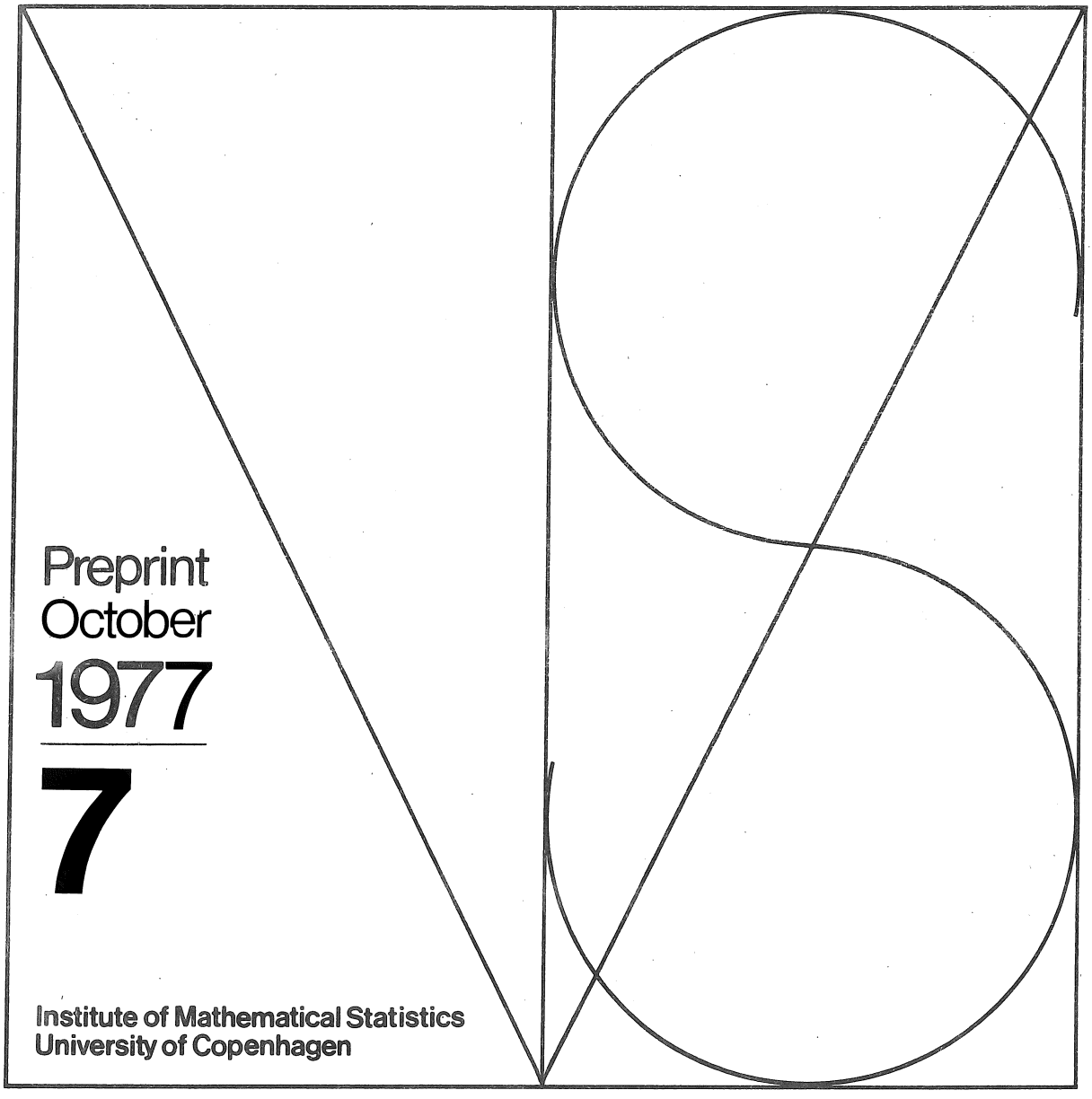
Søren Johansen

The Product Limit Estimator as Maximum Likelihood Estimator



Preprint
October
1977

7

Institute of Mathematical Statistics
University of Copenhagen

Søren Johansen

THE PRODUCT LIMIT ESTIMATOR AS MAXIMUM LIKELIHOOD ESTIMATOR

Preprint 1977 No. 7

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

October 1977

Abstract: It is proved that the product limit estimator of the transition probabilities of a Markov chain can be considered a maximum likelihood estimator in a suitably extended model, that allows fixed points of discontinuity.

Keywords: Product limit estimator. Markov chains. Maximum likelihood estimator. Kaplan Meier estimator.

Running headline:

The product limit estimator.

1. Introduction and summary

Kiefer and Wolfowitz (1956) suggested that for a non-dominated family of probability measures $\mathcal{P}$ one can define a maximum likelihood estimator as follows:

For $\mathbb{P}_1 \in \mathcal{P}$ and $\mathbb{P}_2 \in \mathcal{P}$ let $f(x, \mathbb{P}_1, \mathbb{P}_2) = \frac{d\mathbb{P}_1}{d(\mathbb{P}_1 + \mathbb{P}_2)}(x)$.

Then $\hat{\mathbb{P}}$ is the maximum likelihood estimator if

$$f(x, \hat{\mathbb{P}}, \mathbb{P}) \geq f(x, \mathbb{P}, \hat{\mathbb{P}}) ; \mathbb{P} \in \mathcal{P} \tag{1.1}$$

It is easily seen that in case the family $\mathcal{P}$ is dominated by the σ-finite measure μ, then (1.1) is equivalent to the usual definition of the maximum likelihood estimator.

It is also easily seen, that if $\hat{\mathbb{P}}$ gives positive probability to the observation x, then in order to check (1.1) it is enough to check it for those $\mathbb{P} \in \mathcal{P}$ for which also $\mathbb{P}(x) > 0$. In this case (1.1) reduces to

$$\hat{\mathbb{P}}(x) \geq \mathbb{P}(x) ; \mathbb{P} \in \mathcal{P}. \tag{1.2}$$

Kiefer and Wolfowitz point out that in case we have n observations from a completely unknown distribution, then the empirical distribution function can be considered the maximum likelihood estimator.

In this note it is shown how one can extend the family of continuous time Markov chains in such a way, that the product limit estimator for the transition probabilities can be considered the maximum likelihood estimator.

It is also shown that the product limit estimator considered by Kaplan and Meier (1958) can be considered a maximum likelihood estimator in the sense of (1.1) or (1.2).

2. Estimation in the Markov chain model

We shall first define what is meant by the Markov chain model. The construction we shall use was suggested by Dobrushin (1953),

a similar idea was used by Jacobsen (1972) and an account from the point of view of product integrals is found in Johansen (1977).

Let $B=(B_{ij}, i\in E, j\in E)$ be a finite matrix of finite measures on the Borel sets of $[0,1]$. We assume further for $i\neq j$

$$B_{ij}\geq 0, B_{ii}\leq 0, \Sigma_j B_{ij}=0 \tag{2.1}$$

$$B_{ii}[t]>-1, t\in[0,1]. \tag{2.2}$$

These measures will be called the integrated intensity measures, since the way to think of them in terms of the usual time continuous Markov chains is by the connection

$$B_{ij}[0,t]=\int_0^t q_{ij}(u)\,du,$$

where $q_{ij}(u)=\frac{\partial}{\partial t}p_{ij}(u,t)\big|_{t=u}$ is the intensity at u.

We thus allow a larger class of measures than usually, the point being that we want a family of processes defined that include the continuous time processes that allow fixed points of discontinuity, corresponding to atoms of B.

The transition probabilities are now constructed by the product integrals

$$P(s,t)=\prod_{]s,t]}(I+dB)$$

and from these and an initial distribution p we can again construct the process $(X_t, t\in[0,1])$ via the consistency theorem. The process we get is then called a Markov chain with integrated intensity measure B and initial distribution p.

Notice that conditions (2.1) and (2.2) are needed in order that the matrix P(s,t) becomes a stochastic matrix.

The family of all such measures $\mathbb{P}_{B,p}$ where B satisfies (2.1) and (2.2) is called $\mathcal{P}$ in the following.

In the paper by Aalen and Johansen (1978) only the subfamily was considered where B satisfies the conditions

(2.3) $$B_{ij}(A)=\int_A q_{ij}(t)dt$$

(2.4) $$q_{ij}(t)>0, q_{ii}(t)<0, \Sigma_j q_{ij}(t)=0$$

(2.5) $q_{ij}(\cdot)$ right continuous and has left limits.

The family of probability measures generated this way is called $\mathcal{P}_0$. Note that condition (2.3), (2.4) and (2.5) imply that $\mathcal{P}_0$ is dominated by some finite measure. One may take the $\mathbb{P}$ corresponding to $B(A)=Q\lambda(A)$ where Q is an intensity matrix with $q_{ij}>0$ and λ is Lebesgue measure, and take $p(i)>0$. The maximum likelihood estimator in the family $\mathcal{P}_0$ does not exist, only by enlarging the family to $\mathcal{P}$ do we get the existence. The price paid for this is that the measures in $\mathcal{P}$ are not dominated by some measure μ.

Let now $X^1,\ldots,X^n$ be independent observations of the process $X=(X_t, t\in[0,1])$ with a distribution in the family $\mathcal{P}$. We shall summarize the statistic as follows. Let $t_1\leq\ldots\leq t_m$ denote the points where a jump occurs $t_0=0$, $t_{m+1}=1$, m is then the number of jumps. Let $\Delta n_{ij}(k)$ denote the number of observations that jumped from i to j at time t_k and $n_i(k)$ is the number of observations in i just before t_k, and n_i equal the number of observations starting in i.

The following estimator was considered in Aalen and Johansen (1977) where references to some of its earlier history can be found.

(2.6) $$\hat{B}_{ij}[0,t]=\sum_{t_k\leq t}\frac{1}{n_i(k)}\Delta n_{ij}(k), \hat{B}_{ii}=-\sum_{j\neq i}\hat{B}_{ij}, i\neq j.$$

That is, $\hat{B}$ is a discrete measure with finite support at the points $t_k, k=1,\ldots,m$. The corresponding Markov chain generated by

$$\hat{P}(s,t)=\prod_{]s,t]}(I+d\hat{B})$$

allows only jumps at the points t_k, k=1,...,m and with probabilities equal to the frequencies observed. The measure $\hat{p}$ is just $\hat{p}(i)=n_i/n$.

The object of this section is to prove, that $\hat{\mathbb{P}}=\mathbb{P}_{\hat{B},\hat{p}}$ is the maximum likelihood estimator of $\mathbb{P}$ in the family $\mathcal{P}$ in the sense of (1.1) or rather (1.2) since evidently $\hat{\mathbb{P}}$ or rather $(\hat{\mathbb{P}})^n$ gives positive probability to the observed x.

Let therefore $\mathbb{P}^n$ be any such measure giving positive probability to the occured value $x=(x^1,\ldots,x^n)$. $\mathbb{P}$ thus allows jumps to occur at the points t_k, k=1,...,m but possibly also between.

Let $p_{ij}(k)$ be the $\mathbb{P}$ probability that a particle jumps from i to j at time t_k. Let p(i) denote the initial distribution and let $g_{k,i}$ be the probability that no jump occurs in $]t_{k-1},t_k[$, k=1,2,...,m+1, given that $X_{t_{k-1}}=i$.

We want to derive the probability of the observed x's. Let us start with x^1. Define $i_k=x^1_{t_{k-1}}$, k=1,2,...,m+1. Then x^1 is completely described by being constant on $[t_{k-1},t_k[$ and equal to i_k on this interval. We first find

$$\mathbb{P}\{X_t=i_k,\ t_{k-1}\leq t<t_k,\ X_{t_k}=i_{k+1}\,|\,X_{t_{k-1}}=i_k\}$$

$$= g_{k,i_k}\,p_{i_k,i_{k+1}}(k)$$

and then

$$\mathbb{P}\{X=x^1\}=p(i_1)\prod_{k=1}^{m} g_{k,i_k}\,p_{i_k,i_{k+1}}(k)\,g_{m+1,i_{m+1}}$$

Finally the probability of all observations is given by

$$\mathbb{P}^n\{X^\nu=x^\nu,\nu=1,\ldots,n\}=\prod_\nu p(i_1^\nu)\prod_{k,\nu} g_{k,i_k^\nu}\,p_{i_k^\nu,i_{k+1}^\nu}(k).$$

In terms of the variables introduced above this equals

$$M=\prod_i p(i)^{n_i}\prod_{k=1}^{m+1}(g_{k,i})^{n_i(k)}\prod_{ij}p_{ij}(k)^{\Delta n_{ij}(k)}$$

Although this resembles a proof that $\{n_i, n_i(k), \Delta n_{ij}(k)\}$ is a sufficient statistic, it should be emphasized that firstly we are working with a non-dominated family of measures P and secondly that the reduction is only demonstrated for the special measures in P that give positive probability to the observations $x^1, \ldots, x^n$.

This can be compared with Anderson and Goodman (1957) who studied estimation in discrete time Markov chains, and therefore did not need the factor $\prod_k g_{k,i}^{n_i(k)}$.

It is not hard to verify that M is only maximized if

$$p(i) = \hat{p}(i) = \frac{n_i}{n} \tag{2.7}$$

$$g_{k,i} = 1 \tag{2.8}$$

$$p_{ij}(k) = \frac{\Delta n_{ij}(k)}{n_i(k)} \tag{2.9}$$

Notice that if $n_i(k)=0$, that is, no observations were available in i at time t_k, then the parameters $p_{ij}(k)$ are non identifiable, since they do not occur in M. Since no jumps out of i is actually observed we put $p_{ii}(k)=1$ in this case.

It is seen that the $\mathbb{P}$ corresponding to the choice of (p,B) that maximizes M has the property, that between the points t_k, no jumps can occur. It is thus a discrete time process and it is not difficult to see that it coincides with $\mathbb{P}_{\hat{B},\hat{p}}$.

Thus $\hat{\mathbb{P}} = \mathbb{P}_{\hat{B},\hat{p}}$ is the maximum likelihood estimator in the extended model.

If P_0 is the model it would therefore seem reasonable to smoothe the estimator $\hat{B}$ in such a way that it becomes absolutely continuous and satisfy (2.3), (2.4) and (2.5). In this way the

estimated transition probability would give rise to an estimated probability measure which is in fact in P_0.

This smoothing can be done as follows: for the measures in P_0, we have that the probability that two processes jump at the same time is zero. Hence the variables $\Delta n_{ij}(k)$ are either 0 or 1.

Now let C_{ij} denote the matrix with element (i,j)equal to 1 and element (i,i) equal -1 and the rest zero. Then C_{ij} becomes an extremal intensity matrix in the convex set of intensity matrices $Q=(q_{ij})$ such that $q_{ij}\geq 0$, $q_{ii}\leq 0$, $\Sigma_j q_{ij}=0$.

With this notation we can write

$$\hat{P}(s,t)= \prod_{s<t_k\leq t} (I+n_{i_k}(k)^{-1}C_{i_k j_k}),$$

where the jump at time t_k goes from i_k to j_k.

Now notice that ${C_{ij}}^2=-C_{ij}$ and hence that $\exp(tC_{ij})=I+(1-e^{-t})C_{ij}$, which means that the factors of P(s,t) all have the form $\exp(a_k C_{i_k j_k})$ with $a_k=-\ell n(1-n_{i_k}(k)^{-1})$. If $a_k < \infty$ we now define $\hat{\hat{B}}$ by

$$\hat{\hat{B}}[t_{k-1},t[= \frac{t-t_{k-1}}{t_k-t_{k-1}} a_k C_{i_k j_k}, t_{k-1}\leq t<t_k$$

and $\hat{\hat{P}}(s,t)= \prod_{]s,t]} (I+d\hat{\hat{B}})$, then

$$\hat{\hat{P}}(s,t)=\hat{P}(s,t_{k-1})e^{\hat{\hat{B}}[t_{k-1},t[}$$

$$=\hat{P}(s,t_{k-1})(I+(1-\exp(-a_k((t-t_{k-1})/(t_k-t_{k-1})))C_{i_k j_k}), s<t_{k-1}\leq t<t_k$$

This estimator is seen to be continuous and piecewise continuously differentiable, and at the jump points it coincides with the estimator $\hat{P}$. In this sense the estimator $\hat{P}$ has been smoothed. Some asymptotic properties of $\hat{\hat{P}}$ has been given by Aalen and

Johansen (1978).

Notice that if $n_{i_k}(k) = 1$ then $a_k = \infty$ and the estimator $\hat{\hat{P}}$ no longer is continuous. Hence the smoothing is not possible in this case.

As an application of the above estimator $\hat{B}$ we shall also consider the problem of estimating the waiting time distribution in state i. Let T denote the waiting time. For any probability measure in P_0 it is well known that

$$(2.10) \qquad \mathbb{P}\{T \geq t \mid X_s = i\} = P\{X_u = i, s < u \leq t \mid X_s = i\} = e^{\int_s^t q_{ii}(u)\,du}.$$

A natural estimator of this distribution function is therefore given by

$$(2.11) \qquad e^{\hat{B}_{ii}]s,t]} = \exp - \sum_{s < t_k \leq t} \sum_{j \neq i} \frac{\Delta n_{ij}(k)}{n_i(k)}.$$

The formula (2.10) does not hold for any $\mathbb{P} \in P$ however and has to be replaced by

$$(2.12) \qquad \mathbb{P}\{T \geq t \mid X_s = i\} = \prod_{]s,t]} (1 + dB_{ii}).$$

It is thus seen that the maximum likelihood estimator for the waiting time distribution is then given by

$$(2.13) \qquad \prod_{]s,t]} (1 + d\hat{B}_{ii}) = \prod_{s < t_k \leq t} \left(1 - \sum_{j \neq i} \frac{\Delta n_{ij}(k)}{n_i(k)}\right).$$

This estimator is easily seen to be different from (2.11). It bears a formal resemblance to the Kaplan-Meier estimator considered in the next section but the set up is different.

The Kaplan-Meier estimator is used in a situation where observations can be withdrawn from the system under observation. Thereby the number of observations available is decreased.

In Markov chains, however, more observations can be inserted into the system, i.e. state i, when the other particles jump to state i. Thereby the number of available observations is increased.

3. The Kaplan-Meier estimator

Let $X_1,\ldots,X_n$ be independent identically distributed variables and let $t_1,\ldots,t_n$ be censoring times, such that only

$$M_i=X_i\wedge t_i \text{ and } Y_i=1\{X_i\geq t_i\}$$

are observed. Thus the i'th x value is only observed if it is $\leq t_i$, otherwise we just know that $X_i\geq t_i$.

In the model were the distribution of $X_1,\ldots,X_n$ is completely unspecified we would like to prove that

$$(3.1)\qquad \hat{\mathbb{P}}\{X\geq t\}=\prod_{u_j<t}\left(1-\frac{n_j-s_j}{\sum_{m\geq j} n_m-s_j}\right)$$

is the maximum likelihood estimator. Here $u_1<u_2<\ldots<u_k$ are the different values observed and

$n_j=\sum_{i=1}^{n} 1\{M_i=u_j\}$, $s_j=\sum_{i=1}^{n} 1\{M_i=u_j,Y_i=1\}$. Thus n_j denotes the number of observations taking on the value u_j and s_j denotes how many of these are censored observations.

In case we consider the model specified by the unknown distribution being continuous and that the t_i's are different, then the maximum likelihood estimator does not exist. The estimator $\hat{\mathbb{P}}$ becomes simpler, since for a continuous distribution we know that the probability of finding two observations at the same point is zero. Hence $n_j=1$ and $s_j=y_j$ which is zero or one.

In this case we get

$$(3.2)\qquad \hat{\mathbb{P}}\{X\geq t\}=\prod_{\{u_j<t\}}\left(1-\frac{1-y_j}{n-j+1-y_j}\right)=\prod_{u_j<t}\left(1-\frac{1}{n-j+1}\right).$$

where the product is taken over those $u_j<t$ which were observed to be x values.

Clearly the measure $\hat{\mathbb{P}}$ determined by (3.1) has a finite support contained in the observed values of M_i and therefore gives positive probability to the observed outcome of $(M_i,Y_i)i=1,\ldots,n$.

For any measure $\mathbb{P}$ which gives positive probability to the observation we can prove that

$$\mathbb{P}\{M_i=m_i,Y_i=y_i\}=\mathbb{P}\{X_1=m_i\}^{1-y_i}\,\mathbb{P}\{X_1\geq m_i\}^{y_i}. \tag{3.3}$$

Hence the probability of the observation becomes

$$M=\prod_{i=1}^{n}\mathbb{P}\{X_1=m_i\}^{1-y_i}\,\mathbb{P}\{X_1\geq m_i\}^{y_i}.$$

Let now the different values of $m_1,\ldots,m_n$ be denoted by $u_1<\ldots<u_k$ and let us introduce the notation n_j and s_j as above. Further we let $p_i=\mathbb{P}\{X_1=u_i\}$ and introduce $r_i=\mathbb{P}\{u_i<X_1<u_{i+1}\}i=1,\ldots,k-1,r_0=$ $=\mathbb{P}\{X_1<u_1\},r_k=\mathbb{P}\{u_k<X_1\}$ then $\sum_{i=0}^{k} r_i+\sum_{i=1}^{k} p_i=1$ and

$$M=\prod_{i=1}^{k} p_i^{n_i-s_i}\Big(\sum_{j=i}^{k}(p_j+r_j)\Big)^{s_i} \tag{3.4}$$

In order to maximize (3.4) we first note that without changing p_j+r_j we can increase p_j and thereby M by putting $r_j=0$. We then introduce the failure rates

$$a_i=p_i\Big[\sum_{j\geq i}p_j\Big]^{-1}$$

from which we get

$$p_i=a_i\prod_{j=1}^{i-1}(1-a_j)$$

and

$$\sum_{j\geq i}p_j=\prod_{j=1}^{i-1}(1-a_j)$$

In terms of these parameters we have that the probability of the observations is equal to

$$M=\prod_{i=1}^{k} a_i^{n_i-s_i}\Big(\prod_{j=1}^{i=1}(1-a_j)\Big)^{n_i}$$

$$= \prod_{i=1}^{k} a_i^{n_i - s_i} (1-a_i)^{\sum_{j>i} n_j}$$

With this parametrization we see that M is maximized only if

$$a_i = \frac{n_i - s_i}{\sum_{j \geq i} n_j - s_i} \tag{3.5}$$

It is seen that the choice (3.5) is exactly the measure determined by (3.1).

This argument is essentially that given in the paper by Kaplan and Meier (1958), but it is given here to indicate that it fits in with the general definition (1.1) or (1.2).

With the formalism of the product integral

$$\hat{I\!P}(X \geq t) = \prod_{[0,t[} (1-d\hat{B})$$

where

$$\hat{B}[0,u] = \int_{[0,u]} \hat{I\!P}(x \geq t)^{-1} d\, \hat{I\!P}(x \geq t) = \sum_{u_j \leq u} \frac{n_j - s_j}{\sum_{v \geq j} n_v - s_j}$$

which is the empirical cumulative hazard process. Thus it has also been established that $\hat{B}$ is the maximum likelihood estimator in the unrestricted model, where the underlying distribution is completely unspecified. See the comment by Breslow in the discussion of the paper by Cox (1972).

4. Acknowledgement

It is a pleasure to thank H.D.Brunk and N.Keiding for useful comments and references.

5. References

1. Aalen,O. and Johansen,S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand. J. Statist.
2. Anderson,T.W. and Goodman,L.A. (1957). Statistical inference for Markov chains. Ann. Math. Statist. 28, 89-110.
3. Breslow,N. and Crowley,J. (1974). A large sample study of the life table and product limit estimates under random censoring. Ann. Statist. 2, 437-453.
4. Cox,D.R. (1972). Regression models and life tables. J. Roy. Statist. Soc. Ser. B 34, 187-220.
5. Dobrushin (1953). Generalization of Kolmogorov's equations for Markov processes with a finite number of possible states. Mat. Sb. N.S. 33, 567-596.
6. Jacobsen (1972). A characterization of minimal Markov jump processes. Z. Wahrscheinlichkeitstheorie verw. Gebiete. 23, 32-46.
7. Kaplan,E.L. and Meier,P. (1958). Nonparametric estimation from incomplete observations. J. Am. Statist. Ass. 53, 457-481.
8. Kiefer,J. and Wolfowitz,J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Statist. 27, 887-906.
9. Johansen,S. (1977). Product integrals and Markov processes. Tech. Rep. University of Copenhagen. 13 pp.

Søren Johansen
Institute of Mathematical Statistics
University of Copenhagen
5, Universitetsparken
DK-2100 Copenhagen Ø
Denmark

PREPRINTS 1976

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Becker, Niels: Estimation for an Epidemic Model.

No. 2 Kaplan, Norman: A Generalization of a Result of Erdös and Renyi and a Related Problem.

No. 3 Lauritzen, Steffen & Remmer, Ole: An Investigation of the Asymptotic Properties of Geodetic Least Squares Estimates in Certain Types of Network.

No. 4 Asmussen, Søren: Almost Sure Behavior of Linear Functionals of Supercritical Branching Processes.

No. 5 Hald, Anders & Møller, Uffe: On the SPRT of the Mean of a Poisson Process.

No. 6 Lauritzen, Steffen: Projective Statistical Fields.

No. 7 Keiding, Niels: Inference and Tests for Fit in the Birth-Death-Immigration Process.

No. 8 Møller, Uffe: OC and ASN og the SPRT for the Poisson Process.

No. 9 Keiding, Niels: Population Growth and Branching Processes in Random Environments.

No. 10 Andersen, Per K., Andersen, Søren & Lauritzen, Steffen: The Average Noise from a Poisson Stream of Vehicles.

No. 11 Hald, Anders & Møller, Uffe: Multiple Sampling Plans of Given Strength for the Poisson and Binomial Distributions.

No. 12 Johansen, Søren: Two Notes on Conditioning and Ancillarity.

No. 13 Asmussen, Søren: Some Martingale Methods in the Limit Theory of Supercritical Branching Processes.

No. 14 Johansen, Søren: Homomorphisms and general exponential families.

No. 15 Asmussen, Søren & Hering, Heinrich: Some Modified Branching Diffusion Models.

No. 16 Hering, Heinrich: Minimal Moment Conditions in the Limet theory for General Markov Branching Processes.

PREPRINTS 1977

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Asmussen, Søren & Keiding, Niels: Martingale Central Limit Theorems and Asymptotic Estimation Theory for Multitype Branching Processes.

No. 2 Jacobsen, Martin: Stochastic Processes with Stationary Increments in Time and Space.

No. 3 Johansen, Søren: Product Integrals and Markov Processes.

No. 4 Keiding, Niels & Lauritzen, Steffen L. : Maximum likelihood estimation of the offspring mean in a simple branching process.

No. 5 Hering, Heinrich: Multitype Branching Diffusions.

No. 6 Aalen, Odd & Johansen, Søren: An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations.

No. 7 Johansen, Søren: The Product Limit Estimator as Maximum Likelihood Estimator.