This is to certify that the

dissertation entitled

Filtering For Some Stochastic Processes
With Discrete Observations

presented by

Oleg Makhnin

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics

Co-Major professor
Anatoli Skorokhod

Date July 22, 2002

Co-Major Professor
Habib Salehi

FILTERING FOR SOME STOCHASTIC PROCESSES WITH DISCRETE OBSERVATIONS

By

Oleg V. Makhnin

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

2002

ABSTRACT

FILTERING FOR SOME STOCHASTIC
PROCESSES WITH DISCRETE OBSERVATIONS

By

Oleg V. Makhnin

The processes in question are jump processes and processes with jumping velocity. We estimate the current position of the stochastic process based on past discrete-time observations (non-linear discrete filtering problem). We obtain asymptotic rates for the expected square error of the filter when observations become frequent. These rates are better than those of a linear Kalman filter. For jump process, our method is asymptotically free of the process parameters. Also, estimation of process parameters is addressed.

Acknowledgments

The author would like to thank his advisor, Prof. Skorokhod, for setting up the problem and working with me patiently, providing guidance and insight into the problem.

I also thank my committee members Prof. Gilliland, Prof. LePage and Prof. Salehi whose contributions have been invaluable. Special thanks to Prof. LePage for his help with the manuscript and to Prof. Salehi for his advice and supervision.

I would also like to express my gratitude to Prof. Koul, Prof. Mandrekar, Prof. Page and all other professors at Michigan State University from whom I have learned a lot.

Last, but not least, my gratitude to my wife who supported me throughout this endeavor.

List of Tables

List of Figures

Contents

1 Introduction

This work deals with estimating the position of a stochastic process based on past observations (filtering). With respect to square error, the optimal non-linear filter is the conditional expectation of the current state of the process given the observations. The observations are discrete, and we are interested in the asymptotic behavior of the non-linear filter as these observations become more frequent.

Several factors may affect the asymptotic behavior of a non-linear filter. One of them is the nature of the process itself. The more irregular a process is, the harder it will be to filter. Another factor might be the distribution of observation errors.

The simplest example of such results is the estimation of the mean of a sequence of i.i.d. variables. One can think about this mean as a "process" that remains constant over time. Assume that the variables have a density with respect to Lebesgue measure on $\mathbf{R}$. As pointed out in a book by Ibragimov and Khas'minskii [13], the quality of the estimate depends on whether or not this density is continuous. In the case of a density with discontinuities, the phenomenon of "hyper-efficiency" occurs. One gets different results, for example, in cases of normal distribution and uniform distribution.

1.1 Review of past results

Filtering is a major area of stochastic process theory. This has been progressing rapidly over the last 40+ years, starting with Kolmogorov and Wiener. A great deal of attention has been paid to the filtering with continuous-time observations that typically involves stochastic differential equations. Among the major contributions here are R. Kalman and R. Bucy (1961)[18], A. Shiryaev (1966)[25], T. Kailath (1968)[14], M. Zakai (1969)[27], G. Kallianpur and C. Striebel (1969)[16], G. Kallianpur (1980)[15], B. Rozovskii (1990) [23]. In most of these works, the observation noise is a Wiener process, or, more generally, the observation process satisfies a stochastic differential equation driven by a Wiener process.

Filtering with discrete-time observations was considered by Kalman (1960) [17] and continued in multiple works, including Brémaud (1981)[3]. After the pioneering work by Kalman, a lot of attention has been paid to linear filters, which are linear combinations of observed values. Many works have been devoted to the theory of Kalman filter, for example, Anderson and Moore (1979)[1]. Lately, as the computing facilities have improved greatly, the focus has shifted to non-linear filters, which typically perform much better. Comparison with linear filters is one of topics in this work.

Yashin(1970)[26] derived the optimal non-linear filter for situation when the pro-

cess $X(t)$ is Markov taking values 0 and 1, and the observations are also 0 or 1. This situation is expanded in the book by Elliott, Aggoun and Moore (1995)[10] in the context of Hidden Markov Models. Their approach has become popular recently and involves a change of measure, rendering the observations independent of the process in question. One then arrives at a discrete-time version of Zakai equation, which presents a recursive way to compute the optimal filter. This was used, for example, in Dufour and Kannan (1999)[9] and Kannan(2000)[19].

The recent papers, to mention a few, are Portenko, Salehi and Skorokhod (1997)[21], Ceci, Gerardi and Tardelli (2001)[5], Del Moral, Jacod, Protter (2001)[7]. The latter deals with Monte-Carlo methods for estimating the optimal filter, even in case when no explicit expressions for the filter are available. Monte-Carlo methods are also discussed by Gordon et al. (1993)[12] and Doucet, de Freitas and Gordon (2001)[8].

Relatively little is known about the asymptotic behavior of filtered estimates as observations become frequent. Some results on this are given in [22]. This work considers asymptotics for certain classes of stochastic processes. They include compound Poisson processes and piecewise-linear processes.

The discrete observations are natural in target-tracking, when the process in question is a position of a target, and our observations come from a radar. A special case (with uniform errors) was considered by Portenko, Salehi and Skorokhod (1998) [22],

although they introduce many extra features useful for target-tracking like multi-targets and false targets.

A general exposition of filtering techniques employed can be found in [21]. A detailed overview of the target-tracking from an engineer's prospective is given by Bar-Shalom et al. [2].

1.2 Hyper-efficiency

The results for parameter estimation in i.i.d. case are well desribed in [13]. They can be summarized as follows. Suppose that $\{Y_k\}_{k=1,...,n}$ is a sequence of i.i.d. random variables with density f_θ, depending on parameter θ.

a) Suppose that f_θ is continuously differentiable, with a several additional regularity conditions, including local asymptotic normality for the family $\{f_\theta\}_{\theta\in\Theta}$. Then, both Bayesian and maximum likelihood estimates are asymptotically normal with the rate $\mathbf{E}(\hat{\theta}-\theta)^2 = C/n + o(1/n)$.

b) Suppose that the densities f_θ possess jumps at the finite number of points $x_1(\theta), ..., x_k(\theta)$ and are continuously differentiable elsewhere, plus some identifiability and regularity conditions. The earliest treatment of such a problem known to the author is Chernoff and Rubin [6]. In this case both Bayesian and maximum likelihood estimates have the rate $\mathbf{E}(\hat{\theta}-\theta)^2 = C/n^2 + o(1/n^2)$. That is, the estimates are "hyper-

efficient".

An important special case is when the location parameter is estimated, that is when $f_\theta(x) \equiv f(x-\theta)$. The difference between the above two cases can be illustrated using normal distribution in case (a) and uniform distribution in case (b). For normal distribution, the mean of observed values is a natural estimator with the expected square error $O(1/n)$. For uniform distribution on the interval $[\theta - a, \theta + a]$, the estimator $[max(Y_k) + min(Y_k)]/2$ with the expected square error $O(1/n^2)$ is a better estimator for θ than the mean. Thus, for a density with jumps, the best location estimator is a function of observations near a discontinuity point.

A generalization of these results to multi-dimensional variables and vector parameter θ is published by Ermakov [11]. (This problem also traces to Rubin [24].) When the error density f_θ has discontinuities along a smooth manifold $\mathcal{S}_\theta$, then both Bayesian and maximum likelihood estimates for θ have asymptotic square error of order $1/n^2$. He also has some results on the sequential estimation of such parameters.

The way our problem is different lies in the stochastic-process perspective. We are estimating not a stable, unchanging parameter, but a value of some stochastic process in time.

1.3 Formulation of the problem

Process $X(t)$ is a real-valued stochastic process on the interval $[0,T]$ or $[0,\infty)$, depending on the context. In general, suppose that the apriori distribution of the entire process $X(t,\omega)$ has a density $\pi(X(\cdot))$ with respect to measure ν in a suitable function space $\mathcal{F}[0,T]$. Also let the distribution of observations $\mathcal{L}(observations|X(\cdot))$ have a density f in some space of observations. Keep the notation of $X(\cdot)$ for the entire trajectory of the process X.

In this work, we use two approaches:

a) when estimating the current position of the process $X(t)$ at a given time t, use the information obtained up to this instant ("filtering"), regardless of whether or not the trajectory comprised of resulting estimates $\hat{X}(t)$ belongs to the specific family $\mathcal{F}$, and

b) when estimating the parameters of the process itself, based on observations in the interval $[0,T]$, try to produce a process that belongs to the family $\mathcal{F}$ ("jump process" below), that is reasonably "close" to $X(\cdot)$.

Under certain conditions, prior distribution π_t of X(t), consistent with $\pi(X(\cdot))$, will have a density with respect to Lebesgue measure. As an estimate of process' position, we use the posterior (with respect to π_t) conditional expectation of $X(t)$ given all the observations up to the time t. It is well known that such estimator

minimizes the squared error of estimation.

Consider two varieties of process X:

- "Jump process". Consider a compound Poisson process

$$X(t) = X_0 + \sum_{i:s_i \le t} \xi_i$$

where (s_i, ξ_i) are the events of a 2-dimensional Poisson process on $[0,T] \times \mathbf{R}$. The intensity of this process is given by $\lambda(s,y) = \lambda h(y)$, where $\lambda > 0$ is a constant "time intensity" describing how frequently the jumps of X occur and $h_\theta(y)$ is the "jump density" describing magnitudes of jumps ξ_i of process X. Here $\theta \in \Theta$ is a parameter (possibly unknown) defining the distribution of jumps. In the Bayesian formulation, parameters θ and λ will have a prior density $\pi(\theta, \lambda)$ with respect to Lebesgue measure. Assume that for each $\theta \in \Theta$, $\mathbf{E}\xi_1^2 < \infty$.

Also, assume that starting value X_0 has some prior density $\pi_{X_0}()$.

- "Piecewise Linear Process".

This is a process with jumping velocity V

$$X(t) = X_0 + \int_0^t V(s)ds$$

with V being a compound Poisson process described above.

Assume that the prior distribution of X_0, V_0 is known.

Observations

Our observations $\{Y_j\}$ are always going to be "Process+noise" over a finite grid of values:

$$Y_j = X(j/n) + e_j$$

where the noise variables $\{e_j\}$ are i.i.d. with some density ϕ_θ, possibly depending on a parameter θ and independent of process X. Assume that for each $\theta \in \Theta$, $\mathbf{E}e_1 = 0$, and $\mathbf{E}e_1^2 < \infty$.

A sample path of a Jump Process and observations are given below.

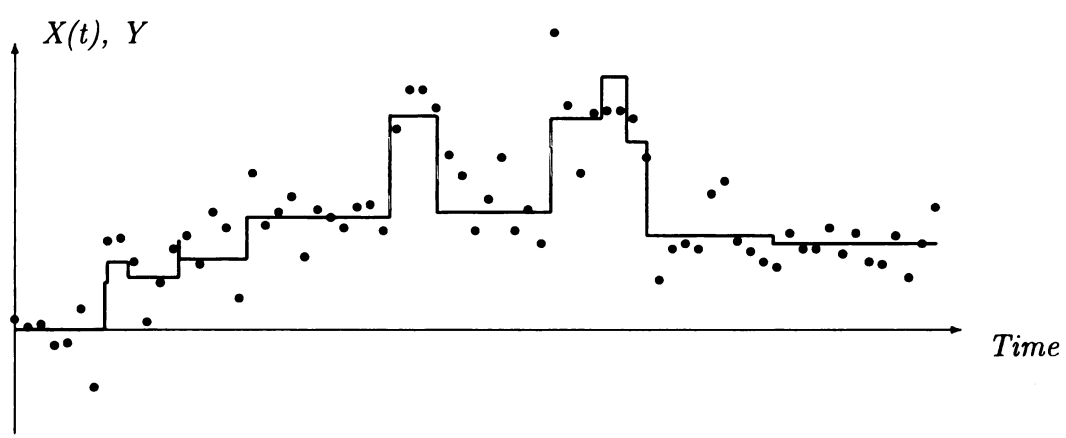


Figure 1

1.4 Main results

We will consider asymptotics when $n \to \infty$: as the observations become frequent, but the process changes slowly (rate of change λ is bounded from above). When the process $X(\cdot)$ is changing fast, its diffusion approximations become appropriate, but

these are not discussed here.

We establish the following results:

- Recursive formulas for conditional density of process position given the observations.

- Asymptotic rates for the estimate of position of jump (compound Poisson) process.

- Asymptotic behavior of the parameters of jump process, as both observation frequency and total time spent observing become large.

- Asymptotic rates for a piecewise-linear process.

- Comparison with linear filters. Simulation results in a small-sample setting.

2 Filtering of a jump process

The filtering and Bayesian estimation problem can be formulated as follows: *find the conditional distribution of the states of the process X and unknown parameters λ, θ given the observations (Y_i), initial distirbution of $X(0) = X_0$ and some prior distribution on the unknown parameters.*

2.1 Recursive formulation

The results in this section are in spirit of Elliott et al. [10]. In the future, use $\tau = 1/n$ as the time between observations.

Denoting $X_k := X(\tau k)$, we have

$$X_{k+1} = X_k + \zeta_{k+1} \tag{1}$$

$$Y_k = X_k + e_k$$

ζ_k is a sum of jumps of X on the interval $[\tau(k-1), \tau k)$:

$$\zeta_k = \sum_{\tau(k-1) \le s_i < \tau k} \xi_i$$

Thus, $\{\zeta_k\}_{k \ge 1}$ are i.i.d. with an atom of mass $e^{-\lambda\tau}$ at 0 and the rest of the mass having (improper) density $\tilde{\psi} = \tilde{\psi}_{\theta,\lambda}$ expressible in terms of the original density of jumps h_θ. To simplify the notation in the sequel, I will call ψ a "density", actually

meaning that $\psi(0)$ is a scaled δ-function, that is for any function g,

$$\int g(x)\psi(x)dx := e^{-\lambda\tau} \cdot g(0) + \int g(x)\tilde{\psi}(x)dx$$

Also, subscripts θ, λ in ϕ_θ, $\psi_{\theta,\lambda}$ will be omitted.

Suppose the priors are given:

$$\theta, \lambda \textit{ have density } \pi(\cdot,\cdot),$$

$$X_0 \textit{ has density } \pi_{X_0}(\cdot).$$

Our goal is to find the posterior conditional distribution

$$\mathcal{L}_\pi(X_k \mid Y_1, ..., Y_k).$$

From (1), we obtain the densities

$$p_{Y_1,...,Y_k}(y_1, ..., y_k \mid X_0, ..., X_k, \theta) = \prod_{j=1}^{k} \phi(y_j - X_j)$$

$$p_{X_0,...,X_k}(x_0, ..., x_k \mid \theta, \lambda) = \pi_{X_0}(x_0) \prod_{j=1}^{k} \psi_{\theta,\lambda}(x_j - x_{j-1}).$$

For briefness, let's denote

$$\mathbf{x}_k := (x_0, ..., x_k), \quad \mathbf{X}_k := (X_0, ..., X_k), \quad \mathbf{Y}_k := (Y_1, ...Y_k).$$

Now, applying Bayes' Theorem, the joint density of $\mathbf{X}_k$, $\mathbf{Y}_k$, θ and λ is

$$p(\mathbf{x}_k, \mathbf{y}_k, \theta, \lambda) = p(\mathbf{y}_k \mid \mathbf{x}_k, \theta, \lambda) \cdot p(\mathbf{x}_k \mid \theta, \lambda) \cdot \pi(\theta, \lambda) =$$

$$= \pi(\theta, \lambda) \cdot \pi_{X_0}(x_0) \cdot \prod_{j=1}^{k} \phi(y_j - x_j) \cdot \prod_{j=1}^{k} \psi(x_j - x_{j-1}), \qquad (2)$$

and the conditional density given the observations

$$p(\mathbf{x}_k, \theta, \lambda \mid \mathbf{Y}_k) = \frac{p(\mathbf{x}_k, \mathbf{Y}_k, \theta, \lambda)}{\int_{\mathbf{R}^k} \int_{\Theta} \int_{\mathbf{R}} p(\mathbf{x}_k, \mathbf{Y}_k, \theta, \lambda) d\mathbf{x}_k \, d\theta \, d\lambda}.$$

Introduce

$$q_k(x, \theta, \lambda) := \int_{\mathbf{R}^k} p(\mathbf{x}_k, \mathbf{Y}_k, \theta, \lambda) \, dx_0 .... dx_{k-1}.$$

It is an unnormalized density of the latest state X_k and parameters θ, λ given the observations $\mathbf{Y}_k$. The normalized density $p_k(x, \theta, \lambda)$ is then given by

$$p_k(x, \theta, \lambda) := \int p(\mathbf{x}_k, \theta, \lambda \mid \mathbf{Y}_k) dx_0 .... dx_{k-1} = \frac{q_k(x, \theta, \lambda)}{\int_{\mathbf{R}} \int_{\Theta} \int_{\mathbf{R}} q_k(x, \theta, \lambda) dx \, d\theta \, d\lambda}.$$

The reason we use this density in an unnormalized form is the recursive relation:

Theorem 1

$$q_0(x, \theta, \lambda) = \pi_{X_0}(x) \cdot \pi(\theta, \lambda),$$

$$q_k(x, \theta, \lambda) = \phi_\theta(Y_k - x) \cdot \int_{\mathbf{R}} \psi_{\theta,\lambda}(x - z) q_{k-1}(z, \theta, \lambda) \, dz = \tag{3}$$

$$= \phi_\theta(Y_k - x) \cdot [e^{-\lambda \tau} q_{k-1}(x, \theta, \lambda) + \int_{\mathbf{R}} \tilde{\psi}_{\theta,\lambda}(x - z) q_{k-1}(z, \theta, \lambda) \, dz].$$

Proof: Straightforward, follows from integrating (2)

□

Remark.

1. In order to use Theorem 1 for the estimation of state X_k, we will compute

$q_j(x,\theta,\lambda), j \le k$ consecutively, then compute marginal unnormalized density $q_k(x) :=$ $\int q_k(x,\theta,\lambda)\, d\theta\, d\lambda$ and then find

$$\hat{X}_k := \mathbf{E}(X_k|\mathbf{Y}_k) = \frac{\int_{\mathbf{R}} x q_k(x) dx}{\int_{\mathbf{R}} q_k(x) dx}. \tag{4}$$

2. Although not derived explicitly, the unnormalized density q has to do with a change of the original probability measure to, say, Q, which makes the observations $Y_1, ..., Y_k$ independent of the process $X(t)$. This way, prior distributions on (θ, λ) and $X(0)$ ensure that the two measures are absolutely continuous with respect to each other. The change of measure approach is used extensively in non-linear filtering.

The recursive formulas for the densities can be used to compute "on-line" updates as new observations are coming in.

2.2 Single-block upper bound for expected square error.

Next, we investigate asymptotic properties of the above filtering estimator $\hat{X}(T) :=$ $\hat{X}_{nT}$ as the observations become frequent ($n \to \infty$). First, we will produce a sub-optimal estimator of $X(T)$ based on a single "block" of observations at time points immediately preceding T.

Assume that the last observation is obtained exactly at the moment T. Denote

$$\ell_T(\tau) := \mathbf{E}(\hat{X}(T) - X(T))^2.$$

The following discussion is based on the well-known fact (e.g. see [3, p. 84])

Lemma 1 *For a square-integrable random variable X, sigma-algebra $\mathcal{F}$ and an $\mathcal{F}$-measurable random variable U,*

$$\mathbf{E}[X - \mathbf{E}(X|\mathcal{F})]^2 \leq \mathbf{E}(X - U)^2$$

□

Setting $\mathcal{F} := \sigma\{Y_1, ..., Y_k\}$, we can see that the filtered estimator $\hat{X}_k$ introduced by (4) has the smallest expected square loss among all possible estimators of X_k based on observations $\mathbf{Y}_k$.

To produce an upper bound on $\ell_T(\tau)$, consider the following sub-optimal estimator of $X(T)$:

$$\overline{Y}_k(\Delta) := \sum_{k-n\Delta<j\leq k} Y_j/(n\Delta),$$

where Δ is the block length to be specified later. Here, $k = k(\tau) = T/\tau$, so that $X(T) = X_k$.

Theorem 2 . Asymptotic upper bound for $\mathbf{E}(\hat{X} - X)^2$

As $\tau \to 0$,

$$\ell_T(\tau) \leq (\lambda \mathbf{E}\xi_1^2 + \mathbf{E}e_1^2)\sqrt{\tau} + o(\sqrt{\tau}) \tag{5}$$

Proof: Consider the estimate $\overline{Y}_k(\Delta)$ introduced above. By Lemma 1, it is no better than $\hat{X}(T)$, that is

$$\ell_T(\tau) \leq \mathbf{E}[X(T) - \overline{Y}_k(\Delta)]^2.$$

Suppose that the process X has m jumps on the interval $(T-\Delta, T]$, with the locations of jumps $\tilde{s}_1, ..., \tilde{s}_m$ and the heights of jumps $\tilde{\xi}_1, ..., \tilde{\xi}_m$.

Denote

$$S_0 := X(T-\Delta),$$

$$S_j := S_{j-1} + \tilde{\xi}_j, \quad 1 \le j \le m$$

consecutive values taken by $X(t)$ for $t \in (T-\Delta, T]$, and $S_m \equiv X(T)$.

Note that

$$\mathbf{E}\, max_j (S_m - S_j)^2 \le \sum_j \mathbf{E}(S_m - S_j)^2 = \frac{m(m+1)}{2} \mathbf{E}\xi_1^2.$$

Therefore,

$$\mathbf{E}[X(T) - \overline{Y}_k(\Delta)]^2 \le \frac{\mathbf{E}e_1^2}{n\Delta} + e^{-\lambda\Delta}\mathbf{E}\xi_1^2 \left[\lambda\Delta + ... \frac{(\lambda\Delta)^m}{m!} \times \frac{m(m+1)}{2} + ...\right]$$

Setting $\Delta = \tau^b$ for some $0 < b < 1$, the above becomes

$$= \mathbf{E}e_1^2 \cdot \tau^{1-b} + (1 - \lambda\tau^b)\mathbf{E}\xi_1^2 \left[\lambda\tau^b + o(\tau^b)\right].$$

Setting $b = 1/2$, we obtain the statement of the Theorem.

□

Remark. Note that since the estimating procedure we used did not depend on θ, the above Theorem is also true when the parameter θ is unknown. In that case, one needs to consider Bayesian loss

$$\ell^B_{T,\pi}(\tau) = \int_\Theta \mathbf{E}_\theta (\hat{X}(T) - X(T))^2 \pi(\theta)\, d\theta$$

and, integrating (5), obtain the bound

$$\ell^B_{T,\pi}(\tau) \le \sqrt{\tau} \int_\Theta (\lambda \mathbf{E}\xi_1^2 + \mathbf{E}e_1^2)\pi(\theta)\, d\theta + o(\sqrt{\tau})$$

To produce finer approximations, we have to assume the knowledge of the error distribution.

2.3 Multiple-block upper bound

Next, we modify our estimating procedure. Starting with time T, we will probe one block of observations after another, stopping whenever we believe that a jump has occurred.

The following results were obtained when the error distribution is considered known. Denote $\sigma_e := \sqrt{\mathbf{E}e_1^2}$.

We use the same idea as before: produce a sub-optimal estimate for $X(T)$ based on $\overline{Y}$ for a suitable interval. The difficulty lies in not knowing where exactly the last jump of process X occurred. Consider the intervals (blocks) $(T_1, T_0], (T_2, T_1], ..., (T_N, T_{N-1}]$, where

$$T_0 := T$$

$$T_j := T_{j-1} - (ln\, n)^j/n, \quad j = 1, ..., N$$

$$T_{N+1} := 0.$$

There is a total of

$$N = \frac{ln\, n}{ln\, ln\, n} - 1$$

blocks; j-th block has length $(ln\, n)^j/n$ and $n_j := (ln\, n)^j$ observations. The last block has length $1/ln\, n$.

Let X_j be the value of the process at the end of j-th block, that is $X_j := X(T_{j-1})$. Let $\overline{Y}_j$ be the average of observations on the block j, that is

$$\overline{Y}_j := n_j^{-1} \sum_k Y_k\, I(T_j < k\tau \leq T_{j-1}).$$

Assumption 1 . *Let*

$$\chi_m := \frac{\sum_{k=1}^m e_k}{\sigma_e \sqrt{m}}$$

be the normalized sum of m errors. Assume that for the distribution of errors e_k the following is true. There exist constants C_1, C_2, C_3 and $K > 0$ such that for all sufficiently large m and all integers j

$$\mathbf{E}[\chi_m^2\, I(|\chi_m| > C_1 \cdot m^{1/Kj})] < C_2\, exp(-C_3\, m^{1/j}).$$

This assumption is satisfied for Normal errors with $K = 2$; in general, it requires e_k to have small tails.

The following is a simpler-looking but more restrictive than Assumption 1:

Assumption 1' . *For χ_m given above, there exist constants $G, \gamma > 0$ such that for*

all sufficiently large m,

$$\mathbf{E}\, exp(\gamma|\chi_m|) \leq G.$$

Proposition 1 *Assumption 1' implies Assumption 1 with* $K = 1$.

Proof:

Suppose that Assumption 1' is satisfied. Let $F_m(.)$ be the distribution function of χ_m. Pick C_1 such that $x^2 < exp(\gamma|x|/2)$ for $|x| > C_1$.

Then for any j,

$$\int_{\mathbf{R}} I\{|x| > C_1 m^{1/j}\}\, x^2\, dF_m(x) \leq \int_{\mathbf{R}} I\{|x| > C_1 m^{1/j}\}\, e^{\gamma|x|/2} dF_m(x) \leq$$

$$\leq exp(-\gamma C_1 m^{1/j}/2) \int_{\mathbf{R}} e^{\gamma|x|} dF_m(x) \;\leq$$

$$\leq \; exp(-\gamma C_1 m^{1/j}/2) \cdot G$$

□

Theorem 3 . Tighter upper bound for $\mathbf{E}(\hat{X}_n(T) - X(T))^2$

Suppose that the error density ϕ *is known and does not depend on the parameter* θ, *and there exists a constant* Λ_0 *such that* $\lambda \leq \Lambda_0$. *Then, under Assumption 1, there exists a constant* C *such that for* $n \to \infty$,

$$\mathbf{E}(\hat{X}_n(T) - X(T))^2 \leq C \frac{ln^M n}{n}$$

with $M = (1 + 2/K) \vee (3 - 2/K)$.

Proof:

Consider $N-1$ blocks as described above. Denote $T^\star$ the point of last jump of X:

$$T^\star = sup\ \{0 \le t \le T : X(t) - X(t-) > 0\}.$$

The idea is to approximate $T^\star$, then take the average of all observations from that moment up to T.

Construct an estimate of $X(T)$ as follows.

Define j_0 as

$$j_0 := inf\{j > 0 : \sqrt{n_j}\,\frac{|\overline{Y}_j - \overline{Y}_{j+1}|}{\sigma_e} > 2C_1 \cdot n_j^{1/Kj}\} \wedge N. \tag{6}$$

Then, as our estimate of $X(T)$, take

$$\tilde{X}(T) := \overline{Y}_{j_0}.$$

We will find an upper bound for the average risk of this estimate, $\ell := \mathbf{E}(\tilde{X}(T) - X(T))^2$. For this, we will need several inequalities, with proofs to follow in the next section.

Case 1 . Block 1 jump

On the event F that the last jump of X occurred on Block 1, $F = \{T_1 < T^\star\}$,

$$\ell_F := \mathbf{E}[(\tilde{X}(T) - X(T))^2 I_F] \le C_3 \frac{ln\, n}{n} \tag{7.1}$$

Case 2 . Correct stopping

In the event S that the last jump of X occurred just before the Block j_0, $S = \{T_{j_0+1} < T^\star \le T_{j_0}\}$

$$\ell_S := \mathbf{E}[(\tilde{X}(T) - X(T))^2 I_S] \le C_4 \frac{ln^2 n}{n} \tag{7.2}$$

Case 3 . Late stopping

In the event L that the last jump of X occurred in Block j, $1 < j \le j_0$,

$L = \{T_{j_0} < T^\star\}$

$$\ell_L := \mathbf{E}[(\tilde{X}(T) - X(T))^2 I_L] \le C_5 \frac{(ln\, n)^{1+2/K}}{n} \tag{7.3}$$

Case 4 . Early stopping

In the event $\mathcal{E}$ that we stopped on the Block j_0 but there was no jump of X, $\mathcal{E} = \{T^\star \le T_{j_0+1}\}$

$$\ell_\mathcal{E} := \mathbf{E}[(\tilde{X}(T) - X(T))^2 I_\mathcal{E}] \le C_6 \frac{(ln\, n)^{3-2/K}}{n} \tag{7.4}$$

Now note that $P(F \cup S \cup L \cup \mathcal{E}) = 1$. Thus, $\ell = \ell_F + \ell_S + \ell_L + \ell_\mathcal{E}$. Also, the estimator $\tilde{X}$ does not depend on λ and particular form of jump density h_θ, as long as the frequency of jumps λ is bounded.

By Lemma 1, the risk of estimate $\hat{X}$ does not exceed the risk of $\tilde{X}$. Combining (7.1) through (7.4), we obtain the proof of the Theorem.

□

2.4 Proofs of inequalities used in Theorem 3

Proof of (7.1)

The probability of jump on the first Block, which has length $ln\, n/n$ is

$P(F) = ln\, n/n + o(ln\, n/n)$, and probability of more than one jump on the first Block is $o(ln\, n/n)$. Therefore,

$$\mathbf{E}[(\tilde{X}(T) - X(T))^2 I_F] \leq (\sigma_e^2/ln\, n + \mathbf{E}\xi_1^2) ln\, n/n \leq C_3 \frac{ln\, n}{n}.$$

□

Proof of (7.2)

Let $j_0(\omega)$ be, as before, the last Block included in the computation of $\tilde{X}(T)$. First, consider the special case $j_0 = N$. Then

$$\mathbf{E}[(\tilde{X}(T) - X(T))^2 I_S I(j_0 = N)] \leq$$

$$\leq \sigma_e^2/(n/ln\, n) \cdot (1/ln\, n + o(1/ln\, n)) \leq const/n$$

Now let $j_0 < N$.

Suppose that the last jump $T^\star(\omega)$ occurred on the Block $j_0 + 1$, that is, $\omega \in S$. Then $X(t) = X(T)$ for $T_{j_0} < t < T$, and the squared loss from estimating $X(T)$ equals the variance of $\overline{Y}_{j_0}$, so that

$$\mathbf{E}[(\tilde{X}(T) - X(T))^2 I_S I(j_0 < N)] \leq \sum_{j=1}^{N} \mathbf{E}[(\overline{Y}_j - X(T))^2 I_S] \leq$$

$$\leq \sum_{j=1}^{N} P(T_{j+1} < T^\star \leq T_j) \cdot \sigma_e^2/n_j$$

by independence of $\{e_k\}$ and process X. Thus,

$$\ell_S \le const/n + \sum_{j=1}^{N} (ln^{j+1}n/n + o(ln^{j+1}n/n)) \cdot \sigma_e^2 \, ln^{-j}n \le C_4 \frac{ln^2 n}{n}.$$

□

Proof of (7.3)

Thanks to (7.1), we can exclude the case when the last jump $T^\star$ happens on Block 1. Therefore, suppose that the last jump happens on Block J, $J > 1$, but we stop the summation only at Block j_0, $j_0 \ge J$.

Denote $N_J := \{last\ jump\ happens\ on\ Block\ J\}$. Our stopping rule (6) implies that for $J \le j \le j_0$,

$$|\overline{Y}_j - \overline{Y}_{j-1}| \le 2\sigma_e C_1 n_{j-1}^{-1/2+1/(j-1)K}$$

Thus,

$$\begin{aligned}
&\mathbf{E}[(\tilde{X}(T) - X(T))^2 I_{N_J}] \le \\
&\le \mathbf{E}(|X(T) - \overline{Y}_{J-1}|^2 + \sum_{j=J}^{j_0} |\overline{Y}_j - \overline{Y}_{j-1}|^2) \cdot P(jump\ on\ Block\ J) \le \\
&\le \left[\frac{ln^J n}{n} + o(\frac{ln^J n}{n})\right] \times \left[\sigma_e^2 ln^{-(J-1)}n + C_7 \sum_{j=J}^{N} n_{j-1}^{-1+2/(j-1)K}\right] \le \\
&\le \sigma_e^2 \frac{ln\, n}{n} + C_7 \frac{(ln\, n)^{1+2/K}}{n} \le C_5 \frac{(ln\, n)^{1+2/K}}{n}
\end{aligned}$$

□

Proof of (7.4)

If the stopping occurred too early then $X(t) = X(T)$ for $T_{j_0+1} < t < T$. Also, the

stopping rule (6) implies that at least one of

$$|\overline{Y}_{j_0+1} - X(T)| > C_1\sigma_e n_{j_0}^{-1/2+1/j_0K},$$

$$|\overline{Y}_{j_0} - X(T)| > C_1\sigma_e n_{j_0}^{-1/2+1/j_0K}$$

is true. Thus,

$$\mathbf{E}[(\tilde{X}(T) - X(T))^2 I_{\mathcal{E}}] \leq$$

$$\leq \mathbf{E}|\overline{Y}_{j_0} - X(T)|^2 \cdot P(|\overline{Y}_{j_0+1} - X(T)| > C_1\sigma_e n_{j_0}^{-1/2+1/j_0K}) +$$

$$+\mathbf{E}\left(|\overline{Y}_{j_0} - X(T)|^2 I(|\overline{Y}_{j_0} - X(T)| > C_1\sigma_e n_{j_0}^{-1/2+1/j_0K})\right) \equiv E_1 + E_2.$$

By Assumption 1, $E_2 \leq NC_2 exp(-n_j^{1/j}) \leq C_2(ln\, n)/n$.

To estimate E_1, consider the Chebyshov-type inequality

$$\left(C_1\sigma_e n_{j_0+1}^{-1/2+1/j_0K}\right)^2 P(|\overline{Y}_{j_0+1} - X(T)| > C_1\sigma_e n_{j_0}^{-1/2+1/j_0K}) \leq$$

$$\leq \mathbf{E}\left(|\overline{Y}_{j_0+1} - X(T)|^2 I(|\overline{Y}_{j_0+1} - X(T)| > C_1\sigma_e n_{j_0+1}^{-1/2+1/j_0K})\right) \leq C_2(ln\, n)/n$$

by Assumption 1. Therefore

$$E_1 \leq \sum_{j=1}^{N-1} \sigma_e^2/n_j \cdot C_2(ln\, n)/n \cdot \left(C_1\sigma_e n_{j+1}^{-1/2+1/jK}\right)^{-2} \leq$$

$$\leq C_6 \sum_j \frac{\sigma_e^2}{ln^j n} \cdot \frac{ln\, n}{n} \cdot (ln\, n)^{(1-2/jK)(j+1)} \leq$$

$$\leq C_6 \sum_j \frac{(ln\, n)^{2-2/K}}{n} \leq C_6 \frac{(ln\, n)^{3-2/K}}{n}$$

□

2.5 Lower bound for expected square error.

Let us, as before, have exactly n observations on the interval $[0,T]$ and the last observation is made at the moment T. Then $X(T) = X_n$.

To estimate the expected squared loss of the Bayesian estimate $\hat{X}_n$ from below, consider the estimator

$$\check{X}_n = \mathbf{E}(X_n|Y_1, ..., Y_n, X_{n-1}, I_n),$$

where $I_n = I(X_n \neq X_{n-1})$ is indicator of the event that some jumps occurred on the last observed interval.

It's easy to see that $\check{X}_n = \mathbf{E}(X_n|Y_n, X_{n-1}, I_n)$ and that $\mathbf{E}(\check{X}_n - X_n)^2 \leq \mathbf{E}(\hat{X}_n - X_n)^2$, since the estimator $\check{X}_n$ is based on a finer sigma-algebra.

Proposition 2 . *The expected square error for* $\check{X}_n$,

$$\mathbf{E}(\check{X}_n - X_n)^2 = C/n + o(1/n),$$

where $C > 0$ *is some constant not depending on* n.

Proof:

Consider random variables

$$Z_n \sim \tilde{\psi}$$

so that $X_n = X_{n-1} + I_n Z_n$, and

$$W_n = Z_n + e_n.$$

Joint distribution of Z_n, W_n does not depend on n and

$$P(Z_n \in dx, W_n \in dy) = \tilde{\psi}(x)\phi(y-x).$$

Also note that on the event $\{I_n = 0\}$, $X_n = X_{n-1}$ and on the event $\{I_n = 1\}$, $Y_n = X_{n-1} + W_n$. Therefore,

$$\check{X}_n = \mathbf{E}(X_n | Y_n, X_{n-1}, I_n) = X_{n-1} + I_n \mathbf{E}(Z_n \mid W_n).$$

Let $\hat{Z}_n := \mathbf{E}(Z_n \mid W_n)$. Then

$$\mathbf{E}(\check{X}_n - X_n)^2 = P(I_n = 1)\mathbf{E}(\hat{Z}_n - Z_n)^2.$$

Clearly, $\mathbf{E}(\hat{Z}_n - Z_n)^2 > 0$ and $P(I_n = 1) = 1 - e^{-\lambda/n} = \lambda/n + o(n^{-1})$. This gives us the statement of Proposition with $C = \lambda \cdot \mathbf{E}(\hat{Z}_n - Z_n)^2$.

□

This proposition shows us that the hyper-efficiency observed in case of estimating a constant mean (different rates for different error distributions) here does not exist, because there's always a possibility of a last-second change in the process. The following informal argument shows us what one can hope for with different error distributions.

Suppose that the number of observations J since the last jump is known. Set

$$\check{X}_n = \mathbf{E}(X_n | Y_1, ..., Y_n, J).$$

Just as before, $\mathbf{E}(\check{X}_n - X_n)^2 \leq \mathbf{E}(\hat{X}_n - X_n)^2$.

The optimal strategy is to use the latest J observations. If the error density ϕ has jumps (e.g. uniform errors) then this strategy yields

$$\mathbf{E}(\check{X}_n - X_n)^2 \simeq n^{-1}(1^2 + \frac{1}{2^2} + ... + \frac{1}{J^2}) \simeq \frac{1}{n}$$

On the other hand, for the continuous error density (e.g. normal errors)

$$\mathbf{E}(\check{X}_n - X_n)^2 \simeq n^{-1}(1 + \frac{1}{2} + ... + \frac{1}{J}) \simeq \frac{ln\, n}{n}$$

3 Estimation of parameters of a jump process

Next, our goal is to estimate the parameters of process itself, that is the time-intensity λ and parameter θ describing jump density h_θ, based on observations $Y(t), 0 \leq t \leq T$. Recursive formula (Theorem 1) will allow us to do it. The question is: how efficient are these estimates?

Assume, as before, that the error density ϕ is known. Without loss of generality, let $\sigma_e = 1$. Also, assume that λ is bounded by some constants: $\Lambda_1 \leq \lambda \leq \Lambda_2$.

When the entire trajectory of the process $X(t, \omega)$ is known, that is, we know exact times $t_1, t_2, ...$ when jumps happened, and exact magnitudes $\xi_i = X(t_i) - X(t_i-), i \geq 1$, the answer is trivial. For example, to estimate intensity, we can just take $\hat{\lambda} := \sum_{i \geq 1} I(t_i \leq T)/T$.

Likewise, inference about h_θ will be based on the jump magnitudes ξ_i. It's clear that these estimates will be consistent only when we observe long enough, that is $T \to \infty$. In fact, we will consider limiting behavior of the estimates as both n and T become large.

Now, when the only observations we have are noisy Y_i, we can try to estimate the locations and magnitudes of jumps of process X. Let n be number of observations on the interval $[0, 1]$. Split the time interval $[0, T]$ into blocks of $m = n^\beta$ observations

each. Let Z_k be the average of observations over Block k,

$$Z_k = \frac{1}{m} \sum_{j=1}^{m} Y_{m(k-1)+j}$$

Consider several cases (see Figure 1). Let $\alpha > 0$ and $\beta > 0$ be specified later.

Case 1. $\sqrt{m}|Z_{k+1} - Z_k| \leq m^\alpha$.

In this case we conclude that no jump occurred on both Block k and Block $k+1$.

Case 2. $\sqrt{m}|Z_{k+1} - Z_k| > m^\alpha$, $\sqrt{m}|Z_{k-1} - Z_k| \leq m^\alpha$, $\sqrt{m}|Z_{k+2} - Z_{k+1}| \leq m^\alpha$.

In this case we conclude that a jump occurred exactly between Block k and Block $k+1$, that is, at time $t = mk/n$. Here, estimate the magnitude of this jump as $\xi^* = Z_{k+1} - Z_k$.

Note: accumulation of errors does not occur when estimating ξ because the estimates are based on non-overlapping intervals.

Case 3. $\sqrt{m}(Z_{k+1} - Z_k) > m^\alpha$ and $\sqrt{m}(Z_k - Z_{k-1}) > m^\alpha$, or

$\sqrt{m}(Z_{k+1} - Z_k) < m^\alpha$ and $\sqrt{m}(Z_k - Z_{k-1}) < m^\alpha$,

In this case we conclude that a jump occurred in the middle of Block k, that is, at time $t = m(k+0.5)/n$. We estimate the magnitude of this jump as $\xi^* = Z_{k+1} - Z_{k-1}$.

Case 4. Jumps occur on the same Block, or on two neighboring Blocks.

The probability that at least two such jumps occur on the interval of a fixed length is asymptotically equivalent to $(m/n)^2 n = m^2/n$. Picking $\beta < 0.5$ we can make this probability small.

Of course, there are errors associated with this kind of detection, we can classify them as:

- Type I Error: we determined that a jump occurred when in reality there was none (this involves Cases 2 and 3).
- Type II Error: we determined that no jump occurred when in reality it did (this involves Cases 1 and 4).
- Placement Error: we determined the location of a jump within a Block or two neighboring Blocks incorrectly.
- Magnitude Error: the error when estimating the value of ξ_i (jump magnitude).

Note that the placement error is small, it is of order m/n. The magnitude error is based on averaging m i.i.d. values, and is therefore of order $m^{-1/2}$.

3.1 Errors in jump detection: Lemmas

Let's estimate the effect of Type I and II errors. Here, as in Section 2.3, we demand that Assumption 1 hold.

Type I errors.

Assume that there are no jumps over the Blocks k and $k+1$, but we detected one according to Case 2 or 3.

Consider

$$P(\sqrt{m}|Z_{k+1} - Z_k| > m^\alpha) = P(|\chi_{m,k+1} - \chi_{m,k}| > m^\alpha),$$

where

$$\chi_{m,k} = \frac{\sum_{j=1}^{m} e_{m(k-1)+j}}{\sigma_e \sqrt{m}}$$

is the sum of normalized errors. Further,

$$P(|\chi_{m,k+1} - \chi_{m,k}| > m^\alpha) \leq 2 \cdot P(|\chi_{m,k}| \geq 0.5\, m^\alpha).$$

From Assumption 1, for any integer $j > 0$

$$\mathbf{E}[\chi_{m,k}^2\, I(|\chi_{m,k}| > C_1 \cdot m^{1/Kj})] < C_2\, exp(-C_3\, m^{1/j}),$$

and the application of Chebyshov's inequality yields

$$P(|\chi_{m,k}| > C_1 \cdot m^{1/Kj}) < C\, exp(-C_3\, m^{1/j}) \cdot m^{-2/Kj}$$

Picking j such that $1/Kj > \alpha > 1/(2Kj)$ and for m large enough, summing up over $Tn\, m^{-1}$ blocks, we obtain

Lemma 2 . *As $n \to \infty$, provided that T grows no faster than some power of n,*

$$P(Type\ I\ error) < C \cdot Tn\, m^{-1} exp(-C_3 m^{2K\alpha}) \to 0$$

□

Type II errors.

Suppose that a jump occurred on Block k, but it was not detected (Case 1), that is

$$|Z_{k-1} - Z_k| \vee |Z_{k+1} - Z_k| \leq m^{\alpha-0.5}$$

Of Blocks $k-1$, $k+1$, pick the closest to the true moment of jump. Without loss of generality, let it be Block $k-1$. Let ξ be the size of the jump. Then averages of $X(t)$ on Blocks k and $k-1$ are different by at least $\xi/2$ and

$$P(|Z_{k-1} - Z_k| \leq m^{\alpha-0.5}) \leq 2P(2|\chi_{m,k}| > |\xi|\sqrt{m}/2 - m^{\alpha}) <$$

$$< C \cdot Tn\, m^{-1} exp(-C_3 m^{2K\varepsilon}) \tag{8}$$

as $n \to \infty$, as long as $|\xi| > m^{-0.5+\alpha+\varepsilon}$, for an arbitrary $\varepsilon > 0$ (use Assumption 1 in a way similar to Lemma 2).

Consider separately

$$P(\bigcup_i \{|\xi_i| > m^{-0.5+\alpha+\varepsilon}\}) \leq C\ (\lambda T + o(T)) m^{-0.5+\alpha+\varepsilon},$$

using the assumption that density of ξ_i is bounded in a neighborhood of 0 and the total number of jumps is $\lambda T + o(T)$. Finally, take into account Case 4 which yields an upper bound $C\lambda T\, m^2/n$. Summing up, we obtain

Lemma 3

$$P(Type\ II\ error) < C \cdot \lambda T(n^{(-0.5+\alpha+\varepsilon)\beta} \vee n^{2\beta-1})$$

□

3.2 Asymptotic behavior of parameter estimates.

For simplicity, determine the behavior of estimates separately, that is consider first an estimate of λ, and then an estimate of θ. Let $\theta \in \Theta$, with Θ being bounded subset of $\mathbf{R}$. Let true values of parameters be λ_0 and θ_0.

Let t_i^* be consecutive jumps of $X(\cdot)$ determined by Cases 2, 3. Estimate the intensity λ by

$$\lambda^* := \frac{1}{T} \sum_{i \geq 1} I(t_i^* \leq T).$$

From the previous discussion it's clear that λ^* is asymptotically equivalent (as $T \to \infty$) to $\hat{\lambda}$ determined from the "true" trajectory of process X. Thus, it possesses the same property, that is asymptotical normality with mean λ_0 and variance C/T for some constant C.

To estimate θ, use the following

Assumption 2 . *Jump magnitude ξ belongs to an* exponential family *with densities with respect to some measure μ,*

$$h_\theta(x) = exp(\theta B(x) - A(\theta))$$

Under this Assumption, $A(\theta) = ln \int exp(\theta B(x)) d\mu(x)$. Also, $A'(\theta) = \mathbf{E}_\theta B(\xi)$ and $I(\theta) := A''(\theta) = Var_\theta[B(\xi)]$ is Fisher information. We follow the discussion in [20,

Example 7.1]. There, the Bayesian estimate with respect to prior $\pi(\theta)$,

$$\tilde{\theta} := \mathbf{E}_\pi(\theta \mid \xi_i, i \geq 1)$$

is asymptotically normal, under some regularity conditions on $\pi(\theta)$.

Assumption 3 .

(a) $I(\theta) > 0$

(b) $h_\theta(x)$ *is bounded in a neighborhood of* 0, *uniformly in* θ.

(c) There is a constant γ, $0 \leq \gamma \leq 1/4$, *such that for large enough* N *there exists* Δ *such that* $P(|\xi_i| > \Delta) = o(N^{-1})$ *and*

$$b_\Delta := sup_{|x| \leq \Delta} \left| \frac{\partial}{\partial x}(ln\, h_\theta(x)) \right| = sup_{|x| \leq \Delta} |\theta B'(x)| = o(N^\gamma)$$

uniformly in θ.

Define the log-likelihood function based on estimated jumps

$$L^*(\theta) = \sum_{i=1}^{N} ln\, h_\theta(\xi_i^*).$$

Theorem 4 . *Let Assumptions 1-3 hold. Then the maximum likelihood estimate*

$$\theta^* = argmax_{\theta \in \Theta}\ L^*(\theta)$$

is asymptotically normal, that is

$$\sqrt{(\lambda_0 T)}\ (\theta^* - \theta_0) \to \mathcal{N}[0, I(\theta_0)^{-1}]$$

in distribution as $T \to \infty$ *no faster than* $T = n^\kappa$, *where* $\kappa < (1/5) \wedge (1 - 4\gamma)$.

Proof:

Pick β, κ, α and ε such that

$$\kappa + 2\beta - 1 < 0$$

$$\kappa + \beta(-0.5 + \alpha + \varepsilon) < 0$$

$$\gamma - \beta/2 < 0.$$

With α, ε arbitrarily small, this is achieved when $2\kappa < \beta < (1-\kappa)/2$, so that $\kappa < (1/5) \wedge (1 - 4\gamma)$.

According to Cases 1-4, the estimated jump magnitudes are

$$\xi_i^* = (\xi_i + \delta_i^0) I_{E^C} + \delta_i I_E,$$

where E is the exceptional set where Type I and II errors occurred, δ_i are the estimates of ξ resulted from these errors, and δ_i^0 are "magnitude errors" discussed in the beginning of this Section.

From Lemmas 2, 3, $P(E) \to 0$ as $n \to \infty$. Therefore we can disregard this set and consider only

$$\xi_i^* = \xi_i + \delta_i^0.$$

Let N be the total number of jumps on $[0, T]$. Consider the log-likelihood function

$$L(\theta) = \sum_{i=1}^{N} ln\, h_\theta(\xi_i).$$

Under the conditions of Theorem, maximum likelihood estimate $\hat{\theta} = argmax_{\theta \in \Theta} L(\theta)$ is asymptotically normal with given variance. Next, we would like to show that the estimate θ^* based on $\{\xi_i^*\}_{1 \leq i \leq N}$ is close to $\hat{\theta}$ based on true values of ξ_i.

Note that both maxima exist because $L''(\theta) = -NA''(\theta)$ and therefore $L(\theta)$ is a convex function, the same is true for $L^*(\theta)$. Furthermore, for any θ in a neighborhood of $\hat{\theta}$,

$$L(\theta) = L(\hat{\theta}) - \frac{(\theta - \hat{\theta})^2}{2} N\, A''(\hat{\theta}) + o(\theta - \hat{\theta})^2$$

Thus, if $L(\theta) - L(\hat{\theta}) = o(1)$ then $(\theta - \hat{\theta})^2 N\, A''(\hat{\theta}) = o(1)$ and therefore $(\theta - \hat{\theta}) = o(N^{-1/2})$.

According to Lemma 4, $|L(\theta) - L^*(\theta)| = o(1)$, and also $|L(\hat{\theta}) - L^*(\hat{\theta})| = o(1)$. Therefore, $(\theta^* - \hat{\theta}) = o(N^{-1/2})$, and the statement of Theorem follows from the similar statement for $\hat{\theta}$.

□

Lemma 4 . *Under the conditions of Theorem 4,*

$$|L(\theta) - L^*(\theta)| = o(1)$$

uniformly in θ, outside some exceptional set E_1 with $P(E_1) = o(1)$.

Proof:

According to Assumption 3.3, the probability of event $E_1 := \bigcup_i \{|\xi_i| > \Delta\}$ is $o(1)$.

Thus, excluding E_1,

$$|L(\theta) - L^*(\theta)| \leq \sum_{i=1}^{N} |ln\, h_\theta(\xi_i) - ln\, h_\theta(\xi_i + \delta_i^0)| \leq \sum_{|\xi_i| \leq \Delta} |\theta B'(\xi_i)| \cdot |\delta_i^0| \leq C b_\Delta n^{-\beta/2}$$

The statement of Lemma follows as $\gamma < \beta/2$.

□

Examples.

Assumption 3(c) is the hardest to verify. It holds, e.g., in following cases:

a) when $B'(x)$ is bounded. For example, exponential distribution with $h_\theta(x) = exp(-\theta x + A(\theta)),\ x \geq 0$.

b) The normal distribution with $h_\theta(x) = exp(-\theta x^2 + A(\theta))$. Here one can pick $\Delta = ln\, N$, and then $b_\Delta = \theta\, ln(N)/2 = o(N^\gamma)$ for arbitrarily small γ.

4 Piecewise-linear process

4.1 Problem formulation

Let's consider the following simplified version of the problem.

Suppose that the velocity of the target is piecewise constant with jumps belonging to the set $\{t_k = k/n; k = 1, ..., n\}$, the probability of jump p_n and distribution of the height of a jump are known. Note that this restriction is not important since for $n \to \infty$ the difference between this and our original process X (with jumps at arbitrary locations) becomes of order $O(n^{-2})$. We will be interested in the case when $p_n = \lambda/n + o(1/n)$. Let $\tau = 1/n$.

Denote V_k the velocity of the target on the interval $[k/n, (k+1)/n]$ and X_k the position of the target at the point k/n.

Suppose the observations Y_k are made at points $1/n, 2/n, ..., (n-1)/n, 1$.

Overall, we have the model

$$\begin{aligned} V_k &= V_{k-1} + \zeta_k \\ X_k &= X_{k-1} + \tau V_{k-1} \\ Y_k &= X_k + e_k, \end{aligned} \tag{9}$$

where e_k are i.i.d. observation errors with a known density ϕ_θ, possibly also depending

on θ, and $\{\zeta_k\}_{k=1,...,n}$ are i.i.d. with distribution

$$\zeta_k = 0 \textit{ with probability } (1 - p_n)$$

$$\zeta_k = \xi_k \textit{ with probability } p_n,$$

and $\{\xi_k\}$ are i.i.d. with a known density h_θ, depending on some parameter θ.

Initial values X_0,V_0 and the parameter θ have some **prior density** $\pi(x_0, v_0, \theta)$. Random variables X_0,V_0, $\{\zeta_k\}$ and $\{e_k\}$ are jointly independent.

The problem of filtering is to find the posterior conditional distribution of the position of process X_n and parameter θ given the observations $Y_1, ..., Y_n$.

4.2 Recursive filtering equations

We will obtain the filtering equations analogous to Section 2.1. But in this case, we have to consider an unobservable variable (velocity), and the resulting filtered density will be two-dimensional.

To simplify notation, in the sequel we will write that ζ_k has "density" $\psi \equiv \psi_\theta$. Consider the joint density of V, θ and Y:

$$\bar{p}_k^\pi(x_0; v_0, ..., v_k; y_1, ..., y_k; \theta) = \pi(x_0, v_0, \theta) \times \prod_{j=1}^{k} \phi_\theta(y_j - x_j) \prod_{j=1}^{k} \psi_\theta(v_j - v_{j-1})$$

where x_k are uniquely determined from (9) as $x_k = x_0 + \tau \sum_{j=0}^{k-1} v_j$.

By changing the variables $\{x_0, v_0, v_1, v_2, ..., v_k\}$ to $\{x_0, x_1, x_2, ..., x_k, v_k\}$, we obtain

the joint density of X, θ, v_k and Y:

$$p_k^\pi(x_0, x_1, ..., x_k, v_k; y_1, ..., y_k; \theta) = \tilde{\pi}_\tau(x_0, x_1, \theta) \times \prod_{j=1}^{k} \phi_\theta(y_j - x_j) \times$$
$$\times \frac{1}{\tau^{k-1}} \prod_{j=1}^{k-1} \psi_\theta(\frac{x_{j+1} - 2x_j + x_{j-1}}{\tau}) \times \psi_\theta(v_k - \frac{x_k - x_{k-1}}{\tau}), \tag{10}$$

where prior density $\tilde{\pi}_\tau(x_0, x_1, \theta)$ can be determined from $\pi(v_0, x_0, \theta)$. The density is "predictive" in its v_k argument: though v_k depends also on Y_{k+1}, we do not include Y_{k+1} into the equation.

Introduce $q_k^\pi(x_k, v_k, \theta)$ **the unnormalized density of** X_k, V_k (position and velocity of the target at the moment k/n) **and** θ **given the observations** $Y_1, ..., Y_k$

$$q_k^\pi(x_k, v_k, \theta) = \int p_k^\pi(x_0, x_1, ..., x_k, v_k; y_1, ..., y_k; \theta)\ dx_0 ... dx_{k-1}. \tag{11}$$

Then the following recursive relation holds

Theorem 5 .

$$q_0^\pi(x, v, \theta) = \pi(x, v, \theta)$$

$$q_k^\pi(x, v, \theta) = \tau\phi_\theta(Y_k - x) \cdot \int \psi_\theta(v - v_{k-1}) q_{k-1}(x - v_{k-1}\tau, v_{k-1}, \theta)\, dv_{k-1} =$$

$$= \tau\phi_\theta(Y_k - x) \Big[e^{-\lambda\tau} q_{k-1}(x - v\tau, v, \theta) + \tag{12}$$

$$+ (1 - e^{-\lambda\tau}) \cdot \int h_\theta(v - v_{k-1}) q_{k-1}(x - v_{k-1}\tau, v_{k-1}, \theta)\, dv_{k-1} \Big]$$

Proof: Replace $v_{k-1} = (x_k - x_{k-1})/\tau$ and rewrite (11) as

$$q_k^\pi(x_k, v_k, \theta) = \int p_{k-1}^\pi(x_0, x_1, ..., x_{k-1}, \frac{x_k - x_{k-1}}{\tau}; y_1, ..., y_{k-1}; \theta) \times$$

$$\times \phi_\theta(Y_k - x_k)\psi_\theta(v_k - \frac{x_k - x_{k-1}}{\tau})\, dx_0 ... dx_{k-1} =$$

$$= \phi_\theta(Y_k - x_k) \int q_{k-1}(x_k, \frac{x_k - x_{k-1}}{\tau}, \theta)\psi_\theta(v_k - \frac{x_k - x_{k-1}}{\tau}) dx_{k-1}$$

Changing the variable x_{k-1} into $x_k - v_{k-1}\tau$, the above yields the statement of the Theorem.

□

4.3 Asymptotics: special case

The question of interest is whether the asymptotic rate of the optimal Bayesian estimate depends on the smoothness of error distribution, that is, whether "hyper-efficiency" takes place. In [22], the asymptotic rate for estimating a linear function $X(t) = X_0 + V_0 t$ was found for uniform errors and was equal to $O(n^{-2})$. Will this rate be changed when we shift to piecewise-linear?

Consider a special case. Let s have the (prior) uniform distribution on $[0, 1]$ and consider the following process instead of X:

$$W(t) = (t - s)^+ = (t - s) \vee 0, \quad 0 \leq t \leq 1$$

with observation errors e_i having uniform distribution on the interval $[-\gamma, \gamma]$ ($\gamma > 0$ is known). The process W in this case is a piecewise linear function with all parameters

known except the turning point s, and we need to estimate s based on observations. Note that estimating the final position $W(1)$ is equivalent to estimating s.

It's clear that the asymptotic rate of $\mathbf{E}(\hat{W}(1) - W(1))^2$ is no worse than the asymptotic rate of X from the previous section. Let's write the likelihood function (the density of distribution $\mathcal{L}(s \mid \textit{observations} \quad Y_1, ..., Y_n)$)

$$f(s) = f(s \mid Y_1, ..., Y_n) = \prod_{0 \le i/n \le 1} \frac{1}{2\gamma} I\{e_i \ge (i/n - s)^+ - \gamma\}$$

Let true value of $s = 1$. Then the likelihood equals to a constant on the interval $[s^*, 1]$, and 0 otherwise, where

$$s^* = min\{u \ : \ (i/n - u)^+ - e_i \le \gamma\}.$$

Then, the Bayesian estimate of s (given its uniform distribution) is

$$\hat{s} = \mathbf{E}(s \mid Y_1, ..., Y_n) = (s^* + 1)/2$$

Consider the step function (see Figure 2)

$$D_u(t) = \frac{1-u}{2} I(t > \frac{u+1}{2}).$$

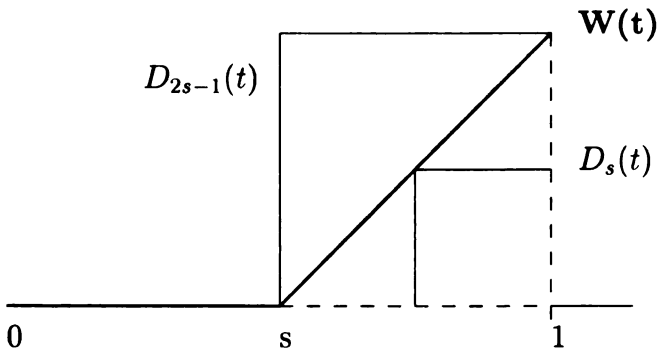


Figure 2

Note that $D_s(t) \leq W(t) \leq D_{2s-1}(t)$. Now consider

$$\tilde{s} = min\{u \ : \ D_u(i/n) - e_i \leq \gamma\}.$$

It's clear that $\tilde{s} > s^*$. Find the asymptotic rate of the estimate $\hat{s}_1 = (\tilde{s} + 1)/2$. We have

$$P(\tilde{s} \leq t) = \prod_{t \leq i/n \leq 1} \frac{1}{2\gamma}(2\gamma - D_s(t)) = (1 - D_s(t))^{(1-t)n}.$$

In case when $t = C\, n^{-1/2}$, $P(\tilde{s} \leq t) \approx exp(-C) \to 0$ when $C \to \infty$. Thus, for s close to 1, $\hat{s}_1$ is $1 - O(n^{-1/2})$.

The same statement can be obtained for $\check{s} = min\{u \ : \ D_{2u-1}(i/n) - e_i \leq \gamma\}$. As a result, we would have $\check{s} < s^* < \tilde{s}$, and it follows that for s close to 1, s^* is $1 - O(n^{-1/2})$.

The case $s = 1$ is the worst, for $s < 1$ the estimate $\hat{s}_1$ is $s - O(n^{-1})$. Finally, this yields the following for the expected square loss (with uniform errors)

Proposition 3 .

$$\mathbf{E}(\hat{W}(1) - W(1))^2 = O((1 - \hat{s})^2 \cdot Cn^{-1/2}) = O(n^{-3/2})$$

□

Keeping in mind that the expected square loss for normal errors cannot be lower than $O(n^{-1})$, we obtain some "hyper-efficiency" in this case, although not as good as when estimating a constant [rate $O(n^{-2})$].

5 Comparison to linear filters

The optimal linear filter for our problem is the well-known Kalman filter. Let's take a look at its asymptotics.

5.1 Jump process

In case of jump process, the model can be re-written as a state-space model (for example, see Brockwell and Davis [4])

$$Y_t = X_t + e_t, \quad t = 1, 2, ...$$

$$X_t = X_{t-1} + \zeta_t$$

and subsequently the optimal linear *predictor* (the estimator of X_t based on all observations up to time $t-1$) is based on second moments of e_t and ζ_t and can be written as

$$\bar{X}_{t+1} = \bar{X}_t + \frac{\Omega_t}{\Omega_t + \sigma_e^2}(Y_t - \bar{X}_t)$$

where $\Omega_t = \mathbf{E}(X_t - \bar{X}_t)^2$ is defined by

$$\Omega_{t+1} = \Omega_t + \mathbf{E}\zeta^2 - \frac{\Omega_t^2}{\Omega_t + \sigma_e^2}$$

and Ω_0 depends on prior distribution of X_0. Furthermore, the optimal linear filter is

$$\hat{X}_{t+1} = \hat{X}_t + \frac{\Omega_t}{\Omega_t + \sigma_e^2}(Y_t - \bar{X}_t) \tag{13}$$

with the error

$$\mathbf{E}(X_t - \hat{X}_t)^2 = \Omega_t - \frac{\Omega_t^2}{\Omega_t + \sigma_e^2}.$$

Note that when the observations become frequent ($n \to \infty$), the optimal predictor and optimal filter become close.

It can be shown that filter (13) is asympotically (when $t \to \infty$) equivalent to the exponential filter

$$\tilde{X}_{t+1} = \tilde{X}_t + \beta(Y_t - \tilde{X}_t)$$

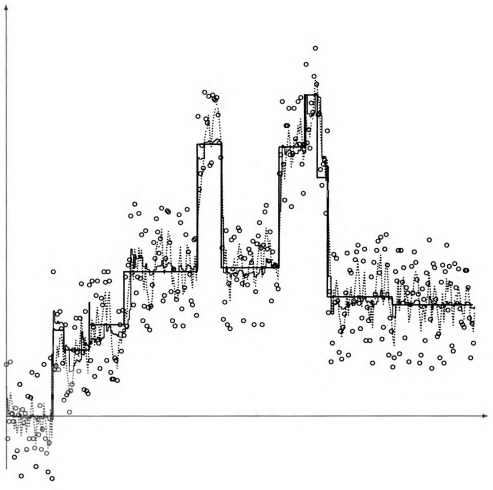
where $\beta = lim_{t\to\infty} \frac{\Omega_t}{\Omega_t + \sigma_e^2}$ and when $n \to \infty$,

$$\beta = \sqrt{\lambda/n} \cdot \sigma_\xi/\sigma_e + o(1/\sqrt{n}).$$

Thus, the asymptotic rate of the best linear filter is of order $n^{-1/2}$:

$$\mathbf{E}(X_t - \hat{X}_t)^2 = \sqrt{\lambda/n} \cdot \sigma_\xi \sigma_e + o(1/\sqrt{n})$$

See Figure 3 for the graphical comparison of linear (exponential) and optimal non-linear filters.



Process position
Linear filter
Non-linear filter
Observations

Figure 3

5.2 Piecewise linear process

In case of piecewise linear process, we can reformulate (1) as a system of state-space equations

$$Y_t = X_t + e_t$$

$$\begin{pmatrix} X \\ V \end{pmatrix}_{t+1} = \begin{pmatrix} 1 & 1/n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ V \end{pmatrix}_t + \begin{pmatrix} \epsilon \\ \zeta \end{pmatrix}_t$$

where X_t is the current position of the process and V_t is the current velocity (note that only position is observed, not velocity); as before, ζ_t is the jump in velocity over the interval $[t/n, (t+1)/n]$, and ϵ_t is the change in X caused by this jump in velocity, which is negligible (of order $O(n^{-2})$) when n is large.

Applying the recursive equations for Kalman filter (see [4]), we obtain the following recursions for the best linear *predictor* $\overline{\begin{pmatrix} X \\ V \end{pmatrix}}_t$:

$$\overline{\begin{pmatrix} X \\ V \end{pmatrix}}_{t+1} = \begin{pmatrix} 1 & 1/n \\ 0 & 1 \end{pmatrix} \overline{\begin{pmatrix} X \\ V \end{pmatrix}}_t + \frac{Y_t - \bar{X}_t}{w_{11} + \sigma_e^2} \begin{pmatrix} w_{11} + w_{21}/n \\ w_{21} \end{pmatrix}$$

with the error matrix

$$\Omega_t = \begin{pmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{pmatrix}_t = \mathbf{E}\left[\overline{\begin{pmatrix} X \\ V \end{pmatrix}}_t - \begin{pmatrix} X \\ V \end{pmatrix}_t\right]\left[\overline{\begin{pmatrix} X \\ V \end{pmatrix}}_t - \begin{pmatrix} X \\ V \end{pmatrix}_t\right]^T$$

defined by recursive relation

$$\Omega_{t+1} = \begin{pmatrix} 1 & 1/n \\ 0 & 1 \end{pmatrix} \Omega_t \begin{pmatrix} 1 & 0 \\ 1/n & 1 \end{pmatrix} + Q_t - \\ - \frac{1}{w_{11} + \sigma_e^2} \begin{pmatrix} (w_{11} + w_{21}/n)^2 & (w_{11} + w_{21}/n)w_{21} \\ (w_{11} + w_{21}/n)w_{21} & w_{21}^2 \end{pmatrix}$$

where

$$Q_t = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_\xi^2 \lambda / n \end{pmatrix} + o(1/n).$$

It follows that the elements of matrix $\Omega = lim_{t\to\infty}\Omega_t$ are

$$w_{11} = \sqrt{2}\,\sigma_e^{3/2}\sigma_\xi^{1/2}\, n^{-3/4} + o(n^{-3/4})$$

$$w_{12} = \sigma_e \sigma_\xi n^{-1/2} + o(n^{-1/2})$$

$$w_{22} = \sqrt{2}\,\sigma_e^{1/2}\sigma_\xi^{3/2}\, n^{-1/4} + o(n^{-1/4})$$

Furthermore, as $n \to \infty$, the optimal predictor $\bar{X}_t$ and optimal filter $\hat{X}_t$ are close. Thus, the asymptotic efficiency of the optimal linear filter $\mathbf{E}(\hat{X}_t - X_t)^2$ is of order $n^{-3/4}$.

A summary of asymptotic behavior of $\sigma^2 := \mathbf{E}(\hat{X}_n - X_n)^2$ for jump and piecewise-linear processes is given in Table 1.

	smooth errors	discontinuous errors	Linear Filter
Jump Process	$\leq ln^M n/n$	$\leq ln^M n/n$	$n^{-1/2}$
Piecewise-Linear Process	$\geq n^{-1}$	$n^{-3/2}$ (Special case)	$n^{-3/4}$
Constant Location Parameter	n^{-1}	n^{-2}	n^{-1}

Table 1. Summary of asymptotic results

One can see that increasing the "smoothness" of a process improves the asymptotics. The behavior of the optimal linear and non-linear filters for a piecewise-linear process is shown in Figure 4. The optimal non-linear filter was evaluated using the sequential Monte-Carlo method described in [8].

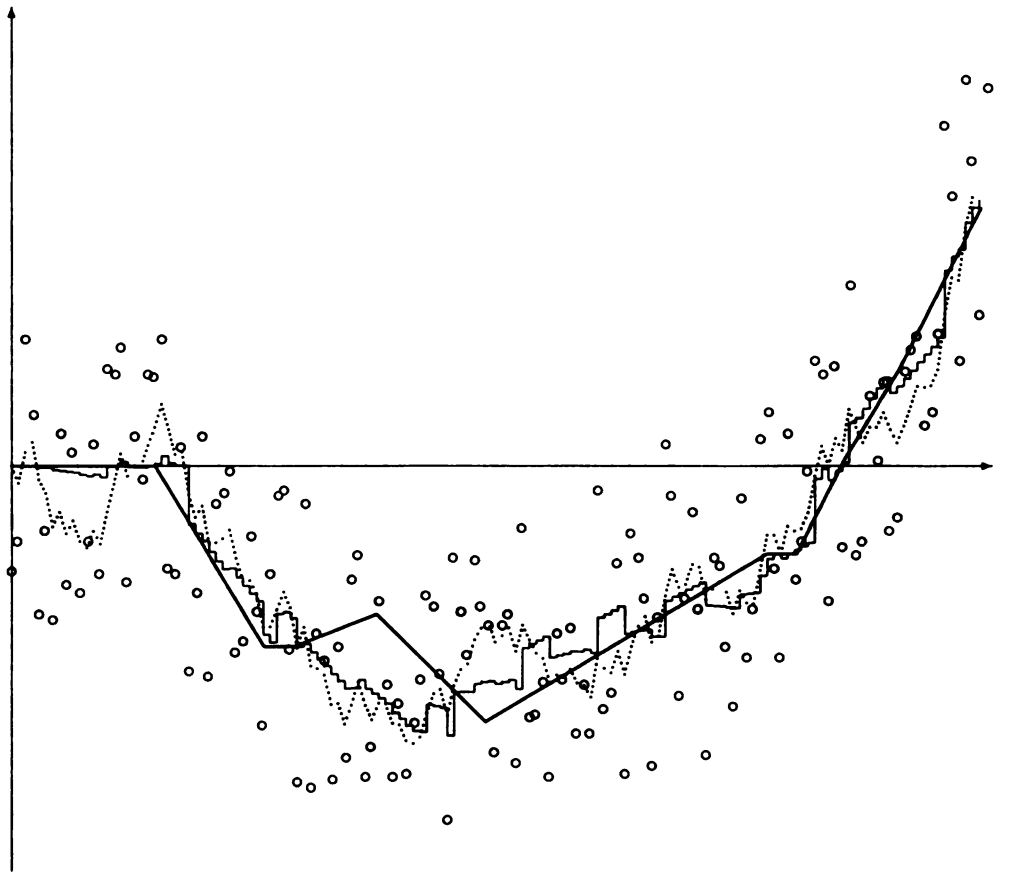


Figure 4

6 Simulation

The author has simulated the optimal filter for the jump process based on Theorem 1. There, the densities $q_k(x)$ are found recursively using the relation (3). We consider a case when the parameters θ, λ and initial state X_0 are known, and the error density is uniform on the interval $[-\gamma, \gamma]$. This causes $q_k(x)$ to be confined to an interval $[Y_k - \gamma, Y_k + \gamma]$, and we keep a discretized version of $q_k(x)$ in memory. Integration required for evaluating (3) was performed numerically.

The results of simulation are given in Table 2. In the case considered, distribution of jumps is uniform on the interval $[-2.5, 2.5]$, with $\gamma = 1$, on the interval $t \in [0, 9]$. Two cases, $\tau = 0.02$ and $\tau = 0.04$ were considered. For each Monte-Carlo sample of the process, the ratio of effectiveness of non-linear filter to the linear one was found:

$$R := \left[\frac{\sum_{t=1}^{9\tau}(\tilde{X}_t - X_t)^2}{\sum_{t=1}^{9\tau}(\hat{X}_t - X_t)^2}\right]^{1/2},$$

where $\hat{X}_t$ is the optimal non-linear filter at time t and $\tilde{X}_t$ is the optimal linear filter. Also, the mean square error MS_{nl} of the optimal non-linear filter is given. In each case, $N = 100$ Monte-Carlo samples were generated.

τ	mean of R	st.dev. of R	MS_{nl}
0.02	2.141	0.535	0.0249
0.04	1.749	0.407	0.0479

Table 2. Simulation results

This and other simulations lead us to believe that the optimal filter becomes more effective relative to the linear filter when:

a) $\tau \to 0$ (we know that from the asymptotics, as well as Table 2).

b) jump magnitudes increase, making the process more "ragged".

Intuitively, a linear filter has to compromise between the periods where the process stays constant (and for the filter to perform better on those intervals, past observations need to carry a greater weight) and the times when jumps happen (to cope with jumps, we need to forget past observations quickly). As a result, greater jumps will upset the performance of a linear filter. Some adaptive filters, for example, IMM filter described in [2], might be more competitive.

7 Open problems

Naturally, it makes sense to extend the results for a piecewise-linear process (of which only the special case is treated in Section 4). This is considerably more difficult than the jump process case, partly due to hyper-efficiency. Another interesting task is to cover the unknown error distribution (in most asymptotic results above, the latter was assumed known). The results on parameter estimation (Section 3) might also be improved.

Also, in view of possible applications, other forms of stochastic process $X(\cdot)$ deserve to be considered. First, two- and three-dimensional processes are obviously of interest. Second, some other types of processes, rather than jump and piecewise-linear, might be more useful. Finally, one needs to consider the situation when parameters of the process are changing themselves, albeit slowly ("parameter tracking"). Such situation is considered in Elliott et al. [10], and no doubt the recursive formulas similar to Theorem 1 could be derived in this case, too.

References

[1] ANDERSON, B.D.O. AND MOORE, J.B. , *Optimal filtering*, Prentice Hall, Englewood Cliffs, N.J., 1979.

[2] Y. BAR-SHALOM, X. RONG LI, AND T. KIRUBARAJAN, *Estimation with applications to tracking and navigation*, Wiley, New York, 2001. Theory, Algorithms and Software.

[3] P. BRÉMAUD, *Point processes and queues*, Springer-Verlag, New York, 1981. Martingale dynamics, Springer Series in Statistics.

[4] BROCKWELL, PETER J.; DAVIS, RICHARD A., *Introduction to time series and forecasting.*, Springer-Verlag, New York, 1996.

[5] C. CECI, A. GERARDI, AND P. TARDELLI, *An estimate of the approximation error in the filtering of a discrete jump process*, Math. Models Methods Appl. Sci., 11 (2001), pp. 181–198.

[6] CHERNOFF, HERMAN; RUBIN, HERMAN, *The estimation of the location of a discontinuity in density.*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, (1954–1955). vol. I, pp. 19–37.

[7] P. DEL MORAL, J. JACOD, AND P. PROTTER, *The Monte-Carlo method for filtering with discrete-time observations*, Probab. Theory Related Fields, 120 (2001), pp. 346–368.

[8] A. DOUCET, N. DE FREITAS, AND N. GORDON (ed.), *Sequential Monte-Carlo methods in practice*, Springer-Verlag, New York, 2001.

[9] F. DUFOUR AND D. KANNAN, *Discrete time nonlinear filtering with marked point process observations*, Stochastic Anal. Appl., 17 (1999), pp. 99–115.

[10] ELLIOTT, ROBERT J.; AGGOUN, LAKHDAR; MOORE, JOHN B., *Hidden Markov models. Estimation and control.*, Springer-Verlag, New York, 1995.

[11] ERMAKOV, M. S. , *Asymptotic behavior of statistical estimates of parameters for a discontinuous multivariate density. (russian)*, Studies in statistical estimation theory, I. Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), (1977). pp. 83–107.

[12] N. GORDON, D. SALMOND, AND A. SMITH, *A novel approach to nonlinear/non-Gaussian Bayesian state estimation*, IEEE Proceedings on Radar and Signal Processing, 140 (1993), pp. 107–113.

[13] IBRAGIMOV, I. A.; KHAS'MINSKII , R. Z. , *Statistical estimation. Asymptotic theory*, Springer-Verlag, New York-Berlin, 1981.

[14] T. KAILATH, *An innovations approach to least-squares estimation, parts I , II*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 646–660.

[15] G. KALLIANPUR, *Stochastic filtering theory*, Springer-Verlag, New York, 1980.

[16] G. KALLIANPUR AND C. STRIEBEL, *Stochastic differential equations occurring in the estimation of continuous parameter stochastic processes*, Teor. Verojatnost. i Primenen, 14 (1969), pp. 597–622.

[17] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, J. Basic Engrg., (1960), pp. 35–45.

[18] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83 (1961), pp. 95–108.

[19] D. KANNAN, *Discrete-time nonlinear filtering with marked point process observations. II. Risk-sensitive filters*, Stochastic Anal. Appl., 18 (2000), pp. 261–287.

[20] LEHMANN,E.L. , *Theory of point estimation*, Wiley, New York, 1983.

[21] PORTENKO, N.; SALEHI, H.; SKOROKHOD, A. , *On optimal filtering of multitarget tracking systems based on point process observations*, Random Oper. Stochastic Equations, 5 (1997). no. 1, 1–34.

[22] ——, *Optimal filtering in target tracking in presence of uniformly distributed errors and false targets*, Random Oper. Stochastic Equations, 6 (1998). no. 3, 213–240.

[23] B. L. ROZOVSKIĬ, *Stochastic evolution systems*, Kluwer Academic Publishers Group, Dordrecht, 1990. Linear theory and applications to nonlinear filtering, Translated from the Russian by A. Yarkho.

[24] RUBIN, HERMAN , *The estimation of discontinuities in multivariate densities, and related problems in stochastic processes*, Proc. 4th Berkeley Sympos. Math. Statist. and Prob., (1961). Vol. I pp. 563–574.

[25] A. N. SHIRYAYEV, *Stochastic equations of non-linear filtering of jump-like Markov processes*, Problemy Peredachi Informacii, 2 (1966), pp. 3–22.

[26] A. I. YASHIN, *Filtering of jump processes*, Avtomat. i Telemeh., (1970), pp. 52–58.

[27] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Gebiete, 11 (1969), pp. 230–243.

Notation

Some special symbols used in this work.

$\mathbf{E}(X)$ expected value of X

$s \vee t$ $max(s,t)$ $s \wedge t$ $min(s,t)$

$\mathbf{R}$ set of real numbers

$C, const$ some constant (often, their exact value is not specified but can be easily obtained)

$I(A)$ or I_A indicator of the set/event A

A^C complement of the set/event A

A^T matrix A transposed

$A \simeq B$ asymptotic equivalence, that is $A/B = const + o(1)$

$A := B$ definition of expression A in terms of expression B

$\mathcal{L}(X), \mathcal{L}(X|...)$ distribution/conditional distribution of X

$argmin_Y\ f(Y)$ the value z such that $f(z) = min_Y\ f(Y)$

$\mathcal{N}(\mu, \sigma^2)$ Normal distribution with specified mean and variance.