Elect. Comm. in Probab. **13** (2008), 210–224

ELECTRONIC COMMUNICATIONS in PROBABILITY

FILTERING AND PARAMETER ESTIMATION FOR A JUMP STOCHASTIC PROCESS WITH DISCRETE OBSERVATIONS

OLEG V MAKHNIN
Department of Mathematics, New Mexico Tech, Socorro NM 87801
email: `olegm@nmt.edu`

Submitted October 31, 2006, *accepted in final form* February 18, 2008

AMS 2000 Subject classification: 60J75, 62M05, 93E11
Keywords: non-linear filtering, Kalman filter, particle filtering, jump processes, target tracking

Abstract
A compound Poisson process is considered. We estimate the current position of the stochastic process based on past discrete-time observations (non-linear discrete filtering problem) in Bayesian setting. We obtain bounds for the asymptotic rate of the expected square error of the filter when observations become frequent. The bounds depend linearly on jump intensity. Also, estimation of process' parameters is addressed.

1 Introduction

Filtering of stochastic processes has attracted a lot of attention. One of the examples is target tracking, when a target is observed over a discrete time grid, corresponding to the successive passes of a radar. Another is signal processing, where the observations are discretized. In both cases, we deal with discrete observations.

We will further refer to the (unobservable) process of interest as "signal" ("state") $X(t)$. Assume that the time grid is uniform. We attempt to establish asymptotic properties of such filtering when the time interval τ between successive observations goes to zero. The signal+noise model takes the form of a state-space model: the process $X(t)$ evaluated at the grid points is a discrete-time Markov process (see (1) below). The observation equation is

$$Y_k = X(k\tau) + e_k, \quad k = 1, 2, ...$$

where e_k are i.i.d. with density ϕ, mean 0 and variance σ_e^2.

The goal of filtering is to estimate the current position $X(t)$ of the process based on all observations up to the moment. The filter given by

$$\mathbb{E}[X(t) \mid Y_1, ...Y_k, \tau k \le t < \tau(k+1)]$$

is optimal with respect to square loss. That is, it minimizes $\mathbb{E}(X(t) - \hat{X}(t))^2$ among all filters $\hat{X}(t)$ based on up-to-the-moment observations (see Lemma 1 below).

We consider a special case of stochastic process: piecewise-constant or pure-jump process. It is 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 $[0, T] \times \mathbb{R}$. 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 $h(y) = h_\theta(y)$ is the "jump density" describing magnitudes of jumps ξ_i of process X. Here $\theta \in \Theta \subset \mathbb{R}^p$ is a parameter (possibly unknown) defining the distribution of jumps and errors.

In the Bayesian formulation, let parameters θ and λ have a prior density $\pi(\theta, \lambda)$ with respect to Lebesgue measure. Assume that for each $\theta \in \Theta$, $\mathbb{E}\xi_1^2 < \infty$. Also, assume that starting value X_0 has some prior density $\pi_0(\cdot)$. The Bayesian approach lends conveniently to the combined parameter and state estimation via the posterior expectation

$$(\widehat{X(t), \theta}, \lambda) := \mathbb{E}_{\pi,\pi_0}[X(t), \theta, \lambda \mid Y_1, ...Y_k, \tau k \le t < \tau(k+1)]$$

Remark 1
a) Generalizations to non-constant jump intensity $\lambda = \lambda(\cdot)$ and more general structure where $\lambda(s, y)$ depends on current position of $X(s)$ are possible, but will not be considered here.

b) A generalization to d-dimensional signal and observations is immediate when the density ϕ of the error distribution takes a special form: $\phi(x_1, ..., x_d) = \prod_{i=1}^d \phi_i(x_i)$. In this case, the optimal filter can be computed coordinate-wise.

More realistic cases in context of target-tracking (like processes with pure-jump velocity or pure-jump acceleration) are more complex and will be discussed elsewhere. A special case of a piecewise-linear process with uniform observational errors was considered in [12], but their suggested asymptotics of $O(n^{-2})$ appear to be incorrect. A general approach to filters for target-tracking was laid out in [11].

There is a substantial body of work dedicated to the filtering of jump processes in continuous noise (see [14] and references therein). In [13], jump processes of much more general type are considered, but the process is assumed to be observed without error.

Recently, considerable interest was devoted to practical methods of estimating the optimal filter. Particle filters described in [6] and [1] are useful, and have generated a great deal of interest in signal-processing community. Still, some convergence problems persist, especially when non-linearity of the model increases. Typically, the optimal filter is infinite-dimensional and requires great computational effort.

On the other hand, the popular Kalman filter (which is an optimal *linear* filter, see, e.g., [4]) is fast but is based on the assumption of normality for both process and errors and may not be very efficient. In target-tracking literature, several low-cost improvements for Kalman filter have been suggested. In particular, interactive multiple model (IMM) filters were described in [2].

See Fig. 1 for an example of the process considered and a graphical comparison of optimal

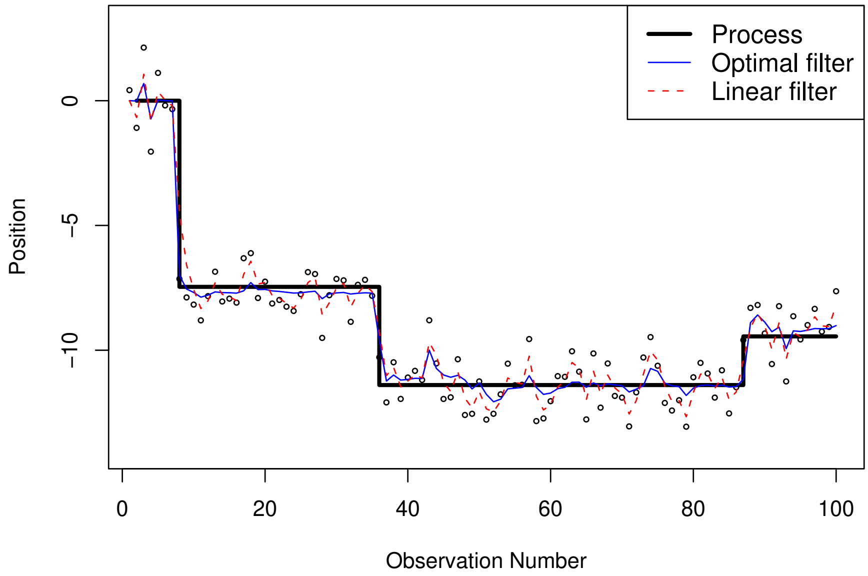

Figure 1: optimal non-linear and linear filters

and Kalman filters. Parameters are $\tau = 0.2$, $\lambda = 0.2$, Normal jumps with mean 0 and variance $= 5^2$, and $\sigma_e = 1$.

One of the questions is how much improvement does the (optimal) non-linear filter bring compared to (linear) Kalman filter. Another is how does the error distribution affect this improvement.

To clarify the role of the error distribution, consider a special case well-known in statistics: when the signal is constant over time. Then the problem reduces to estimating a location parameter for the density. Asymptotic efficiency (as number of observations, or sampling rate $n \to \infty$) of such estimator depends on the type of density [5]. For example, a continuous (e.g. Normal) density allows the expected squared error of $O(n^{-1})$ which is typical for statistical estimation problems; but for a discontinuous (e.g. Uniform) density the expected squared error improves to $O(n^{-2})$. A point of discontinuity can be used to estimate the location parameter much more efficiently than, say, the mean of the distribution. This phenomenon was called hyper-efficiency in [5]. These questions are addressed in detail in [8]. This paper presents some results derived in [10]. In Section 2 we consider the asymptotics for the optimal filter, and compare it with Kalman filter. Section 3 deals with estimating parameters λ, θ, and Section 4 contains the results of a simulation study.

2 Filtering of a jump process

2.1 Recursive formulation

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

$$\begin{aligned} X_k &= X_{k-1} + \zeta_k \\ Y_k &= X_k + e_k \end{aligned} \tag{1}$$

Here, ζ_k is a sum of jumps of X on the interval $[\tau(k-1), \tau k)$. Thus, $\{\zeta_k\}_{k\geq 1}$ 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, I will call ψ a "density", actually meaning that $\psi(0)$ is a scaled δ-function, that is for any function g,

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

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

The joint density of $X_{0:k} := (X_0, ..., X_k)$, $Y_{1:k} := (Y_1, ..., Y_k)$, θ and λ is found as

$$\begin{aligned} p(x_{0:k}, y_{1:k}, \theta, \lambda) &= p(y_{1:k} \mid x_{0:k}, \theta, \lambda) \cdot p(x_{0:k} \mid \theta, \lambda) \cdot \pi(\theta, \lambda) = \\ &= \pi(\theta, \lambda) \cdot \pi_0(x_0) \cdot \prod_{j=1}^{k} \phi(y_j - x_j) \cdot \prod_{j=1}^{k} \psi(x_j - x_{j-1}) \end{aligned} \tag{2}$$

Let $q_k(x, \theta, \lambda) := \int_{\mathbb{R}^k} p(x_{0:k}, Y_{1:k}, \theta, \lambda)\, dx_0 \ldots dx_{k-1}$ be the unnormalized density of the latest state X_k and parameters θ, λ given the observations $Y_{1:k}$. The following relation is well-known (see e.g. [6]).

Theorem 1.

$$\begin{aligned} q_0(x, \theta, \lambda) &= \pi_{X_0}(x) \cdot \pi(\theta, \lambda), \\ q_k(x, \theta, \lambda) &= \phi_\theta(Y_k - x) \cdot \int_{\mathbb{R}} \psi_{\theta,\lambda}(x - z) q_{k-1}(z, \theta, \lambda)\, dz = \\ &= \phi_\theta(Y_k - x) \cdot \left[e^{-\lambda\tau} q_{k-1}(x, \theta, \lambda) + \int_{\mathbb{R}} \tilde{\psi}_{\theta,\lambda}(x - z) q_{k-1}(z, \theta, \lambda)\, dz \right] \qquad \square \end{aligned} \tag{3}$$

Remark 2

a) In order to use Theorem 1 for the estimation of state X_k, one should compute $q_j(x, \theta, \lambda), j \leq k$ consecutively, then compute marginal unnormalized density $q_k(x) := \int q_k(x, \theta, \lambda)\, d\theta\, d\lambda$ and then find

$$\hat{X}_k := \mathbb{E}(X_k | Y_{1:k}) = \frac{\int_{\mathbb{R}} x q_k(x) dx}{\int_{\mathbb{R}} q_k(x) dx}. \tag{4}$$

b) 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 $X(t)$, see [7]. 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 often used in non-linear filtering.

The recursive formulas for the densities may be used to compute "on-line" updates as new observations are coming in. Unfortunately, they generally cannot be integrated in closed form, and approximations such as particle filters are used in practice.

2.2 Multiple-block upper bound for expected square error

Next, we investigate asymptotic properties of the above filtering estimator $\hat{X}(T) := \hat{X}_{\lfloor nT \rfloor}$ as the observations become frequent ($1/\tau = n \to \infty$). The following results were obtained when the error distribution is considered known. Without loss of generality, assume $T = 1$ (since we can always rescale time, and use the new jump intensity λT). We assume that n is integer, so that the last jump occurs at exactly 1. If n is not integer, then the estimation of $X(1)$ is based on $Y_1, ..., Y_{\lfloor n \rfloor}$, and the expected loss added on the interval $(\tau\lfloor n \rfloor, 1]$ is of order $1/n$, so the overall asymptotics will not change.

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,*

$$\mathbb{E}[X - \mathbb{E}(X|\mathcal{F})]^2 \le \mathbb{E}(X - U)^2 \quad \square$$

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 $Y_{1:k}$. We will produce a sub-optimal estimate for $X(1)$ based on the mean of observations for a suitable interval. The difficulty lies in not knowing where exactly the last jump of process X occurred.

Consider the intervals (blocks) going backwards $(T_1, T_0], (T_2, T_1], ..., (T_N, T_{N-1}]$, where

$$\begin{aligned} T_0 &:= 1 \\ T_j &:= T_{j-1} - (\ln n)^j/n, \qquad j = 1, ..., N \end{aligned}$$

and the total number of blocks is $N := \lfloor \frac{\ln n}{\ln \ln n} \rfloor - 1$; j-th block has length $(\ln n)^j/n$ and n_j observations. The exact integer value of n_j depends on the positioning of blocks, but it differs from $(\ln n)^j$ by no more than 1. Thus, in the future calculations we would sometimes use $(\ln n)^j$ instead of n_j, as long as it will not change the asymptotic values of expressions.
Note that the total length of all blocks, for n large enough, is no greater than

$$\frac{\ln n}{n} \cdot \frac{(\ln n)^{\frac{\ln n}{\ln \ln n} - 1} - 1}{\ln n - 1} = \frac{n - \ln n}{n(\ln n - 1)} < 1$$

Starting at the time 1, we will probe back one block of observations after another, stopping whenever we believe that a jump has occurred.

Let $\overline{Y}_j$ be the average of observations from the block j, that is

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

Assumption 1. *Let*

$$\chi_m := \frac{\sum_{k=1}^m e_k}{\sigma_e \sqrt{m}} \tag{5}$$

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 and $K > 0$ such that for all sufficiently large m and uniformly for all positive integers j,

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

This assumption is satisfied for Normal errors with $K = 2$; in general, it requires e_k to have small tails. The following is 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, $\mathbb{E}\,\exp(\gamma|\chi_m|) \le G$.*

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 large enough, so that $x^2 < \exp(\gamma|x|/2)$ for $|x| > C_1$, and $\gamma C_1/2 > 1$.
Then for any j,

$$\int_{\mathbb{R}} I\{|x| > C_1 m^{1/j}\}\, x^2\, dF_m(x) \le \int_{\mathbb{R}} I\{|x| > C_1 m^{1/j}\}\, e^{\gamma|x|/2} dF_m(x) \le$$
$$\le \exp(-\gamma C_1 m^{1/j}/2) \int_{\mathbb{R}} e^{\gamma|x|} dF_m(x) \ \le \exp(-\gamma C_1 m^{1/j}/2) \cdot G \quad \square$$

Theorem 2. Upper bound for $\mathbb{E}(\hat{X}_n(1) - X(1))^2$
Suppose that the error density ϕ is known and does not depend on the parameter θ. Then, under Assumption 1, there exists a constant C such that for large enough n,

$$\mathbb{E}(\hat{X}_n(1) - X(1))^2 \le C\lambda \frac{\ln^M n}{n}$$

with $M = max\{1 + \frac{2}{K}, 2\}$.

Proof:
Consider N blocks as described above. Denote $T^\star$ the point of last jump of X: $T^\star = \sup\{0 \le t \le 1 : X(t) - X(t-) > 0\}$. Let also $J^\star =$ the number of Block where the last jump happened. The idea is to approximate $T^\star$, then take the average of all observations from that Block.

Construct an estimate of $X(1)$ as follows. Define j_0 as

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

Then, as our estimate of $X(1)$, take $\tilde{X}(1) := \overline{Y}_{j_0}$. We will find an upper bound for the average risk of this estimate, $\ell := \mathbb{E}(\tilde{X}(1) - X(1))^2$. For this, we will need several inequalities, with proofs to follow in the next section.

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

$$\ell_S := \mathbb{E}[(\tilde{X}(1) - X(1))^2 I_S] \le C_4\, \lambda \frac{\ln^2 n}{n} \tag{7.1}$$

Case 2: Late stopping
In the event $L = \{T_{j_0} < T^\star\} = \{J^\star \le j_0\}$

$$\ell_L := \mathbb{E}[(\tilde{X}(1) - X(1))^2 I_L] \le C_5\, \lambda \frac{(\ln n)^{(1+2/K)\vee 2}}{n} \tag{7.2}$$

Case 3: Early stopping
In the event $\mathcal{E}$ that we stopped on the Block j_0 but there was no jump of X, $\mathcal{E} = \{T^\star \le T_{j_0+1}\}$

$$\ell_{\mathcal{E}} := \mathbb{E}[(\tilde{X}(1) - X(1))^2 I_{\mathcal{E}}] \le C_6 \frac{(\ln n)^{2-2/K}}{n} \tag{7.3}$$

Now note that $P(S \cup L \cup \mathcal{E}) = 1$. Thus, $\ell = \ell_S + \ell_L + \ell_{\mathcal{E}}$. Also, the bound for the asymptotic rate of the estimator $\tilde{X}$ does not depend on the particular form of jump density h_θ.

By Lemma 1, the risk of estimate $\hat{X}$ does not exceed the risk of $\tilde{X}$. Even though the rate in (7.3) does not depend on λ, it is of a smaller order. Combining (7.1) through (7.3), we obtain the proof of the Theorem. $\square$

2.3 Proofs of inequalities used in Theorem 2

Proof of (7.1)
If $T^\star \le T_{j_0}$, then $X(t) = X(1)$ for $T_{j_0} < t < 1$, and the squared loss from estimating $X(1)$ equals the variance of $\overline{Y}_{j_0}$, so that

$$\mathbb{E}[(\tilde{X}(1) - X(1))^2 I_S)] = \textstyle\sum_{j=1}^{N-1} \mathbb{E}[(\overline{Y}_j - X(1))^2 I_S I(J^\star = j+1)] \le$$

$$\textstyle\sum_{j=1}^{N-1} \mathbb{E}[(\overline{Y}_j - X(1))^2 I(J^\star = j+1)] \le \sum_{j=1}^{N-1} P(\text{a jump on Block } j+1) \cdot \sigma_e^2 / n_j$$

by independence of $\{e_k\}$ and process X. [Note that we used $\sum_{j=1}^{N-1}$ in order to get rid of randomness of j_0 and therefore possible dependence on the $\overline{Y}_{j_0}$ value.] Thus,

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

Proof of (7.2)
The stopping rule (6) implies that for $J^\star \le j \le j_0$, $|\overline{Y}_j - \overline{Y}_{j-1}| \le 2\sigma_e C_1 n_{j-1}^{-\frac{1}{2}+\frac{1}{K(j-1)}}$. Thus, using the inequality $(A-B)^2 \le 2A^2 + 2B^2$,

$$\mathbb{E}[(\tilde{X}(1) - X(1))^2 I_L] = \mathbb{E}[(\overline{Y}_{j_0} - X(1))^2 I_L] \le$$

$$2\,\mathbb{E}\left[\left(\textstyle\sum_{j=J^\star}^{j_0} |\overline{Y}_j - \overline{Y}_{j-1}|\right)^2 I_L\right] + 2\,\mathbb{E}\left[(\overline{Y}_{J^\star-1} - X(1))^2 I_L\right] \equiv 2\,E_1 + 2\,E_2$$

Now,

$$E_1 \le \textstyle\sum_{j=J^\star}^{j_0} P(\text{a jump on Block } J^\star)\, n_{j-1}^{-1+\frac{2}{K(j-1)}} \le$$

$$const \cdot \frac{\lambda}{n} \left(\ln^{J^\star} n + o(\ln^{J^\star} n)\right) (\ln n)^{-(J^\star-1)+2/K} \le const \cdot \frac{\lambda}{n} (\ln n)^{1+2/K}$$

(the inequalities hold regardless of j_0, $J^\star$), and

$$E_2 \le \mathbb{E}[(\overline{Y}_{J^\star-1} - X(1))^2] \le \textstyle\sum_{j=2}^{N} \mathbb{E}[(\overline{Y}_{j-1} - X(1))^2 I(J^\star = j)] \le$$

$$\textstyle\sum_{j=2}^{N} \mathbb{E}[(\overline{Y}_{j-1} - X(T_{j-2}))^2 I(\text{no jump on Block } j-1, \text{a jump on Block } j)] \le$$

$$\sum_{j=2}^{N} \sigma_e^2 (\ln n)^{-(j-1)} \cdot \frac{\lambda}{n} \left(\ln^j n + o(\ln^j n)\right) \le const \cdot \lambda \frac{\ln^2 n}{n}$$

Consider separately the case $J^\star = 1$. Then, the summation for E_1 has lower limit of $J^\star + 1 = 2$, and E_2 changes to $\mathbb{E}[(\overline{Y}_1 - X(1))^2 I(J^\star = 1)]$, so

$$E_2 \le (\sigma_e^2/\ln n + \mathbb{E}\xi_1^2 + o(1)) \cdot \lambda \frac{\ln n}{n} \le const \cdot \lambda \frac{\ln n}{n}$$

because the probability of more than one jump on Block 1 is $o(\ln n/n)$ $\square$

Proof of (7.3)
If the stopping occurred too early then $X(t) = X(1)$ for $T_{j_0+1} < t < 1$. Also, the stopping rule (6) implies that at least one of

$$|\overline{Y}_{j_0+1} - X(1)| > C_1 \sigma_e n_{j_0}^{-\frac{1}{2}+\frac{1}{j_0 K}}, \quad |\overline{Y}_{j_0} - X(1)| > C_1 \sigma_e n_{j_0}^{-\frac{1}{2}+\frac{1}{j_0 K}}$$

is true. Thus,

$$\mathbb{E}[(\tilde{X}(1) - X(1))^2 I_{\mathcal{E}}] \le \mathbb{E}|\overline{Y}_{j_0} - X(1)|^2 \cdot P\left(|\overline{Y}_{j_0+1} - X(1)| > C_1 \sigma_e n_{j_0}^{-\frac{1}{2}+\frac{1}{j_0 K}}\right) +$$
$$\mathbb{E}\left(|\overline{Y}_{j_0} - X(1)|^2 I\left[|\overline{Y}_{j_0} - X(1)| > C_1 \sigma_e n_{j_0}^{-\frac{1}{2}+\frac{1}{j_0 K}}\right]\right) \equiv E_3 + E_4.$$

By Assumption 1, $E_4 \le \sum_{j=1}^N n_j^{-1} \mathbb{E}\left(|\chi_{n_j}|^2 I\left[|\chi_{n_j}| > C_1 \sigma_e n_j^{-\frac{1}{2}+\frac{1}{jK}}\right]\right) \le$
$C_2 \sum_{j=1}^N n_j^{-1} \exp(-n_j^{1/j}) \le C_2/n$. This bound does not depend on λ!

To estimate E_3, note that $P\left(|\overline{Y}_{j_0+1} - X(1)| > C_1 \sigma_e n_{j_0}^{-\frac{1}{2}+\frac{1}{j_0 K}}\right)$ would increase if $n_{j_0}^{-\frac{1}{2}+\frac{1}{j_0 K}}$ is replaced by $n_{j_0+1}^{-\frac{1}{2}+\frac{1}{(j_0+1)K}}$, and consider the Chebyshev-type inequality

$$D^2 P(|\overline{Y}_{j_0+1} - X(1)| > D) \le \mathbb{E}\left[|\overline{Y}_{j_0+1} - X(1)|^2 I(|\overline{Y}_{j_0+1} - X(1)| > D)\right] \equiv E_5$$

Then plug in $D = C_1 \sigma_e n_{j_0+1}^{-\frac{1}{2}+\frac{1}{(j_0+1)K}}$ and, similarly to E_4, obtain $E_5 \le C_2/n$ by Assumption 1. Therefore

$$E_3 \le \sum_{j=1}^{N-1} \sigma_e^2/n_j \cdot C_2/n \cdot \left(C_1 \sigma_e n_{j+1}^{-\frac{1}{2}+\frac{1}{(j+1)K}}\right)^{-2} \le$$
$$\le C_6 \sum_j (\ln n)^{-j} \cdot n^{-1} \cdot (\ln n)^{(j+1)(1-\frac{2}{(j+1)K})} = C_6 \sum_j \frac{(\ln n)^{1-2/K}}{n} \le C_6 \frac{(\ln n)^{2-2/K}}{n} \quad \square$$

2.4 Lower bound for expected square error

Let us, as before, have exactly n observations on the interval $[0, T]$, and we only need to consider $T = 1$. Then $X(1) = X_n$.

To get the lower bound for the expected squared loss of the Bayesian estimate $\hat{X}_n$, consider the estimator

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

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

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

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

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

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

Proof:
Consider random variables $Z_m,\ m = 1, 2, ...$ having density $\tilde{\psi}$ so that $X_m = X_{m-1} + I_m Z_m$, and let $W_m = Z_m + e_m$. Joint distribution of Z_m, W_m does not depend on n and

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

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,

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

Let $\hat{Z}_n := \mathbb{E}(Z_n \mid W_n)$. Then $\mathbb{E}(\check{X}_n - X_n)^2 = P(I_n = 1)\mathbb{E}(\hat{Z}_n - Z_n)^2$. Clearly, $\mathbb{E}(\hat{Z}_n - Z_n)^2 > 0$ and $P(I_n = 1) = 1 - e^{-\lambda/n} = \lambda/n + o(n^{-1})$. This gives us the statement of Proposition with $C = \mathbb{E}(\hat{Z}_n - Z_n)^2$. $\square$

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-instant change in the process. The following informal argument shows what one can hope for with different error distributions.

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

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

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

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

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

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

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

2.5 Comparison to Kalman filter

The optimal *linear* filter for our problem is the well-known Kalman filter (see e.g. [4]). Its asymptotic rate is of order $n^{-1/2}$:

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

This rate is much worse than $\ln^M n/n$ given by Theorem 2. See Fig. 1 for a graphical comparison of linear (Kalman) and optimal non-linear filters.

3 Parameter estimation for the jump process

Next, our goal is to estimate the parameters of the process itself, that is, the time-intensity λ and parameter θ describing jump density h_θ, based on observations $Y_j, 0 \leq j\tau \leq T$. 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: $0 < \Lambda_1 \leq \lambda \leq \Lambda_2$.

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 the 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_j, 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 $m := \lfloor n^\beta \rfloor$ observations each. Let the total number of Blocks $b = \lfloor \frac{Tn}{m} \rfloor$ (we would get a slightly suboptimal estimate by ignoring some observations).

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. 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 jumps occurred on any of Blocks $k, k+1$.

Case 2. $\sqrt{m}|Z_{k+1} - Z_k| > m^\alpha$, $\sqrt{m}|Z_k - Z_{k-1}| \leq m^\alpha$, $\sqrt{m}|Z_{k+2} - Z_{k+1}| \leq m^\alpha$.
In this case we conclude that a jump occurred exactly between Block k and Block $k+1$, that is, at the time $t = mk/n$. Let us 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 has occurred in the middle of Block k, that is, at the 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 an interval of length $2m/n$ is asymptotically equivalent to $\lambda^2(m/n)^2 b \approx \lambda^2 Tm/n$. We will pick T, β to 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 $\delta_i := \xi_i^* - \xi_i$ is based on the difference of averages of 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 2.2, 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 $\chi_{m,k} = m^{-0.5} \sum_{i=1}^m e_{m(k-1)+i}$ is the sum of normalized errors. Further,

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

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

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

and an application of Chebyshev's inequality yields

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

Picking j such that $\frac{1}{Kj} < \alpha$ and for m large enough, summing up over $\lfloor Tn\, m^{-1} \rfloor$ blocks, we obtain

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

$$P(\textit{Type I error}) < const \cdot Tn\, m^{-1} \exp(-m^{1/j}) \to 0. \quad \square$$

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

$$max\{|Z_{k-1} - Z_k|, |Z_{k+1} - Z_k|\} \le m^{\alpha - 0.5}$$

Of Blocks $k-1$, $k+1$, pick the one 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, the averages of $X(t)$ on Blocks k and $k-1$, $\overline{X}_k$ and $\overline{X}_{k-1}$, differ by at least $\xi/2$ and, conditional on ξ,

$$P(|Z_{k-1} - Z_k| \le m^{\alpha-0.5}) \le 2P(|Z_k - \overline{X}_k| > |\xi|/4 - m^{\alpha-0.5}) \le$$
$$2P(2|\chi_{m,k}| > |\xi|\sqrt{m}/4 - m^\alpha) < 2P\left[|\chi_{m,k}| > const \cdot m^\alpha(m^\varepsilon/4 - 1)\right],$$

as long as $|\xi| > m^{-0.5+\alpha+\varepsilon}$, for an arbitrary $\varepsilon > 0$. Picking j and using Assumption 1 in a way similar to Lemma 2, we obtain, after summing up over $\lfloor Tn\,m^{-1} \rfloor$ blocks, the upper bound on Type II errors $const \cdot Tn\,m^{-1}\exp(-m^{1/j})$. This bound will go to 0 faster than any power of n and can be ignored.

We still need to consider the case of "small jumps"

$$P\left(\bigcup_i \left\{|\xi_i| \le m^{-0.5+\alpha+\varepsilon}\right\}\right) \le const\,(\lambda T + o(T))\,m^{-0.5+\alpha+\varepsilon},$$

using the assumption 3(c) below 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 $const \cdot \lambda^2 Tm/n$. Summing up, we obtain

Lemma 3.

$$P(\textit{Type II error}) < const \cdot \lambda T\, max\{n^{(-0.5+\alpha+\varepsilon)\beta}, \lambda n^{\beta-1}\} \quad \square$$

3.2 Asymptotic behavior of parameter estimates

Since the jump magnitudes are independent of locations, we can determine the behavior of estimates separately, that is consider first an estimate of λ, and then an estimate of θ. Suppose that the true values of parameters are λ_0 and θ_0. Let t_i^* be consecutive jumps of $X(\cdot)$ determined by Cases 2, 3. Estimate the intensity λ by

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

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 h_θ 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, $A'(\theta) = \mathbb{E}_\theta B(\xi)$ and $I(\theta) := A''(\theta) = Var_\theta[B(\xi)]$ is Fisher information. We follow the discussion in [9, Example 7.1].

Assumption 3.
a) $\theta \in \Theta$, where Θ is a bounded subset of $\mathbb{R}$
b) $I(\theta) \ge I_{min} > 0$
c) $h_\theta(x)$ is bounded in a neighborhood of 0, uniformly in θ.
d) There is a constant $\gamma > 0$ such that for large enough M there exists $\Delta = \Delta(M)$ such that $P(|\xi_i| > \Delta) = o(M^{-1})$ and

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

Define the log-likelihood function based on N^* estimated jumps (determined in Cases 2, 3)

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

Theorem 3. *Let Assumptions 1-3 hold. Then the maximum likelihood estimate $\theta^* = argmax_{\theta\in\Theta}\, L^*(\theta)$ is asymptotically normal, that is*

$$\sqrt{\lambda_0 T}\ (\theta^* - \theta_0) \to \mathcal{N}[0, I(\theta_0)^{-1}] \tag{8}$$

in distribution as $n \to \infty$, and $T \to \infty$ no faster than $T = n^\kappa$, where $\kappa < (3+2\gamma)^{-1}$.

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

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

where E is the exceptional set where Type I and II errors occurred, ξ_i^0 are the estimates of ξ resulting from these errors, and δ_i are "magnitude errors" discussed in the beginning of Section 3.

Pick β, κ, α and ε such that

$$\kappa + \beta - 1 < 0, \quad \kappa + \beta(-0.5 + \alpha + \varepsilon) < 0, \quad \kappa(1+\gamma) < \beta/2$$

With α, ε arbitrarily small, this is achieved when $2\kappa(1+\gamma) < \beta < (1-\kappa)$, so that $\kappa < (3+2\gamma)^{-1}$. This choice, applied to Lemmas 2 and 3, ensures that $P(E) \to 0$ as $n \to \infty$. Therefore we can disregard this set and consider only $\xi_i^* = \xi_i + \delta_i$. Let N be the total number of jumps on $[0, T]$, $N \to \lambda_0 T$ almost surely. 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 the given mean and variance. Next, we would like to show that the estimate θ^* based on $\{\xi_i^*\}_{1\le i\le N^*}$ 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 concave function, the same is true for $L^*(\theta)$. Furthermore, for any θ in a neighborhood of $\hat\theta$,

$$L(\theta) = L(\hat\theta) - \frac{(\theta - \hat\theta)^2}{2} N\, A''(\hat\theta) + o(\theta - \hat\theta)^2 \tag{9}$$

According to Lemma 4 below, $L(\theta) - L^*(\theta) = o(1)$ uniformly in θ, so that their maximal values $L(\hat\theta) - L^*(\theta^*) = o(1)$ also. Therefore, $L(\hat\theta) - L(\theta^*) = o(1)$, and from (9), it follows that $(\theta^* - \hat\theta)^2 N\, A''(\hat\theta) = o(1)$. Hence, $(\theta^* - \hat\theta) = o(N^{-1/2})$, and the statement of Theorem follows from the well-known similar statement for $\hat\theta$. □

Lemma 4. *Under the conditions of Theorem 3, $|L(\theta) - L^*(\theta)| = o(1)$ uniformly in θ, outside some exceptional set E_1 with $P(E_1) = o(1)$.*

Proof:
Let $M = \lambda_0 T$ and pick Δ according to Assumption 3(d), so that probability of event $E_2 := \bigcup_i \{|\xi_i| > \Delta\}$ is $o(1)$. Also, for a $\nu > 0$, consider $E_3 := \bigcup_i \{|\delta_i| > n^{-\beta/2+\nu}\}$. Since ξ_i are based on sums of $m = \lfloor n^\beta \rfloor$ i.i.d. terms, they are asymptotically normal, and $P(E_3) \to 0$ exponentially fast for any $\nu > 0$. Let another exceptional set $E_4 := \{N > \lambda_0 T + (\lambda_0 T)^\zeta\}$, $P(E_4) \to 0$ as long as $\zeta > 1/2$. Finally, $N = N^*$ outside the exceptional set E.

Thus, excluding $E_1 := E_2 \cup E_3 \cup E_4 \cup E$, for large enough n,

$$|L(\theta) - L^*(\theta)| \le \sum_{i=1}^N |\ln h_\theta(\xi_i) - \ln h_\theta(\xi_i + \delta_i)| \le \sum_{|\xi_i|\le\Delta} |\theta B'(\xi_i)| \cdot |\delta_i| <$$

$$b_\Delta N n^{-\beta/2+\nu} \le const \cdot (\lambda_0 T + o(T))^{1+\gamma}\, n^{-\beta/2+\nu} \le const \cdot \lambda_0^{1+\gamma}\, n^{-\beta/2+\nu+\kappa(1+\gamma)}$$

The statement of Lemma follows as $\kappa(1+\gamma) < \beta/2$. $\square$

Assumption 3(d) 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 $\Delta(M) = \ln M$, and then $b_\Delta = \theta \ln(M)/2 = o(M^\gamma)$ for arbitrarily small γ.

4 Simulations

The author has simulated the optimal filter for the jump process using the particle filtering method described in [6]. The results of simulation are given in Table 1. The parameters are $\lambda = 0.2$, Normal jumps with variance $= 10^2$ and $\sigma_e = 1$, and $n = 1/\tau$ varies.
For each sample path of the process, the ratio of effectiveness of non-linear filter to the linear one was found:

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

where $\hat{X}_t$ is the optimal non-linear filter at time t and $\tilde{X}_t$ is the Kalman filter. Also, the estimate of mean square error $\mathbb{E}(\hat{X}_t - X_t)^2$ of the optimal non-linear filter is given, along with its standard error. In each case, $N = 25$ sample paths were generated.

n	5	10	15	20	25	30	50
R	0.353	0.252	0.244	0.232	0.229	0.211	0.190
MSE(Opt.)	0.277	0.197	0.149	0.141	0.132	0.126	0.088
St.Error (MSE Opt.)	0.032	0.016	0.013	0.009	0.011	0.012	0.009

Table 1. 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 1).
b) Noise variance σ_e^2 increases.
c) jump magnitudes increase, making the process more "ragged".

Acknowledgments

The author would like to thank Prof. Anatoly Skorokhod for proposing the problem, Prof. Raoul LePage and the anonymous referee for valuable suggestions.

References

[1] C. Andrieu, M. Davy, and A. Doucet, *Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions*, IEEE Trans. Signal Process., 51 (2003), pp. 1762–1770. MR1996962

[2] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation*, Wiley, New York, 2001.

[3] P. Brémaud, *Point processes and queues*, Springer-Verlag, New York, 1981. MR0636252

[4] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*, Springer-Verlag, New York, 2002. MR1894099

[5] H. Chernoff and H. Rubin, *The estimation of the location of a discontinuity in density*, in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I, Univerity of California Press, Berkeley and Los Angeles, 1956, pp. 19–37. MR0084910

[6] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte-Carlo methods in practice*, Springer-Verlag, New York, 2001. MR1847783

[7] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models*, vol. 29 of Applications of Mathematics (New York), Springer-Verlag, New York, 1995. MR1323178

[8] I. A. Ibragimov and R. Z. Has′minskiĭ, *Statistical estimation. Asymptotic theory*, Translated from the Russian by Samuel Kotz, Springer-Verlag, New York, 1981. MR0620321

[9] E. L. Lehmann, *Theory of point estimation*,Wiley, New York, 1983. MR0702834

[10] O. V. Makhnin, *Filtering for some stochastic processes with discrete observations*, PhD thesis, Michigan State University, 2002.

[11] N. Portenko, H. Salehi, and A. Skorokhod, *On optimal filtering of multitarget tracking systems based on point processes observations*, Random Oper. Stochastic Equations, 5 (1997), pp. 1–34. MR1443419

[12] N. Portenko, H. Salehi, and A. Skorokhod, *Optimal filtering in target tracking in presence of uniformly distributed errors and false targets*, Random Oper. Stochastic Equations, 6 (1998), pp. 213–240. MR1630999

[13] Y. Shimizu and N. Yoshida, *Estimation of parameters for diffusion processes with jumps from discrete observations*, Statistical Inference for Stochastic Processes, 9 (2006), pp. 227–277. MR2252242

[14] M. Vellekoop and J. Clark, *A nonlinear filtering approach to changepoint detection problems: Direct and differential-geometric methods*, SIAM Review, 48 (2006), pp. 329–356. MR2240595