Daumer, Falk, Beyer:

On-line monitoring using Multi-Process Kalman Filtering

Sonderforschungsbereich 386, Paper 54 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

On-line monitoring using Multi-Process Kalman Filtering

Preliminary Discussion Paper

Martin Daumer, Markus Falk, Ulrich Beyer

Institut für medizinische Statistik und Epidemiologie
Klinikum rechts der Isar - Technische Universität München

Contents

1 Introduction

On-line monitoring of time series becomes more and more important in different areas of application like medicine, biometry and finance. In medicine, on-line monitoring of patients after transplantation of renals [1] is an easy and prominent example. In finance, fast end reliable recognition of changes in level and trend of intra-daily stock market prices is of obvious interest for ordering and purchasing. In this project, we currently consider monitoring of surgical data like heart-rate, blood pressure and oxygenation.
From a statistical point of view, on-line monitoring can be considered as on-line detection of changepoints in time series. That means, changepoints have to be detected in real time as new observations come in, usually in short time intervals. Retrospective detection of changepoints, after the whole batch of observations has been recorded, is nice but useless in monitoring patients during an operation.
There are various statistical approaches conceivable for on-line detection of changepoints in time series. Dynamic or state space models seem particularly well suited because "filtering" has historically been developed exactly for on-line estimation of the "state" of some system. Our approach is based on a recent extension of the so-called multi-process Kalman filter for changepoint detection [2]. It turned out, however, that some important issues for adequate and reliable application have to be considered, in particular the (appropriate) handling of outliers and, as a central point, adaptive on-line estimation of control- or hyper-parameters. In this paper, we describe a filter model that has this features and can be implemented in such a way that it is useful for real time applications with high frequency time series data.
Recently, simulation based methods for estimation of non-Gaussian dynamic models have been proposed that may also be adapted and generalized for the purpose of changepoint detection. Most of them solve the smoothing problem, but very recently some proposals have been made that could be useful also for filtering and, thus, for on-line monitoring [4, 5, 6]. If these approaches are a useful alternative to our development needs a careful comparison in future and is beyond the scope of this paper.

2 The Dynamic Linear Model

Throughout this paper we will use the notation as in [3].

2.1 Definition of the Dynamic Linear Model

Let $Y_t \in R$ be the observation at time $t \in [0, 1 \ldots T\}$. Then for each timepoint t the dynamic linear model is defined by

$$\begin{aligned} \text{Observation equation:} \quad & Y_t = F_t'\theta_t + v_t \ , \qquad v_t \sim N(0, V_t) \\ \text{System Equation:} \quad & \theta_t = G_t\theta_{t-1} + w_t \ , \quad w_t \sim N(0, W_t) \end{aligned}$$

where F_t, G_t are known design matrices describing the deterministic part of the observation process and of the system evolution. Both processes are disturbed by the Gaussian noise terms $v(t)$ (observation variance) and $w(t)$ (evolution variance), which are assumed to be mutually independent with variances V_t and W_t.

The model is initialized by a known prior for the initial state vector θ_0, usually taken to be

$$(\theta_0 | D_0) \sim N(m_0, C_0),$$

where generally $D_t = \{Y_t, \ldots, Y_0\} = \{Y_t, D_{t-1}\}$ represents the information set at time t, such that D_0 represents the initial information.

The dynamic linear model with design matrices F_t, G_t and variances V_t, W_t may symbolically be written as $M_t = \{F, G, V, W\}_t$.

2.2 Estimating the state vector θ_t

The following updating equations are used in estimating θ_t (see also [3]):

(a) Given posterior information at time $t-1$

$$(\theta_{t-1} | D_{t-1}) \sim N\left(m_{t-1}, C_{t-1}\right)$$

we arrive at the

(b) Prior at time t

$$(\theta_t | D_{t-1}) \sim N\left(a_t, R_t\right),$$

where $a_t = G_t m_{t-1}$, $R_t = G_t C_{t-1} G_t' + W_t$.

Next we can forecast Y_t, thereby using information up to time $t-1$.

(c) One-step forecast

$$(Y_t | D_{t-1}) \sim N(f_t, Q_t),$$

where $f_t = F_t' a_t$, $Q_t = F_t' R_t F_t + V_t$.

Eventually we obtain the

(d) Posterior at time t

$$(\theta_t | D_t) \sim N(m_t, C_t),$$

where $m_t = a_t + A_t e_t$, $C_t = R_t - A_t A_t' Q_t$

with $A_t = R_t F_t Q_t^{-1}$, $e_t = Y_t - f_t$.

3 A Multi-Process Model for the On-line Monitoring Problem

Combinations of different filters are called Multi-Process Models. Let $\mathcal{A}$ be some index set and for $\alpha \in \mathcal{A}$ let $M_t(\alpha)$ be the model corresponding to α (for some F, G, V, W depending on t and α).
In the simplest case there is some fixed (though maybe unknown) α such that the model $M_t(\alpha)$ holds for all t and this is what turns out to be general enough to handle the on-line problem.
For the estimation of α we use Bayes' theory. Given an initial prior $p(\alpha|D_0)$, inferences about α can be done sequentially by $p(\alpha|D_t) \propto p(\alpha|D_{t-1}) p(Y_t|\alpha, D_{t-1})$.
Our multi-process model for the on-line monitoring problem—including multiple changepoints—is based on the dynamic changepoint model developed in [2], covering situations with at most one changepoint. In the following we give a brief description of the latter model.

3.1 The dynamic changepoint model

The structural component model of [2] describes a system without a changepoint by a simple random walk. Changepoints are incorporated by a "switch," which adds at some fixed but unknown time τ a (possibly noisy) drift to the system equation. Thus the observation and system equations are:

$$\begin{array}{rll} \text{Observation equation:} & Y_t = \mu_t + v_t \ , & v_t \sim N(0, \sigma_y^2) \\ \text{System Equation:} & \mu_t = \mu_{t-1} + z_t^{(\tau)} \beta_{t-1} + w_{1t} \ , & w_{1t} \sim N(0, \sigma_\mu^2) \\ & \beta_t = \beta_{t-1} + w_{2t} \ , & w_{2t} \sim N(0, \sigma_\beta^2) \end{array}$$

where $z_t^{(\tau)}$ is an indicator variable with

$$z_t^{(\tau)} = \begin{cases} 0 & : \quad t < \tau \\ 1 & : \quad t \geq \tau \end{cases}$$

We shall use the following notation: the "0-filter" refers to a filter with $z_t^{(\tau)} = 0$ and the "1-filter" refers to filter with $z_t^{(\tau)} = 1$.
Every $\tau \in \{0, 1, \ldots, T\}$ defines a different model. The collection of all these single-process models labeled by τ is called the multi-process model $M_t(\tau)$.
In matrix notation:

$$Y_t = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{F'} \theta_t + v_t$$

$$\underbrace{\begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix}}_{\theta_t} = \underbrace{\begin{bmatrix} 1 & z_t(\tau) \\ 0 & 1 \end{bmatrix}}_{G_t^\tau} \underbrace{\begin{pmatrix} \mu_{t-1} \\ \beta_{t-1} \end{pmatrix}}_{\theta_{t-1}} + \underbrace{\begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}}_{w_t}$$

with

$$Var(v_t) =: V_t \ , \ Var(w_t) =: W_t$$

We discuss the problem of choosing V_t and W_t in section 4.2. The updating equations given a changepoint $\tau = j$ are described in section 2.2.

3.2 The estimation of changepoints

The posterior distributions of the changepoints $P(\tau = j | D_t)$, $j = [1, \ldots, T]$, can be calculated by Bayes with:

$$P(\tau = j|D_t) \propto P(Y_t|D_{t-1}, \tau = j)\, P(\tau = j|D_{t-1})$$

These probabilities must be initialized. If π denotes the probability that a changepoint occurs until time T, a reasonable initial prior is the uniform prior:

$$P(\tau = j|D_0) = \frac{\pi}{T}, j = 1 \dots T$$

For the estimation of τ it is only necessary to consider models up to time t, since all conditional models with $\tau \geq t + 1$ are identical:

$$P(Y_t|D_{t-1}, \tau = j) = P(Y_t|D_{t-1}, \tau = t + 1) \;\;, j \geq t + 1.$$

Hence the posterior distribution of the changepoint τ at time t is given by

$$P(\tau = j|D_t) \;=\; \begin{cases} c_t \Phi(y_t; f_t^j, Q_t^j) \cdot P(\tau = j|D_{t-1}) & j \leq t \\ c_t \Phi(y_t; f_t^{t+1}, Q_t^{t+1}) \cdot P(\tau = j|D_{t-1}) & t < j \leq T \end{cases}$$

$$P(\tau > T|D_t) \;=\; c_t \Phi\left(y_t; f_t^{t+1}, Q_t^{t+1}\right) \cdot P(\tau > T|D_{t-1})$$

where $f_t^j = E(Y_t|D_{t-1}, \tau = j)$ and $Q_t^j = Var(Y_t|D_{t-1}, \tau = j)$ are the mean and variance of the one-step forecast density (see section 2.2) and Φ is the density of the normal distribution, c_t being the normalization constant.

The dynamic linear changepoint model seems to be an appropriate model, which allows to detect on-line deviations from an assumed course of a monitored variable. But there exist still some unsolved problems:

- Outliers can have an important influence on the probability of a changepoint.
- Long observation periods entail the need for handling many models simultaneously such that the algorithm becomes too slow for real time applications.
- The original model allows only to detect at most one changepoint during the observation period.

- The variances V_t and W_t are in many practically important cases unknown.

The next chapter shows how these problems can be solved.

4 Towards an On-line Monitoring Alert System

4.1 Introduction of a time window for the $1 - filters$

The computational time increases rapidly with the increasing number of *1-filters* to be processed such that the speed may easily drop below the limit for real time applications. To overcome this problem we introduce a window $[t-b, t]$ for these *1-filters*, with some positive b depending on the computational power and the specific problem. Then the probabilities for the changepoints are

$$P(\tau = j|D_t) = \begin{cases} c_t\Phi(y_t; f_t^j, Q_t^j) \cdot P(\tau = j|D_{t-1}) & t-b \leq j \leq t \\ c_t\Phi(y_t; f_t^{t+1}, Q_t^{t+1}) \cdot P(\tau = j|D_{t-1}) & t < j \leq T \\ 0 & j < t-b \end{cases}$$

$$P(\tau > T|D_t) = c_t\Phi\left(y_t; f_t^{t+1}, Q_t^{t+1}\right) \cdot P(\tau > T|D_{t-1})$$

Only the *0-filter* and the *t-b+1* *1-filters* are considered in the calculations of the changepoint probabilities. The result is a constant calculation speed over time. An additional advantage is that the model is now able to deal with more than one changepoint. Since a changepoint before time $t-b$ is no longer respected, we estimate the posterior distribution of the actual changepoint using information only from within this time window. However, the window $t-b$ is dynamic. One $1 - filter$ is added for the new observation and in the same moment we drop the $1 - filter$ for observation $t-b$. Hence, in moving the window over time we are able to detect sequential changepoints.

4.2 Hierarchical Multiprocess Models

Let $P(M_t^{(\alpha)}|D_t)$ be the probability that the model $M_t^{(\alpha)}$, for some $\alpha \in \mathcal{A}$, holds at time t. Then we define a hierarchical model by the probability

$$P(M_t^{(\alpha)}, M_t^{(\beta)}|D_t) = P(M_t^{(\alpha)}|M_t^{(\beta)}, D_t)P(M_t^{(\beta)}|D_t),$$

where $\beta \in \mathcal{B}$, and $\mathcal{A}, \mathcal{B}$ are disjoint parameter sets.

If one is interested in marginal probabilities one may calculate them via
$P(M_t^{(\alpha)}|D_t) =$
$\sum_{i \in \mathcal{B}} P(M_t^{(\alpha)}|M_t^{(\beta_i)}, D_t)P(M_t^{(\beta_i)}|D_t)$.

This definition should not be confused with a multiprocess model of class II, in which one will not distinguish between $\mathcal{A}$ and $\mathcal{B}$. A hierarchical model is the combination of two ore more multiprocess models of class I. So one is able to follow a decision tree within the set of different filters. We will use it to build an estimation procedure for the unknown V_t and W_t, as well as for the outlier detection.

Before we propose the estimation procedures for the unknown variances we make some basic considerations. Until now we did not distinguish between different filters and their variances. However, this will become important, when we are going to estimate this variances on-line. The fundamental approach to the On-line monitoring problem using the dynamic linear changepoint model is, that the new observation Y_t is explained by two types of models (the 0- and 1-filters). A changepoint is detected when the 1-filters are better in predicting Y_t with than the 0- filter.
Since the system equations of the 0- and $1-filters$ are different ($\mu_t = \mu_{t-1} + w_t^{(0)}$ for the $0-filter$ and $\mu_t = \mu_{t-1} + \beta_{t-1} + w_t^{(\tau)}$ for the $1-filters$) one will have different evolution variances $W_t^{(0)}, W_t^{(\tau)}, \tau = 1, .., t$.
The only difference between the $0-$ and $1-filters$ is the slope parameter β. Hence, in adding a slope parameter to the system equation a part of the evolution variance estimated for the $0-filter$, is now explained by the slope itself, and therefore $W_t^{(\tau)} \leq W_t^{(0)}$. However, the observation variance V_t, which has the interpretation of measurement error, is identical for both models, because the observation equations are identical too.
These considerations lead to the following estimation concept. Since we are not able to find a closed estimation procedure, in terms of a single multiprocess filter, to estimate V_t and W_t simultaneously we have to estimate these variances separately and independent from the estimation of a changepoint. Therefore we will introduce, for estimation of the unknown variances, a new multiprocess

filter consisting of the 0-filter and the 1-filter (with $\tau = 1$). This leads to what we call a *hierarchical* on-line estimation procedure.

Online Estimation of the unknown Variances V and W

4.3 Estimation of V

To estimate the unknown observational variance V_t treated here as constant over time, we adapted a conjugate sequential updating procedure, described in West, Harrison (1989, 118ff). Since V becomes now a random quantity the normal distribution changes into a t-distribution and we will obtain the following system:

$$\begin{aligned} \text{Observation equation:} \quad & Y_t = F'\theta_t + v_t \ , && v_t \sim N(0, V) \\ \text{System Equation:} \quad & \theta_t = G_t^\tau \theta_{t-1} + w_t \ , && w_t \sim T_{n_{t-1}}(0, W_t), \tau \in \{T+1, 1\} \end{aligned}$$

where $T_{n_{t-1}}(\mu, \sigma^2)$ denotes the noncentral T-distribution with mean μ, variance σ^2 and n_{t-1} degrees of freedom. The expression $\tau \in \{T+1, 1\}$ indicates two filters, one for the $0-filter$ and one for the $1-filter$, that started from the beginning. The updating equations will take now the form:

(a) Posterior at $t-1$: $(\theta_{t-1}|D_{t-1}, \tau) \sim T_{n_{t-1}}(m_{t-1}, C_{t-1})$

(b) Prior at t: $(\theta_t|D_{t-1}, \tau) \sim T_{n_{t-1}}(a_t, R_t)$

$$\begin{aligned} a_t^j &= G_t^j m_{t-1}^j \\ R_t^j &= G_t^j C_{t-1}^j G_t'^j + W_t \end{aligned}$$

(c) one-step forecast: $(Y_t|D_{t-1}, \tau) \sim T_{n_{t-1}}(f_t, Q_t)$

$$\begin{aligned} f_t^j &= F'a_t^j \\ Q_t^j &= S_{t-1}^j + F'R_t^j F \end{aligned}$$

(d) Posterior at t: $(\theta_t|D_t, \tau) \sim T_{n_{t-1}}(m_t, C_t)$

$$m_t^j = a_t^j + A_t^j e_t^j$$

$$C_t^j = S_t^j/S_{t-1}^j \left[R_t^j - A_t^j A'^j_t Q_t^j\right]$$
$$S_t^j = \frac{d_t}{n_t}$$

where
$n_t = n_{t-1} + 1, \; d_t = d_{t-1} + S_{t-1}^j e^{j2}{}_t/Q_t^j$ and $A_t = R_t^j F/Q_t^j$

Under the assumption, that the estimated variances $V^{(0)}$ and $V^{(\tau=1)}$ are now known quantities, we can combine the two filters in a multiprocess model and use this to get an estimate of V simultaneously.
Using Bayes we get

$$P(V^j|D_t) \propto P(Y_t|V^j, D_{t-1})P(V^j|D_{t-1}), j \in \{T+1, 1\}$$

and we can get an estimate of V by

$$\hat{V} = \sum_j V^j P(V^j|D_t).$$

As initial probabilities $P(V^j|D_0)$ one can use the probabilities $\pi, 1-\pi$, which were used to initialize $P(\tau|D_0)$ in the changepoint estimation procedure. The single estmates of V^j will be passed to the hierarchical changepoint model.

4.4 Estimating W_t

Similar to the previous section we will build an estimation procedure to calculate W_t. In a first step we transform the problem of estimating W_t to a problem were we have to estimate a discounted variance. As proposed by [3] we introduce a discounting factor δ, with $0 < \delta \leq 1$. By definition we can set

$$W_t = P_t(1-\delta)/\delta$$

with

$$P_t = G_t^j C_{t-1}^j G_t'^j.$$

One advantage is now that in contrast to W_t, δ is scale free. Furthermore δ is related to the signal to noise ratio $r = W_t/V_t = (1-\delta)^2/\delta$. In the literature values like $\delta = 0.7$ or 0.9 are chosen to be fixed and the usual updating equations are used to estimate the state vector θ_t. Hence, a possible strategy could be to analyze several data sets with defined changepoints and to look for the best value of δ, where not more but the maximum of the defined changepoints can be detected. But this would not be an On-line estimation of the evolution variance W_t. Another possibility is to estimate the unknown discounting factor δ similarly to the observation variance V. Our proposal is to do the following: As mentioned in the beginning of this chapter we need a δ for the $0-filter$ and the $1-filter$. So we have to build two different multiprocess models. First let $\tau \in \{T+1, 1\}$. Then

(a) choose a discrete set $[\delta_1, \delta_2, \ldots, \delta_k]$ of values for δ (k appropriately chosen)

(b) calculate at each step the probabilities of δ using

$$P(\delta^{(\tau)}|D_t) \propto P(Y_t|\delta^{(\tau)}, D_{t-1}, \tau)P(\delta^{(\tau)}|D_{t-1}, \tau)$$

one may estimate $\delta^{(\tau)}$ using

$$\hat{\delta^{(\tau)}} = \sum_{i=1}^{k} \delta_i P(\delta_i|D_t, \tau)$$

It seems to be natural to use the uniform distribution $P(\delta_j|D_0) = \frac{1}{k}, j = 1, .., k$ for the initial probabilities. This method appears to be a good estimation strategy for the unknown $\delta^{(\tau)}$. Onece again the estimated parameters are passed to the hierarchical changepoint model.

4.5 Respecting Outliers

To detect outliers we used the ideas of [1]. An outlier can be interpreted as a sudden perturbation of the observation equation. To include this possibility we could enlarge the multiprocess model by an extra filter for outliers (which we call "N-filter"), which is exactly the *0-filter* with an enlarged observation variance. Since, V and W_t are estimated on-line this will not work. Instead of including the extra $N-filter$ into the changepoint model we introduce an extra multiprocess model for outliers. This model will become the first level of our hierarchical multiprocess model. Let

$$\begin{aligned} M_t(\kappa) \quad : \quad & Y_t = F'\theta_t + v_t, \qquad v_t \sim N(0, \kappa V) \\ & \theta_t = G_t^{(T+1)}\theta_{t-1} + w_t,\ w_t \sim N(0, W_t), \end{aligned}$$

where κ is chosen sufficiently large, say $\kappa = 100$. Now we may estimate the probability of an outlier by

$$P(Outlier|D_t) \propto P(Y_t|Outlier, D_{t-1})P(Outlier|D_{t-1})$$

.

4.6 Initialization

We show now how the prior distributions of the state vectors must be specified. We initialize the $0-filter$ and $1-filter$ similar to [2] with a data driven prior. The variances $C_0^{(\tau)}$ follow hereby a diffuse prior.
The $0-filter$ $(\theta_0|D_0, T+1) \sim N(m_0^{(T+1)}, C_0^{(T+1)})$ is initialized by

$$m_0^{(T+1)} = (Y_1, 0)' \text{ and } C_0^{(T+1)} = \begin{pmatrix} V_0 + W_0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Prior information of the $1-filters$ $(\theta_0|D_0, \tau \leq t) \sim N(m_0^{(\tau\leq t)}, C_0^{(\tau\leq t)})$ is recursively defined by

$$m_0^{(\tau=t)} = (m_{t-1,1}^{(\tau=t-1)}, Y_{t-1} - m_{t-1,1}^{(\tau=t-1)})',$$
$$C_0^{(\tau=t)} = \begin{pmatrix} C_0^{(\tau=t-1)} & -C_0^{(\tau=t-1)} \\ -C_0^{(\tau=t-1)} & C_0^{(\tau=t-1)} + V_t + W_t \end{pmatrix}.$$

Furthermore, we have to specify starting values for the variances V, W and we have to choose a discrete set for the discounting factor δ. Since we are going to estimate all parameters on-line we need a good guess for the signal to noise ratio r of the underlying process. Otherwise the estimation procedure will not converge to the true values.
For the approximation of r we will use the first $2k$ observations (chose k appropriately) and we define the following quantities:

$$a = V(Y_i - Y_{i-1}), i = 1, .., 2k,$$
$$b = V(Y_i + Y_{2k-i+1}), i = 1, .., k.$$

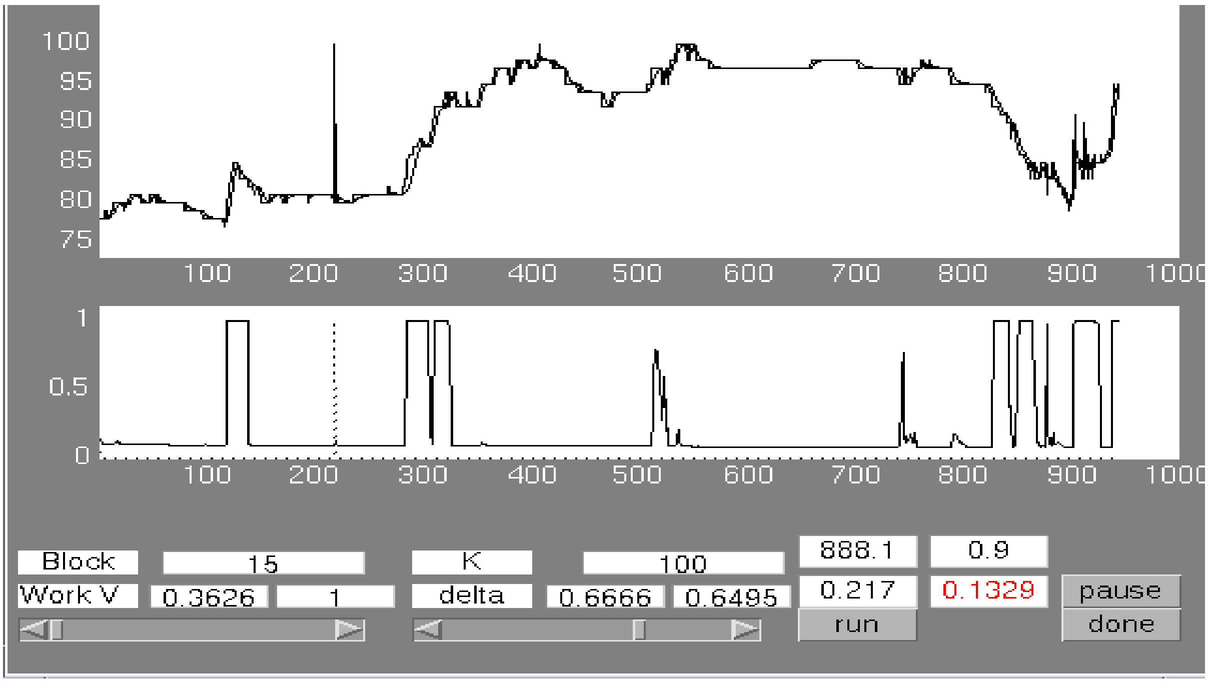
Then

$$r = 2\frac{a-b}{(2k+1)b-a}.$$

With this we are able to choose V and W such that $r = W/V$. Furthermore, we can now choose a discrete set for δ. Since $r = (1-\delta)^2/\delta$ it is convenient to choose a discrete interval of δ about r. This interval can then be updated as new observations are made using the same approximation as before.

5 Example

The following data are the ECG measurements, taken every five seconds, from a patient undergoing a skin transplantation. Monitoring did start when the first steps in preparing the patient were finished and anesthesia was completed. The first window will show the ECG measurements with the filtered values of the $0-filter$. The second window displays the estimated cumulative probability that a changepoint did occur during the observation window. Furthermore we will display the probability of an outlier at the actual timepoint.



We see from the figure, that at observation 120 an alert is given. This coincides with the beginning of the first skin cut. At 250 we did introduce an outlier, who was detected by the $N-filter$. At 285 the operation starts. Since the patient did react to this, the anesthesiologist did intervene. The weak changepoint at 506 was in this stabilization period. From 540 to 752 we have a stable phase. At 752 we do observe a weak changepoint. This was to the end of the operation and the anesthesiologist did begin the weak up phase. At 821 the patient did awake.

6 Acknowledgments

For helpful comments we thank especially Prof. Dr. Ludwig Fahrmeir, Prof. Dr. Albrecht Neiss and Dr. Stefan Wagenpfeil.

References

[1] Smith, A.F.M. and West, M., Monitoring renal transplants: an application of the multi-process Kalman filter. Biometrics, 39, 867-878 (1983)

[2] Whittaker, J., and Frühwirth-Schnatter, *A Dynamic Changepoint Model for Detecting the Onset of the Growth in Bacteriological Infections*, Appl. Statist. **43**, No. 4, (1994)

[3] West, M., and Harrison, P.J., Bayesian Forecasting and Dynamical Models, Springer, New York (1989)

[4] Kitagawa, G., Monte Carlo smoother for non-gaussian nonlinear state space models, Journal of Computation and Graphical Statistics 5, 1-25 (1996)

[5] Kitagawa, G., and Matsumoto, N., Detection of Coseismic Changes of Underground Water Level, Jasa 91, 521-528 (1996)

[6] Kim,S., Shephard, N., Chib, S.,Stochastic Volatility: Likelihood Inference and Comparison with Arch Models, Technical Report (1996)

[7] Berzuini, C., Best, N.G., Gilks, W.R., Larizza, C., Dynamic Conditional Independence Models and Markov Chain Monte Carlo Methods, Technical Report (1996)