SIAM REVIEW
Vol. 48, No. 2, pp. 329–356

A Nonlinear Filtering Approach to Changepoint Detection Problems: Direct and Differential-Geometric Methods*

M. H. Vellekoop†
J. M. C. Clark‡

Abstract. A benchmark change detection problem is considered which involves the detection of a change of unknown size at an unknown time. Both unknown quantities are modeled by stochastic variables, which allows the problem to be formulated within a Bayesian framework. It turns out that the resulting nonlinear filtering problem is much harder than the well-known detection problem for *known* sizes of the change, and in particular that it can no longer be solved in a recursive manner. An approximating recursive filter is therefore proposed, which is designed using differential-geometric methods in a suitably chosen space of unnormalized probability densities. The new nonlinear filter can be interpreted as an adaptive version of the celebrated Shiryayev–Wonham equation for the detection of a priori known changes, combined with a modified Kalman filter structure to generate estimates of the unknown size of the change. This intuitively appealing interpretation of the nonlinear filter and its excellent performance in simulation studies indicates that it may be of practical use in realistic change detection problems.

Key words. change detection, nonlinear filtering, differential geometry

AMS subject classification. 60G35

DOI. 10.1137/050647438

1. Introduction.

Many problems in engineering necessitate the quick and accurate detection of sudden changes in dynamical systems. When one tries to track a certain object (such as an airplane) on radar, there may be a change of flight path, and quick detection of such a change is crucial if one wants to filter out noise from the radar observations. When analyzing seismic data to predict earthquakes or to locate possible oil wells, it is of obvious importance to detect whether certain changes in the collected data are significant or not. Complex biomedical signals, such as the EEG, can be analyzed only by segmentation, which requires change detection procedures that can be applied automatically to the large quantities of data that are generated

*Published electronically May 2, 2006. This paper originally appeared in *SIAM Journal on Control and Optimization*, Volume 42, Number 2, 2003, pages 469–494. A preliminary short version of this paper appeared as "Changepoint Detection Using Nonlinear Filters" in Proceedings of the 4th European Control Conference, Brussels, 1997.
http://www.siam.org/journals/sirev/48-2/64743.html

†Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands (m.h.vellekoop@math.utwente.nl).

‡Department of Electrical and Electronic Engineering (also Centre for Process Systems Engineering), Imperial College, Exhibition Road, London SW7 2BT, United Kingdom (j.m.c.clark@imperial.ac.uk).

in the process. And in chemical plants, a safe operation environment can often only be guaranteed by close monitoring of many signals since immediate action is required once an undesired change in operation conditions has been detected.

The problem of detecting parameter changes in dynamical systems on the basis of noisy observations has therefore attracted a lot of attention in the last thirty years, and the literature dealing with it is extensive. For good surveys of the field and further references, the reader is referred to the papers by Basseville [2], Iserman [13], Lai [21], and Willsky [26], and the excellent book by Basseville and Nikiforov [3]. As is pointed out in [2], the basic method proposed in most of the literature on change detection consists of two steps. First, the problem is transformed into a standard problem by generating certain *residuals*: change indicating signals which are ideally close to zero when no change occurs. Then, in a separate second step, sophisticated statistical methods are developed to solve the resulting detection problem in terms of these residuals. In this paper we will provide a contribution to the second step; the first step will very much depend on the particular application that one wishes to consider, and it is therefore not treated here.

The statistical tools used in this second step usually originate in the field of sequential hypothesis testing, and a wide variety of results concerning their use in change detection problems is now available [21]. Typically these tests compare a certain functional of the observations with a threshold, and an alarm is raised as soon as this threshold is reached. Important examples of such schemes include the celebrated CUSUM and generalized likelihood ratio (GLR) schemes.

In this paper we want to propose a different approach, in which change detection is considered to be an on-line estimation problem in which a dynamical system possesses certain parameters which may exhibit sudden changes that need to be detected [8]. In our Bayesian formulation of the problem we assume that *both* the time *and* the size of the change are unknown a priori, thus acknowledging the fact that in many practical situations the behavior of the residual after the change is not completely known and detection is thus necessarily linked to estimation. In many practical detection problems, one does not only want to know *that* a change has occurred; one also wants to obtain on-line estimates of relevant statistics *after* the change.

We do not consider the problem in which one tries to detect changes off-line, or where one tries to estimate the *time* of the change. GLR methods and maximum likelihood estimators have been defined for such problems; see, for example, the analysis in [20]. Results concerning such off-line methods have been derived [6] under the assumption that one does not exactly know the correct model after the change (although the assumed model should be "close" to the correct model in a predescribed sense). Those results on the off-line detection problem are in that sense complementary to the methods we will propose here, but since their goal and assumptions differ from ours, we redirect the reader to the reference given above for further information.

In a continuous time framework, we can define a basic change detection problem concerning a simple jump process, which is equal to zero up to a certain random time τ, then jumps to a random value X, after which it stays constant again. We assume that such a signal can be observed in white noise, and the purpose is to study the conditional distribution of the signal given all the observations up to the current time t and relevant statistics generated by this conditional distribution.

If the value after the change X is known a priori, one can find an explicit stochastic differential equation for the Bayesian a posteriori probability that a change has occurred—the celebrated Shiryayev–Wonham equation [22, 28]. In fact, this statistic

can still be calculated recursively if X is known to belong to a finite set of possible values. The problem can then be solved using the theory of hidden Markov models [10].

However, if there exists an infinity of possible jump sizes X, then the problem becomes much harder, since the detection and estimation problems now become closely interrelated. The problem of finding on-line estimates for an unknown constant X which is observed in Gaussian noise can be solved using a Kalman filter, which consists of a finite number of explicit stochastic differential equations in which the observations need to be fed to generate the estimates. But this filter will not perform well when the value of X suddenly changes. The Kalman filter uses the conditional variance of its estimate in the estimation process, and when X changes, the filter will for a long time "refuse to believe" that something happened. Where the Shiryayev–Wonham filter can be used for *known* values of X to detect the unknown time of the change τ and the Kalman filter can be used to estimate X for a known value of τ, when both X and τ are unknown, neither of the schemes will produce good results.

We need to be a bit more specific about what we mean by "good results" when trying to detect sudden changes. In every change detection problem there is a tradeoff between *detection speed* (if a change occurs, how long does it take to notice the change?) and the *probability of false alarm* (if no change occurs, how often will the system still raise an alarm?). When the variance of the noise process goes to zero it should become easier to detect the jumps, so the detection speed should go to zero. It turns out that the rate at which this detection speed goes to zero can be characterized explicitly, and this *asymptotic detection delay* is therefore an important characteristic of a change detection scheme. Since we will calculate the conditional probabilities that a jump has occurred, we can control the probability of false alarm rather easily by choosing our threshold for the alarm appropriately.

In this paper we will thus formulate and study an approximation to the optimal filter for processes containing a jump of unknown size and show its excellent performance in terms of this asymptotic detection delay. The conditional estimates we are interested in can be characterized using the nonlinear filtering theory for discontinuous stochastic processes, and the optimal nonlinear filter for this case has been derived in [11]. As is often the case in nonlinear filtering problems [7], this filter does not admit a finite-dimensional recursive implementation, such as the Kalman filter or Shiryayev–Wonham filter we mentioned earlier. However, since the conditional probability distribution of the process based on the noisy observations can be derived explicitly, this may be used as a starting point for approximations which are suboptimal yet can be implemented recursively.

Such approximations can be interpreted as a *projection* of a trajectory in an uncountably infinite-dimensional space of probability densities onto a finite-dimensional manifold in that space. Our approach extends a powerful statistical projection technique, which was introduced by Brigo, Hanzon, and LeGland in order to filter nonlinear diffusions [4, 5, 12], and which is based on differential-geometric methods in statistical information theory [1, 17, 18]. We will show that the resulting filter can be parametrized as a modified Kalman filter which feeds an adaptive version of the Shiryayev–Wonham filter for *known* changes that we mentioned earlier. This interpretation may help to explain its excellent performance when compared to other detection schemes, as will be illustrated in a number of simulation studies.

The structure of this paper is as follows. In the next section we introduce the stochastic change detection model and derive the nonlinear filter equations for such

models. In sections 3 and 4 we formulate two recursive filtering algorithms, which are based on information-theoretic approximations and approximation of conditional moments, respectively. In section 5 we discuss the relationship between these two filters. In section 6 we introduce and analyze a three-dimensional nonlinear filter based on the results derived in earlier sections, and we illustrate the performance of this filter in some simulation studies in section 7. We finish with conclusions and suggestions for further research in the last section.

2. The Change Detection Model and Optimal Filter Equations. In this section we define the optimal filter estimates as generated by the conditional probability distributions given all the information in the observations up to the current time t. To define a notion of "available information" we will set up the abstract framework in terms of σ-algebras generated by observations, since this allows us to characterize the conditional distributions explicitly in Theorem 2.1. Readers who are not familiar with this setup may find it helpful to take a look at the derivation for the analogous discrete time case, which is given directly after the proof of Theorem 2.1.

Let $(\Omega, \mathcal{F}, P)$ be the complete canonical probability space for Brownian motion, i.e., $\Omega = C([0,\infty[)$, the set of all scalar continuous functions on $\mathbb{R}^+$, $\mathcal{F}$ the usual σ-algebra generated by the topology of uniform convergence on compact sets, and P the Wiener measure on $\mathcal{F}$. Let $\{\mathcal{F}_t, t \geq 0\}$ be a filtration satisfying the usual conditions, i.e., an increasing family of σ-algebras which is right-continuous and such that $\mathcal{F}_0$ contains all P-null sets. We will use $\mathbb{P}(A)$ as a shorter notation for $P(\{\omega \in \Omega : A(\omega)\})$ in this paper, where $A(\omega)$ is a condition on ω, and we will denote the expectation operator by $\mathbb{E}$, so for a stochastic variable Z we use $\mathbb{E}Z$ to denote $\int_\Omega Z(\omega) dP(\omega)$.

Consider the signal

$$S_t = \begin{cases} 0, & 0 \leq t < \tau, \\ X, & t \geq \tau, \end{cases} \tag{2.1}$$

where $X \in \mathbb{R}$ and $\tau \in \mathbb{R}^+$ are two independent finite random variables on Ω with distribution functions F, G, respectively. We will assume that X and τ have probability densities f and g, so $\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du$ for all $x \in \mathbb{R}$ and $\mathbb{P}(\tau \leq r) = G(r) = \int_0^r g(u)du$ for all $r \in \mathbb{R}^+$. We assume that f and g are both strictly positive on their domains $\mathbb{R}$ and $\mathbb{R}^+$ in this paper, unless we explicitly state otherwise. We will use $E_t = \mathbf{1}_{\{t \geq \tau\}}$ to denote a unit jump process, so $S_t = XE_t$ for all $t \geq 0$.

We will suppose that the signal S_t can be observed in additive white noise. We therefore define a scalar observation process $\{Y_t^\epsilon, t \geq 0\}$ by

$$dY_t^\epsilon = S_t\, dt + \epsilon\, dW_t, \qquad Y_0^\epsilon = 0, \tag{2.2}$$

where $\{(W_t, \mathcal{F}_t), t \geq 0\}$ is a standard Brownian motion process on $(\Omega, \mathcal{F}, P)$, which is independent of both X and τ, and where ϵ is a real positive parameter representing the noise intensity.

Let $\mathcal{S}_t$ be a second filtration which is contained in $\mathcal{F}_t$ and satisfies the usual conditions as well, such that both X and $\mathbf{1}_{\{t \geq \tau\}}$ are $\mathcal{S}_t$-measurable for all $t \geq 0$, i.e., X is $\mathcal{S}_0$-measurable and τ is a stopping time with respect to $\mathcal{S}_t$. The σ-algebra $\mathcal{S}_t$ then represents the *state information* up to time t. Likewise, we define $\mathcal{Y}_t^\epsilon$ as the σ-algebra generated by the observation process up to time t:

$$\mathcal{Y}_t^\epsilon \stackrel{\text{def}}{=} \sigma(\{Y_s^\epsilon, 0 \leq s \leq t\}) \subset \mathcal{F}_t.$$

We are interested in the analysis of the conditional laws of the signal S_t, given the observations record up to time t. In particular, we would like to estimate the magnitude of the jump X at time t and the probability that the jump has already occurred before time t:

$$\mathbb{E}[X \mid \mathcal{Y}_t^\epsilon], \qquad \mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon).$$

Since such statistics can be calculated from the conditional distribution of S_t given the observation record $\mathcal{Y}_t^\epsilon$, we will study this conditional law of the signal on a fixed finite time interval $[0, T]$. Our results concerning the postjump time period are therefore conditioned on the set $\{\omega \in \Omega : \tau(\omega) \leq T\}$. We will assume that there exists a $\delta > 0$ such that

$$\mathbb{E}\, \exp[\, \delta X^2\,] < \infty \tag{2.3}$$

throughout the rest of the paper, to make sure that certain conditional expectations that we wish to calculate do indeed exist.

THEOREM 2.1. *Under the assumptions mentioned above, the conditional probability density of the signal S_t, given the observations $\{Y_s^\epsilon, 0 \leq s \leq t\}$, is given by*

$$(\rho_t^\epsilon)^{-1}\, [\, (1 - G(t))\, \delta_0(x) + q_t^\epsilon(x)\,], \tag{2.4}$$

where

$$q_t^\epsilon(x) = f(x) \int_0^t g(r)\, \exp\left[\frac{x}{\epsilon^2}\, (Y_t^\epsilon - Y_r^\epsilon) - \frac{x^2}{2\epsilon^2}\, (t - r) \right] dr,$$

$$\rho_t^\epsilon = 1 - G(t) + \int_{\mathbb{R}} q_t^\epsilon(x)\, dx.$$

Proof of Theorem 2.1. We derive an expression for the conditional distribution of the signal through the Kallianpur–Striebel formula and Girsanov's theorem. One may show that the necessary conditions for these methods to be applicable are indeed satisfied [14] because of (2.3). We then find for the conditional distribution of S_t given the observations [14, 27]

$$\mathbb{P}(S_t \in B \mid \mathcal{Y}_t^\epsilon) = (\rho_t^\epsilon)^{-1} \int_{\mathbb{R}} \int_0^\infty \mathbf{1}_B(x \mathbf{1}_{\{t \geq r\}})\, e^{\frac{Z(x,r,t)}{\epsilon^2}}\, dG(r) dF(x), \tag{2.5}$$

where B is a Borel-measurable set, $\mathbf{1}_B$ is the indicator function for the set B, ρ_t^ϵ is a normalization factor equal to the double integral of the right-hand side of this expression for $B = \mathbb{R}$, and

$$\begin{aligned} Z(x, r, t) &= \int_0^t x\, \mathbf{1}_{\{s \geq r\}}\, dY_s^\epsilon - \frac{1}{2} \int_0^t (x\, \mathbf{1}_{\{s \geq r\}})^2\, ds \\ &= \left[x\, (Y_t^\epsilon - Y_r^\epsilon) - \frac{x^2}{2}\, (t - r) \right] \mathbf{1}_{\{r \leq t\}}. \end{aligned}$$

After decomposing the inner integral in (2.5) into the intervals $[0, t[$ and $[t, \infty[$ we find

$$\mathbb{P}(S_t \in B \mid \mathcal{Y}_t^\epsilon) = (\rho_t^\epsilon)^{-1} \int_B \int_0^t e^{\frac{Z(x,r,t)}{\epsilon^2}}\, dG(r) dF(x) + (\rho_t^\epsilon)^{-1}\, (1 - G(t)) \int_B \delta_0(x)\, dx.$$

Here and in what follows we will allow the slight abuse of notation which represents the Dirac measure with its unit mass in the origin as an integral over a Dirac density $\delta_0(x)$,

i.e., $\int_B \delta_0(x)dx = \mathbf{1}_{\{0\in B\}}$. For $B = \mathbb{R}$ we obtain an expression for the normalization factor:

$$\rho_t^\epsilon = \int_{\mathbb{R}} \int_0^t e^{\frac{Z(x,r,t)}{\epsilon^2}} \, dG(r)dF(x) + (1 - G(t)),$$

and this proves the result. □

Using Îto's differentiation rule, one may easily check that the density $q_t^\epsilon(x)$ satisfies the following Îto stochastic differential equation:

$$dq_t^\epsilon(x) = f(x)g(t)\,dt + \frac{x}{\epsilon^2} q_t^\epsilon(x)\,dY_t^\epsilon\,, \tag{2.6}$$

with initial value $q_0^\epsilon(x) = 0$ for all $x \in \mathbb{R}$. This is the Duncan–Mortensen–Zakai equation of nonlinear filtering for the conditional distribution outside the origin, and it may be derived directly using the infinitesimal generator of the Markov process $\{S_t,\, t \geq 0\}$.

To get some intuition for this continuous time result, we now briefly look at a discrete time analogue. Define for $n \in \mathbb{N}$

$$S_n(\omega) = \begin{cases} 0, & n < \bar{\tau}(\omega), \\ X(\omega), & n \geq \bar{\tau}(\omega), \end{cases}$$

with X as before, and where the discrete *jump time* $\bar{\tau} : \Omega \to \mathbb{N}^+$ is a stochastic variable on the positive integers with $\mathbb{P}(\bar{\tau} = k+1) = g_k > 0$ for $k \in \mathbb{N}$ and $\mathbb{P}(\bar{\tau} = 0) = 0$. We collect discrete observations in the set $\mathcal{Y}_n^\epsilon = \{Y_n^\epsilon, n = 0, 1, \ldots, N\}$ according to

$$Y_n^\epsilon - Y_{n-1}^\epsilon = S_n + \epsilon W_n, \qquad Y_0 = 0,$$

where the $\{W_n, n \geq 0\}$ are independent and identically distributed Gaussian variables with mean zero and variance one, which we assume to be independent of both X and the jump time $\bar{\tau}$.

We now want to find the conditional distribution of S_n given the observations in $\mathcal{Y}_n^\epsilon$. It will be convenient to use the stochastic process defined by

$$Z_n = Y_n^\epsilon - Y_{n-1}^\epsilon = S_n + \epsilon\, W_n, \qquad Z_0 = 0.$$

Since the process $\{Y_k^\epsilon,\, 0 \leq k \leq n\}$ can be reconstructed from $\{Z_k,\, 0 \leq k \leq n\}$ and vice versa, we have that

$$p_{S_n|Y_1^\epsilon, Y_2^\epsilon, \ldots, Y_n^\epsilon} = p_{S_n|Z_1, Z_2, \ldots, Z_n}.$$

We find

$$\begin{aligned} p_{S_n|\mathcal{Y}_n^\epsilon} &= p_{S_n|Z_1,\, Z_2, \ldots, Z_n} \\ &= \sum_{k=1}^{n} \mathbb{P}(\bar{\tau} = k \mid Z_1, \ldots, Z_n)\, p_{S_n|Z_1,\ldots,Z_n,\,\bar{\tau}=k} \\ &\qquad + \mathbb{P}(\bar{\tau} > n \mid Z_1, \ldots, Z_n)\, p_{S_n|Z_1,\ldots,Z_n,\,\bar{\tau}>n}. \end{aligned} \tag{2.7}$$

We use the notation

$$\phi_\epsilon(x) = \frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{1}{2}x^2/\epsilon^2}$$

for the Gaussian density with standard deviation ϵ. For the first factor we find that

$$\mathbb{P}(\bar{\tau} = k \mid Z_1, \ldots, Z_n) = \frac{\mathbb{P}(Z_1, \ldots, Z_n \mid \bar{\tau} = k)\,\mathbb{P}(\bar{\tau} = k)}{\mathbb{P}(Z_1, \ldots, Z_n)} = \frac{\mathbb{P}(Z_1, \ldots, Z_n \mid \bar{\tau} = k) g_{k-1}}{\mathbb{P}(Z_1, \ldots, Z_n)}. \tag{2.8}$$

We now use the fact that conditioned on the event $\{\omega \in \Omega : \bar{\tau}(\omega) = k\}$, the first k variables $\{Z_i,\ i = 1, \ldots, k-1\}$ are independent of $\{Z_i,\ i = k, \ldots, n\}$ and $N(0, \epsilon^2)$ distributed:

$$\mathbb{P}(Z_1, \ldots, Z_n \mid \bar{\tau} = k) = \left[\prod_{i=1}^{k-1} \phi_\epsilon(Z_i)\right] \cdot \mathbb{P}(Z_k, \ldots, Z_n, \mid \bar{\tau} = k),$$

and express the distribution of the later stochastic variables $\{Z_i,\ i = k, \ldots, n\}$ in terms of the probability density f:

$$\begin{aligned}\mathbb{P}(Z_k, \ldots, Z_n, \mid \bar{\tau} = k) &= \int_{-\infty}^{\infty} \mathbb{P}(Z_k, \ldots, Z_n \mid \bar{\tau} = k,\ X = u)\, f(u)\, du \\ &= \int_{-\infty}^{\infty} f(u) \left[\prod_{i=k}^{n} \phi_\epsilon(Z_i - u)\right] du.\end{aligned}$$

Substituting all this in (2.8) gives the first probability on the right-hand side of (2.7). For the second probability we find

$$\begin{aligned}\mathbb{P}(\bar{\tau} > n \mid Z_1, \ldots, Z_n) &= \frac{\mathbb{P}(Z_1, \ldots, Z_n \mid \bar{\tau} > n)\,\mathbb{P}(\bar{\tau} > n)}{\mathbb{P}(Z_1, \ldots, Z_n)} \\ &= \frac{1}{\mathbb{P}(Z_1, \ldots, Z_n)} \left[\prod_{i=1}^{n} \phi_\epsilon(Z_i)\right] \sum_{k=n+1}^{\infty} g_{k-1}.\end{aligned}$$

We must now determine the distribution functions $p_{S_n \mid Z_1, \ldots, Z_n,\, \bar{\tau} = k}$ and $p_{S_n \mid Z_1, \ldots, Z_n,\, \bar{\tau} > n}$ in (2.7). The last one is trivial since we know that $S_n = 0$ if $\bar{\tau} > n$:

$$p_{S_n \mid Z_1, \ldots, Z_n,\, \bar{\tau} > n} = \delta_0(S_n).$$

To find the first one, we note that

$$p_{S_k \mid Z_1, \ldots, Z_k,\, \bar{\tau} = k} = \frac{f(S_k)\,\phi_\epsilon(Z_k - S_k)}{\int_{-\infty}^{\infty} f(u)\,\phi_\epsilon(Z_k - u) du},$$

while for $j \geq k$ we have, conditioned on $\bar{\tau} = k$, that $S_j = S_k$ so

$$p_{S_n \mid Z_1, \ldots, Z_n,\, \bar{\tau} = k} = \frac{f(S_n) \prod_{i=k}^{n} \phi_\epsilon(Z_i - S_n)}{\int_{-\infty}^{\infty} f(u) \prod_{i=k}^{n} \phi_\epsilon(Z_i - u)\, du}.$$

We have now determined all terms in (2.7) and find

$$\begin{aligned} p_{S_n|Z_1,\dots,Z_n} &= \sum_{k=1}^{n} g_{k-1} \left[\prod_{i=1}^{k-1} \phi_\epsilon(Z_i)\right] \left[\prod_{i=k}^{n} \phi_\epsilon(Z_i - S_n)\right] f(S_n)/c \\ &\quad + \sum_{k=n+1}^{\infty} g_{k-1} \left[\prod_{i=1}^{n} \phi_\epsilon(Z_i)\right] \delta_0(S_n)/c \\ &= \sum_{k=0}^{n-1} \frac{g_k}{(\epsilon\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{k}\left(\frac{Z_i}{\epsilon}\right)^2 - \frac{1}{2}\sum_{i=k+1}^{n}\left(\frac{Z_i - S_n}{\epsilon}\right)^2\right] f(S_n)/c \\ &\quad + \left(1 - \sum_{k=0}^{n-1} g_k\right) \frac{1}{(\epsilon\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{Z_i}{\epsilon}\right)^2\right] \delta_0(S_n)/c, \end{aligned}$$

where c is an appropriately chosen normalization constant. Using the fact that $\sum_{i=k+1}^{n} Z_i = Y_n^\epsilon - Y_k^\epsilon$ and noting that the factor $\exp[-\sum_{i=1}^{n} \frac{Z_i^2}{2\epsilon^2}]/c(\sqrt{2\pi}\epsilon)^n$ does not depend on S_n, we rewrite this as follows:

$$\begin{aligned} p_{S_n|\mathcal{Y}_n^\epsilon}(x) &= (N_n^\epsilon)^{-1} \left[q_n^\epsilon(x) + \left(1 - \sum_{k=0}^{n-1} g_k\right) \delta_0(x)\right], \\ q_n^\epsilon(x) &= f(x) \sum_{k=0}^{n-1} g_k \exp\left[\frac{x}{\epsilon^2}(Y_n^\epsilon - Y_k^\epsilon) - \frac{x^2}{2\epsilon^2}(n-k)\right], \\ N_n^\epsilon &= \int_{\mathbb{R}} q_n^\epsilon(x)\, dx + 1 - \sum_{k=0}^{n-1} g_k, \end{aligned}$$

where q_n^ϵ can now be interpreted as the density of S given that a jump has occurred. But after expressing q_{n+1}^ϵ in terms of q_n^ϵ, we then find the following analogue of the Zakai equation for the continuous time nonlinear filtering problem (2.6):

$$q_{n+1}^\epsilon(x) = (\, f(x)\, g_n + q_n^\epsilon(x)\,) \exp\left[\tfrac{x}{\epsilon^2}\,(Y_{n+1}^\epsilon - Y_n^\epsilon) - \tfrac{x^2}{2\epsilon^2}\right]. \tag{2.9}$$

We can interpret this equation to update q_n^ϵ as the result of two separate steps. Between observations the term $f(x)g_n$ is added to $q_n^\epsilon(x)$, and the result is then combined in a Bayesian way with the new observation, by multiplication with the exponential term $\exp\left[\tfrac{x}{\epsilon^2}\,(Y_{n+1}^\epsilon - Y_n^\epsilon) - \tfrac{x^2}{2\epsilon^2}\right]$.

The Duncan–Mortensen–Zakai equation (2.6) for our original continuous time problem suggests that to calculate the optimal filter estimates we have to solve a stochastic partial differential equation on-line. It can indeed be shown that no finite-dimensional sufficient statistic exists for this problem. Since we need such a finite-dimensional statistic, which can be updated on-line in a recursive manner for practical implementation, we will propose and analyze finite-dimensional approximations to the infinite-dimensional optimal filter objects in the following sections.

3. Differential-Geometric Approximations. The first finite-dimensional approximation that we wish to consider uses projection operators in a space of unnormalized probability densities to map the infinite-dimensional optimal filtering objects onto fixed finite-dimensional structures. The appropriate framework for this approximation method is given by the differential-geometrical theory of statistical information

and in particular the theory of statistical manifolds. For an excellent introduction to these relatively new fields, the reader is referred to the book by Amari [1] for the general theory and to the papers by Kulhavý [17, 18, 19] for its application to parameter estimation problems. Most important for the approach we wish to take here is the application of differential-geometric methods to the filtering problem for nonlinear diffusions [4, 5, 12]. Our analysis forms an extension of the work reported in these papers, and we have therefore tried to keep our notation consistent with them whenever possible.

The main idea of our approach will be that we define a finite-dimensional statistical manifold $\mathcal{H}^{1/2}$ in the infinite-dimensional space of unnormalized probability densities. A basis will be derived for the tangent space in every point of this manifold, and we can use these to define a local projection operator which maps the infinitesimal increments generated by the nonlinear filtering equations onto such tangent spaces. The resulting stochastic vector field on $\mathcal{H}^{1/2}$ then defines our nonlinear filter.

It may help the reader to compare $\mathcal{H}^{1/2}$ to a curved manifold in finite-dimensional Euclidean space, where the definition of a manifold in terms of its coordinates and projection onto tangent planes (such as in Figure 1) are much more intuitive.

In order to use a Hilbert space structure, we will work in the space of square roots of unnormalized probability densities. Let $\mathcal{M}$ be the set of all (not necessarily normalized) finite nonnegative measures κ on $\mathbb{R}$ which are absolutely continuous with respect to Lebesgue measure and have Radon–Nikodým derivatives p which are strictly positive Lebesgue almost everywhere. Then we have that the function $\sqrt{p} : x \mapsto \sqrt{p(x)}$ is an element of $\mathcal{L}^2$, the Hilbert space of Lebesgue-square integrable functions from $\mathbb{R}$ to $\mathbb{R}^+ \setminus \{0\}$. Denote the subspace of $\mathcal{L}^2$ consisting of such square roots of strictly positive densities by $\mathcal{R}$. We define on it a metric $d_{\mathcal{R}}$ induced by the norm $\|\cdot\|_{\mathcal{L}^2}$, which in turn defines the *Hellinger metric* $d_{\mathcal{M}}$ on the set of measures κ we started with:

$$d_{\mathcal{M}}(\kappa_1, \kappa_2) = d_{\mathcal{R}}(\sqrt{p_1}, \sqrt{p_2}) = \|\sqrt{p_1} - \sqrt{p_2}\|_{\mathcal{L}^2} = \sqrt{\int_{\mathbb{R}} (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 \, dx}.$$

To find a recursive approximation for the infinite-dimensional optimal filter, we have to define finite-dimensional structures in the infinite-dimensional space $\mathcal{R}$. We will therefore consider an $(m+1)$-dimensional manifold N (with $m \in \mathbb{N}$), which as a

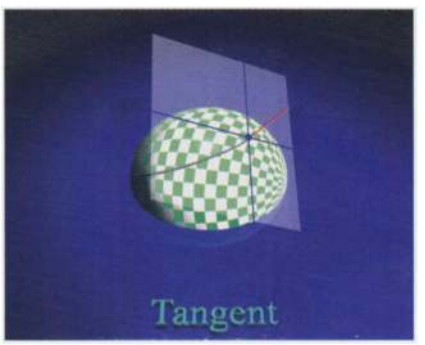


Fig. 1 *Manifold coordinates and tangent space in finite-dimensional space.*

subset of $\mathcal{R}$ is imbedded in the larger Hilbert space $\mathcal{L}^2$. This means that N is locally homeomorphic to $\mathbb{R}^{m+1}$ and is thus described locally by a *chart*: if $\sqrt{p} \in N$, then there exists an open neighborhood $\mathcal{H}^{1/2}$ of $\sqrt{p}$ in N and a homeomorphism $\varphi : \mathcal{H}^{1/2} \to \Theta$ onto an open and convex subset Θ of $\mathbb{R}^{m+1}$. We will assume that there exists in fact one global and smooth coordinate chart for the entire manifold, so we will consider manifolds $\mathcal{H}^{1/2}$ defined by

$$\mathcal{H}^{1/2} = \{\sqrt{p(\cdot,\theta)} : \ \theta = (\theta_0, \theta_1, \ldots, \theta_m) \in \Theta\} = \varphi^{-1}(\Theta),$$

where

$$\left\{ \frac{\partial \varphi^{-1}(\theta)}{\partial \theta_0}, \frac{\partial \varphi^{-1}(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \varphi^{-1}(\theta)}{\partial \theta_m} \right\}$$

is assumed to be a set of linearly independent vectors in $\mathcal{L}^2$ for all $\theta \in \Theta$. To find the differential-geometric structure of such manifolds $\mathcal{H}^{1/2}$ around a point $\sqrt{p} \in \mathcal{H}^{1/2}$, we consider smooth maps $\alpha :]-\nu, \nu[\to \mathcal{H}^{1/2}$ $(\nu > 0)$ such that $\alpha(0) = \sqrt{p}$. The Fréchet derivative of α in zero $D\alpha(0)$, defined by

$$\lim_{t \to 0} \frac{\|\alpha(t) - \alpha(0) - D\alpha(0) \cdot t\|_{\mathcal{L}^2}}{t} = 0,$$

can be interpreted as a tangent vector to the curve α on the manifold $\mathcal{H}^{1/2}$. We therefore define the tangent vector space $\mathcal{T}_{\sqrt{p}}\mathcal{H}^{1/2}$ in $\sqrt{p}$ to $\mathcal{H}^{1/2}$ as the set of *all* possible Fréchet derivatives $D\alpha(0)$ for all such maps α:

$$\mathcal{T}_{\sqrt{p}}\mathcal{H}^{1/2} = \{\, D\alpha(0) : \ \alpha \text{ smooth map }]-\nu, \nu[\to \mathcal{H}^{1/2} \text{ with } \alpha(0) = \sqrt{p} \,\}.$$

This is a linear subspace of $\mathcal{L}^2$, which we may calculate more explicitly. Let $\alpha = \varphi^{-1} \circ \bar{\alpha}$, where $t \to \bar{\alpha}(t)$ is a smooth map from $]-\nu, \nu[$ to Θ with $\bar{\alpha}(0) = \theta$ for a fixed $\theta \in \Theta$. Then we may apply the chain rule to $\alpha : t \to \sqrt{p(\cdot, \bar{\alpha}(t))}$ to find

$$\begin{aligned} D\alpha(0) &= D\sqrt{p(\cdot, \bar{\alpha}(t))}\Big|_{t=0} = \sum_{k=0}^{m} \frac{\partial \sqrt{p(\cdot,\theta)}}{\partial \theta_k} \bar{\alpha}_k'(0) \\ &= \sum_{k=0}^{m} \frac{1}{2\sqrt{p(\cdot,\theta)}} \frac{\partial p(\cdot,\theta)}{\partial \theta_k} \bar{\alpha}_k'(0), \end{aligned}$$

which shows that

$$\mathcal{T}_{\sqrt{p(\cdot,\theta)}}\mathcal{H}^{1/2} = \operatorname{span} \bigcup_{k=0}^{m} \{B_k(\cdot,\theta)\}, \qquad B_k(\cdot,\theta) = \frac{1}{2\sqrt{p(\cdot,\theta)}} \frac{\partial p(\cdot,\theta)}{\partial \theta_k}. \tag{3.1}$$

The functions $B_k(\cdot,\theta)$ are linearly independent since φ was assumed to be a chart, so they form a basis for the $(m+1)$-dimensional tangent space in the point $\sqrt{p(\cdot,\theta)}$ on the manifold. The inner products of the basis elements in $\mathcal{L}^2$ generate a matrix function $H(\theta)$:

$$\langle B_i(\cdot,\theta), B_j(\cdot,\theta) \rangle_{\mathcal{L}^2} = \int_{\mathbb{R}} \frac{1}{4p(x,\theta)} \frac{\partial p(x,\theta)}{\partial \theta_i} \frac{\partial p(x,\theta)}{\partial \theta_j} \, dx \ \stackrel{\text{def}}{=} \ \tfrac{1}{4} H_{ij}(\theta).$$

In all points of the manifold $\sqrt{p(\cdot,\theta)} \in \mathcal{H}^{1/2}$ where this matrix is invertible, we can define an orthogonal projection operator Π_θ which maps linear subspaces of $\mathcal{L}^2$

containing the finite-dimensional tangent vector space (3.1) onto this tangent vector space, using the formula

$$v \overset{\Pi_\theta}{\longmapsto} \sum_{i=0}^{m} \left[\sum_{j=0}^{m} 4\,[H(\theta)]^{-1}_{ij} \langle v,\, B_j(\cdot,\theta)\rangle_{\mathcal{L}^2} \right] B_i(\cdot,\theta). \tag{3.2}$$

In this paper we will use a special class of parametrized families of densities, the finite-dimensional *unnormalized exponential families.* An unnormalized exponential family is given by

$$\mathcal{H}^{1/2} = \{\sqrt{p(\cdot,\theta)},\ \theta \in \Theta\}, \qquad p(x,\theta) = f(x) \exp\left[\sum_{k=0}^{m} \theta_k c_k(x) \right], \tag{3.3}$$

where m is a strictly positive integer, $\{c_0, \ldots, c_m\}$ is a set of linearly independent scalar functions on $\mathbb{R}$, and f is the probability density of the jump size X, as introduced in the previous section. The parameter vector $\theta = (\theta_0, \theta_1, \ldots, \theta_m)$ is restricted to lie in the parameter set Θ, which is an open nonempty convex subset of $\mathbb{R}^{m+1}$ satisfying

$$\Theta \subseteq \Theta_0 \overset{\text{def}}{=} \left\{ \theta \in \mathbb{R}^{m+1} : \int_{\mathbb{R}} f(x) \exp\left[\sum_{k=0}^{m} \theta_k c_k(x) \right] dx < \infty \right\}.$$

Throughout this paper we will use the manifold generated by $c_k(x) = x^k$ for $k = 0, 1, \ldots, m$, with m an even strictly positive integer, and $\Theta = \{\theta \in \mathbb{R}^{m+1},\ \theta_m < 0\}$. On such manifolds, the differential-geometric structure turns out to be a particularly transparent one. The basis vectors of the tangent space in $\sqrt{p(\cdot,\theta)}$ are given by

$$B_k(x,\theta) = \frac{1}{2\sqrt{p(x,\theta)}} \frac{\partial p(x,\theta)}{\partial \theta_k} = \frac{1}{2} x^k \sqrt{p(x,\theta)} \tag{3.4}$$

for $k = 0, 1, \ldots, m$, and if we define

$$\eta^k(\theta) \overset{\text{def}}{=} \int_{\mathbb{R}} c_k(x) p(x,\theta)\, dx = \int_{\mathbb{R}} x^k p(x,\theta)\, dx,$$

we find that the earlier defined inner product matrix $H(\theta)$ for the basis elements of the tangent space in a point $\sqrt{p(\cdot,\theta)}$ on the manifold is equal to

$$H_{ij}(\theta) = 4\ \langle B_i(\cdot,\theta), B_j(\cdot,\theta)\rangle_{\mathcal{L}^2} = \eta^{i+j}(\theta). \tag{3.5}$$

The matrix $H(\theta)$ will be differentiable with respect to θ for all $\theta \in \Theta$ if all finite order moments of the jump size X exist, since

$$\frac{\partial \eta^i}{\partial \theta_j} = \int_{\mathbb{R}} x^i \frac{\partial p(x,\theta)}{\partial \theta_j}\, dx = \int_{\mathbb{R}} x^{i+j}\, p(x,\theta)\, dx = \eta^{i+j}(\theta) = H_{ij}(\theta). \tag{3.6}$$

For $\theta \in \Theta$ the matrix $H(\theta)$ will also be invertible, because if $H(\theta)y = 0$ for some vector $y \in \mathbb{R}^{m+1}$, then

$$\begin{aligned} 0 &= \sum_{i=0}^{m} \sum_{j=0}^{m} y_i H_{ij}(\theta) y_j = \sum_{i=0}^{m} \sum_{j=0}^{m} \int_{\mathbb{R}} y_i x^{i+j} y_j p(x,\theta)\, dx \\ &= \int_{\mathbb{R}} \left(\sum_{i=0}^{m} y_i\, x^i \right)^2 p(x,\theta)\, dx, \end{aligned}$$

which implies that y is the zero vector in $\mathbb{R}^{m+1}$ since p is strictly positive Lebesgue almost everywhere. We remark that the matrix $H(\theta)$ coincides with the Fisher information matrix for our class of problems, since we can write it as

$$H_{ij}(\theta) = \int_{\mathbb{R}} \frac{\partial \ln p(x,\theta)}{\partial \theta_i} \frac{\partial \ln p(x,\theta)}{\partial \theta_j} p(x,\theta)\, dx.$$

The most important structural property is (3.6). It is exploited repeatedly in [4, 5], and it will play a central role in our analysis as well. A density from the exponential family may be characterized in terms of the θ-coordinate system, or the η-coordinate system, and on Θ the two are related by a diffeomorphism $\eta = \eta(\theta)$, which has the Fisher information matrix H as its Jacobian. In terms of Amari [1], the pair (θ, η) forms a *dual coordinate system.* However, our particular choice for this exponential family is not just motivated by this important property but also by other information-theoretic considerations, since it may be shown that it is in fact the class of densities which maximize the *entropy* of a density with respect to Lebesgue measure once its $m+1$ moments $\{\eta_0, \ldots, \eta_m\}$ have been specified.

The difference between our problem and the nonlinear filtering problem for diffusions treated in [5] lies mainly in the fact that our state equation does not evolve smoothly (in fact, not even continuously) and that its evolution depends on two stochastic variables (the jump size X and the jump time τ). We have seen in the previous section that the conditional distribution of the signal $\{S_t,\, t \geq 0\}$ consists of a Dirac measure in the origin and a smooth density outside the origin, and for reasons which will become clear later, we do not want to project that part of the conditional distribution which is represented by the Dirac measure. It will therefore be more convenient to apply the projection method to the Duncan–Mortensen–Zakai equation (2.6) for the absolutely continuous part of the density $q_t^\epsilon(x)$ which we defined in Theorem 2.1:

$$dq_t(x) = f(x)g(t)\, dt + \frac{x}{\epsilon^2} q_t(x)\, dY_t^\epsilon, \tag{3.7}$$

with initial condition $q_0(x) = 0$ for all $x \in \mathbb{R}$. Note that we will suppress the ϵ-dependency of this conditional density in our notation from now on.

Our definition of the exponential family also differs from the manifolds used for diffusion processes in the sense that the densities in our manifold are *not normalized.* In fact the differential-geometric structure takes the form of a cone: all scalar multiples of a certain density on the manifold also lie on the manifold because of the introduction of the extra parameter θ_0. This is important, since the Duncan–Mortensen–Zakai equation (3.7) provides an unnormalized version of the conditional density outside the origin, and the normalization constant turns out to have a particular significance in our case. Indeed,

$$\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon) = \mathbb{P}(S_t \neq 0 \mid \mathcal{Y}_t^\epsilon) = \frac{\int_{\mathbb{R}} q_t(x)\, dx}{\int_{\mathbb{R}} q_t(x)\, dx + 1 - G(t)}, \tag{3.8}$$

so the normalization constant is linked to the probability that a jump has occurred, and estimation of its value using the parameter θ_0 will thus be essential.

Note that alternatively we could have directly defined a projection filter without these modifications, when using projections on measures consisting of convex combinations of the Dirac delta measure and members of the family of exponential

distributions

$$\tilde{p}(dx,\theta) = \gamma\,\delta_0(dx) + (1-\gamma)\frac{f(x)e^{\theta_1 c_1(x)+\cdots+\theta_k c_k(x)}}{\int_{\mathbb{R}} f(u)e^{\theta_1 c_1(u)+\cdots+\theta_k c_k(u)}du}\,dx,$$

where the parameter γ replaces the old parameter θ_0:

$$\gamma = \frac{1-G(t)}{1-G(t)+e^{\theta_0}\int_{\mathbb{R}} f(u)e^{\theta_1 c_1(u)+\cdots+\theta_k c_k(u)}du} \in [0,1].$$

However, our present formulation involves only measures which are absolutely continuous with respect to Lebesgue measures, and since this allows us to work directly with density functions, it is slightly more convenient.

To simplify the calculations on our statistical manifold we will work with the Stratonovich form of the Duncan–Mortensen–Zakai equation:

$$dq_t(x) = \left[f(x)g(t) - \frac{x^2}{2\epsilon^2}q_t(x)\right]dt + \frac{x}{\epsilon^2}q_t(x)\circ dY_t^\epsilon,$$

and the differential equation for $\sqrt{q_t} \in \mathcal{L}^2$ thus becomes

$$d\sqrt{q_t(x)} = \left[\frac{f(x)g(t)}{2\sqrt{q_t(x)}} - \frac{x^2}{4\epsilon^2}\sqrt{q_t(x)}\right]dt + \frac{x}{2\epsilon^2}\sqrt{q_t(x)}\circ dY_t^\epsilon.$$

To simplify notation we rewrite this as

$$d\sqrt{q_t} = \mathcal{P}_1(\sqrt{q_t})\,dt + \mathcal{P}_2(\sqrt{q_t})\circ dY_t^\epsilon,$$

with the nonlinear operators $\mathcal{P}_i$ $(i=1,2)$ on $\mathcal{L}^2$ defined in an obvious way. To make sure that these operators do indeed map back into $\mathcal{L}^2$ when we apply them to our approximate densities $p(\cdot,\theta)$, we need the following condition:

For all $\theta \in \Theta$ we have that

$$\text{(A)} \qquad \int_{\mathbb{R}} x^4\,p(x,\theta)\,dx < \infty \qquad \text{and} \qquad \int_{\mathbb{R}} \frac{f(x)^2}{p(x,\theta)}\,dx < \infty.$$

The first part of this condition is rather mild, and the second part will be satisfied if the tails of the density f vanish rapidly enough. We will see that both parts of condition (A) are *not necessary* to formulate our approximate filter, but they are needed if one wants to interpret the filter as the result of a projection in $\mathcal{L}^2$.

The operators $\Pi_\theta \circ \mathcal{P}_i$, with Π_θ as defined in (3.2), now generate a stochastic vector field on the manifold $\mathcal{H}^{1/2}$:

$$(3.9) \qquad d\sqrt{p(\cdot,\theta_t)} = \left[\Pi_{\theta_t}\circ\mathcal{P}_1(\sqrt{p(\cdot,\theta_t)})\right]dt + \left[\Pi_{\theta_t}\circ\mathcal{P}_2(\sqrt{p(\cdot,\theta_t)})\right]\circ dY_t^\epsilon.$$

Note that we will always use the notation q for the real unnormalized conditional density outside the origin, and p for its projection.

Our aim is now to describe the evolution of the density in terms of our parameter vector θ_t; i.e., we want to find a stochastic differential equation for the result of the inverse mapping from the trajectory of projected densities on the manifold $\mathcal{H}^{1/2}$ into our parameter set $\Theta \subseteq \mathbb{R}^{m+1}$. It turns out that we can easily extend the analysis that was carried out in [4] for diffusion processes.

THEOREM 3.1. *Let the conditions of the previous section and condition* (A) *be satisfied. Then the parameter vector* θ_t *describing the filter* (3.9) *on the manifold generated by the exponential family* (3.3) *with* $c_k(x) = x^k$ *satisfies the Stratonovich stochastic differential equation*

(3.10)
$$d\theta_t = g(t)\ [H(\theta_t)]^{-1} \begin{pmatrix} 1 \\ \mathbb{E}X \\ \mathbb{E}X^2 \\ \vdots \\ \mathbb{E}X^m \end{pmatrix} dt - \frac{1}{2\epsilon^2} \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} dt + \frac{1}{\epsilon^2} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \circ dY_t^\epsilon,$$

with the matrix function H *defined as before by*

$$H_{ij}(\theta_t) = \int_{\mathbb{R}} x^{i+j} p(x, \theta_t)\, dx = \theta_{i+j}.$$

This stochastic differential equation has a unique solution up to the (possibly infinite) almost surely strictly positive exit time $\inf\{t \geq 0 : \theta_t \notin \Theta\}$.

Proof of Theorem 3.1. We deal with the two terms in (3.9) separately. For the first one we find, using (3.2),

$$\begin{aligned}
&\Pi_{\theta_t} \circ \mathcal{P}_1(\sqrt{p(\cdot, \theta_t)}) \\
&= \sum_{i=0}^m \sum_{j=0}^m 4\,[\,H(\theta_t)\,]^{-1}_{ij} \left[\int_{\mathbb{R}} \mathcal{P}_1(\sqrt{p})(x)\, B_j(x, \theta_t)\, dx \right] B_i(\cdot, \theta_t) \\
&= \sum_{i=0}^m \sum_{j=0}^m 4\,[H(\theta_t)]^{-1}_{ij} \left[\int_{\mathbb{R}} \left(\frac{f(x)g(t)}{2\sqrt{p(x,\theta_t)}} - \frac{x^2}{4\epsilon^2} \sqrt{p(x,\theta_t)} \right) \frac{1}{2} x^j \sqrt{p(x,\theta_t)}\, dx \right] B_i(\cdot, \theta_t) \\
&= \sum_{i=0}^m \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \left[g(t)\mathbb{E}X^j - \frac{\eta_t^{j+2}}{2\epsilon^2} \right] B_i(\cdot, \theta_t).
\end{aligned}$$

Analogously, the second vector field can be shown to satisfy

$$\Pi_{\theta_t} \circ \mathcal{P}_2(\sqrt{p(\cdot, \theta_t)}) = \sum_{i=0}^m \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \left[\frac{\eta_t^{j+1}}{\epsilon^2} \right] B_i(\cdot, \theta_t).$$

But since

$$d\sqrt{p(\cdot, \theta_t)} = \sum_{i=0}^m \left[\frac{1}{2\sqrt{p(\cdot,\theta)}} \frac{\partial p(\cdot, \theta)}{\partial \theta_i}\bigg|_{\theta=\theta_t} \circ d(\theta_t)_i \right] = \sum_{i=0}^m B_i(\cdot, \theta_t) \circ d(\theta_t)_i,$$

equating the coefficients in front of the basis vectors $B_i(\cdot, \theta_t)$ of the tangent space in $\sqrt{p(\cdot, \theta_t)}$ then gives that

$$\begin{aligned}
(d\theta_t)_i &= g(t) \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \mathbb{E}X^j\, dt - \frac{1}{2\epsilon^2} \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \eta_t^{j+2}\, dt \\
&\quad + \frac{1}{\epsilon^2} \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \eta_t^{j+1} \circ dY_t^\epsilon \\
&= g(t) \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \mathbb{E}X^j\, dt - \frac{1}{2\epsilon^2} \mathbf{1}_{\{i=2\}}\, dt + \frac{1}{\epsilon^2} \mathbf{1}_{\{i=1\}} \circ dY_t^\epsilon,
\end{aligned}$$

because of (3.5). Existence and uniqueness of a solution of this equation up to the almost surely positive exit time $\inf\{t > 0 : \theta_t \notin \Theta\}$ is guaranteed since we showed before that $[H(\theta)]^{-1}$ exists for all $\theta \in \Theta$, and since $H(\theta)$ is infinitely many times differentiable with respect to θ on this set, its inverse certainly satisfies a local Lipschitz condition. □

Some care must be taken when defining the initial conditions for the stochastic differential equation for θ_t. At time $t = 0$ the density outside the origin is equal to $q_0(x) = 0$ for all x, which would mean that $\theta_0 = -\infty$ and that the other values in the θ-vector can be chosen arbitrarily. We can overcome this problem by looking at the moments vector η instead of θ. We have remarked before that on the domain Θ the θ-vectors and η-vectors are related by a diffeomorphism. If we look at a small time $\delta > 0$, we see from the Duncan–Mortensen–Zakai equation (3.7) that $q_\delta(x)$ approximately equals $f(x)g(0)\delta$. By (3.6), we have $H(\theta_t) \circ d\theta_t = d\eta_t$, so rewriting (3.10) in terms of moments gives

$$d\eta_t = g(t) \begin{pmatrix} 1 \\ \mathbb{E}X \\ \mathbb{E}X^2 \\ \vdots \\ \mathbb{E}X^m \end{pmatrix} dt - \frac{1}{2\epsilon^2} [H(\theta_t)] \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} dt + \frac{1}{\epsilon^2} [H(\theta_t)] \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \circ dY_t^\epsilon,$$

so the moments η_δ at time δ are approximately equal to $g(0)\delta$ times the moments of X, as the expression for q_δ confirms. These moments will then uniquely determine the value of the parameter vector θ_δ, which may then be used as the initial condition for the stochastic differential equation for θ_t. Note that this is the only place where our assumption that $g(0)$ be strictly positive is explicitly needed, and if one is prepared to formulate alternative initial conditions for the approximate filter, this assumption can be weakened.

Equation (3.10) for the evolution of θ_t has a remarkably simple structure. In particular, since it has a constant diffusion coefficient, the Stratonovich and Îto forms of the stochastic differential equation coincide, and every Euler scheme to find numerical approximations to its solution will coincide with a Milstein scheme, guaranteeing strong convergence of order 1 [16]. Moreover, it is quite easy to give a clear interpretation of the stochastic differential equation. Since q_t approximates the conditional density of S_t outside the origin, i.e., the conditional density of X, we can interpret the stochastic differential equation for θ_t as the sum of two separate vector fields. The first one keeps the conditional density of X close to the prior density of X: since $d\eta_t = H(\theta_t) \circ d\theta_t$, the solution of $d\theta_t = g(t)\,[H(\theta_t)]^{-1}\,[1 \;\; \mathbb{E}X \;\; \ldots \;\; \mathbb{E}X^m]^T\,dt$ would simply be $G(t) = \mathbb{P}(t < \tau)$ times that density on the manifold which has the same first m moments as X. The second vector field $d\theta_t = -\frac{1}{2\epsilon^2}[0\ 0\ 1\ \ldots\ 0]^T\,dt + \frac{1}{\epsilon^2}[0\ 1\ 0\ \ldots\ 0]^T \circ dY_t^\epsilon$ describes the evolution of the Kalman filter for a Gaussian distributed random variable X observed in white noise of intensity ϵ^2. Before the jump, $Y_t^\epsilon = \epsilon W_t$, and the influence of the stochastic increment dY_t^ϵ will be small, while after the jump it will become significant due to the nonzero drift in Y_t^ϵ.

The fact that the diffusion coefficient vector in the stochastic differential equation for θ_t is a constant vector is a consequence of our choice of the basis functions $\{c_0(x), \ldots, c_m(x)\}$ which generate the exponential family. The diffusion coefficient vector will always be constant if the function j in the observation equation $dY_t^\epsilon = j(S_t)\,dt + \epsilon\,dW_t$ (in our case simply $j(x) = x$) and its square (in our case $j(x)^2 = x^2$) are both in the linear space spanned by the functions $\{c_0(x), \ldots, c_m(x)\}$.

A proof is given in [4] for nonlinear filtering problems where the signal S_t is a diffusion instead of a jump process, and this result carries over directly to our case.

4. Statistical Approximations. In the previous section, the conditional probability distribution of our original signal process $\{S_t,\, t \geq 0\}$ was approximated by a member of a finite-dimensional family of distributions. Another possible approximation to the optimal filter can be found by applying the Kushner–Stratonovich equation of nonlinear filtering. This equation describes the evolution of conditional statistics in time by means of a stochastic differential equation driven by the observation process $\{Y_t^\epsilon, t \geq 0\}$. We will use it to find such stochastic differential equations for the evolution of the moments of our conditional density and then use these equations to define another approximate filter. To do so, we first state the Kushner–Stratonovich equation (for the special case where the state noise and observation noise are independent) and then derive a stochastic differential equation for the process $\{S_t,\, t \geq 0\}$ which makes it possible to apply it to our particular filtering problem.

Let $\{V_t,\, t \in [0,T]\}$ be a scalar stochastic process such that V_0 is $\mathcal{S}_0$-measurable, with $\mathbb{E}|V_0| < \infty$ and

$$dV_t = D_t\, dt + dM_t, \tag{4.1}$$

$$dY_t^\epsilon = S_t\, dt + \epsilon\, dW_t. \tag{4.2}$$

We assume the following (see section 2 for the definition of the state filtration $\mathcal{S}_t$):

- $\{M_t,\, t \geq 0\}$ is a right-continuous square integrable $\mathcal{S}_t$-martingale with left-hand limits, which is independent of the Wiener process $\{W_t,\, t \geq 0\}$;
- $\{D_t,\, t \geq 0\}$ is an $\mathcal{S}_t$-adapted process with $\mathbb{E}\int_0^T D_u^2\, du < \infty$; and
- $\{V_t,\, t \geq 0\}$ is such that $\mathbb{E}\int_0^T (S_u V_u)^2\, du < \infty$.

We will use the notation $\widehat{\alpha}_t = \mathbb{E}[\alpha_t \mid \mathcal{Y}_t^\epsilon]$ for the conditional expectation of stochastic processes $\{\alpha_t,\, t \geq 0\}$ with respect to the observations σ-algebra $\mathcal{Y}_t^\epsilon$. The Kushner–Stratonovich equation then states that for $t \in [0,T]$ we have [14, 27]

$$d\widehat{V}_t = \widehat{D}_t\, dt + \frac{1}{\epsilon^2}\left(\widehat{S_t V_t} - \widehat{S}_t\, \widehat{V}_t\right) d\nu_t^\epsilon, \tag{4.3}$$

with initial condition $\widehat{V}_0 = \mathbb{E}V_0$. The process

$$\nu_t^\epsilon = Y_t^\epsilon - \int_0^t \widehat{S}_u\, du \tag{4.4}$$

is called the *innovation process*, and under the conditions stated, it is a Brownian motion with respect to the observations filtration $\{\mathcal{Y}_t^\epsilon,\, t \geq 0\}$.

In order to be able to apply the Kushner–Stratonovich equation to our problem, we will now derive a description for the signal $\{S_t,\, t \geq 0\}$ of the form (4.1). Let $E_t = \mathbf{1}_{\{t \geq \tau\}}$ denote, as before, the right-continuous $\mathcal{S}_t$-measurable process which jumps from zero to one at time τ. The probability that the jump occurs in the time interval $[t, t+dt]$ given that it has not occurred before time t equals $\lambda(t)dt + o(dt)$, where $\lambda(t)$ is the *hazard rate* at time t, defined by

$$\lambda(t) = \frac{g(t)}{1 - G(t)}.$$

Define the process M_t as E_t minus the integral of this hazard rate up to time $t \wedge \tau$ (where we introduce the usual notation $a \wedge b$ for the minimum of a and b):

$$M_t = E_t - K_t, \qquad K_t = \int_0^{t\wedge\tau} \lambda(s)\, ds = -\ln(1 - G(t \wedge \tau)).$$

Tedious but straightforward calculations show that M_t is an $\mathcal{S}_t$-martingale (for details, see, e.g., [9]). But since

$$t \wedge \tau = \int_0^t (1 - E_u)\, du,$$

we have that

$$M_t = E_t + \ln\left[1 - G\left(\int_0^t (1 - E_u) du\right)\right],$$

and we thus find the following representation for E_t:

$$\begin{aligned} dE_t &= \lambda(t \wedge \tau)\,(1 - E_t)\, dt + dM_t \\ &= \lambda(t)\,(1 - E_t)\, dt + dM_t, \end{aligned} \tag{4.5}$$

where we have used the fact that $\lambda(t \wedge \tau)(1 - E_t) = \lambda(t)(1 - E_t)$, since if $t \wedge \tau = \tau$, then $1 - E_t = 0$. Our original process may now be represented as $S_t = XE_t$, so it satisfies

$$dS_t = \lambda(t)\,(X - S_t)\, dt + X\, dM_t, \tag{4.6}$$

and in fact for arbitrary $k \in \mathbb{N} \setminus \{0\}$

$$d(S_t)^k = \lambda(t)\,(X^k - (S_t)^k)\, dt + X^k\, dM_t. \tag{4.7}$$

We can now apply the Kushner–Stratonovich equation to this representation of our signal process, but we first prove a lemma that will be used to simplify the equations which it generates.

LEMMA 4.1. *For all $t \in [0,T]$ and $k \in \mathbb{N} \setminus \{0\}$, we have that, almost surely,*

$$\mathbb{E}\,[X^k - (S_t)^k \mid \mathcal{Y}_t^\epsilon] = (1 - \widehat{E}_t)\,\mathbb{E}X^k. \tag{4.8}$$

Proof of Lemma 4.1. Let B be any set in $\mathcal{Y}_t^\epsilon$. Then by definition,

$$\begin{aligned} \int_B \mathbb{E}\,[X^k - (S_t)^k \mid \mathcal{Y}_t^\epsilon](\omega)\, dP(\omega) &= \int_B (\, X^k(\omega) - (S_t)^k(\omega)\,)\, dP(\omega) \\ &= \int_B X^k(\omega)\, \mathbf{1}_{\{t<\tau(\omega)\}}\, dP(\omega) \\ &= \int_{B \cap \{\omega : t < \tau(\omega)\}} X^k(\omega)\, dP(\omega). \end{aligned}$$

But we have that the σ-algebra generated by sets of the form $B \cap \{\omega : t < \tau(\omega)\}$ (with $B \in \mathcal{Y}_t^\epsilon$) is independent of sets in the σ-algebra generated by X^k, since $Y_t^\epsilon = \epsilon W_t$ on $\{\omega : t < \tau(\omega)\}$ and the process $\{W_t,\, t \geq 0\}$ is independent of X, so

$$\begin{aligned} \int_{B \cap \{\omega : t < \tau(\omega)\}} X^k(\omega)\, dP(\omega) &= \int_{B \cap \{\omega : t < \tau(\omega)\}} \mathbb{E}X^k\, dP(\omega) \\ &= \left(\mathbb{E}X^k\right) \int_B \mathbf{1}_{\{t<\tau(\omega)\}}\, dP(\omega) \\ &= \left(\mathbb{E}X^k\right) \int_B \mathbb{E}\,[1 - \mathbf{1}_{\{t \geq \tau\}} \mid \mathcal{Y}_t^\epsilon](\omega)\, dP(\omega) \\ &= \left(\mathbb{E}X^k\right) \int_B (1 - \widehat{E}_t(\omega))\, dP(\omega), \end{aligned}$$

and we may now conclude that (4.8) holds by the almost sure uniqueness property of conditional expectations. □

THEOREM 4.2. *Let the random variables X and τ and the stochastic processes $\{S_t,\, t \geq 0\}$ and $\{Y_t^\epsilon,\, t \geq 0\}$ be defined as in section 2, and let X and τ satisfy all conditions mentioned in that section. Then the optimal filter estimate $\widehat{S}_t = \mathbb{E}\,[S_t \mid \mathcal{Y}_t^\epsilon]$ and higher order moments for $t \in [0,T]$ are generated by the following Îto stochastic differential equations ($k \in \mathbb{N} \setminus \{0\}$):*

$$d\widehat{E}_t = \lambda(t)(1-\widehat{E}_t)\,dt + \frac{1}{\epsilon^2}\widehat{S}_t(1-\widehat{E}_t)\,d\nu_t^\epsilon, \tag{4.9}$$

$$\begin{aligned} d\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon] &= \lambda(t)\,\mathbb{E}X^k(1-\widehat{E}_t)\,dt \\ &\quad + \frac{1}{\epsilon^2}\Big(\,\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon] - \widehat{S}_t\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]\Big)\,d\nu_t^\epsilon, \end{aligned} \tag{4.10}$$

with initial conditions $\widehat{E}_0 = \mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]_{t=0} = 0$ for all $k \in \mathbb{N} \setminus \{0\}$, and where the innovation process $\{\nu_t^\epsilon,\, t \in [0,T]\}$ is defined by (4.4).

Proof of Theorem 4.2. The conditions for application of the Kushner–Stratonovich equation are obviously satisfied for the process E_t since

$$\mathbb{E}\int_0^T [\lambda(u)(1-E_u)]^2 du \;\leq\; \int_0^T \lambda(u)^2 du < \infty$$

(note that $\lambda(t)$ is finite for all $t \geq 0$ and continuous since we assumed that $g(t)$ is continuous and $G(t) < 1$ for all $t \geq 0$) and $\mathbb{E}\int_0^T |E_u S_u| du < T \cdot \mathbb{E}|X| < \infty$. Here and in the rest of the proof we use the fact that all finite order moments of X exist because of condition (2.3), which implies that $\mathbb{E}|X|^k < \infty$ for all $k \geq 0$.

Since $\{W_t,\, t \geq 0\}$ was assumed to be independent of X and τ, it is independent of $\{M_t,\, t \geq 0\}$. The Kushner–Stratonovich equation applied to (4.5) thus results in

$$\begin{aligned} d\widehat{E}_t &= \lambda(t)(1-\widehat{E}_t)\,dt + \frac{1}{\epsilon^2}(\,\widehat{S_t E_t} - \widehat{S}_t\widehat{E}_t\,)\,d\nu_t^\epsilon \\ &= \lambda(t)(1-\widehat{E}_t)\,dt + \frac{1}{\epsilon^2}\widehat{S}_t(1-\widehat{E}_t)\,d\nu_t^\epsilon, \end{aligned} \tag{4.11}$$

where we have used the fact that $S_t E_t = S_t$. The initial condition is $\widehat{E}_0 = \mathbb{E}(E_0) = 0$. To find the conditional moments $\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]$ for $k \in \mathbb{N} \setminus \{0\}$, we use (4.7). Since X is $\mathcal{S}_t$-measurable and independent of τ, and $\mathbb{E}|X|^{2k} < \infty$, the process $\{X^k M_t,\, t \geq 0\}$ is again a square integrable $\mathcal{S}_t$-martingale which is independent of $\{W_t,\, t \geq 0\}$. The two other conditions for the Kushner–Stratonovich formula are satisfied as well, since

$$\mathbb{E}\int_0^T (X^k - (S_u)^k)^2 \lambda(u)^2 du \leq \mathbb{E}|X|^{2k}\int_0^T \lambda(u)^2 du < \infty,$$

$$\mathbb{E}\int_0^T |(S_u)^k S_u| du \leq T \cdot \mathbb{E}|X|^{k+1} < \infty.$$

We therefore have that for $k \in \mathbb{N} \setminus \{0\}$

$$\begin{aligned} d\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon] &= \lambda(t)\,\mathbb{E}\,[X^k - (S_t)^k \mid \mathcal{Y}_t^\epsilon]\,dt \\ &\quad + \frac{1}{\epsilon^2}\Big(\,\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon] - \widehat{S}_t\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]\Big)\,d\nu_t^\epsilon, \end{aligned} \tag{4.12}$$

with initial condition $\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]_{t=0} = \mathbb{E}[(S_t)^k]_{t=0} = 0$.

Using the result of Lemma 4.1, we see that (4.12) can be simplified to

$$d\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon] = \lambda(t)\,\mathbb{E}X^k(1-\widehat{E}_t)\,dt + \frac{1}{\epsilon^2}\left(\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon] - \widehat{S}_t\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]\right) d\nu_t^\epsilon. \tag{4.13}$$

This proves Theorem 4.2. □

Note that $\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]$, the conditional moment of order k, depends on the conditional moment of order $k+1$, $\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon]$, so (4.9)–(4.10) do not form a closed set of equations. If we want to use these equations to define a finite-dimensional approximation to the optimal filter, we need to use an appropriate closure formula to approximate higher order moments in terms of lower order moments. One possible closure formula, which was proposed in [11], assumes the third order central moment to be zero at all times, i.e., $\mathbb{E}\,[(S_t - \widehat{S}_t)^3 \mid \mathcal{Y}_t^\epsilon] = 0$ for all $t \in [0,T]$.

This closing of the infinite set of moment equations that has now been generated, by expressing higher order moments in terms of lower order moments, means that we restrict our densities to belong to a specific family of distributions. As was pointed out earlier in [4], the *a priori assumption* that the conditional density will belong to this family at every time instant is often incorrect. But it was shown in the same paper that a sound mathematical basis can be given for this so-called *assumed density principle* in some cases which involve the filtering of nonlinear diffusions, by showing that the resulting filter is equivalent to a *projection* in probability density space, like the one we described in the preceding section. In the next section we will show that this idea can be applied to our change detection problem as well, and that we can gain considerable insight into the nature of such problems in doing so.

5. The Assumed Density Principle. To formulate our differential-geometric approximate filter of section 3 in terms of the conditional moments it generates, we define, bearing in mind the interpretation of the normalization constant given in (3.8), the following statistics (where $\approx$ means approximates):

$$\begin{aligned}
\check{E}_t &= \frac{\int_{\mathbb{R}} p(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx + 1 - G(t)} \approx \mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon), \\
\check{X}_t &= \frac{\int_{\mathbb{R}} xp(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx} \approx \mathbb{E}[X \mid \mathcal{Y}_t^\epsilon], \\
\check{S}_t^n &= \frac{\int_{\mathbb{R}} x^n\, p(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx + 1 - G(t)} \approx \mathbb{E}[(S_t)^n \mid \mathcal{Y}_t^\epsilon],
\end{aligned} \tag{5.1}$$

with $n \in \mathbb{N}$, so $\check{S}_t^0 = E_t$ and $\check{S}_t^1 = \check{S}_t$.

We remark that this implies that $\check{S}_t = \check{X}_t\check{E}_t$, i.e., that the conditional estimate of the signal $\check{S}_t$ naturally splits into two statistics $\check{E}_t$ and $\check{X}_t$, which approximate the conditional probability of a jump having occurred and the best estimate of the jump size, respectively. We have shown in (4.8) that the optimal filter estimates satisfy, for $\widehat{X}_t = \mathbb{E}\,[X \mid \mathcal{Y}_t^\epsilon] \neq 0$,

$$\widehat{S}_t = \widehat{X}_t\left(1 - (1-\widehat{E}_t)\frac{\mathbb{E}X}{\widehat{X}_t}\right),$$

so the *optimal* filter estimates will in general *not* satisfy the equation $\widehat{S}_t = \widehat{X}_t\widehat{E}_t$.

However, this does not imply that the estimate $\check{S}_t$ which is generated by the differential-geometric approximation is different from the filter estimate $\widehat{S}_t$ generated by closing the Kushner–Stratonovich equations, as we did in the previous section. We now show that they are in fact the same if we use (4.9) and (4.10) to calculate the first $m+1$ moments and then close the equations *in an appropriate way*, by choosing $\check{S}_t^{m+1}$ appropriately.

THEOREM 5.1. *Let the conditions of Theorem* 3.1 *be satisfied, and let the process* $\{\theta_t,\ t \geq 0\}$ *be defined as in* (3.10). *Define* $\check{E}_t$, $\check{S}_t$, *and* $\check{S}_t^n$ *as in* (5.1) *for* $n = 0, \dots, m+1$. *Then* $\check{E}_t = \check{S}_t^0$, $\check{S}_t = \check{S}_t^1$, *and* $\check{S}_t^n$ $(n = 2, \dots, m)$ *satisfy*

$$d\check{S}_t^n = \lambda(t)\,(1 - \check{E}_t)\,\mathbb{E}X^n\,dt + \frac{1}{\epsilon^2}\,(\,\check{S}_t^{n+1} - \check{S}_t\,\check{S}_t^n\,)\,(\,dY_t^\epsilon - \check{S}_t\,dt\,),$$

with initial conditions $\check{S}_0^n = 0$ *for all* $n = 0, \dots, m$.

Proof of Theorem 5.1. To find the stochastic differential equations for the $\check{S}_t^n$ $(n = 0, \dots, m)$ we must first find the equations for the approximated conditional moments η_t^k $(k = 0, \dots, m)$, but this is relatively simple since (3.6) implies that

$$d\eta_t^k = [\,\eta_t^k \quad \eta_t^{k+1} \quad \cdots \quad \eta_t^{k+m}\,] \circ d\theta_t,$$

and the result of Theorem 3.1 then gives

$$d\eta_t^k = g(t)\,\mathbb{E}X^k\,dt - \frac{\eta_t^{k+2}}{2\epsilon^2}\,dt + \frac{\eta_t^{k+1}}{\epsilon^2} \circ dY_t^\epsilon. \tag{5.2}$$

Using the Îto form of (5.2),

$$d\eta_t^k = g(t)\,\mathbb{E}X^k\,dt + \frac{\eta_t^{k+1}}{\epsilon^2}\,dY_t^\epsilon,$$

we find by Îto's differentiation rule that for all $k = 0, \dots, m$,

$$\begin{aligned}
d\check{S}_t^k &= \frac{d\eta_t^k}{\eta_t^0 + 1 - G(t)} - \frac{\eta_t^k\,d(\eta_t^0 + 1 - G(t))}{(\eta_t^0 + 1 - G(t))^2} - \frac{d\eta_t^k\,d(\eta_t^0 + 1 - G(t))}{(\eta_t^0 + 1 - G(t))^2} \\
&\quad + \frac{\eta_t^k\,d(\eta_t^0 + 1 - G(t))d(\eta_t^0 + 1 - G(t))}{(\eta_t^0 + 1 - G(t))^3} \\
&= \frac{\mathbb{E}X^k g(t)dt + \eta_t^{k+1}dY_t^\epsilon/\epsilon^2}{\eta_t^0 + 1 - G(t)} - \frac{\eta_t^k\,\eta_t^1 dY_t^\epsilon/\epsilon^2}{(\eta_t^0 + 1 - G(t))^2} - \frac{\eta_t^{k+1}\eta_t^1 dt/\epsilon^2}{(\eta_t^0 + 1 - G(t))^2} \\
&\quad + \frac{\eta_t^k\,(\eta_t^1)^2 dt/\epsilon^2}{(\eta_t^0 + 1 - G(t))^3} \\
&= \frac{g(t)}{1 - G(t)}\left(1 - \frac{\eta_t^0}{\eta_t^0 + 1 - G(t)}\right)\mathbb{E}X^k\,dt \\
&\quad + \frac{1}{\epsilon^2}\left(\frac{\eta_t^{k+1}}{\eta_t^0 + 1 - G(t)} - \frac{\eta_t^k\eta_t^1}{(\eta_t^0 + 1 - G(t))^2}\right)\left(dY_t^\epsilon - \frac{\eta_t^1}{\eta_t^0 + 1 - G(t)}\,dt\right) \\
&= \lambda(t)(1 - \check{E}_t)\,\mathbb{E}X^k\,dt + \frac{1}{\epsilon^2}\,(\,\check{S}_t^{k+1} - \check{S}_t\,\check{S}_t^k\,)\,(\,dY_t^\epsilon - \check{S}_t\,dt\,),
\end{aligned}$$

which proves the theorem. □

These equations are precisely the same as the ones we derived for the filter of the previous section, (4.9) and (4.10), if we replace $\mathbb{E}\,[S_t^k \mid \mathcal{Y}_t^\epsilon]$ by $\check{S}_t^k$. It thus follows

that if we close these equations by choosing $\check{S}_t^{m+1}$ appropriately, then the two filters generate the same estimates almost surely. However, some care must be taken in finding the appropriate closure formula. For example, in the Gaussian case ($m = 2$), we must *not* choose the third central moment to be equal to zero, as we proposed at the end of section 4. Since $p(x,\theta_t)/\int_{\mathbb{R}} p(x,\theta_t)dx$ is assumed to be Gaussian, and since a Gaussian variable A satisfies $\mathbb{E}A^3 = [\mathbb{E}A]\cdot[3\mathbb{E}A^2 - 2(\mathbb{E}A)^2]$, we have

$$(5.3)\qquad \frac{\eta_t^3}{\eta_t^0} = \frac{\eta_t^1}{\eta_t^0}\left(3\frac{\eta_t^2}{\eta_t^0} - 2\left(\frac{\eta_t^1}{\eta_t^0}\right)^2\right) \quad \Rightarrow \quad \frac{\check{S}_t^3}{\check{E}_t} = \frac{\check{S}_t}{\check{E}_t}\left(3\frac{\check{S}_t^2}{\check{E}_t} - 2\left(\frac{\check{S}_t}{\check{E}_t}\right)^2\right).$$

Only when this more complicated closure formula for $\mathbb{E}[(S_t)^3 \mid \mathcal{Y}_t^\epsilon]$ in terms of the lower order moments $\mathbb{E}[(S_t)^2 \mid \mathcal{Y}_t^\epsilon]$, $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$, and $\mathbb{E}[E_t \mid \mathcal{Y}_t^\epsilon]$ is used will the estimates generated by the Kushner–Stratonovich equation be the same, almost surely, as those generated by our differential-geometric approximation.

6. A Three-Dimensional Filter. Although the filter derived in section 3 using differential-geometric methods is thus equivalent to the filter derived in section 4 *when the correct closure formula is used*, there are certain advantages of the first parametrization. We already mentioned the fact that better schemes can be used to calculate numerical approximations of (3.10). Another advantage is the much more intuitive structure of the filter. If we define the a priori moments of the jump size X as $P^n = \mathbb{E}(X - \mathbb{E}X)^n$ and the approximate filter estimates

$$\check{P}_t^n = \frac{\int_{\mathbb{R}}(x - \check{X}_t)^n p(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx} \approx \mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{Y}_t^\epsilon])^n \mid \mathcal{Y}_t^\epsilon],$$

then one may show by a tedious but straightforward exercise in Stratonovich calculus [23] that for $n = 2,\dots,m$,

$$(6.1)\qquad d\check{E}_t = \lambda(t)(1-\check{E}_t)\,dt + \check{E}_t(1-\check{E}_t)\frac{\check{X}_t}{\epsilon^2}\,(dY_t^\epsilon - \check{S}_t\,dt),$$

$$(6.2)\qquad d\check{X}_t = \lambda(t)\frac{1-\check{E}_t}{\check{E}_t}\,(\mathbb{E}X - \check{X}_t)\,dt + \frac{\check{P}_t^2}{\epsilon^2}\,(dY_t^\epsilon - \check{X}_t\,dt),$$

$$\begin{aligned} d\check{P}_t^n = \lambda(t)\frac{1-\check{E}_t}{\check{E}_t}\Bigg[& P^n - \check{P}_t^n + n(P^{n-1} - \check{P}_t^{n-1})\,(\mathbb{E}X - \check{X}_t) \\ & + \sum_{k=0}^{n-2}\binom{n}{k}P^k(\mathbb{E}X - \check{X}_t)^{n-k}\Bigg]\,dt \\ & - \frac{n\check{P}_t^2}{\epsilon^2}\left[\check{P}_t^n - \tfrac{1}{2}(n-1)\,\check{P}_t^2\check{P}_t^{n-2}\right]dt \\ & + \frac{1}{\epsilon^2}\left[\check{P}_t^{n+1} - n\check{P}_t^2\check{P}_t^{n-1}\right](dY_t^\epsilon - \check{X}_t\,dt). \end{aligned}$$

These stochastic differential equations can be interpreted as the sum of vector fields which drive the conditional density to the a priori density of X (and these dominate before the jump when $\check{E}_t$ will be close to zero), vector fields which resemble those of the Kalman filter for a constant signal (which dominate after the jump when $\check{E}_t$ will be close to one), and some extra terms which make sure we do not leave the manifold that we project upon.

Note that the terms involving the innovation process in the equations for $\widehat{P}_t^n$ ($n = 2, \ldots, m$) will all be zero if and only if the central moments satisfy the equation $\widehat{P}_t^{n+1} = n\widehat{P}_t^2 \widehat{P}_t^{n-1}$ for $n = 0, \ldots, m$. Since we have that $\widehat{P}_t^0 = 1$ and $\widehat{P}_t^1 = 0$ for all $t \geq 0$, this will be the case if for all $n \in \mathbb{N}$,

$$\widehat{P}_t^{2n} = (\widehat{P}_t^2)^n \cdot 2^{-n} \frac{(2n)!}{n!}, \qquad \widehat{P}_t^{2n+1} = 0;$$

i.e., the first $m + 1$ central moments should be the same as those of a Gaussian distribution. This suggests that the equations will become even simpler if both X and our manifold are Gaussian, which is exactly the exponential family we get if we take $m = 2$ and the parameter set $\Theta = \{(\theta_0, \theta_1, \theta_2) : \theta_2 < 0\}$. In the rest of this section we will analyze the detection and estimation scheme when such a manifold of unnormalized Gaussian densities is used.

If we substitute the relation $\check{S}_t = \check{X}_t \check{E}_t$ into the stochastic differential equation for $\check{E}_t$ given by (6.1), we see the close connection with the Shiryayev–Wonham detector for *known* jump sizes. As we remarked in section 2, if we assume that the jump size X is known a priori, say, $X = a$, then the conditional probability that the jump has occurred, $\pi_t = \mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$, is finite-dimensionally computable. In fact, it follows from the Shiryayev–Wonham equation [22, 28] that

$$d\pi_t = \lambda(t)(1 - \pi_t)\, dt + \pi_t(1 - \pi_t)\frac{a}{\epsilon^2}(dY_t^\epsilon - a\pi_t\, dt), \qquad \pi_0 = 0. \tag{6.3}$$

Our estimate of the probability that a jump has occurred satisfies a modified version of this Shiryayev–Wonham equation, where a *known* jump size $X = a$ in the equation is replaced by a time-varying *estimated* jump size $\check{X}_t$. For $m = 2$ our differential-geometric approximation thus becomes a mixture of modified Kalman filter equations and this *adaptive* Shiryayev–Wonham equation:

$$\begin{aligned}
\check{S}_t &= \check{E}_t\, \check{X}_t, \\
d\check{E}_t &= \lambda(t)(1 - \check{E}_t)\, dt + \check{E}_t(1 - \check{E}_t)\frac{\check{X}_t}{\epsilon^2}\,(\, dY_t^\epsilon - \check{X}_t\check{E}_t\, dt\,), \\
d\check{X}_t &= \lambda(t)\frac{1 - \check{E}_t}{\check{E}_t}\,(\mathbb{E}X - \check{X}_t)\, dt + \frac{\check{P}_t^2}{\epsilon^2}\,(\, dY_t^\epsilon - \check{X}_t\, dt\,), \\
d\check{P}_t^2 &= \lambda(t)\frac{1 - \check{E}_t}{\check{E}_t}\,[(\mathbb{E}X - \check{X}_t)^2 + \operatorname{Var} X - \check{P}_t^2\,]\, dt - \frac{(\check{P}_t^2)^2}{\epsilon^2}\, dt,
\end{aligned}$$

with $\check{E}_0 = 0$, $\check{P}_0^2 = \operatorname{Var} X$, and $\check{X}_0 = \mathbb{E}X$.

In Figure 2, a block diagram of the filter is given that highlights the decomposition of the problem in a detection and an estimation part, which communicate through the jump size estimate $\check{X}_t$ and the conditional probability ratio $(1 - \check{E}_t)/\check{E}_t$. We remark that the original optimal detection and estimation problem as we formulated it cannot be solved recursively because we want to perform detection and estimation *simultaneously*. If the estimation problem was trivial (i.e., if we knew the jump size X immediately after the jump), the detection problem could be solved recursively, since we could then use a Shiryayev–Wonham filter tuned at $a = X$. If the detection

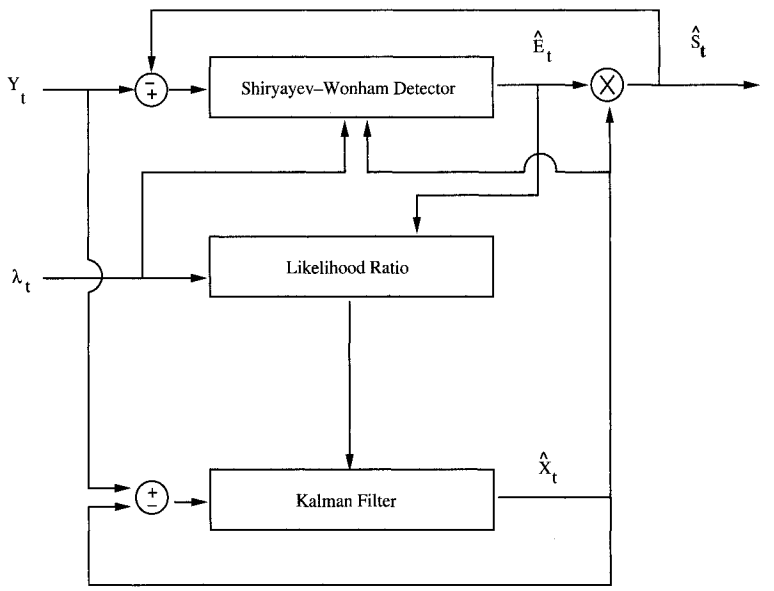


Fig. 2 *Structure of the three-dimensional approximating filter.*

problem was trivial (i.e., we knew immediately after time τ that a jump had occurred), then the estimation problem could be solved recursively, since we could simply start a Kalman filter at time τ.

In our combined problem, however, we must make sure that the Kalman filter does not start filtering too early, since it would then filter the zero signal for some time while its conditional variance $\check{P}_t^2$ would decrease, making its reaction too slow when the jump does indeed occur. The likelihood ratio term in the equation for $\check{P}_t^2$, which pulls the conditional variance $\check{P}_t^2$ back to the variance of X as long as $\check{E}_t$ is small, prevents this from happening. After the jump has occurred, a good estimate of X should quickly become available, and the stochastic differential equation for $\check{E}_t$ will then resemble the Shiryayev–Wonham equation. We therefore expect the estimate for the conditional probability that a change has occurred to converge to one quite quickly after that. We will see in the simulation studies of the next section that this will indeed be the case.

7. Simulation Results. In this section we will investigate the performance of the approximating filters that we defined in previous sections, by means of simulation studies. To do so, we first have to establish that there exists a finite, nonexploding solution to our filter equations which is unique up to equivalence and almost surely continuous. Such a proof can indeed be given using the theory of stochastic Lyapunov functions, but we will omit it here and refer the interested reader to the original paper, where a full proof is given.

In the simulation studies performed here we compare the optimal filter and our approximations. We can find the optimal filter estimates by solving the Duncan–Mortensen–Zakai equation (2.6), but this requires a lot of computational time. To do this, we used a grid which divided the interval $[-3.0, 3.0]$ for possible values of X in 1500 equidistant points. In Figure 3 we plot an example of the evolution of the conditional density over time, and also of the conditional mean, which represents

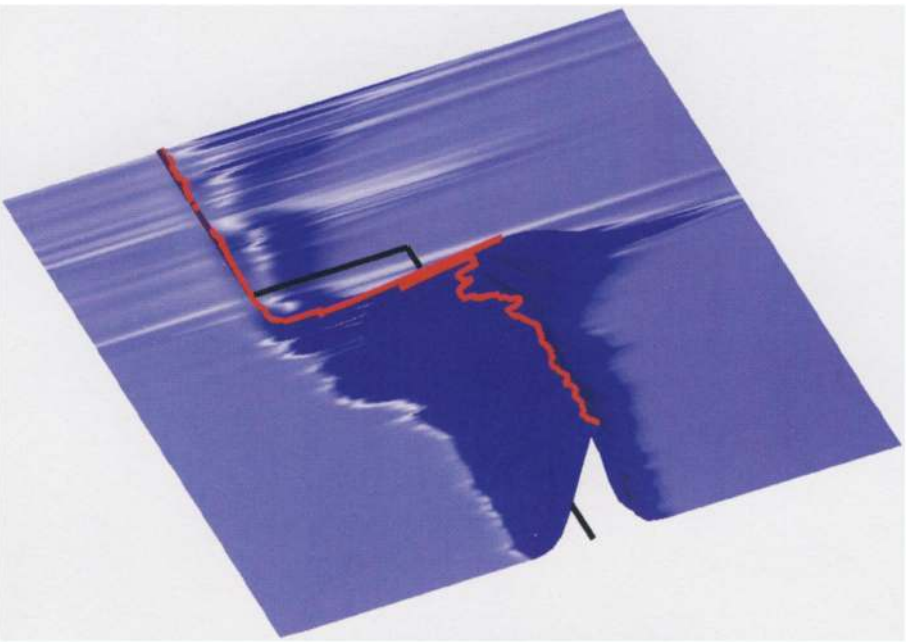


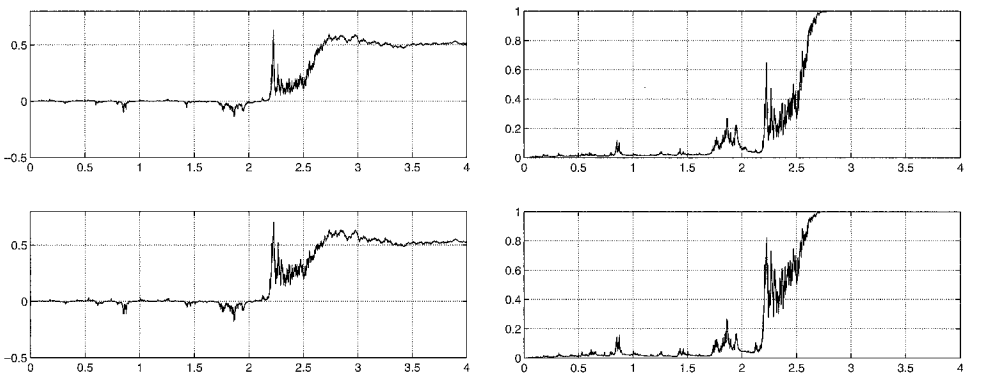
Fig. 3 *Conditional density and optimal estimate.*

the optimal estimate of the signal S at the time. The delta measure in zero (which represents most of the probability mass before the jump takes place) has been omitted from the plot. We can clearly see the detection delay in this figure and the long tails in the distribution before the jump has happened.

Alternatively, we can use our simple three-dimensional filter to get approximations for these optimal filter estimates on-line. In the figures in this section we show estimates for both the value of the signal, $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$, and for the conditional probability that a jump has occurred, $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$. The top graphs refer to the optimal estimates, and the bottom graphs to our approximations in all cases.

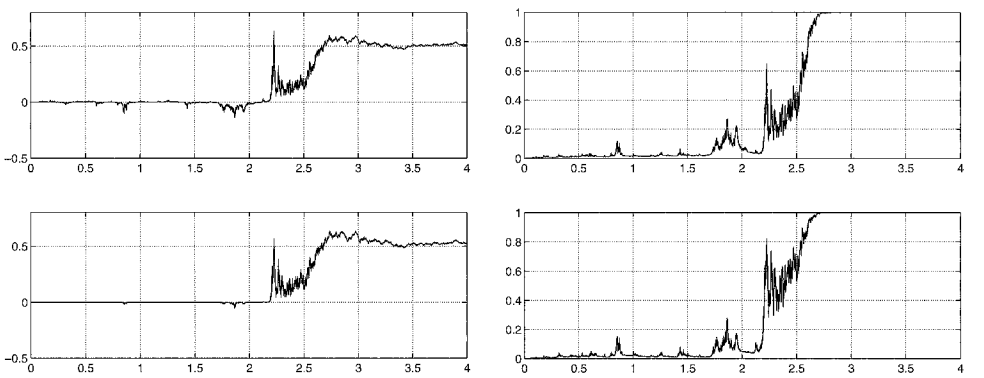
Experiment 1. For the first simulation study we took τ to be an exponentially distributed stochastic variable with mean 15.0, and the jump size X to be normally distributed with zero mean and unit variance. We let the actual jump take place at $\tau = 2.0$, and the jump size was taken to be $X = 0.5$ exactly. The noise parameter ϵ was taken as 0.10. All filters estimates were calculated on a time interval $t \in [0.0, 4.0]$, using an Euler scheme with step size $4.0 \cdot 10^{-5}$.

We first simulated the differential-geometric approximation as formulated in the previous section; i.e., we took the dimension of the filter $m + 1 = 3$, which means we project upon a manifold of Gaussian densities. In Figures 4 and 5 the results are shown for two different implementations of our filter. In Figure 4 the filter was implemented by (3.10), the stochastic differential equation for the parameter vector θ_t, while in Figure 5 the direct equations for the moments which we derived in the previous section were used. There are some small differences between the two, which should be attributed to inaccuracies in the calculation of $[H(\theta_t)]^{-1}$ in (3.10) and in the numerical method we use. However, in both cases the filter estimates show excellent behavior both before and after the change point. Both implementations slightly



(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$ (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$

Fig. 4 *Comparison between optimal and approximate filter, using* (3.10).

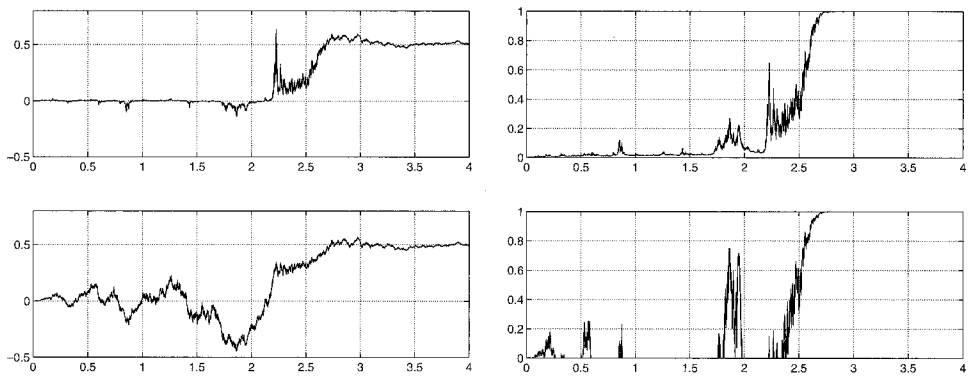


(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$ (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$

Fig. 5 *Comparison between optimal and approximate filter, using moments.*

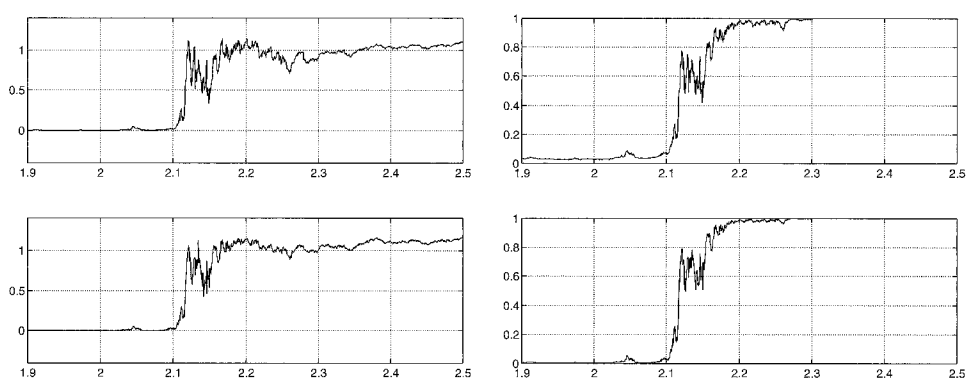
overestimate the conditional probability of a jump having occurred, but only after the jump. Around $t = 2.6$ the approximate and the exact conditional signal estimates are already indistinguishable. More importantly, the small delay in detection of the optimal filter (seen to be approximately 0.10 here) is the same for the approximate filter. For an extensive analysis of such detection delays in the optimal filter and its suboptimal approximations, the reader is referred to [11] and [23].

For comparison, Figure 6 shows a simulation of the same model setup for the approximating filter which we derived in section 4, where conditional moments are generated by using the Kushner–Stratonovich equation and the assumption that the third order central conditional moment is equal to zero. We showed in (5.3) that this filter, which was proposed in [11], is not equivalent to our filter, and its behavior is seen to be a lot worse. Although it will estimate both the signal and the conditional probability correctly in the long run, its behavior before the change is totally unacceptable. Indeed, the conditional probability is negative most of the time, and the estimates of the signal before the change are not close to the true value zero at all.



(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$. (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$.

Fig. 6 *Comparison between optimal filter and filter of section 4.*



(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$ (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$

Fig. 7 *Comparison between optimal and approximate filter.*

Experiment 2. To show that the excellent results for our differential-geometric approximation are not just a consequence of X being Gaussian, we performed a second simulation in which X was taken to be uniformly distributed on $[0, 2]$. The jump time was given the same distribution as in the first set of experiments, and the actual jump time was again taken to be $\tau = 2.0$. The jump size was taken equal to $X = 1.0$, and $\epsilon = 0.10$.

Figure 7 shows the estimates generated by our approximate filter, implemented by (3.10). The detection delay of 0.10 is almost exactly the same as for the optimal filter, and good filter estimates are produced almost directly after that. Apparently the algorithm works quite well for a jump size X with a uniform distribution, even though this distribution cannot be approximated very well on the exponential manifold that we project upon. In practice this is not important, since after the jump the behavior in the center of the state space can be shown to be asymptotically Gaussian in a large deviations sense. We again refer to [11] and [23] where an exact statement of this result is given, which helps to explain the good performance of our filter.

8. Conclusions. In this paper, we have argued that nonlinear filtering theory can be used to characterize and approximate relevant conditional statistics in those change detection problems where the size of the change is not known a priori. We have shown that a simple three-dimensional nonlinear filter can be defined which may be shown to have a global and unique solution under mild conditions and which performs well in simulation studies. Apart from an interpretation in terms of information geometry and in terms of an assumed density principle, we may view the equations for this filter as an adaptive version of the Shiryayev–Wonham equation, fed by estimates from a modified Kalman filter.

Some interesting problems are still open at the moment. These include, for example, the design of adaptive change detectors for discrete time problems, the design of detectors for more complicated signal changes such as the changing slope process [25]

$$R_t = \begin{cases} 0, & 0 \le t < \tau, \\ X(t-\tau), & t \ge \tau, \end{cases}$$

and the derivation of further theoretical properties of the filters that we defined in this paper.

REFERENCES

[1] S. AMARI, *Differential-Geometric Methods in Statistics*, Lecture Notes in Statist. 28, Springer-Verlag, Berlin, 1985.

[2] M. BASSEVILLE, *Detecting changes in signals and systems, a survey*, Automatica, 24 (1988), pp. 309–326.

[3] M. BASSEVILLE AND I. V. NIKIFOROV, *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[4] D. BRIGO, B. HANZON, AND F. LEGLAND, *A differential geometric approach to nonlinear filtering: The projection filter*, IEEE Trans. Automat. Control, 43 (1998), pp. 247–252.

[5] D. BRIGO, B. HANZON, AND F. LEGLAND, *Approximate filtering by projection on the manifold of exponential densities*, Bernoulli, 5 (1999), pp. 495–534.

[6] F. CAMPILLO, Y. KUTOYANTS, AND F. LEGLAND, *Small noise asymptotics of the GLR test for off-line change detection in misspecified diffusion processes*, Stochastics Stochastics Rep., 70 (2000), pp. 109–129.

[7] M. CHALEYAT-MAUREL AND D. MICHEL, *Des résultats de non existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.

[8] M. H. A. DAVIS, *The application of nonlinear filtering to fault detection in linear systems*, IEEE Trans. Automat. Control, 20 (1975), pp. 257–259.

[9] M. H. A. DAVIS, *Markov Models and Optimization*, Chapman and Hall, London, 1993.

[10] R. J. ELLIOTT, L. AGGOUN, AND J. B. MOORE, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1995.

[11] P. G. FOTOPOULOS, *Estimation and Detection of Jump Processes with Small Observation Noise*, Ph.D. Thesis, Imperial College, London, 1994.

[12] B. HANZON, *A differential-geometric approach to nonlinear filtering*, in Geometrization of Statistical Theory, C. T. J. Dodson, ed., ULDM Publications, University of Lancaster, Lancaster, UK, 1987, pp. 219–233.

[13] R. ISERMAN, *Process fault detection based on modeling and estimation methods—A survey*, Automatica, 20 (1984), pp. 387–404.

[14] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.

[15] R. Z. KHAS'MINSKII, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.

[16] P. E. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer-Verlag, Berlin, 1992.

[17] R. KULHAVÝ, *Recursive nonlinear estimation: A geometric approach*, Automatica, 26 (1990), pp. 545–555.

[18] R. KULHAVÝ, *Recursive nonlinear estimation: Geometry of a space of posterior densities*, Automatica, 28 (1992), pp. 313–323.

[19] R. KULHAVÝ, *System identification: From matching data to matching probabilities*, in Plenary Lectures and Minicourses, G. Bastin and M. Gevers, eds., European Control Conference, Brussels, 1997, pp. 131–160.

[20] Y. A. KUTOYANTS, *Parameter Estimation for Stochastic Processes*, Heldermann Verlag, Berlin, 1984.

[21] T. L. LAI, *Sequential changepoint detection in quality control and dynamical systems (with discussion)*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 613–658.

[22] A. N. SHIRYAYEV, *On optimum methods in quickest detection problems*, Theory Probab. Appl., 8 (1963), pp. 22–46.

[23] M. H. VELLEKOOP, *Rapid Detection and Estimation of Abrupt Changes by Nonlinear Filtering*, Ph.D. Thesis, Imperial College, London, 1997.

[24] M. H. VELLEKOOP AND J. M. C. CLARK, *Asymptotic behaviour of the optimal filter of jump and slope jump processes*, in Proceedings of the 35th IEEE Conference on Decision and Control, Vol. 2, IEEE Press, Piscataway, NJ, 1996, pp. 1163–1168.

[25] M. H. VELLEKOOP AND J. M. C. CLARK, *Changepoint detection using nonlinear filters*, in Proceedings of the 4th European Control Conference, Brussels, 1997 (CD-ROM).

[26] A. S. WILLSKY, *A survey of design methods for failure detection in dynamic systems*, Automatica, 12 (1976), pp. 601–611.

[27] E. WONG AND B. HAJEK, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1984.

[28] W. M. WONHAM, *Some applications of stochastic differential equations to optimal nonlinear filtering*, J. Soc. Indust. Appl. Math. Ser. A Control, 2 (1965), pp. 347–369.