LECTURE 7

In the approximated NP-PP problem:

$$Y_i \sim \mathcal{P}\left(\lambda(\hat{t}_i \mid \mathcal{H}_{\hat{t}_i}, \vartheta)\,\Delta t\right)$$

$$\log \lambda(t \mid \mathcal{H}_t, \vartheta) = g(t, \mathcal{H}_t, \vartheta)$$

$$i = 1, 2, 3, \ldots, m$$

the parameter vector ϑ is time-invariant and must estimated from the entire realization $(y_1, y_2, \ldots, y_m)$ of the sequence of RVs $Y_1, Y_2, \ldots, Y_m \Rightarrow$ What if $\vartheta = \vartheta_t$, i.e., the parameter vector is a function of time and perhaps past history? The case is pertinent since λ is a function of t and $\mathcal{H}_t \Rightarrow$ We consider the new problem:

$$Y_i \sim \mathcal{P}\left(\lambda(\hat{t}_i \mid \mathcal{H}_{\hat{t}_i}, \vartheta_i)\,\Delta t\right)$$

$$\log \lambda(t \mid \mathcal{H}_t, \vartheta_t) = g(t, \mathcal{H}_t, \vartheta_t)$$

$$\vartheta_i \triangleq \vartheta_{\hat{t}_i} \qquad i = 1, 2, 3, \ldots, m$$

In this case, we have one new vector ϑ_i for each new observation $y_i \Rightarrow$ A ML method could not be viable $\Rightarrow$ An efficient and recursive solution can be obtained under the additional assumptions:

i) $\vartheta_{i+1} = F_i \vartheta_i + \eta_i \qquad i = 0, 1, 2, \ldots, m$

where: F_i - known matrix $\forall i$

$\eta_i \sim N(0, Q_i)$ $\forall i$ - white noise

$Q_i \triangleq$ covariance matrix

ii) $\vartheta_0 \triangleq$ initial condition, $\vartheta_0 \sim N(\mu_0, W_0)$

μ_0, W_0 - known

$\Rightarrow$

Assumption (i) means that there is an evolution model for ϑ_i $\forall i$

Assumptions (i) and (ii) mean that each vector ϑ_i is a Gaussian random vector with its own mean and covariance matrix

Since ϑ_i is a random vector $\forall i$, we can consider the conditional probability $P(\vartheta_i | Y_i, \mathcal{H}_{\hat{t}_i})$ and the correspondent pdf: $f_{\vartheta_i}(\vartheta_i | Y_i, \mathcal{H}_{\hat{t}_i})$.

For sake of simplicity, let us use the notation: $\mathcal{H}_i = \mathcal{H}_{\hat{t}_i}$ and $f_\vartheta = f_{\vartheta_i}$ (the latter makes sense since the random vectors ϑ_i are all Gaussian). By applying the Bayes' theorem, we have:

$$P(\vartheta_i | Y_i, \mathcal{H}_i) = \frac{P(\vartheta_i, Y_i, \mathcal{H}_i)}{P(Y_i, \mathcal{H}_i)} = \frac{P(Y_i | \mathcal{H}_i, \vartheta_i) P(\vartheta_i | \mathcal{H}_i)}{P(Y_i | \mathcal{H}_i)}$$

We can note:

- $P(Y_i | \mathcal{H}_i, \vartheta_i) \overset{\text{Formulation of the NP-PP problem}}{\cong} \exp\left\{ y_i \log\left(\lambda(\hat{t}_i | \mathcal{H}_i, \vartheta_i)\Delta t\right) - \lambda(\hat{t}_i | \mathcal{H}_i, \vartheta_i)\Delta t \right\}$

- $f_\vartheta(\vartheta_i | Y_i, \mathcal{H}_i) = \frac{P(Y_i | \mathcal{H}_i, \vartheta_i)}{P(Y_i | \mathcal{H}_i)} \cdot f_\vartheta(\vartheta_i | \mathcal{H}_i) \quad (*)$

- $P(Y_i | \mathcal{H}_i)$ is a normalization factor that assures the integral of f_ϑ over ϑ_i is equal to 1

- $f_\vartheta(\vartheta_i | \mathcal{H}_i) = \int \underset{\text{joint-probability function}}{f(\vartheta_i, \vartheta_{i-1} | \mathcal{H}_i)} \, d\vartheta_{i-1} = \int f_\vartheta(\vartheta_i | \vartheta_{i-1}, \mathcal{H}_i) f_\vartheta(\vartheta_{i-1} | \mathcal{H}_i) \, d\vartheta_{i-1}$

$$= \int f_\vartheta(\vartheta_i | \vartheta_{i-1}, \mathcal{H}_i) f_\vartheta(\vartheta_{i-1} | Y_{i-1}, \mathcal{H}_{i-1}) \, d\vartheta_{i-1}$$

$f_\theta(\vartheta_i \mid Y_i, \mathcal{H}_i) \triangleq$ Posterior density of ϑ_i at time t_i

$f_\theta(\vartheta_i \mid \mathcal{H}_i) \triangleq$ Prediction density of ϑ_i at time t_i with observations up to t_{i-1}

$$\Downarrow$$

$$f_\theta(\vartheta_i \mid Y_i, \mathcal{H}_i) \propto P(Y_i \mid \mathcal{H}_i, \vartheta_i) \int f_\theta(\vartheta_i \mid \vartheta_{i-1}, \mathcal{H}_i)\, f_\theta(\vartheta_{i-1} \mid Y_{i-1}, \mathcal{H}_{i-1})\, d\vartheta_{i-1}$$

↑ posterior density at time t_i

↑ It is related to the prediction density

↑ posterior density at time t_{i-1}

Assumptions (i) and (ii) also determine the following facts:

a) $P(\vartheta_i \mid \vartheta_{i-1}, \mathcal{H}_i) \underset{\text{evolution model}}{=} P(F_i \vartheta_{i-1} + \eta_i \mid \vartheta_{i-1}, \mathcal{H}_i) \underset{\text{for any given value } \vartheta_{i-1}, \mathcal{H}_i}{=} P(\eta_i)$ since η_i is white noise

$\Rightarrow f_\theta(\vartheta_i \mid \vartheta_{i-1}, \mathcal{H}_i)$ - Gaussian function of the var.: $\vartheta_i - F_i \vartheta_{i-1}$

b) Because of (a), if $f_\theta(\vartheta_{i-1} \mid Y_{i-1}, \mathcal{H}_{i-1})$ is Gaussian, then $f_\theta(\vartheta_i \mid Y_i, \mathcal{H}_i)$ is Gaussian too, i.e., the posterior density function preserves its characteristics in time $\Rightarrow$ It is therefore plausible to focus on the quantities:

$$\vartheta_{i|i-1} \triangleq \underset{\theta}{E}[\vartheta_i \mid \mathcal{H}_i] \qquad W_{i|i-1} \triangleq \underset{\theta}{E}[(\vartheta_i - \vartheta_{i|i-1})^2 \mid \mathcal{H}_i] = \text{var}[\vartheta_i \mid \mathcal{H}_i]$$

$$\vartheta_{i|i} \triangleq \underset{\theta}{E}[\vartheta_i \mid Y_i, \mathcal{H}_i] \qquad W_{i|i} \triangleq \underset{\theta}{E}[(\vartheta_i - \vartheta_{i|i})^2 \mid Y_i, \mathcal{H}_i] = \text{var}[\vartheta_i \mid Y_i, \mathcal{H}_i]$$

Note: - $\vartheta_{i|i-1}$ and $W_{i|i-1}$ are mean and covariance matrix of the one-step prediction density $f_\theta(\vartheta_i \mid \mathcal{H}_i)$

- $\vartheta_{i|i}$ and $W_{i|i}$ are mean and covariance matrix of the posterior density function $f_\theta(\vartheta_i \mid Y_i, \mathcal{H}_i)$

④

Because of the definition, we have:

1) $$\theta_{i|i-1} = \underset{\theta}{E}\left[\theta_i \mid \mathcal{H}_i\right] = F_i \underset{\theta}{E}\left[\theta_{i-1} \mid \mathcal{H}_i\right] = F_i\, \theta_{i-1|i-1}$$

2) $$W_{i|i-1} = \text{var}\left[\theta_i \mid \mathcal{H}_i\right] = \text{var}\left[F_i\theta_{i-1} + \eta_i \mid \mathcal{H}_i\right] = \text{var}\left[F_i\,\theta_{i-1} \mid \mathcal{H}_i\right] + \text{var}\left[\eta_i\right]$$

since η_i is white noise

$$= F_i\, W_{i-1|i-1}\, F_i^T + Q_i$$

Because of assumptions (i) and (ii) we also have:

$$f_\theta(\theta_i \mid \mathcal{H}_i) = \int f_\theta(\theta_i \mid \theta_{i-1}, \mathcal{H}_i)\, f_\theta(\theta_{i-1} \mid Y_{i-1}, \mathcal{H}_{i-1})\, d\theta_{i-1}$$

$$= f(\eta_i) \int f_\theta(\theta_{i-1} \mid Y_{i-1}, \mathcal{H}_{i-1})\, d\theta_{i-1} = f(\eta_i) \text{ - Gaussian function}$$

Therefore, if we replace $\eta_i = \theta_i - F_i\,\theta_{i-1}$ with the prediction error $\varepsilon_i \triangleq \theta_i - \theta_{i|i-1}$ we have:

$$f_\theta(\theta_i \mid \mathcal{H}_i) \cong \frac{1}{\sqrt{(2\pi)^r \det(W_{i|i-1})}} \exp\left\{-\frac{1}{2}(\theta_i - \theta_{i|i-1})^T W_{i|i-1}^{-1} (\theta_i - \theta_{i|i-1})\right\}$$

where $r \triangleq$ size of vector θ_i

Therefore, the conditional pdf $f_\theta(\theta_i \mid Y_i, \mathcal{H}_i)$ in (*) can be approximated with a function that is proportional to:

$$\exp\left\{y_i \log\left(\lambda(\hat{t}_i \mid \mathcal{H}_i, \theta_i)\,\Delta t\right) - \lambda(\hat{t}_i \mid \mathcal{H}_i, \theta_i)\,\Delta t + \right.$$

$$\left. - \frac{1}{2}(\theta_i - \theta_{i|i-1})^T W_{i|i-1}^{-1} (\theta_i - \theta_{i|i-1})\right\} \qquad (**)$$

On the other hand, because of (6), $f_\theta(\theta_i | Y_i, \mathcal{H}_i)$ is Gaussian too, and therefore it should be:

$$f_\theta(\theta_i | Y_i, \mathcal{H}_i) \propto \exp\left\{-\frac{1}{2}(\theta_i - \theta_{i|i})^T W_{i|i}^{-1}(\theta_i - \theta_{i|i})\right\} \quad (***)$$

Beside a constant scaling factor to be calculated, we can therefore combine (**) and (***) and write:

$$\exp\left\{y_i \log\left(\lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)\Delta t\right) - \lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)\Delta t - \frac{1}{2}(\theta_i - \theta_{i|i-1})^T W_{i|i-1}^{-1}(\theta_i - \theta_{i|i-1})\right\}$$

$$= A \exp\left\{-\frac{1}{2}(\theta_i - \theta_{i|i})^T W_{i|i}^{-1}(\theta_i - \theta_{i|i})\right\}$$

A: coefficient to be determined

$\Downarrow$

$$y_i \log\left(\lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)\Delta t\right) - \lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)\Delta t - \frac{1}{2}(\theta_i - \theta_{i|i-1})^T W_{i|i-1}^{-1}(\theta_i - \theta_{i|i-1})$$

$$= -\frac{1}{2}(\theta_i - \theta_{i|i})^T W_{i|i}^{-1}(\theta_i - \theta_{i|i}) + \log A$$

$\Downarrow$ By differentiating with respect to θ_i

c) $$W_{i|i}^{-1}(\theta_i - \theta_{i|i}) = W_{i|i-1}^{-1}(\theta_i - \theta_{i|i-1}) - \left(\frac{y_i}{\lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)} - \Delta t\right)\nabla_\theta \lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)$$

where: $\nabla_\theta \lambda \triangleq \left[\frac{\partial \lambda}{\partial \theta_{i,1}}, \frac{\partial \lambda}{\partial \theta_{i,2}} \dots \frac{\partial \lambda}{\partial \theta_{i,r}}\right]^T$ - gradient and $\theta_i \triangleq [\theta_{i,1}\ \theta_{i,2} \dots \theta_{i,r}]^T$

Note also the relationship that is usually used:

$$\nabla_\theta \log \lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)\left(y_i - \lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)\Delta t\right) = \left(\frac{y_i}{\lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)} - \Delta t\right)\nabla_\theta \lambda(\hat{t}_i | \mathcal{H}_i, \theta_i)$$

Under assumptions (i), (ii), and (6), formula (c) should hold $\forall \theta_i \Rightarrow$ By choosing $\theta_i = \theta_{i|i-1}$ we have:

$$W_{i|i}^{-1}\left(\theta_{i|i-1} - \theta_{i|i}\right) = -\nabla_\theta\left(\log\lambda(\hat{t}_i|\mathcal{H}_i, \theta_{i|i-1})\right)\left(y_i - \lambda(\hat{t}_i|\mathcal{H}_i, \theta_{i|i-1})\Delta t\right)$$

$$\Downarrow$$

3) $$\theta_{i|i} = \theta_{i|i-1} + W_{i|i}\,\nabla_\theta\left(\log\lambda(\hat{t}_i|\mathcal{H}_i, \theta_{i|i-1})\right)\left(y_i - \lambda(\hat{t}_i|\mathcal{H}_i, \theta_{i|i-1})\Delta t\right)$$

Moreover, if we differentiate (c) with respect to θ_i:

d) $$W_{i|i}^{-1} = W_{i|i-1}^{-1} - \nabla_\theta\left\{\nabla_\theta\left(\log\lambda(\hat{t}_i|\mathcal{H}_i, \theta_i)\right)\left(y_i - \lambda(\hat{t}_i|\mathcal{H}_i, \theta_i)\Delta t\right)\right\}$$

$$= W_{i|i-1}^{-1} - H_\theta \log\lambda\left(y_i - \lambda\Delta t\right) + \left(\nabla_\theta \log\lambda\right)^T\left(\nabla_\theta \log\lambda\right)\lambda\Delta t$$

where $H_\theta \log\lambda \triangleq$ Hessian of $\log\lambda(\hat{t}_i|\mathcal{H}_i, \theta_i)$ - Evaluation of formula (d) for $\theta_i = \theta_{i|i-1}$ leads to the equation:

4) $$W_{i|i}^{-1} = W_{i|i-1}^{-1} - \left[H_\theta \log\lambda\left(y_i - \lambda\Delta t\right) + \left(\nabla_\theta \log\lambda\right)^T\left(\nabla_\theta \log\lambda\right)(\lambda\Delta t)\right]_{\theta_i = \theta_{i|i-1}}$$

The combination of equations (1)-(4) defines an adaptive filter that recursively updates the estimation of the expected value of parameter vector θ_i at each time step t_i:

(1) + (2)

$$\theta_{i|i-1} = F_i\,\theta_{i-1|i-1}$$

$$W_{i|i-1} = F_i\,W_{i-1|i-1}\,F_i^T + Q_i$$

} Prediction Step

(3)+(4)

$$W_{i|i}^{-1} = W_{i|i-1}^{-1} + I(\theta_{i|i-1})$$
$$\theta_{i|i} = \theta_{i|i-1} + W_{i|i}\,\nabla_\theta(\log\lambda)\,(y_i - \lambda\Delta t)$$

} Correction Step

where: $I(\theta_{i|i-1}) \triangleq -H_\theta \log\lambda\,(y_i - \lambda\Delta t) + (\nabla_\theta \log\lambda)^T(\nabla_\theta\log\lambda)(\lambda\Delta t)$

$\lambda \triangleq \lambda(\hat{t}_i \mid \mathcal{H}_i, \theta_{i|i-1})$

Note: $(y_i - \lambda\Delta t)$ provides the difference between the probability of having an event in $(t_{i-1}, t_i]$ given the model and whether the event actually occurs ⇒ It is the INNOVATION of the filter

↗ It is the error signal used to update the model parameters

The vector $W_{i|i}\nabla_\theta(\log\lambda)$ can be seen as a learning gain in eq (3) ⇒ It is the equivalent of the Kalman gain for standard Kalman filters and it depends on the statistical properties of θ_i and the point process

□

The adaptive filter (1)-(4) is relevant for the following reasons:

* If $\eta_i = 0\ \forall i$ (i.e., θ_i is a deterministic process) ⇒ $W_{i|i-1} = F_i W_{i-1|i-1} F_i^T$ ⇒

$W_{i|i}^{-1} = \left(F_i W_{i-1|i-1} F_i^T\right)^{-1} + I(\theta_{i|i-1})$ ⇒ The filter is analogous to the recursive least-square filter (RLS) for deterministic models

$\{\theta_i\}$ is a Gaussian Random Walk ↖

* If $F_i = I_r\ \forall i$ (identity matrix), $W_{i|i} = \Sigma_\eta$ - constant } ⇒ $\theta_{i|i} = \theta_{i-1|i-1} + \Sigma_\eta \nabla_\theta(\log\lambda)(y_i - \lambda\Delta t)$

with $\lambda = \lambda(\hat{t}_i \mid \mathcal{H}_i, \theta_{i-1|i-1})$

$\Rightarrow$ The filter is analogous to the steepest descendent formula

$*$ The formulation of the filter (1)-(4) exploits the fact that ϑ_i $i=1,2,3,\dots,m$ are (Gaussian) RVs and that an evolution model is given (assumptions i-ii)

$\Downarrow$

ϑ_i could be replaced by a vector of explanatory variables $X_i \triangleq [X_{1,i}\ X_{2,i}\ \dots\ X_{q,i}]^T$ and the value of X_i at each time step could be recursively estimated from the realization of the NP-PP

$\Downarrow$

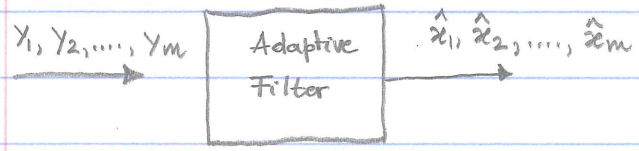
The filter (1)-(4) can be used for solving the following problem:

DECODING PROBLEM

$$? \ (x_1, x_2, \dots, x_m) : \quad Y_i \sim P\left(\lambda(\hat{t}_i \mid \mathcal{H}_i, X_i, \vartheta)\Delta t\right)$$

$$\log \lambda(t \mid \mathcal{H}_t, X, \vartheta) = g(t, \mathcal{H}_t, X, \vartheta)$$

$$X_i = x_i \quad i=1,2,3,\dots,m$$



Ex.: (scalar case) Let us assume that: $\log \lambda = \alpha_0 + \left\{\sum_{j=1}^{k} \beta_j \Delta N_{(t-j\Delta t,\, t-(j-1)\Delta t]}\right\} x$

$\Downarrow$

$$\nabla_x \log \lambda = \sum_{j=1}^{k} \beta_j \Delta N_{(t-j\Delta t,\, t-(j-1)\Delta t]} \ \text{- const. for a given time } t$$

Let us have: $x_i = x_{i-1} + \eta_i \Rightarrow x_{i|i-1} = x_{i-1|i-1}$

$\eta_i \sim N(0, \sigma^2) \Rightarrow W_{i|i-1} = W_{i-1|i-1} + \sigma^2$

$$W_{i|i}^{-1} = W_{i-1|i-1}^{-1} + (\nabla_x \log\lambda)^2 \lambda(\hat{t}_i | \mathcal{H}_i, x_{i-1|i-1}, \vartheta)\,\Delta t$$

$$x_{i|i} = x_{i-1|i-1} + W_{i|i}\,(\nabla_x \log\lambda)\,(y_i - \lambda(\hat{t}_i | \mathcal{H}_i, x_{i-1|i-1}, \vartheta)\,\Delta t)$$

where $\vartheta \triangleq [\alpha_0\ \beta_1\ \beta_2\ \dots\ \beta_K]^T$ □

Ex.: Let us assume that: $\log\lambda = \alpha_0 + \sum_{j=1}^{K} x_{j,t}\,\Delta N_{(t-j\Delta t, t-(j-1)\Delta t]}$

(vector, log-linear case)

$$\Downarrow$$

$$X_i = [x_{1,i}\ \ x_{2,i}\ \dots\ x_{K,i}]^T$$

$$\nabla_x \log\lambda = \left[\Delta N_{(t-\Delta t, t]}\ \ \Delta N_{(t-2\Delta t, t-\Delta t]}\ \dots\right]^T \triangleq \Gamma_i$$

$$(\nabla_x \log\lambda)^T(\nabla_x \log\lambda) = \begin{bmatrix} \Delta N^2_{(t-\Delta t, t]} & \Delta N_{(t-\Delta t, t]}\Delta N_{(t-2\Delta t, t-\Delta t]} & \cdots \\ \vdots & \ddots & \\ & & \end{bmatrix} \rightarrow \Lambda_i$$

This matrix is updated at each time step t_i based on the previous history

Let us assume: $X_i = X_{i-1} + \eta_i$, $\eta_i \sim N(0, \Sigma_i)$ ⇒

$$X_{i|i-1} = X_{i-1|i-1}$$
$$W_{i|i-1} = W_{i-1|i-1} + \Sigma_i$$
$$W_{i|i}^{-1} = W_{i|i-1}^{-1} + \Lambda_i\,\lambda_i\,\Delta t$$
$$X_{i|i} = X_{i-1|i-1} + W_{i|i}\,\Gamma_i\,(y_i - \lambda_i\,\Delta t)$$

where $\lambda_i \triangleq e^{\alpha_0} \cdot e^{\Gamma_i^T X_{i-1|i-1}}$ □

Ex.: Let us assume that $X_{t+\tau}$ is the vector of kinematic variables (e.g., position and velocity variables) of a point in the Cartesian 3D space at time $t+\tau$, with $\tau > 0$ - fixed and known

(kinematics decoding)

Let us also suppose that the vector X_t is an explanatory variable for $M>1$ point processes (e.g., M spike trains from individual neurons that are collected simultaneously), each one with its own CIF. $\lambda_w(t|\mathcal{H}_t^w, X_{t+\tau}, \theta^w)$

$\theta^w \triangleq$ parameter vector for the w-th point process
$\mathcal{H}_t^w \triangleq$ history up to time t for the w-th point process

In this case, we have to replace the factor $P(Y_i|\mathcal{H}_i, \theta_i)$ in the original formula (*) with the joint probability distribution:

$$P\left(Y_{i,1}=y_{i,1}, Y_{i,2}=y_{i,2}, \ldots, Y_{i,M}=y_{i,M} \,\middle|\, \mathcal{H}_i^1, \mathcal{H}_i^2, \ldots, \mathcal{H}_i^M, X_{\hat{t}_i+\tau}, \theta^1, \theta^2, \ldots, \theta^M\right)$$

where $Y_{i,w}$ is the Bernoulli RV associated with the w-th point process in the interval $(t_{i-1}, t_i]$, $w=1,2,3,\ldots,M \Rightarrow$ Because dependences are encompassed in the histories $\mathcal{H}_t^w$, $w=1,2,3,\ldots,M$, we can write:

$$P\left(Y_{i,1}=y_{i,1}, Y_{i,2}=y_{i,2}, \ldots\right) = \prod_{w=1}^{M} P\left(Y_{i,w} \,\middle|\, \mathcal{H}_i^w, X_{\hat{t}_i+\tau}, \theta^w\right)$$

$$\cong \exp\left\{ \sum_{w=1}^{M} y_{i,w} \log\left(\lambda_w(\hat{t}_i|\mathcal{H}_i^w, X_{\hat{t}_i+\tau}, \theta^w)\Delta t\right) + \right.$$

$$\left. - \sum_{w=1}^{M} \lambda_w(\hat{t}_i|\mathcal{H}_i^w, X_{\hat{t}_i+\tau}, \theta^w)\Delta t \right\} \qquad (\nabla\Delta)$$

Also, if $X_{t+\tau}$ is a continuous process (e.g., representation of kinematic variables), it is usually common to assume:

$$X_{t+\tau} = \mu_X + F X_{t+\tau-1} + \eta_{t+\tau} \qquad (\nabla\Delta\nabla)$$

where μ_x provides the average position and velocity and F is time-invariant $\Rightarrow$ By replacing $(\nabla\Delta)$ in $(**)$ and using $(\nabla\Delta\tau)$ we have the multivariate version of the filter:

1') $\tilde{X}_{i|i-1} = \mu_x + F\tilde{X}_{i-1|i-1}$

2') $W_{i|i-1} = F W_{i-1|i-1} F^T + Q_i$

3') $\tilde{X}_{i|i} = \tilde{X}_{i|i-1} + W_{i|i} \sum_{w=1}^{M} \nabla_x (\log \lambda_w)\,(y_{i,w} - \lambda_w \Delta t)$

4') $W_{i|i}^{-1} = W_{i|i-1}^{-1} - \sum_{w=1}^{M} H_x \log \lambda_w \,(y_{i,w} - \lambda_w \Delta t) + \sum_{w=1}^{M} (\nabla_x \log \lambda_w)^T (\nabla_x \log \lambda_w)\,\lambda_w \Delta t$

where: $\tilde{X}_i \triangleq X_{\hat{t}_i+\tau} \Rightarrow \tilde{X}_{i|i-1} = X_{\hat{t}_i+\tau|\hat{t}_{i+\tau-1}}$

$\tilde{X}_{i|i} = X_{\hat{t}_i+\tau|\hat{t}_i+\tau}$

$\lambda_w \triangleq \lambda_w(\hat{t}_i \mid \mathcal{H}_i^w, \tilde{X}_{i|i-1}, \theta^w)$

$H_x \triangleq$ hessian matrix with respect to the components of $X_{t+\tau}$

$\nabla_x \triangleq$ gradient with respect to the components of $X_{t+\tau}$ □

* General Properties of the Adaptive Filter

Filters (1)-(4) and (1')-(4') and their variations share a few key properties:

- The original problem is formulated as:

$$\left.\begin{aligned} r_i &\sim P(\eta_i(\theta_i)) \\ \theta_i &\sim N(\mu_i, \Sigma_i) \end{aligned}\right\} \Rightarrow$$

The first RV is conditioned on the second RV, which is conditioned on parameters μ_i, Σ_i coming from the evolution model

where η_i is obtained from combining $\lambda(\cdot)$ and $g(\cdot)$

$\Rightarrow$ It is a Conditional HIERARCHICAL Model and it requires the calculation of $f_Y(\eta|\vartheta)$ ($\Rightarrow$ conditional pdf given by the CIF in our case) and $f_\Theta(\vartheta|\mu,\Sigma)$ ($\Rightarrow$ we implicitly used when we computed the prediction steps)

- The original problem has an observable RV (i.e., Y_i) and a latent RV (i.e., ϑ_i) which cannot be observed and whose value must be inferred from the value of the observable RV $\Rightarrow$ It is a LATENT VARIABLE Estimation problem

- The adaptive filter can be envisioned as a generalization of the Kalman filter to the Point Processes (but with important differences too):

	Kalman Filter	Adaptive Filter
Evolution Model	$\vartheta_i = F\vartheta_{i-1} + \eta_i$	$\vartheta_i = F_i\vartheta_{i-1} + \eta_i$
	$Y_i = B\vartheta_i + \varepsilon_i$	$Y_i \sim \mathcal{P}(\lambda(t_i\|\mathcal{H}_i,\vartheta_i)\Delta t)$
	$\eta_i \sim N(0,Q)$ $\varepsilon_i \sim N(0,R)$	$\eta_i \sim N(0,Q_i)$
	$\vartheta_0 \sim N(\mu_0, Q)$ $\forall i$	$\vartheta_0 \sim N(\mu_0, Q_0)$
Prediction Step	$\vartheta_{i\|i-1} = F\vartheta_{i-1\|i-1}$	$\vartheta_{i\|i-1} = F_i\vartheta_{i-1\|i-1}$
	$W_{i\|i-1} = FW_{i-1\|i-1}F^T + Q$	$W_{i\|i-1} = F_iW_{i-1\|i-1}F_i^T + Q_i$
Correction Step	$W_{i\|i} = W_{i\|i-1} + W_{i\|i-1}B^T \cdot \ldots$ $(BW_{i\|i-1}B^T+R)^{-1}BW_{i\|i-1}$	$W_{i\|i}^{-1} = W_{i\|i-1}^{-1} + I(\vartheta_{i\|i-1})$
	$\vartheta_{i\|i} = \vartheta_{i\|i-1} + W_{i\|i-1}B^T \cdot \ldots$ $(BW_{i\|i-1}B^T+R)^{-1}(Y_i - B\vartheta_{i\|i-1})$	$\vartheta_{i\|i} = \vartheta_{i\|i-1} + W_{i\|i}\nabla_\vartheta(\log\lambda)\cdot$ $(Y_i - \lambda\Delta t)$

↖ Error signals used to drive the estimation ↗

- The formulation of the filter is based on formula (*):

$$f_\theta(\vartheta_i|Y_i,\mathcal{H}_i) = \frac{P(Y_i|\mathcal{H}_i,\vartheta_i)}{P(Y_i|\mathcal{H}_i)} f_\theta(\vartheta_i|\mathcal{H}_i)$$

where $P(Y_i|\mathcal{H}_i,*)$ substitutes for $f_Y(Y_i|\mathcal{H}_i,*)$ given the discrete nature of Y_i $\Rightarrow$ Let us drop the dependency on $\mathcal{H}_i$ and note:

$f_\theta(\vartheta_i|Y_i) = f_{\theta|Y}(\vartheta_i|Y_i)$ - Posterior density of ϑ_i conditioned on Y_i

$f_Y(Y_i|\vartheta_i) = f_{Y|\theta}(Y_i|\vartheta_i)$ - Prediction density of the observable variable Y_i

$f_Y(Y_i) = \int f_{Y|\theta}(Y_i|\vartheta) f_\theta(\vartheta) d\vartheta$ - Marginal probability density of Y_i

$f_\theta(\vartheta_i) = \int f_{\theta|\theta}(\vartheta_i|\vartheta_{i-1}) f_\theta(\vartheta_{i-1}) d\vartheta_{i-1}$ - We expressed the a priori distribution of ϑ_i at each step in term of the a priori distribution at the previous step and the prediction density

$\Downarrow$

The original problem is a Bayesian Estimation problem where:
- likelihood function of the point process determines the prediction density of Y_i
- evolution model and prediction density of ϑ_i determine the a priori density of ϑ_i
- evolution model provides a formula for the posterior density of ϑ_i

- In general, the formula:

$$f_{\theta|Y}(\vartheta|Y) = \frac{f_{Y|\theta}(Y|\vartheta) f_\theta(\vartheta)}{\int f_{Y|\theta}(Y|\vartheta) f_\theta(\vartheta) d\vartheta}$$

can be used to estimate the posterior density regardless of whether $\vartheta \sim N(\mu, \Sigma)$ for any pair (μ, Σ) or not

For instance, we can have:

$f_\vartheta(\vartheta) = 1$ (Uniform)

$f_{Y|\vartheta}(Y|\vartheta) = \vartheta^Y (1-\vartheta)^{n-Y}$ (Binomial)

$$\Rightarrow f_{\vartheta|Y}(\vartheta|Y) = \frac{\vartheta^y (1-\vartheta)^{n-y}}{\int \vartheta^y (1-\vartheta)^{n-y} d\vartheta}$$

$$\Rightarrow f_{\vartheta|Y}(\vartheta|Y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}$$ - Beta distribution $B(\alpha,\beta)$

with: $\alpha \triangleq Y+1$ $\beta \triangleq n-Y+1$ $\Gamma(\omega) \triangleq \int_0^\infty x^{\omega-1} e^{-x} dx$

Because $U(0,1) = B(1,1)$ we have: $f_\vartheta(\vartheta) \sim B \Rightarrow f_{\vartheta|Y} \sim B$ - The a priori distribution is called a "conjugate" prior distribution. Because the same happens in case of Gaussian priors, we have that the adaptive filter is developed by exploiting conjugate priors. □

References:

Eden et al. (2004), Neural Comput., vol. 16, pp. 971-998 ⇒ A copy is on Husky CT

Truccolo et al. (2005), J. Neurophysiol., vol. 93, pp. 1074-89 ⇒ A copy is on Husky CT

Textbook: ch. 16 (sections. 16.1 - 16.1.1 - 16.1.2 - 16.1.3 - 16.2 - 16.2.1 - 16.2.4 - 16.2.5)