $10^{50}$ terms, whereas we find a reasonable approximation of the likelihood function with $r = 8$ terms.

For notational simplicity we will in the sequel consider point processes in two-dimensional space, although the methods can be generalized to $d$ dimensions in a fairly straightforward manner. We will also assume that the density function $k(\,\cdot\,|X_i)$ for point displacement from $X_i$, cf. (3), is a two-dimensional normal density with mean vector $X_i + \mu$, with a systematic displacement $\mu = (\mu_1, \mu_2)$, and with variances and correlation, $\sigma_1^2$, $\sigma_2^2$ and $\rho$, that do not depend on $X_i$. Further, we assume the homogeneous noise model from § 3·3.

Let us call a combination $s = (M_1, N_1, \pi)$, that we sum over in (2), a state. The crucial issue in our approximate likelihood computation is to find states $(M_1, N_1, \pi)$ such that the corresponding terms $T(M_1, N_1, \pi)$ give large contributions to (2). This we will try to achieve by use of a deterministic, iterative algorithm which consists of a starting procedure for finding an initial set of states and local maximizations over suitably chosen neighbourhoods of states until no further improvement is obtained. In the following sections we will study:

(i) two approximations of the integral over $A^c$ in (3) with emphasis on points close to the boundary of $A$,

(ii) neighbours of a state $(M_1, N_1, \pi)$ to be considered in the search for terms that give large contributions to (2),

(iii) an iterative deterministic procedure for finding an optimal set of $r$ states for the likelihood approximation and simultaneously finding approximate maximum likelihood estimates of parameters, and

(iv) a starting procedure to find an initial set of $r$ candidate states for the iterative procedure.

### 4·2 *Two approximations for points close to boundaries*

For most points $X_i$ it turns out that they are so far from the boundary of the observation area $A$ that we safely can use the simplifying approximation

$$\int_{A^c} k(y|X_i)\,dy = 0. \tag{4}$$

In the first approximation of (3) we assume that all $X$-points are so far from the boundary that we can use the approximation (4), and thus replace (3) by

$$T(M_1, N_1, \pi) = p^{|M_1|}\ (1-p)^{|M \setminus M_1|}\ \lambda^{|N \setminus N_1|}\ \exp\left((1-\lambda)|A|_d\right)\ \left\{\prod_{i \in M_1} k(Y_{\pi(i)}|X_i)\right\}. \tag{5}$$

For $s = (M_1, N_1, \pi)$ we note that (5) is maximized as a function of the parameter vector

$$\theta = (p, \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \tag{6}$$

by $\hat{\theta}(s) = (\hat{p}, \hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\rho})$, where $\hat{p} = |M_1| \,/\, |M|$, $\hat{\lambda} = |N \backslash N_1|/|A|_d$, and $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\rho})$ are the standard maximum likelihood estimates of the parameters in a two–dimensional normal distributions based on the sample $(Y_{\pi(i)} - X_i, i \in M_1)$.

In the second approximation of (3) we assume that the boundary, viewed locally from $X_i$, can be approximated by a straight line. More precisely, let $d_i$ denote the distance from $X_i$ to the boundary, and introduce local coordinates $(V_{i1}, V_{i2})$ in the direction of the nearest point on the boundary and orthogonal to it. We may use the approximation

$$\int_{A^c} k(y|X_i)\, dy = P_\theta(V_{i1} > d_i), \tag{7}$$

which can be expressed in terms of the last five parameters of the parameter vector $\theta$ in (6), the standard one-dimensional normal distribution function and $d_i$. For most points $X_i$ the right hand side of (7) will still be effectively zero, but for points close to the boundary it gives a better approximation than (4). With the approximation (7) we do not get closed form expressions for the parameter vector that maximizes (3) as we got above with the approximation (4).

### 4·3 *Neighbours of a state*

For a state $(M_1, N_1, \pi)$ we will say that $(M_1', N_1', \pi')$ is a neighbour state if it is obtained by adding a pair of $X$- and $Y$-points, by removal of a pair, by 'swapping' one of the components in a pair with an unpaired point, or by exchanging correspondences for two pairs of states — more precisely, if $(M_1', N_1', \pi')$ is obtained from $(M_1, N_1, \pi)$ in one of the following five ways:

(a) Addition of a pair: $M_1' = M_1 \cup \{i'\}$, where $i' \in M \setminus M_1$, $N_1' = N_1 \cup \{j'\}$, where $j' \in N \setminus N_1$, $\pi'(i) = \pi(i), i \in M_1$, and $\pi'(i') = j'$. The number of such neighbours are $|M \setminus M_1|\,|N \setminus N_1|$.

(b) Removal of a pair: $M_1' = M_1 \setminus \{i'\}$, where $i' \in M_1$, $N_1' = N_1 \setminus \{j'\}$, where $j' \in N_1$, $\pi'(i) = \pi(i), i \in M_1'$, and $\pi(i') = j'$. This can be done in $|M_1| = |N_1|$ ways.

(c) Swapping an $X$-point: $M_1' = (M_1 \setminus \{i'\}) \cup \{i''\}$, where $i' \in M_1$ and $i'' \in M \setminus M_1$, $N_1' = N_1$, $\pi'(i) = \pi(i), i \in M_1 \setminus \{i'\}$, and $\pi'(i'') = \pi(i')$. There are $|M_1|\,|M \setminus M_1|$ such neighbours.

(d) Swapping a $Y$-point: $M_1' = M_1$, $N_1' = (N_1 \setminus \{j'\}) \cup \{j''\}$, where $j' \in N_1$ and $j'' \in N \setminus N_1$, $\pi'(i) = \pi(i), i \in M_1 \setminus \{i'\}$, where $\pi(i') = j'$, and $\pi'(i') = j''$. Swapping a $Y$-point can be done in $|N_1|\,|N \setminus N_1|$ ways.

(e) Exchange among two pairs: $M_1' = M_1$, $N_1' = N_1$, $\pi'(i) = \pi(i), i \in M_1 \setminus \{i', i''\}$, where $i' \in M_1$ and $i'' \in M_1$, $i' \neq i''$, $\pi'(i') = \pi(i'')$, and $\pi'(i'') = \pi(i')$. The number of such neighbours are $|M_1|\,(|M_1| - 1)/2$.

We note that being a neighbour is a reciprocal relation.

In our computations in the example below we will consider a reduced set of neighbours of a state $s = (M_1, N_1, \pi)$ which is obtained as follows:

(a) Addition of a pair: for an added pair with $X_{i'}$ as $X$-point we only consider that unpaired $Y$-point which is closest to $X_{i'}$; there are $|M \setminus M_1|$ such neighbours, or zero if all $Y$-points are already paired.

(b) Removal of a pair: no restriction; there are $|M_1|$ such neighbours.

(c) Swapping an $X$-point: when swapping an $X$-point $X_{i'}$ in a pair we consider only replacing it with the unpaired $X$-point which is closest to the corresponding $Y$-point; there are $|M_1|$ such neighbours, or zero if all $X$ points are already paired.

(d) Swapping a $Y$-point: when swapping a $Y$-point $Y_{j'}$ in a pair we consider only replacing it with the unpaired $Y$-point which is closest to the corresponding $X$-point; there are $|N_1|$ such neighbours, or zero if all $Y$-points are already paired.

(e) Exchange among two pairs: for a pair with $X$-point $X_{i'}$ we consider only the swap involving the pair with $X$-point closest to $X_{i'}$; there are at most $|M_1|$ such neighbours.

Note that the total number of elements in the reduced set of neighbours of a state is at most $|M| + 3|M_1|$.

### 4·4 *An iterative likelihood maximization procedure*

Let $T(s, \theta)$ denote the term (5) with $s = (M_1, N_1, \pi)$ and $\theta$ given by (6). Consider a set $S$ of states and the truncated likelihood

$$L(S, \theta) = \sum_{s \in S} T(s, \theta). \tag{8}$$

The iterative procedure that we suggest for finding a good approximation (8) of (2) with a small fixed number $r$ of states in $S$ runs as follows:

(i) Find an initial candidate set $S_0$ with $r$ states and a corresponding maximizing parameter vector $\theta_0$ as described below in § 4·5.

(ii) Let $S_t$ be our candidate set of states at stage $t$, $t = 0, 1, \ldots$, with a corresponding maximizing parameter vector $\theta_t$. To find the updated set $S_{t+1}$ we proceed as follows: Consider all states which are either contained in $S_t$ or are members of the reduced neighbour sets (as described in the previous section ) of any state in $S_t$. This augmented set of states is denoted $S'_t$. Let $S_{t+1}$ consist of those $r$ states $s$ in $S'_t$ that have the largest values of $T(s, \theta_t)$. Let $\theta_{t+1}$ be the $\theta$-value that maximizes $L(S_{t+1}, \theta)$. We find $\theta_{t+1}$ by a quasi-Newton method with $\theta_t$ as starting value.

(iii) If $S_{t+1} = S_t$ or, considering finite numerical accuracy, if $L(S_{t+1}, \theta_{t+1}) \leq L(S_t, \theta_t) + \epsilon$ for a given small $\epsilon > 0$, cf. (10) below, we stop and choose

$$L(S_{t+1}, \theta) = \sum_{s \in S_{t+1}} T(s, \theta) \tag{9}$$

as our likelihood approximation, and $\theta_{t+1}$ as our approximate maximum likelihood parameter estimate. Otherwise the previous step is iterated.

From the description (ii) above of the basic iteration step it follows that $L(S_{t+1}, \theta_t) \geq L(S_t, \theta_t)$ and $L(S_{t+1}, \theta_{t+1}) \geq L(S_{t+1}, \theta_t)$. This implies

$$L(S_{t+1}, \theta_{t+1}) \geq L(S_t, \theta_t), \tag{10}$$

that is, the criterion function at stage $t + 1$ is at least as good as the criterion function at stage $t$.

### 4·5 *Starting procedure*

To start our procedure for finding a small number of large terms we choose an initial set of $r$ states by use of a 'greedy' algorithm similar to the one used in Dralle & Rudemo (1997). It can be described as follows:

> For each $X$-point we let a circle grow. Circles grow simultaneously at the same speed from all $X$-points. When a circle hits a $Y$-point that has not been hit

before by any other growing circle it is paired with the corresponding $X$-point, and the growth of this circle is stopped. After the $k$th hit we assemble the $k$ pairs thus obtained in a state $s_k = (M_{1k}, N_{1k}, \pi_k)$, for each $k = 0, \ldots, k_1$, where $k_1 = \min(m, n)$. Assume first that $k_1 \geq r$. As initial set $S_0$ of states we choose those $r$ states $s_k$ that have the largest values of $\sup_\theta T(s_k, \theta)$, $k = 0, \ldots, k_1$. Let $\theta_0$ be the $\theta$-value that maximizes $L(S_0, \theta)$. We find $\theta_0$ by a quasi-Newton method with $\hat{\theta}(s^*)$ as starting value, where $\hat{\theta}(s)$ is defined immediately after (6), and $s^*$ is the state among $s_k$, $k = 0, \ldots, k_1$, that maximizes $\sup_\theta T(s_k, \theta)$.

In case $r > k_1$ we adjust the algorithm in a natural way. We start with the available $k_1$ states from the starting procedure, and then expand the number of states in each step of the algorithm in § 4·4, until we have $r$ states among the neighbouring states.

The starting procedure finds pairs of $X$- and $Y$-points close together. From an intuitive point of view this is reasonable — provided that the displacements typically are small compared to the distances between neighbouring $X$-points, and that the intensity of the ghost points is not so high that they are likely to enter between the true and the displaced points.

## 5. Example: Tree top positions from an aerial photograph

### 5·1 *Data: 'true' tree top points and points from template matching*

Digital analysis of high resolution aerial photographs has recently been shown to be a promising method for acquiring detailed information on individual trees in forest stands. Panchromatic images obtained from a flight 560 m above a thinning experiment in Norway spruce (*Picea abies* (L.) Karst.) have been studied in Dralle & Rudemo (1996, 1997) by use of kernel smoothing, which proved to be a useful method for images obtained close to the nadir. For off-nadir images a template matching method turned out to be more effective, see Larsen & Rudemo (1998), where optimal templates were obtained for different geometries of image acquisition. The upper part of Figure 1 shows an image with sidelighted trees with a ground-projected pixel size of $0.15 \times 0.15\,\mathrm{m}^2$, and this image is the source of the $Y$-points analyzed in the present paper.

FIGURE 1 ABOUT HERE

The 'true' $X$-points were obtained in the following way. For all trees in the subplot delineated in the upper part of Figure 1 tree base positions measured in the field were extrapolated to estimated tree top height and superimposed on the image to yield an initial estimate of the true tree top positions as described in Dralle & Rudemo (1997). These positions were adjusted by manual inspection of the images to correct for errors introduced by deviations in the tree height estimates, variations due to wind and inaccuracy in the image rectification. The resulting tree top positions, $X_1, \ldots, X_m$, $m = 171$, coincide with the centres of the circles shown in the lower part of Figure 1.

FIGURE 2 ABOUT HERE

The $Y$-points were obtained by matching a template constructed from a single light-reflection model (indicated in the left part of Figure 2) adapted to the positions of the camera and light sources: both the sun as a point source and a set of minor sources distributed over the hemisphere and representing diffuse light. The resulting template, shown in the right part of Figure 2 bounded by an ellipse, was moved pixelwise over the image. Local maxima of the correlation between template and image pixel grey levels were considered as candidate positions of tree tops. The optimal translation along the tree axis, size and eccentricity of the bounding ellipse was obtained in Larsen & Rudemo (1998) and the resulting optimal boundary is shown in the right part of Figure 2. The (ground projected) size of the half axes of the optimal ellipse are 1.58 m and 1.42 m. Altogether there were 570 positive local correlation maxima. The histogram of the correlation values of these maxima is shown in Figure 3. The histogram indicates a bimodal distribution where, roughly, large maxima correspond to tree tops and small maxima correspond to superimposed noise. Three sets of $Y$-values, $Y_1, \ldots, Y_n$, were studied corresponding to the correlation maxima above or equal to varying limits $r_0$. These limits were chosen to obtain approximately $n = m$, $n = 1.2m$ and $n = 1.4m$. Using the limits $r_0 = 148/255$, $r_0 = 127/255$ and $r_0 = 110/255$ we obtained $n = 171$, $n = 206$ and $n = 243$, respectively. The $Y$-points in the data set with $n = 206$ points are shown as black dots in the lower part of Figure 1.

FIGURE 3 ABOUT HERE

### 5·2 *Results from approximate likelihood analyses*

We will use the above described set of $X$-points and the set with 206 $Y$-points as the basic datasets in this section. The points are shown in the lower part of Figure 1 within the polygonal border of the set $A$. To describe the data we consider the model in § 3·3 with the approximation (4) implying that each term in the likelihood function has the form (5). Thus we assume that the thinning probabilities and the intensity of the Poisson noise are homogeneous on the set $A$, that the displacement distribution $k$ is a two-dimensional normal distribution, and that the probability of censoring of a displaced point is negligible.

FIGURE 4 ABOUT HERE

Figure 4 shows the value of the term $\log\sup_\theta T(s_k,\theta)$ as a function of the number $k$ of pairs in the state $s_k$ for the starting procedure in § 4·5. It is seen that the size of the terms considered in the starting procedure can vary a lot. The sudden decrease for large $k$ appears when the algorithm starts to pair $X$- and $Y$-points far apart. The subsequent iterative algorithm (§ 4·4) tends to select terms among the neighbours of the largest and second largest terms of the starting procedure, and exhibits after that fast convergence, typically in two or three steps.

FIGURE 5 ABOUT HERE

Our algorithm does not prescribe how to choose the number $r$ of states in $S$ used in the approximation of the likelihood function. We have run the algorithm with $r = 1,\ldots,16$ terms and Figure 5 shows the value of the approximation of the likelihood function as a function of the number of terms in the approximation. The improvement of adding a term is largest when few terms are already included. We have chosen $r = 8$ terms in the approximation in the following computations as this seems (from Figure 5) to give an adequate approximation.

TABLE 1 ABOUT HERE

Table 1 shows the value of $T(s^{(i)},\hat{\theta})/T(s^{(1)},\hat{\theta})$, $i = 1,\ldots,8$, of the terms evaluated in the likelihood approximation and the number of pairs in each of the corresponding states. The terms are sorted in order of decreasing contribution to the likelihood function. The differences between the states of the eight terms are small. Compared to the state of the leading term the other seven terms are obtained by swapping a $Y$-point, removal of a pair, addition of a pair, or a combination of two such operations.

Table 2 shows the estimated parameters based on the eight terms as well as their estimated standard deviations found by numerical differentiation of the approximate likelihood function. The estimate $\hat{p} = 0.941$ means that approximately $0.941 \cdot 171 = 160.9$ out of the 171 $X$-points are matched.

TABLE 2 ABOUT HERE

The size of the estimated standard deviations seem quite reasonable. We would, for instance, expect the standard deviation of the estimate of $\mu_1$ to be around $\{1.004/(0.941 \cdot 171)\}^{\frac{1}{2}} = 0.079$ if the estimates for the displacement parameters were obtained in a model with independent identically distributed normal observations.

As we have an approximation of the likelihood function we are also able to perform approximate likelihood ratio tests. Testing, for instance, the hypothesis $(\mu_1, \mu_2) = (0, 0)$ we find the test variable $-2 \log Q = 17.85$ with an approximate $p$-value of 0.00013 from a $\chi^2$-distribution with two degrees of freedom, and the hypothesis is rejected. Further testing shows that we can accept $\mu_2 = 0$ ($p$-value 0.47) but not $\mu_1 = 0$ in agreement with Table 2.

TABLE 3 ABOUT HERE

Table 3 shows parameter estimates obtained when the number of $Y$-points are varied as described in the end of § 5·1. The change in $\hat{\lambda}$ is not so interesting *per se* as it is essentially determined by the surplus number of $Y$-points. Note that the $\hat{p}$ estimate increases with an increasing number of $Y$-points, and that this increase is largest when the number of $Y$-points increases from 171 to 206. The increase in $\hat{p}$ is accompanied with an increase in the variance of the displacement, but generally the parameters of the displacement distribution seem fairly stable when the number of $Y$-points varies.

## 6. Discussion

### 6·1 *Spatial point processes with noise*

Models for random processes with noise is an important area of statistics which mainly has been focussed on additive models, which are particularly tractable for Gaussian processes. Models for point processes in one dimension have also been thoroughly treated, for instance by martingale methods. However, incompletely observed spatial point processes, which are the main topic of this paper, have received less interest. Point processes from images, as in the example of the present paper, are potentially important sources of spatial point processes observed with noise, cf. Young et al. (1998) for template matching of cells in digital microscopy, and van Lieshout (1994) for a Bayesian analysis of a scene with a random set of objects generated by a point process.

### 6·2 *Computational considerations*

Only deterministic parameter search methods have been discussed in the present paper. Random search methods are interesting alternatives which are particularly useful if our estimation problem is considered as a missing data problem where the matching between original and displaced points is unobserved. Possible algorithms are e.g. the stochastic EM algorithm, Diebolt & Celeux (1993), MCMC approximations of the likelihood function, Geyer (1996), and a Bayesian approach with missing data, Smith & Roberts (1993). By regarding the matchings as missing data in a simulation type algorithm the evaluation of the large sum in the likelihood (1) is replaced by sampling from a (large) number of states in a simulation.

Related problems with matching of point sets arise in image analysis of multiple images from different perspectives. An example is given by Cross & Hancock (1998) where an EM algorithm is used to match graphs.

If the displacement density $k$ has bounded support the approximation $\int_{A^c} k(y|X_i)\,dy = 0$ used in the example in this paper holds exactly for points $X_i$ such that the corresponding support lies within $A$, and it is a good approximation when $A$ is large compared to, say, the 95% contour curves for the displacement distribution.

The initial step used in the present paper seems to function well, and our conclusion from some experimentation is that the precise implementation of the initial step does not appear to be crucial.

In general the suggested model and estimation algorithm work best when (i) the random displacements are relatively small compared to the distances between the points in the original point process $X$ and (ii) the thinning probabilities and the intensity of the superimposed Poisson process is relatively small. If these conditions are not satisfied the model might still hold, but one could expect that we would then need a large number of terms in the likelihood approximation and that many iterations would be necessary in the likelihood maximization, in contrast to our example. The risk of finding a local optimum instead of a global optimum would also be increased. Stochastic search algorithms would then be a natural alternative. Our example is, however, realistic for high quality aerial photographs of forests, and the view analysed with sidelighted trees is the most difficult of the three views studied in Larsen & Rudemo (1998).

### 6·3 *Extensions*

In this paper we have considered the case where both the $X$- and the $Y$-processes are observed, which enables us to estimate parameters of the disturbance mechanisms. When we have such estimates it is possible to consider estimation of the $X$-process based on observation of the $Y$-process alone. A natural procedure is to use a Bayesian approach based on a prior point process model for the $X$-process.

## Acknowledgements

The research reported in this paper was supported by Dina, Danish Informatics Network in the Agricultural Sciences and the Gothenburg Stochastic Centre. We are further indebted to Morten Larsen, Geoff Nicholls, Antti Penttinen, Helle Sørensen, and Elke Thönnes for valuable comments.

## Appendix

### *Proof of Theorem 1*

The likelihood function $L(Y|X)$ is the Radon-Nikodym derivative of the measure described in § 1·2 with respect to the Poisson process on $A$ with intensity 1. We find the density by considering the point process $Y$ on a sequence of partitions.

A partition $\mathcal{B}$ of $A$ is a set of disjoint subsets of $A$ that have $A$ as their union, and the norm of $\mathcal{B}$ is defined as the maximal mesh-size, i.e.

$$\|\mathcal{B}\| = \max_{B\in\mathcal{B}} \sup_{x,y\in B} \|x-y\|,$$

where $\|x-y\|$ is the Euclidean distance between $x$ and $y$. For a partition $\mathcal{B}$ we let $Y^{\mathcal{B}} = (Y(B), B \in \mathcal{B})$ denote the restriction of $Y$ to $\mathcal{B}$. Let us consider a sequence of partitions $\mathcal{B}_n, n \geq 1$, of $A$. Assume that every $B \in \mathcal{B}_n$ can be written as a union of sets in $\mathcal{B}_{n+1}$, that $\|\mathcal{B}_n\| \to 0$ and that $|B|_d > 0$ for all $B \in \mathcal{B}_n$.

Let $P$ be the distribution of $Y$ described in § 1·2 and let $P_0$ correspond to the Poisson process on $A$ with intensity 1. Further, let $P^n$ and $P_0^n$ denote the distribution of $Y^{\mathcal{B}_n}$ under $P$ and $P_0$. Hoffmann-Jørgensen (1994, Sec. 11.11) now assures that if $P^n \ll P_0^n$, then $P \ll P_0$ and the Radon-Nikodym derivative $L_n = \frac{dP^n}{dP_0^n}$ converges $P_0$-a.s. to the Radon-Nikodym derivative $L = \frac{dP}{dP_0}$. (For this to hold, we need $\sup_n L_n < \infty$ a.s., but this condition is satisfied if we have convergence towards a limit $L < \infty$ a.s.)

In our case $P^n$ and $P_0^n$ are distributions of a finite dimensional vector of integer-valued random variables, and it follows that $P^n \ll P_0^n$. The Radon-Nikodym derivative is obtained by dividing the two point probabilities with each other. We proceed to determine $L_n$.

Let $y$ be a possible realization of $Y$ (without multiple points). Then

$$P_0^n(Y^{\mathcal{B}_n} = y^{\mathcal{B}_n}) = \prod_{B \in \mathcal{B}_n} \exp(-|B|_d) \frac{|B|_d^{y(B)}}{y(B)!} = \exp(-|A|_d) \prod_{B \in \mathcal{B}_n} \frac{|B|_d^{y(B)}}{y(B)!}.$$

Let us now turn to $P^n(Y^{\mathcal{B}_n} = y^{\mathcal{B}_n}|X)$. To evaluate this probability we sum over the finitely many ways in which the configuration $y = \{y_j, j \in N\}$ can arise. Let $M_1$ and $N_1$ denote subsets of $M$ and $N$ with $|M_1| = |N_1|$ and let $\pi \in \mathcal{P}(M_1, N_1)$. The set $M_1$ corresponds to $X$-points that are neither thinned nor censored but displaced to a position within $A$. The displaced $X$-points correspond to $y_j$, $j \in N_1$, via the transformation $\pi$ such that $j = \pi(i)$. However, when we consider the distribution of the discrete variable $Y^{\mathcal{B}_n}$, rather than the full information $Y$, we should only consider into which sets in $\mathcal{B}_n$ the displaced points fall. For $z \in A$ we further let $B(z)$ be the unique set $B \in \mathcal{B}_n$ such that $z \in B$.

The probability of a point $Z_i$ generated from the distribution $k(\,\cdot\,|X_i)$, $i \in M \setminus M_1$, not being observed (considering also thinning) is

$$p(X_i) \int_{A^c} k(z|X_i)\, dz + 1 - p(X_i).$$

This is seen from calculations like $P(\text{not observed}) = P(\text{not thinned}) P(\text{censored}|\text{not thinned}) + P(\text{thinned})$. Similarly $P(Z_i \in B) = P(\text{not thinned}) P(Z_i \in B|\text{not thinned})$ for $i \in M_1$, so that

$$P(Z_i \in B) = p(X_i) \int_B k(z|X_i)\, dz.$$

Let us consider $n$-values so large that $y(B) \in \{0,1\}$ for $B \in \mathcal{B}_n$. Then we find

$$P^n(Y^{\mathcal{B}_n} = y^{\mathcal{B}_n}|X) = \sum_{\substack{M_1 \subseteq M \\ N_1 \subseteq N \\ |M_1|=|N_1|}} \sum_{\pi \in \mathcal{P}(M_1,N_1)} F_1 F_2 F_3,$$

where the three factors $F_1$, $F_2$, and $F_3$, correspond to displaced points, thinned or censored points, and ghost points:

$$\begin{aligned} F_1 &= \prod_{i \in M_1} p(X_i) \int_{B(y_{\pi(i)})} k(z|X_i)\, dz, \\ F_2 &= \prod_{i \in M \setminus M_1} \left( p(X_i) \int_{A^c} k(z|X_i)\, dz + 1 - p(X_i) \right), \\ F_3 &= \prod_{B \in \mathcal{B}_n} \exp\left( - \int_B g(z|X)\, dz \right) \left\{ \int_B g(z|X)\, dz \right\}^{y(B) - |\{i \in M_1 : y_{\pi(i)} \in B\}|} \\ &= \exp\left( - \int_A g(z|X)\, dz \right) \left\{ \prod_{j \in N \setminus N_1} \int_{B(y_j)} g(z|X)\, dz \right\}. \end{aligned}$$

We now get the Radon-Nikodym derivative $L_n(y^{\mathcal{B}_n}|X) = \frac{dP^n}{dP_0^n}(y^{\mathcal{B}_n})$ as

$$L_n(y^{\mathcal{B}_n}|X) = \frac{P^n(y^{\mathcal{B}_n}|X)}{P_0(y^{\mathcal{B}_n})} = \exp\left( - \left( \int_A g(z|X)\, dz - |A|_d \right) \right) \sum_{\substack{M_1 \subseteq M \\ N_1 \subseteq N \\ |M_1|=|N_1|}} \sum_{\pi \in \mathcal{P}(M_1,N_1)} \tilde{F}_1 F_2 \tilde{F}_3$$

with $F_2$ as above,

$$\tilde{F}_1 = \prod_{i \in M_1} p(X_i) \frac{\int_{B(y_{\pi(i)})} k(z|X_i)\, dz}{|B(y_{\pi(i)})|_d} \quad \text{and} \quad \tilde{F}_3 = \prod_{j \in N \setminus N_1} \frac{\int_{B(y_j)} g(z|X)\, dz}{|B(y_j)|_d}.$$

Now, letting $n \to \infty$, $\|\mathcal{B}_n\| \to 0$ and replacing $y$ with $Y$ we find that the three factors $\tilde{F}_1$, $F_2$, and $\tilde{F}_3$ converge towards $L_1$, $L_2$, and $L_3$, respectively. □

## References

Cross, A. D. J. & Hancock, E. R. (1998). Graph matching with a dual step EM algorithm. *IEEE Trans. Pat. Anal. Mach. Intel.* **20**, 1236–1253.

Daley, D. J. & Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* New York: Springer-Verlag.

Diebolt, J. & Celeux, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixture proportion. *Stochastic Models* **9**, 599–613.

Dralle, K. & Rudemo, M. (1996). Stem number estimation by kernel smoothing of aerial photos. *Canad. J. Forest Res.* **26**, 1228–1236.

Dralle, K. & Rudemo, M. (1997). Automatic estimation of individual tree positions from aerial photos. *Canad. J. Forest Res.* **27**, 1728–1736.

Geyer, C. J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice,* Ed. W. R. Gilks, S. Richardsson and D. J. Spiegelhalter, pp. 241–258. London: Chapman & Hall.

Hoffmann-Jørgensen, J. (1994). *Probability with a View toward Statistics*, volume 2. London: Chapman & Hall.

Larsen, M. & Rudemo, M. (1998). Optimizing templates for finding trees in aerial photographs. *Pattern Recognition Lett.* **19**, 1153–1162.

McLachlan, G. J. & Krishnan, T. (1997). *The* EM *Algorithm and Extensions.* New York: Wiley.

Penttinen, A., Stoyan, D. & Henttonen, H. M. (1992). Marked point processes in forest statistics. *Forest Science* **38**, 806–824.

Smith, A. F. M. & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* **B 55**, 3–23.

Stoyan, D., Kendall, W. S. & Mecke, J. (1995). *Stochastic Geometry and its Applications,* 2nd ed. New York: Wiley.

Titterington, D., Smith, A. & Markov, U. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

van Lieshout, M. N. M. (1994). Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Adv. Appl. Prob.* **26**, 281–300.

Young, D., Glasbey, C. A., Gray, A. J. & Martin, N. J. (1998). Towards automatic cell identification in DIC microscopy. *J. Microscopy* **192**, 186–193.

## Tables and figures

Table 1: *The value of* $T(s^{(i)},\hat{\theta})/T(s^{(1)},\hat{\theta})$, $i = 1, \ldots, 8$, *for the terms evaluated at the approximate maximum likelihood estimate* $\hat{\theta}$ *and the number of pairs in each of the corresponding states*

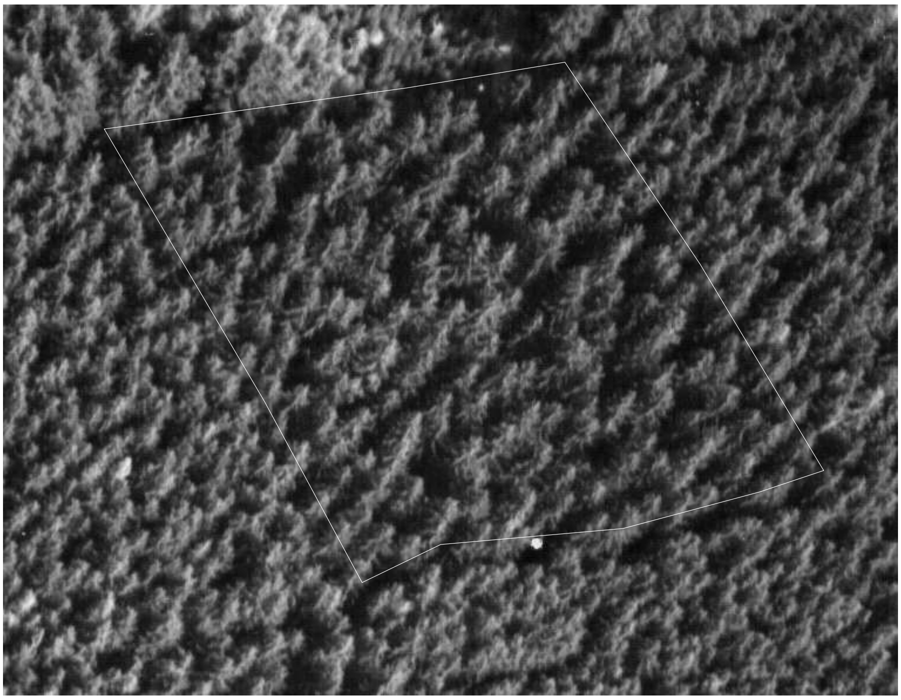
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $T(s^{(i)},\hat{\theta})/T(s^{(1)},\hat{\theta})$ | 1 | 0.319 | 0.141 | 0.045 | 0.027 | 0.017 | 0.008 | 0.008 |
| # pairs | 161 | 161 | 160 | 160 | 162 | 160 | 162 | 160 |

Table 2: *Parameter estimates and estimated standard deviations based on eight terms in the likelihood sum*

| parameter | $p$ | $\lambda$ | $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $\text{cov}_{1,2}$ |
|---|---|---|---|---|---|---|---|
| estimate | 0.941 | 0.00028 | -0.342 | 0.082 | 1.004 | 2.028 | -0.049 |
| std. dev. | 0.018 | 0.00004 | 0.080 | 0.113 | 0.115 | 0.235 | 0.115 |

Table 3: *Parameter estimates based on eight terms in the likelihood sum for three different sets* $Y_1, \ldots, Y_n$ *with* $n$*=171, 206 and 243, respectively*

| parameter | $p$ | $\lambda$ | $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $\text{cov}_{1,2}$ |
|---|---|---|---|---|---|---|---|
| $n$=171 | 0.894 | 0.00011 | -0.330 | 0.059 | 0.993 | 1.864 | -0.092 |
| $n$=206 | 0.941 | 0.00028 | -0.342 | 0.082 | 1.004 | 2.028 | -0.049 |
| $n$=243 | 0.952 | 0.00050 | -0.335 | 0.047 | 1.001 | 2.160 | -0.035 |



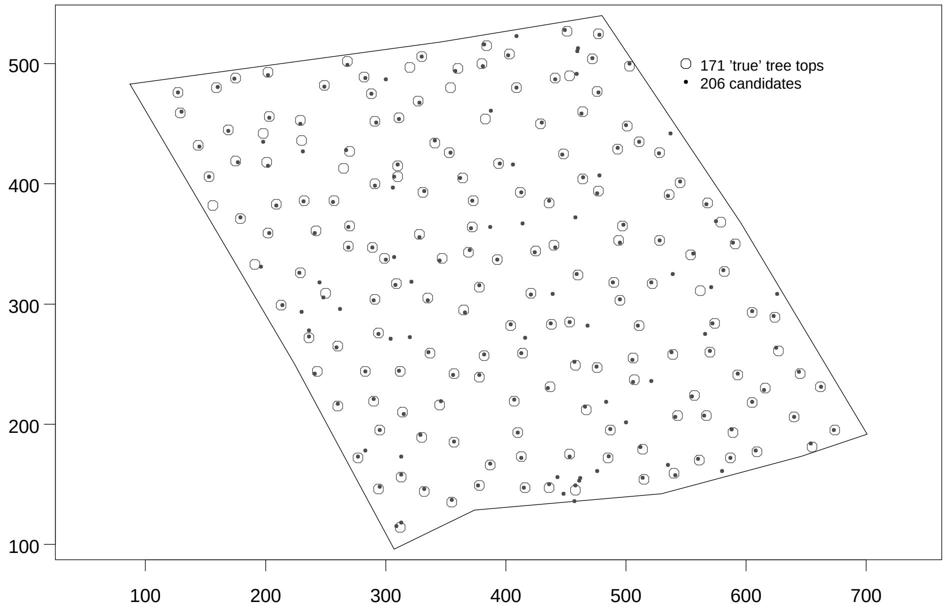


Figure 1: Image with sidelighted trees, and, in the lower part, 171 $X$-points (centres of circles) corresponding to 'true' tree tops and 206 $Y$-points (dots) corresponding to template matching. The area of the delineated subplot is 4 454 m$^2$, and the unit of the axes in the lower part is linear pixel size, 0.15 m.

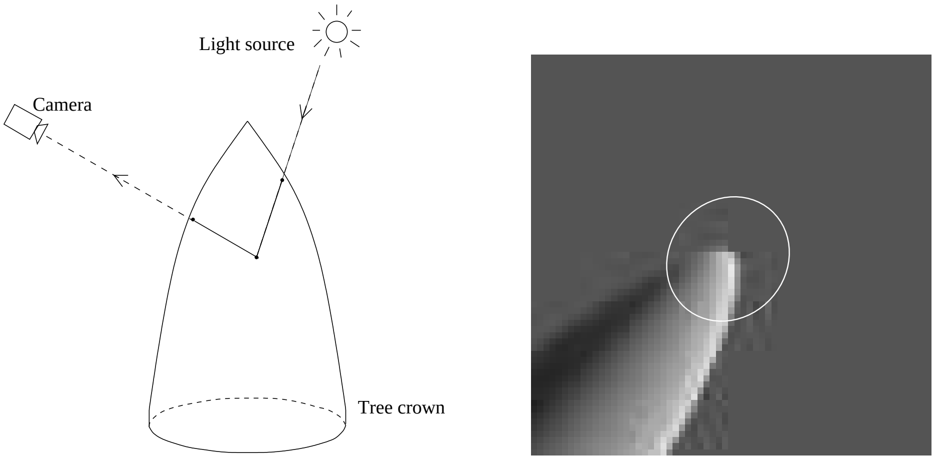


Figure 2: Model tree and, in the right part, template with optimal bounding ellipse.

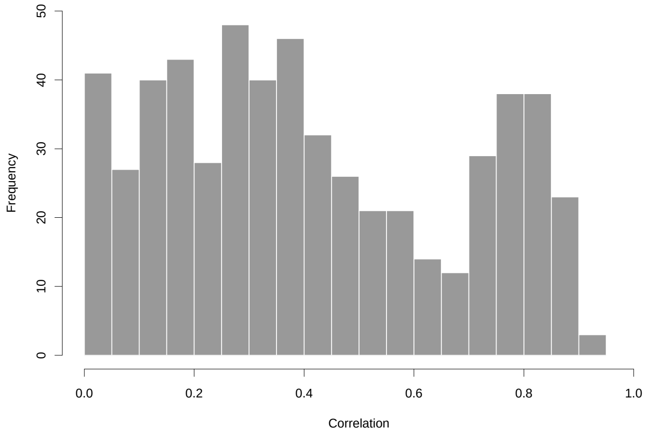


Figure 3: Histogram of the values of 570 positive local maxima for the correlation between image and template.

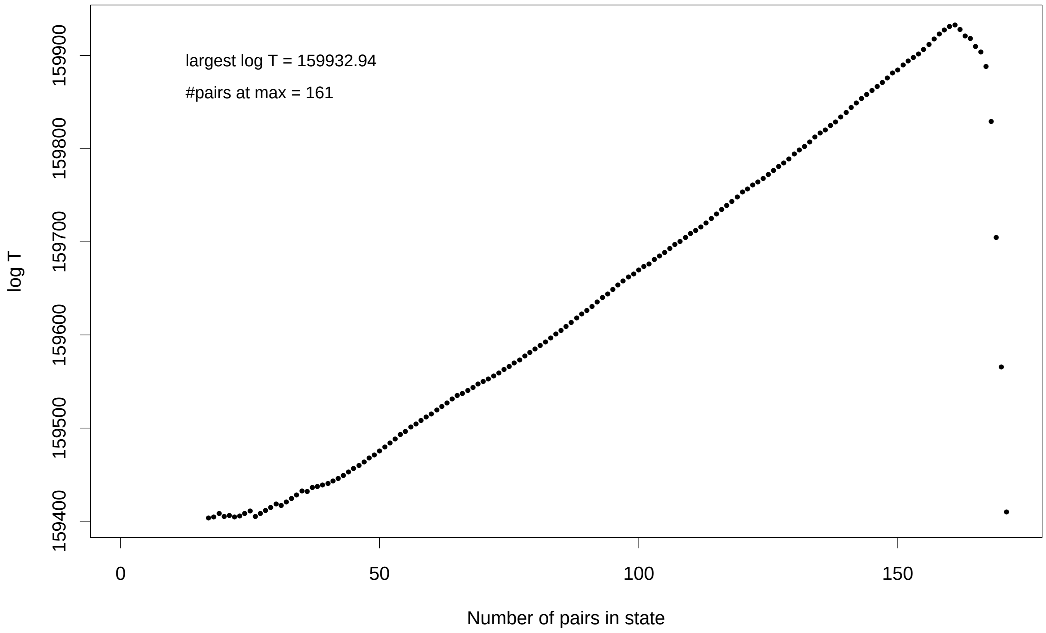


Figure 4: The logarithm of the terms obtained in the starting procedure for the approximate likelihood maximization.

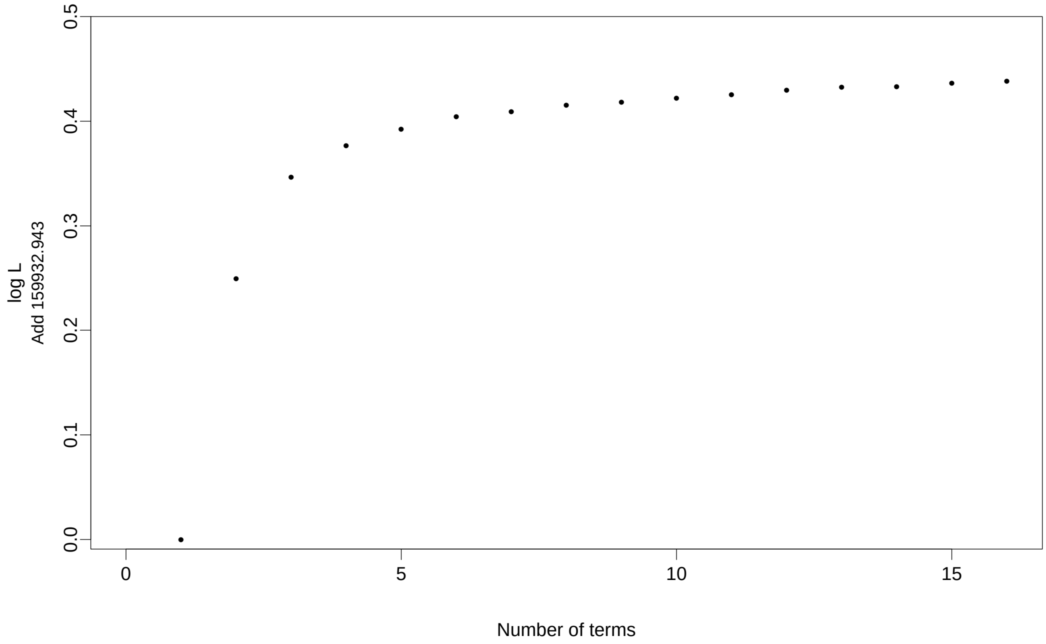


Figure 5: Value of the conditional log likelihood as a function of the number of terms included in the likelihood sum.