of M_n and the leading bias term B_{2n} of $G[m_n^{(1)}, m_n^{(2)}]$ for $c = 0.99$ are

$$B_{1n} = h^2 + h^4, \quad B_{2n} = \sqrt{152.76}\, h^4.$$

This shows that if $h^2 > 1/(\sqrt{152.76} - 1)$ the mse of $G[m_n^{(1)}, m_n^{(2)}]$ dominates the mse of m_n. Similar conditions can be found by varying $m^{(2)}$ and $m^{(4)}$.

In a practical situation a choice of R that avoids a situation of this kind, described in the example, seems to be impossible. Such a selection of R has to take into account the unknown values $m^{(2)}(t)$ and $m^{(4)}(t)$. It is therefore impossible in a practical solution to compute the parameter regions where $G[m_n^{(1)}, m_n^{(2)}]$ actually improves ordinary kernel regression estimate m_n.

We also compared the leading terms of the mse $G[m_n^{(1)}, m_n^{(2)}](t)$ and of the mse $m_n(t)$ of a fixed regression curve in Table II. Shown are the ratios of the two leading terms for different values of h, h_1, and c with $n = 100$ and $\sigma^2 = 1$. The regression curve $m(t) = \sin t$ was selected, and the mse at $t = \pi/4$ was evaluated with $K \in \mathcal{R}_2$ as before. A bandwidth h, being roughly about 0.3, would minimize the mse of $m_n(t)$; therefore only combinations are shown with $h, h_1 \in \{0.2, 0.3, 0.4\}$. The use of $G[m_n^{(1)}, m_n^{(2)}]$ may result in an mse nearly twice as high as the corresponding mse of m_n as can be seen from the entry $(h, h_1, c) = (0.3, 0.3, 0.9)$ in Table II.

References

[1] B. Efron, "The jackknife, the bootstrap and other resampling plans," SIAM publication CBMS-NSF, 1982.

[2] V. A. Epanechnikov, "Nonparametric estimation of a multivariate probability density," *Theory Prob. Appl.*, vol. 14, pp. 153–158, 1969.

[3] A. A. Georgiev, "Nonparametric system identification by kernel methods," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 356–358, 1984.

[4] H. L. Gray and W. R. Schucany, *The Generalized Jackknife Statistic*. New York: Marcel Dekker, 1972.

[5] L. Györfi, "The rate of convergence of k-NN regression estimate and classification," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 500–509, 1981.

[6] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[7] M. B. Priestley and M. T. Chao, "Non-parametric function fitting," *J. Roy. Statist. Soc. B*, vol. 34, pp. 385–392, 1972.

[8] W. R. Schucany and J. P. Sommers, "Improvement of kernel-type density estimators," *J. Amer. Statist. Ass.*, vol. 72, pp. 420–423, 1977.

An On-Line Parameter Estimation Algorithm for Counting Process Observations

PETER SPREIJ

Abstract—**The parameter estimation problem for counting process observation is considered. It is assumed that the intensity of the counting process is adapted to the family of σ-algebras generated by the counting process itself and that the intensity depends linearly on some deterministic constant parameters. An on-line parameter estimation algorithm is then presented for which convergence is proved by using a stochastic approximation type lemma.**

I. Introduction

Counting processes frequently occur as observations in mathematical models for industrial processes and in biology, software engineering, and nuclear medicine. Usually, such a counting process can be considered as the output process of some stochastic system. The underlying state process then influences the counting process. A problem is then to estimate this state, given the observations. This is known as the filtering problem and has been investigated extensively [1].

The solution of this problem requires knowledge of all parameters needed to describe the stochastic system, which means that one can compute the solution to the filtering problem only if one knows the correct parameter values. Unfortunately, in many cases these are not known and therefore need to be estimated. This may happen before the processes start running, using related additional information and/or observations. In the former case some asymptotic results for off-line maximum likelihood estimation are available [3], [4].

The purpose of the present correspondence is to make a contribution to the on-line parameter estimation problem in a specific case. The approach has proven to be fruitful in discrete time ARMAX processes [7] or continuous time Gaussian AR processes [6].

The correspondence is organized as follows. In Section II we give some basic results for counting processes. In Section III we give a heuristic derivation of our parameter estimation algorithm. Section IV contains the convergence proof of the algorithm.

II. Preliminary Results

We assume that we are given a complete probability space $(\Omega, \mathcal{F}, P)$, a time set $T = [0, \infty)$, and a filtration $\{\mathcal{F}_t\}_{t \geq 0}$ satisfying the usual conditions of [2]. All stochastic processes in the sequel are defined on $\Omega \times T$ and adapted to $\{\mathcal{F}_t\}_{t \geq 0}$. We study the case that we are given: an observed process, which is a counting process, that is a map n: $\Omega \times T \to \mathbb{N}_0$, which has only jumps of magnitude $+1$. Then it is known [1], [2] that n is a submartingale and therefore admits the so-called Doob–Meyer decomposition (with respect to $\{\mathcal{F}_t\}_{t \geq 0}$)

$$n_t = \Lambda_t + m_t \tag{2.1}$$

where Λ: $\Omega \times T \to \mathbb{R}$ is a predictable increasing process and m a local martingale. Now assume that Λ is an absolutely continuous process, say $\Lambda_t = \int_0^t \lambda_s\, ds$; then we can rewrite (2.1) as

$$dn_t = \lambda_t\, dt + dm_t. \tag{2.2}$$

The process λ is called the intensity process.

Often a major problem for counting process observations is to identify the intensity process λ. This problem can be set up in two stages. In the first stage we have to solve a filtering problem. To be precise we have to determine $\hat{\lambda}_t = E(\lambda_t | \mathcal{F}_t^n)$, where $\mathcal{F}_t^n = \sigma\{n_s, s \leq t\}$. Then $\hat{\lambda}_t$ is the optimal (in the sense of mean squared error) estimate given the observations during $[0, t] \subset T$ and given the values of deterministic parameters. We can then replace (2.2) by the minimal decomposition of n (i.e., with respect to $\{\mathcal{F}_t^n\}$)

$$dn_t = \hat{\lambda}_t\, dt + d\bar{m}_t \tag{2.3}$$

where $\bar{m}$ is a local martingale adapted to $\{\mathcal{F}_t^n\}_{t \geq 0}$. In the second stage one looks for estimates of remaining unknown deterministic parameters. If one adopts the maximum likelihood criterion, (2.3) and the computation of $\hat{\lambda}_t$ appear to be crucial. The likelihood functional in this case is known [1, p. 174] to be

$$L_t = \exp\left[-\int_0^t (\hat{\lambda}_s - 1)\, ds + \int_0^t \log \hat{\lambda}_{s^-}\, dn_s \right]. \tag{2.4}$$

The Model

From here on we assume that $\hat{\lambda}$ has a special structure

$$\hat{\lambda}_t = p^T \phi_t \tag{2.5}$$

where $p \in \mathbb{R}^m$ is the vector of unknown parameters and ϕ: $\Omega \times T \to \mathbb{R}^m$ is a process adapted to $\{\mathcal{F}_t^n\}_{t \geq 0}$ and thus known. Indeed (2.5) imposes a restrictive condition on the intensity

Manuscript received August 1, 1984; revised July 12, 1985. This work was presented in part at the Fourteenth Conference on Stochastic Processes and Their Applications, Gothenburg, Sweden, June 12–16, 1984.

The author is with the Center for Mathematics and Computer Science, P.O. Box 4079, 1009 AB, Amsterdam, The Netherlands.

IEEE Log Number 8406623.

process λ. Self-exciting processes exist for which the intensity is of the form (2.5); see the example below. If ϕ_t comes from a recursive filter, it cannot be expected that (2.5) is satisfied. For these (adaptive filtering) problems other algorithms are needed.

The minimal decomposition (2.3) now becomes

$$dn_t = p^T\phi_t\,dt + d\bar{m}_t. \tag{2.6}$$

Plugging (2.5) into (2.4) and writing $L_t(p)$ instead of L_t in order to express the dependence of the likelihood functional on p, we get

$$L_t(p) = \exp\left[-p^T\int_0^t \phi_s\,ds + t + \int_0^t \log\left(p^T\phi_{s-}\right)dn_s\right]. \tag{2.7}$$

III. Derivation of the Algorithm

In this section we state a parameter estimation algorithm for the model (2.5), (2.6). The proof that the parameter estimates given by this algorithm indeed converge to the true parameter value will be given in Section IV. The algorithm is constructed in such a way that the estimates $\hat{p}_t$ of p approximately maximize the likelihood functional (2.7), or equivalently, minimize $J_t(\cdot)$ given by

$$J_t(p) = p^T\int_0^t \phi_s\,ds - \int_0^t \log\left(\phi_{s-}^T p\right)dn_s. \tag{3.1}$$

After posing the algorithm we present a heuristic derivation.

A. Algorithm

Consider the model (2.5), (2.6). An approximate maximum likelihood parameter estimation algorithm is given by

$$d\hat{p}_t = R_t\phi_{t-}\left(dn_t - \phi_t^T\hat{p}_t\,dt\right), \tag{3.2}$$

$$dR_t = -R_t\phi_t\phi_t^T R_t\,dt, \tag{3.3}$$

with initial conditions $\hat{p}_0$ and R_0, respectively.

The interpretation is that for each t $\hat{p}_t$ approximately minimizes $J_t(\cdot)$ as stated earlier and that R_t is up to a multiplicative scalar factor an approximation of the second derivative of $J_t(\cdot)$. Thus (3.2), (3.3) can be considered as a quasi-Newton scheme for minimizing the family of functions $\{J_t(\cdot)\}_{t\geq 0}$. Observe that R_t stays positive definite when the initial value R_0 is chosen to be symmetric and positive definite, since $dR_t^{-1} = \phi_t\phi_t\,dt$.

B. Heuristic Derivation

To understand the algorithm (3.2), (3.3) it is useful to consider first a nonstochastic situation. Let J: $\mathbb{R}_+ \times \mathbb{R}^m \to \mathbb{R}$, $J \in C^2(\mathbb{R}_+ \times \mathbb{R}^m, \mathbb{R})$ such that $J(t,\cdot)$: $\mathbb{R}^m \to \mathbb{R}$ has a unique minimum, attained for say $x(t)$. Under some regularity conditions it then follows from the implicit function theorem that the function $t \to x(t)$ satisfies the differential equation

$$dx(t) = -\left[\frac{\partial^2}{\partial x^2}J(t,x(t))\right]^{-1}\frac{\partial^2}{\partial x\,\partial t}J(t,x(t))\,dt. \tag{3.4}$$

Let us now return to our estimation problem, that is, finding the value $\hat{p}_t$ that for each t minimizes (3.1). For an evolution equation for $\hat{p}_t$ one tries to find an equation like (3.4). However, the functional J of (3.1) does not satisfy the desired smoothness conditions, and therefore one has to look for something related to (3.4). Our choice is

$$d\hat{p}_t = -\left[J_t''(\hat{p}_{t-})\right]^{-1}\partial_t J_t'(\hat{p}_t) \tag{3.5}$$

where prime denotes partial differentiation with respect to p and ∂_t means the partial forward differential operator with respect to t. In order to specify the algorithm fully we also need recursive expressions for $J_t''(\hat{p}_t)$ and $J_t'(\hat{p}_t)$. Later on we will establish almost sure convergence of the family $\{\hat{p}_t\}_{t\geq 0}$ to the true parameter value p_0.

From (3.1) we get by formal differentiation

$$J_t'(p) = \int_0^t \phi_s\,ds - \int_0^t \frac{\phi_{s-}}{p^T\phi_{s-}}\,dn_s, \tag{3.6}$$

hence

$$\partial_t J_t'(p) = \phi_t\,dt - \frac{\phi_{t-}}{p^T\phi_{t-}}\,dn_t. \tag{3.7}$$

Define $k_t = \phi_t/\hat{p}_t^T\phi_t$ and $Q_t = [J_t''(\hat{p}_t)]^{-1}$. Using these expressions and (3.7), we can rewrite (3.5) as

$$d\hat{p}_t = Q_{t-}k_{t-}\left(dn_t - \phi_t^T\hat{p}_t\,dt\right). \tag{3.8}$$

The next problem is the finding of a recursion for Q_t. It turns out that an exact equation for Q_t cannot be obtained for $p \in \mathbb{R}^m$ with $m \geq 2$ and that certain approximations are not satisfactory in that these cause problems in analyzing the convergence properties of the algorithm.

On the other hand, the case of $p \in \mathbb{R}^1$ is easy to handle, and it will be illustrative for the multivariable case. In this case (3.6) reads

$$J_t'(p) = \int_0^t \phi_s\,ds - \frac{n_t}{p}, \tag{3.9}$$

hence

$$J_t''(p) = \frac{n_t}{p^2}. \tag{3.10}$$

Therefore, Q_t becomes $\hat{p}_t^2/n_t$ and with $k_t = 1/\hat{p}_t$ (3.8) reads

$$d\hat{p}_t = \frac{\hat{p}_{t-}}{n_{t-}}\left(dn_t - \phi_t\hat{p}_t\,dt\right). \tag{3.11}$$

Observe that $\hat{p}_t = n_t/\Phi_t$, where $\Phi_t = \int_0^t \phi_s\,ds$ satisfies (3.11), and this value for $\hat{p}_t$ is also found by directly minimizing (3.1). One can prove that $\hat{p}_t$ given by (3.11) converges to the true parameter value, using the method of Section IV.

Applying the stochastic calculus to $Q_t = \hat{p}_t^2/n_t$ one can verify that Q_t satisfies

$$dQ_t = -2Q_t^2k_t\phi_t\,dt + Q_{t-}^2k_{t-}^2\,dn_t. \tag{3.12}$$

Returning to the multivariate case $p \in \mathbb{R}^m$, $m \geq 2$ one would like to extend (3.12) in order to obtain an evolution equation for Q_t. This suggests

$$dQ_t = -2Q_tk_t\phi_t^TQ_t\,dt + Q_{t-}k_{t-}k_{t-}^TQ_{t-}\,dn_t. \tag{3.13}$$

One hopes that (3.8) together with (3.13) constitutes the desired algorithm. Although (3.8), (3.13) yield some appealing properties suggested by the case $p \in \mathbb{R}^1$, such as $\hat{p}_t = Q_t\Phi_t$, $\hat{p}_t^TQ_t^{-1}\hat{p}_t = n_t$ and $\Phi_t^T\hat{p}_t = n_t$, we were not able to prove the desired convergence properties. The major bottleneck was the verification of the technical condition (see (4.5))

$$\int_0^\infty Q_{t-}k_{t-}k_{t-}^TQ_{t-}\,dn_t < \infty, \tag{3.14}$$

which is a trivial exercise if $p \in \mathbb{R}^1$. The main cause of this technical problem was the term $\phi_t^T\hat{p}_t$ in the denominator of k_t. Therefore, we tried to incorporate this term in Q_t so that $Q_tk_t = R_t\phi_t$, for some matrix valued process R_t; the idea was then to find an equation for R_t.

Inspection of the case $p \in \mathbb{R}^1$, neglecting the derivatives of ϕ and using $Q_tk_t = R_t\phi_t$ then leads from (3.13) to

$$dR_t = -R_t\phi_t\phi_t^TR_t\,dt. \tag{3.3}$$

IV. Convergence Proof

In this section we present a convergence proof for the algorithm (3.2), (3.3) which establishes almost sure convergence of the parameter estimates to the true parameter value. The proof is completely in the spirit of the proofs in [6], [7]. We begin by stating an important technical lemma, which is a simple version of a more general result in [6], that in turn can be considered as the continuous time counterpart of a result in discrete time stochastic approximation [5].

Lemma 1: Let x, a, b be nonnegative stochastic processes and m be a local martingale such that $x = a - b + m$, and assume

that

1) a and b are increasing processes with $a_0 = b_0 = 0$,
2) $\exists c \in \mathbb{R}_+$ such that $\forall t$: $\Delta a_t = a_t - a_{t^-} \le c$ almost surely,
3) $\lim_{t\to\infty} a_t < \infty$ almost surely.

Then

a) $\lim_{t\to\infty} x_t$ exists and is finite almost surely,
b) $\lim_{t\to\infty} b_t$ is finite almost surely.

Here is our main result.

Theorem 1: Consider the algorithm (3.2), (3.3). Let p_0 be the true parameter value. Let $\tilde{p}_t = \hat{p}_t - p_0$ and let $\psi_t = \phi_t^T\phi_t$, $\Psi_t = \int_0^t \psi_s\, ds + \operatorname{tr}(R_0^{-1})$.

Assume

1) $\lim_{t\to\infty} \Psi_t = \infty$ almost surely,
2) $\int_0^\infty \Psi_t^{-2}\psi_t\phi_t\, dt < \infty$ almost surely,
3) $\lim_{t\to\infty} \Psi_t^{-1}\int_0^t \phi_s\phi_s^T\, ds = C$, where $C \in \mathbb{R}^{m\times m}$ is positive definite almost surely.

Then

a) $\lim_{t\to\infty} \hat{p}_t = p_0$ almost surely,
b) $\lim_{t\to\infty} \Psi_t^{-1}\int_0^t (\phi_s^T\tilde{p}_s)^2\, ds = 0$ almost surely.

Proof: From (3.2), (3.3) it follows that

$$d\tilde{p}_t = R_t\phi_{t^-}\left(dn_t - \phi_t^T\hat{p}_t\, dt\right) = R_t\phi_{t^-}\left(d\bar{m}_t - \phi_t^T\tilde{p}\, dt\right) \quad (4.1)$$

$$dR_t^{-1} = \phi_t\phi_t^T\, dt. \quad (4.2)$$

Observe that $\Psi_t = \operatorname{tr}(R_t^{-1})$. Define the Lyapunov process

$$w_t = \Psi_t^{-1}\left(\tilde{p}_t R_t^{-1}\tilde{p}_t + \int_0^t \left(\tilde{p}_s^T\phi_s\right)^2 ds\right), \quad (4.3)$$

then

$$dw_t = -\Psi_t^{-1}w_t\psi_t\, dt + \phi_t^T R_t\phi_t\Psi_t^{-1}p_0^T\phi_t\, dt + dm_{1t}, \quad (4.4)$$

where m_1 is a local martingale. Next we apply Lemma 1 to (4.4). Because w, Ψ are positive, we then see that the only thing we have to check is assumption 3 of Lemma 1:

$$\int_0^\infty \phi_t^T R_t\phi_t\Psi_t^{-1}p_0^T\phi_t\, dt < \infty. \quad (4.5)$$

To that end, let $\rho_t = \operatorname{tr} R_t$. Let γ_{it} be one of the eigenvalues of R_t^{-1}, then $\lim_{t\to\infty}\Psi_t^{-1}\gamma_{it} = c_i > 0$ by assumption 3 of the theorem. Hence $\gamma_{it} = c_i\Psi_t(1 + o(1))$, $(t \to \infty)$. Now γ_{it}^{-1} is an eigenvalue of R_t, $\gamma_{it}^{-1} = c_i^{-1}\Psi_t^{-1}(1 + o(1))$, $(t \to \infty)$. Hence $\rho_t = \Psi_t^{-1}(\sum c_i^{-1} + o(1))$, $(t \to \infty)$, or $\rho_t = o(\Psi_t^{-1})$, $(t \to \infty)$. Recall that for a positive definite matrix A, $x^TAx \le x^Tx \cdot \operatorname{tr}(A)$ and $x^TA^2x \le x^Tx(\operatorname{tr}(A))^2$. Then

$$\begin{aligned}
&\int_0^\infty \phi_t^T R_t\phi_t\Psi_t^{-1}p_0^T\phi_t\, dt \\
&\quad = \int_0^\infty \phi_t^T R_t R_t^{-1}R_t\phi_t\Psi_t^{-1}p_0^T\phi_t\, dt \\
&\quad \le \int_0^\infty \phi_t^T R_t^2\phi_t p_0^T\phi_t\, dt \le \int_0^\infty \phi_t^T\phi_t\rho_t^2 p_0^T\phi_t\, dt \\
&\quad = p_0^T\int_0^\infty \psi_t\rho_t^2\phi_t\, dt = p_0^T\int_0^\infty \phi_t\psi_t O\left(\Psi_t^{-2}\right) dt < \infty,
\end{aligned}$$

by assumption 2.

Then from Lemma 1 we conclude that w and $\int_0^\infty w_s\Psi_s^{-1}\psi_s\, ds$ almost surely converge. We claim that $\lim_{t\to\infty} w_t = 0$ almost surely. If not, a subset of Ω with positive probability and an $\epsilon > 0$ exists such that $\lim_{t\to\infty} w_t \ge 2\epsilon$ on this subset. However, then we also have on the same subset

$$\int_0^\infty \Psi_t^{-1}w_t\psi_t\, dt \ge \epsilon\int_0^\infty \Psi_t^{-1}\psi_t\, dt = \left[\log(\Psi_t)\right]_0^\infty = \infty,$$

by assumption 1. This contradicts the second assertion of lemma 1. Since w is the sum of two positive quantities we have both

$$\lim_{t\to\infty} \Psi_t^{-1}\int_0^t \left(\tilde{p}_s^T\phi_s\right)^2 ds = 0 \quad \text{almost surely} \quad (4.6)$$

and

$$\lim_{t\to\infty} \tilde{p}_t^T\frac{R_t^{-1}}{\Psi_t}\tilde{p}_t = 0 \quad \text{a.s.} \quad (4.7)$$

Because of assumption 3 we know that $\lim_{t\to\infty}\Psi_t^{-1}R_t^{-1} = C > 0$, hence $\lim_{t\to\infty}\tilde{p}_t = 0$ almost surely.

V. Examples

1) If ϕ: $T \to \mathbb{R}^2$, $\phi(t) = [1, \sin t + 1]$, then the conditions of the theorem are satisfied. The matrix C in assumption 3 becomes $\frac{1}{5}\begin{bmatrix} 2 & 2 \\ 2 & 3 \end{bmatrix}$.

2) Let ϕ: $T \times \Omega \to \mathbb{R}^2$, $\phi_t = (1, 1 + (-1)^{n_t})$ and $p = (a, b) \in \mathbb{R}_+^2$. By analogy, the second component of ϕ jumps like a random telegraph process. Conditions 1 and 2 of Theorem 1 are easily verified. To check condition 3 let us first define

$$X_t = t^{-1}\int_0^t (-1)^n\, ds.$$

Then

$$\Psi_t^{-2}\int_0^t \phi_s\phi_s^T\, ds = \left(3 + t^{-1}\operatorname{tr}\left(R_0^{-1}\right) + 2X_t\right)^{-1} \times\begin{bmatrix} 1 & 1 + X_t \\ 1 + X_t & 2 + 2X_t \end{bmatrix}.$$

We now proceed to compute a.s.—$\lim_{t\to\infty} X_t$. Since $n_t = (a + b)t + btX_t + m_t$, we find that

$$X_t = b^{-1}\left(t^{-1}n_t - t^{-1}m_t - a - b\right).$$

The quadratic variation process $\langle m\rangle_t = (a + b)t + btX_t \le (a + 2b)t$. It then follows from the strong law of large numbers for martingales that $t^{-1}m_t \to 0$ almost surely. Finally, we must evaluate the asymptotic behavior of $t^{-1}n_t$. Define $T_k = \inf\{t \ge 0\colon n_t = k\}$. Then

$$\sum_{k=0}^\infty \frac{k}{T_{k+1}}1_{\{T_k \le t < T_{k+1}\}} \le t^{-1}n_t \le \sum_{k=0}^\infty \frac{k}{T_k}1_{\{T_k \le t < T_{k+1}\}}.$$

Consequently,

$$\text{a.s.}-\lim_{t\to\infty} t^{-1}n_t = \text{a.s.}-\lim_{t\to\infty}\frac{k}{T_k}.$$

Let $\tau_j = T_j - T_{j-1}$, $j = 1, 2, \cdots$. Then $\{\tau_j\}$ is a sequence of independent random variables, and $E\tau_{2l} = a^{-1}$, $E\tau_{2l+1} = (a + 2b)^{-1}$. Now the strong law of large numbers for independent random variables applies, and we get

$$\text{a.s.}-\lim_{t\to\infty}\frac{T_k}{k} = \text{a.s.}-\lim_{t\to\infty}\frac{1}{k}\sum_{j=1}^k \tau_j = \frac{1}{2}\left(\frac{1}{a} + \frac{1}{a + 2b}\right) = \frac{a + b}{a(a + 2b)}.$$

Collecting the foregoing results we find

$$\text{a.s.}-\lim_{t\to\infty} X_t = \frac{1}{b}\left[\frac{a(a + 2b)}{a + b} - a - b\right] = -\frac{b}{a + b}.$$

The conclusion is that

$$\text{a.s.}-\lim_{t\to\infty}\Psi_t^{-1}\int_0^t \phi_s\phi_s^T\, ds = \frac{1}{3a + b}\begin{bmatrix} a + b & a \\ a & 2a \end{bmatrix} > 0.$$

VI. Remarks

Clearly, condition 3 of the theorem is sufficient to identify all the components of p_0, but it seems that one cannot do without it. The strict positive definiteness of C is lost in either of the following situations that are worked out for $p_0 \in \mathbb{R}^2$. Let $\phi = [\phi_1, \phi_2]$ and let $\lim_{t\to\infty}\phi_{1t}/\phi_{2t} = 0$. Let $p_0 = [p_{01}, p_{02}]^T$. Then

one can expect to identify p_{01}. For suppose $dn_{it} = p_{oi}\phi_{it}\,dt + dm_{it}$, $i = 1,2$, and let $n_t = n_{1t} + n_{2t}$. Then eventually all the observations of n_t are almost entirely those of n_{2t}, which does not yield much information about p_{o1}. Indeed C now becomes $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$. Similarly, if $\lim_{t\to\infty}\phi_{1t}/\phi_{2t} = c \in (0,\infty)$, one can only expect to identify $cp_{01} + p_{02}$.

It might be difficult to check assumptions 2 and 3 of Theorem 2. Assumption 1 will in general be easy to verify. A sufficient condition for assumptions 1 and 2 to hold is, for example, $\phi_t \sim t^a$ $(a > -1/2)$. A necessary condition for assumption 3 is that the eigenvalues of $\int_0^t \phi_s\phi_s^T\,ds$ are of the same order as $t \to \infty$. Assumption 3 is similar to the notion of persistence of excitation that appears in identification problems for ARMAX systems.

Condition 3 of the theorem appears as a technical condition, necessary for the proof of Theorem 2. It seems, however, to be related to

$$\lim_{t\to\infty} \frac{1}{p_0^T\Phi_t}\int_0^t \frac{\phi_s\phi_s^T}{p_0^T\phi_s}\,ds > 0 \quad \text{almost surely} \qquad (6.1)$$

where $\Phi_t = \int_0^t \phi_s\,ds$. Here (6.1) has an appealing interpretation. To see this, define a normalized version of (3.1) by

$$H_t(p) = \frac{1}{p_0^T\Phi_t} J_t(p). \qquad (6.2)$$

Then minimization of $H_t(\cdot)$ is equivalent with minimization of $J_t(\cdot)$. One can easily check that for large t $H_t''(p)|_{p=p_0}$ can be approximated by (6.1). Hence (6.1) says that for $t \to \infty$ p_0 is indeed a minimum point of $H_t(\cdot)$.

We have not discussed the asymptotic distribution of the estimates $\hat{p}_t$ generated by (3.2) and (3.3). This issue will be addressed in another publication.

References

[1] P. Brémaud, *Point Processes and Queues*. New York: Springer, 1981.

[2] C. Dellacherie and P.A. Meyer, *Probabilités et potentiel*. Paris, France: Hermann, 1980, chs. V–VIII.

[3] Yu. Lin'kov, "Estimates of parameters of counting processes," *Prob. Inform. Transmission*, vol. 18, pp. 63–76, 1982.

[4] Y. Ogata, "The asymptotic behavior of maximum likelihood, estimators for stationary point processes," *Ann. Inst. Statist. Math.*, vol. 30, pt. A, pp. 243–261, 1978.

[5] H. Robbins and D. Siegmund, "A convergence theorem for nonnegative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, J.S. Rustagi, Ed. New York; Academic, 1971, pp. 233–256.

[6] J.H. van Schuppen, "Convergence results for continuous-time adaptive stochastic filtering algorithms," *J. Math. Anal, Appl.*, vol. 96, pp. 209–225, 1983.

[7] V. Solo, "The convergence of AML," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 958–962, 1979.

Statistical and Computational Performance of a Class of Generalized Wiener Filters

MARK G. KARPOVSKY, SENIOR MEMBER, IEEE, AND LAZAR A. TRACHTENBERG, MEMBER, IEEE

***Abstract*—A class of suboptimal Wiener filters is considered, and their computational and statistical performances (and the trade-off between the two) are studied and compared with those for known classes of suboptimal Wiener filters. A general model of a suboptimal Wiener filter over a group is defined, which includes, as special cases, the known filters based on the discrete Fourier transform (DFT) in the case of a cyclic group and the Walsh–Hadamard transform (WHT) in the case of a dyadic group. Statistical and computational performances of various group filters are investigated. The cyclic and the dyadic group filters are known to be computationally the best ones among all the group filters. However, they are not always the best ones statistically and other (not necessarily Abelian) group filters are studied. Results are compared with those for the cyclic group filters (DFT), and the general problem of selecting the best group filter is posed. That problem is solved numerically for small-size signals (≤ 64) for the first-order Markov process and random sine wave corrupted by white noise. For the first-order Markov process with the covariance matrix $B^{(s,l)} = \rho^{|s-l|}$ as ρ increases, the use of various non-Abelian groups results in improved statistical performance of the filter as compared to the DFT. Similarly, for the random sine wave with covariance matrix $B^{(s,l)} = \cos\lambda(s-l)$ as λ decreases, non-Abelian groups result in a better statistical performance of the filter than the DFT does. However, that is compensated for by the increased number of computations to perform the filtering.**

Manuscript received October 31, 1983; revised August 8, 1985.

This work was supported in part by the National Science Foundation under Grants DCR-8317763 and ECS-8512748. This work was presented in part at the International Workshop on Fault Detection and Spectral Techniques, Boston, MA, October 12–14, 1983.

M. G. Karpovsky is with the College of Engineering, Boston University, Boston, MA 02215.

L. A. Trachtenberg is with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104.

IEEE Log Number 8406619.

I. Introduction

In recent years interest has grown in utilizing orthogonal transforms in digital signal processing in order to improve statistical or computational performance to permit a trade-off between these two criteria by utilizing a certain chosen orthogonal transform [1], [3], [7], [14].

A common quality shared by many fast transforms which enables their classification (see, e.g., [4], [5]) is that they can be represented as Kronecker products of matrices which may or may not be sparse or structured. By virtue of this Kronecker product representation new transforms can be generated from old ones simply by using the Kronecker product. In a given problem, such as Wiener filtering with given statistical characteristics of a signal and noise, one can *select* a computationally good approximating transform to a statistically optimal transform and the selection can be done out of the family of known fast transforms with a Kronecker product representation. (See [10], where a good reference list can be found, and [1].)

Another approach to the same problem of Wiener filtering would be to *construct* a computationally good approximation to a given statistically optimal transform. A possibility of solving that problem analytically for classes of signals defined by their covariance matrices (e.g., for signals whose covariance matrices are Toeplitz) has been pointed out in [12], [18], [19], [28] and this approach deserves further elaboration. Yet another approach is to *construct experimentally* a computationally good approximating transform to a transform which is known to be good statistically. For example, the discrete cosine transform (DCT) has a nearly optimal statistical performance for highly correlated Markov signals (see [24]), and it has recently been approximated by computationally convenient transforms [8]. Here even for small n (up to 32 vector-components of a signal) the problem is difficult, involves tedious trial and error procedures, and requires artistry rather than clear-cut methods. Another disadvantage is that a success with approximating one transform (as DCT) for some n (say $n = 16$, 32) cannot be generalized to be used to approximate other transforms [8], [26].

A number of researchers [15], [1], [3], [11], [17] have selected a family of fast transforms which are group theoretic by their nature; i.e., they are based on group characters of corresponding Abelian groups: examples are the discrete Fourier transform (DFT) in the case of a cyclic group and the Walsh–Hadamard transform (WHT) (or simply the Walsh transform) in the case of a dyadic group [1], [3], [11], [15]–[17], [27]. The use of non-Abelian groups was discussed in [13], [20].

These transforms exist for any number n, are computed analytically by formulas, and possess Kronecker product representations (which guarantee speed of computation for nonprime

0018-9448/86/0300-0303$01.00