Estimation des variables latentes par des récursivités de Kalman généralisées

Latent variable estimation by generalized Kalman recursions

Sadeq A.Kadhim[1] Joseph Ngatchou-Wandji [2,3]

[1]Department of Statistics, Mustansiriyah University, Iraq. sadiq2061@gmail.com
[2] Institut Élie Cartan de Lorraine, Université de Lorraine, France.
[3] École des Hautes Études en Santé Publique, France. joseph.ngatchou-wandji@univ-lorraine.fr

RÉSUMÉ. Cet article étudie des modèles à espace d'état pour des observations longitudinales multi-catǵorielles et des états (latents) caractérisés par les modèles dits CHARN (Conditional Heteroskedastic AutoRegressive Nonlinear). Ces des derniers sont estimés via des récursivités de Kalman généralisées, basées sur des filtres particulaires et l'algorithme EM. Nos résultats généralisent les travaux existants. Ils sont illustrés par des simulations numériques et sont appliqués aux données de patientes opérées d'un cancer du sein.

ABSTRACT. This paper discusses state-space models with multi-categorical longitudinal observations and states characterized by the so-called Conditional Heteroskedastic AutoRegressive Nonlinear (CHARN) models. The latter are estimated via generalized Kalman recursions based on particle filters and EM algorithm. Our findings generalize the literature. They are illustrated by numerical simulations and applied to data from patients surged for breast cancer.

MOTS-CLÉS. Récursivités de Kalman généralisées, Modèles à espace d'état non-linéaires , Données multicatégorielles longitudinales , Variables latentes , Filtre particulaires, Algorithme EM

KEYWORDS. Generalized Kalman recursions, Generalized state space models, Multicategorical longitudinal data, Latent variables, Particle filters, EM algorthim

Introduction

State-space models usually link an observed time series to an unobserved time series by a two-equations system defining a relationship between the terms of the first series and those of the second series. The values of the unobserved series are the *states* and those of the observed series are the *observations*.

In this paper, we consider generalized state-space models associated with models that can be non-linear with noises possibly non-Gaussian. Based on multi-categorical and longitudinal observations, we estimate the states using generalized Kalman recursions, particle filters and EM algorithm. Our main purpose is to estimate, from answers to a questionnaire, some latent variables in fields as quality of life, economics, industry or many others of interest. The answers to the questionnaire are the observations and the latent variables are the states, which can be among others, the patient health, the business confidence, the morale of customers, the level of anxiety of machine or robot users in factories.

More precisely, let $X_i(t)$ be the latent variables generated by a person $i, (i = 1, \cdots, n)$, at time $t, (t = 1, \cdots, T)$. Let $Y_i(t)$ be the variable observed at the place of $X_i(t)$. We aim at estimating the $X_i(t)$'s based on the observation of the $Y_i(t)$'s. In the present work, the $Y_i(t)$'s are the individuals' responses to a questionnaire, while the $X_i(t)$'s represent unobserved numerical quantities owned by the individuals, and that one would like to estimate from the responses of the questionnaire. They are assumed here to be drawn from a very general class of nonlinear autoregressive models termed CHARN models by Härdle et al. (1988).

This work is mainly motivated by the desire to estimate the latent trait in quality of life. Bousseboua

and Mesbah (2010) gave a natural progression of the Rasch model to longitudinal data. But they studied a special case of a dichotomous response option for each question (yes-no, agree-disagree, etc.) They considered a Gaussian latent Markov process and a Gaussian latent auto-regressive AR(1) process. Bartolucci et al. (2014) proposed a model for longitudinal categorical data. The latent process was designed with different means and regression coefficients but with similar variance, by an AR(1) process. They discussed a class of longitudinal data models where a series of discrete latent variables following a Markov chain defines the non-observed individual interest characteristics. Fahrmeir and Wagenpfiel (1997) and Fahrmeir and Tutz (2013) studied non-linear time series or discrete longitudinal observations. They developed an inference method based on the posterior mode. They obtained effective smoothings using the working Kalman filtering and smoothing. Durbin and Koopman (2000) discussed non-Gaussian state-space models with non-Gaussian time series data from both classical and Bayesian approaches. They suggested an approach based entirely on sampling significance and antithetic variables. Czado and Song (2008) proposed a new class of state-space models from longitudinal data, in which the observation equation contains both deterministic and random linear predictors defined in an additive form. They developed an algorithm for the Markov Chain Monte Carlo (MCMC) to make statistical inferences for binary and binomial responses models. They illustrated the applicability of their model in both simulation studies and data examples. Dunsmuir and Scott (2015) created the **R** Package **glarma**. They considered the generalized state-space models for non-Gaussian time series (GLARMA) described in Brockwell and Davis (1996) and in Durbin and Koopman(2000).

Nearly all the above cited papers considered the latent variables from linear models. In the present work they are from possibly non-linear models, while the observations are from a new class of multi-categorical longitudinal multivariate processes. Indeed, we assume the data are from a longitudinal study, where the individuals participate to an interview. It is well known that interview in the quality of life studies aims at measuring the individuals' health at regular intervals. Typically, it includes filling out a questionnaire in which multiple choice questions are answered, the questionnaire being conceived to measure perceived health of individuals at the current moment.

This paper is organized as follows : Section 2 discusses the theoretical framework of generalized linear models. This includes the state-space equations and their parameters estimation. Section 3 discusses the posterior distribution, while Section 4 discusses estimating latent variables by the posterior mode via the generalized Kalman recursions. Section 5 presents the results of the simulation experiment done with various scenarios, and a medical application. Section 6 concludes our work.

1. Generalized state-space models

A generalized linear stat-space model consists in two equations : the observation equation and the state equation. In this section, we present those used in our study.

1.1. *The observation equation*

1. The conditional probability of $Y_{ik}(t)$ given $X_i(t)$ is a multinomial distribution. That is, for all $i = 1,\cdots,n,\ k = 1,\cdots,q,\ c_k \geq 1,\ t = 1,\cdots,T$, the conditional probability of $Y_{ik}(t)$ given $X_i(t)$ can be

written as follows

$$\mathrm{P}[Y_{ik}(t) = (y_{ik}^1(t), \cdots y_{ik}^{c_k}(t)) \mid X_i(t) = x_i(t)] = \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{y_{ik}^s(t)}, \quad [1]$$

where

$$\begin{aligned} \pi_{ik}^s(t) &= \frac{\exp[\eta_{ik}^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]},\ s < c_k \\ \pi_{ik}^{c_k}(t) &= \frac{1}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]} \end{aligned} \quad [2]$$

with $\sum_{s=1}^{c_k} y_{ik}^s(t) = 1$ and $\sum_{s=1}^{c_k} \pi_{ik}^s(t) = 1$. The link function $\eta_{ik}^s(t)$ is defined with the logit function as follows

$$\begin{aligned} \eta_{ik}^s(t) = \mathrm{logit}(\pi_{ik}^s(t)) &= \log\left[\frac{\pi_{ik}^s(t)}{\pi_{ik}^{c_k}(t)}\right] \\ &= \log\left[\frac{\pi_{ik}^s(t)}{1-\sum_{j=1}^{c_k-1}\pi_{ik}^j(t)}\right] = \mathbf{u}_i^\top(t)\beta_k^s + X_i(t), \end{aligned}$$

and

$$\eta_{ik}^{c_k}(t) = \log\left[\frac{\pi_{ik}^{c_k}(t)}{\pi_{ik}^{c_k}(t)}\right] = \log(1) = 0.$$

$\mathbf{u}_i(t) = (u_{i1}(t), \cdots, u_{ir}(t))^\top$ is the independent covariate r-dimensional vector. For $k = 1, \cdots, q$, the $\beta_k^s = (\beta_{k1}^s, \cdots, \beta_{kr}^s)^\top$'s are the vectors of unknown regression parameters.

2. The vectors $Y_{ik}(1), \cdots, Y_{iq}(T)$ are conditionally independent given the latent variable vectors $(X_i(1), \cdots, X_i(T)) = (x_i(1), \cdots, x_i(T))$:

$$\begin{aligned} \mathrm{P}[Y_{ik}(1) &= y_{ik}(1), \cdots, Y_{iq}(T) = y_{iq}(T) \mid X_i(1) = x_i(1), \cdots, X_i(T) = x_i(T)] \\ &= \prod_{t=0}^{T} \mathrm{P}[Y_{iq}(t) = y_{iq}(t) \mid X_i(t) = x_i(t)]. \end{aligned}$$

3. The vectors $Y_{i1}(t), \cdots, Y_{iq}(t)$ are conditionally independent given the latent variables $X_i(t)$:

$$\begin{aligned} \mathrm{P}[Y_{i1}(t) &= y_{i1}(t), \cdots, Y_{iq}(t) = y_{iq}(t) \mid X_i(t) = x_i(t)] \\ &= \prod_{k=1}^{q} \mathrm{P}[Y_{ik}(t) = y_{ik}(t) \mid X_i(t) = x_i(t)]. \end{aligned}$$

1.2. *The state equation*

The state equation is defined for any $i = 1, \ldots, n$ and $t = 0, \ldots, T$ by

$$X_i(t) = F[X_i(t-1), \mathbf{u}_i(t), \gamma] + H[X_i(t-1), \mathbf{u}_i(t), \delta]\varepsilon_i(t), \quad [3]$$

where γ, δ are the model parameters, $F(\cdot,\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^r \times \mathbb{R}^l \longrightarrow \mathbb{R}$ and $H(\cdot,\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^r \times \mathbb{R}^l \longrightarrow \mathbb{R}$ are nonlinear smooth functions. The sequence $(\varepsilon_i(t))$ denoting the noise process for the state process satisfies

$$E[\varepsilon_i(t)] = 0, \mathrm{Var}[\varepsilon_i(t)] = R_t > 0$$

with density function

$$expf(\nu_i(t), \phi_i(t); z) = \exp\left\{\frac{z\nu_i(t) - b[\nu_i(t)]}{\phi_i(t)} + c[z, \phi_i(t)]\right\} \quad [4]$$

in which $\nu_i(t)$ is a canonical parameter or the link function, $\phi_i(t)$ denotes the dispersion or the scale parameter, and $b[\nu_i(t)]$ and $c[z,\phi_i(t)]$ are functions which can take different forms. The latent process $(X_i(t) : 1 \leq t \leq T)$ is Markovian and the density p of the conditional distribution $X_i(t)$ given $X_i(t-1)$ satisfies, for any $x_i(t) \in \mathbb{R}$:

$$p(x_i(t) \mid X_i(t-1)) \sim expf(\nu_i(t), \phi_i(t); x_i(t)).$$

Note that

$$\mu_i(t) = E[X_i(t) \mid X_i(t-1), \mathbf{u}_i(t)] = F[X_i(t-1), \mathbf{u}_i(t), \gamma]$$
$$V_i(t) = \mathrm{Var}[X_i(t) \mid X_i(t-1), \mathbf{u}_i(t)] = H^2[X_i(t-1), \mathbf{u}_i(t), \delta] R_t.$$

Then the joint law g_i of the vectors $\mathbf{X}_i = (X_i(0), X_i(1), \cdots, X_i(T))^\top$ deduced easily by conditioning is given by

$$\begin{aligned} g_i(\mathbf{X}_i) &= \prod_{t=1}^{T} p(X_i(t) \mid X_i(t-1)) p(X_i(0)) \\ &= \prod_{t=0}^{T} expf(\nu_i(t), \phi_i(t); X_i(t)), \\ &= \exp\left\{ \sum_{t=0}^{T} \left[\frac{X_i(t)\nu_i(t) - b[\nu_i(t)]}{\phi_i(t)} + c_i[X_i(t), \phi_i(t)] \right] \right\}. \end{aligned} \quad [5]$$

1.3. *The marginal likelihood*

For any $i = 1, \ldots, n$ and $T \geq 1$, define the following vectors

$$\begin{aligned} \mathbf{Y}_i &= (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top \\ \mathbf{Y}_i(t) &= (Y_{i1}^\top(t) \cdots, Y_{iq}^\top(t))^\top, t = 0, 1, 2, \cdots, T, \end{aligned}$$

where $Y_{ik}(t) = (Y_{ik}^{(1)}(t), \cdots, Y_{ik}^{(c_k)}(t))^\top$, $k = 1, \cdots, q$, $c_l \geq 1$. We denote by $\theta = (\beta, \gamma, \delta)$, the vector of model parameters, where $\beta = (\beta_1^\top, \cdots, \beta_q^\top)^\top$, with the β_k's standing for $c_k \times r$ matrices, where c_k is the category of the item $k, k = 1, \cdots, q$ and r denotes the number of covariates.

Let $\mathbf{y}_i = (\mathbf{y}_i^\top(0), \mathbf{y}_i^\top(1), \cdots, \mathbf{y}_i^\top(T))^\top$ with $\mathbf{y}_i(t) = (y_{i1}^\top(t) \cdots, y_{iq}^\top(t))^\top$. The joint density function of the observations can be written as

$$\begin{aligned} & P(\mathbf{Y}_1^\top = \mathbf{y}_1^\top, \mathbf{Y}_2^\top = \mathbf{y}_2^\top, \cdots, \mathbf{Y}_n^\top = \mathbf{y}_n^\top) = \prod_{i=1}^{n} \int \cdots \int P(\mathbf{y}_i \mid \mathbf{x}_i; \theta) g_i(\mathbf{x}_i) d\mathbf{x}_i \\ = & \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} P(Y_{ik}(t) = y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t) \mid x_i(t)) g_i(\mathbf{x}_i) d\mathbf{x}_i \\ = & \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{y_{ik}^s(t)} g_i(\mathbf{x}_i) d\mathbf{x}_i. \\ = & \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \prod_{s=1}^{c_k} \left[\frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \right]^{y_{ik}^s(t)} g_i(\mathbf{x}_i) d\mathbf{x}_i, \end{aligned}$$

where $g_i(\mathbf{x}_i)$ is given by equation (5).

From the above equality, one easily obtains the following likelihood

$$p(\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \cdots, \mathbf{Y}_n^\top) = \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \prod_{s=1}^{c_k} \left[\frac{\exp[\mathbf{u}_i^\top(t)\beta_k^s + x_i(t))]}{1 + \sum_{j=1}^{c_k} \exp[\mathbf{u}_i^\top(t)\beta_k^j + x_i(t)]} \right]^{Y_{ik}^s(t)} \times g_i(\mathbf{x}_i) d\mathbf{x}_i, \quad [6]$$

1.4. *The EM algorithm*

In incomplete data situations one usually uses the EM algorithm for computing the maximum likelihood estimators of the parameters. For a general presentation of the EM algorithm see Dempster et al. (1977) and for its use in the current context, Moussedek and Mesbah (2010). The EM algorithm works as follows : if $\theta^{(0)}$ denotes an initial value for θ, for $m = 0, 1, \cdots$, the $(m+1)$-th iteration of the EM algorithm works as follows :

1.4.1. *Expectation -step :*

calculate the expectation $Q(\theta \mid \theta^{(m)})$:

$$\begin{aligned} Q(\theta \mid \theta^{(m)}) &= E\{\log[f(\mathbf{Y}, \mathbf{X}; \theta)] \mid \mathbf{Y}, \theta^{(m)}\} \\ &= \sum_{i=1}^{n} \int \cdots \int [\log\{g_i(\mathbf{x}_i, \theta_i)\} + \log\{p(\mathbf{Y}_i \mid \mathbf{x}_i)\}] \\ &\quad \times p(\mathbf{x}_i \mid \mathbf{Y}_i, \theta^{(m)}) d\mathbf{x}_i, \end{aligned}$$

where $f(\mathbf{Y}, \mathbf{X}; \theta)$ is the likelihood of θ $(\mathbf{Y}, \mathbf{X})$.

With this, the E-step for first-order Markov latent process from Exponential family distribution is :

$$Q(\theta \mid \theta^{(m)}) = G_1 + G_2 \quad [7]$$

where

$$\begin{aligned} G_1 &= \sum_{i=1}^{n} \int \cdots \int \sum_{t=0}^{T} \left[\frac{x_i(t)\nu_i(t) - b[\nu_i(t)]}{\phi_i(t)} + c[x_i(t), \phi_i(t)] \right] \\ G_2 &= \sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=0}^{T} \sum_{s=1}^{c_k} y_{ik}^s(t) \int \cdots \int \log[\pi_{ik}^s(t)] \times p(\mathbf{x}_i \mid \mathbf{Y}_i, \theta^{(m)}) d\mathbf{x}_i. \end{aligned} \quad [8]$$

and $p(\mathbf{x}_i \mid \mathbf{Y}_i, \theta^{(m)})$ is the conditional density of latent vector $\mathbf{X}_i$ given the observation vector $\mathbf{Y}_i$ (we further use the particle filtering algorithm to find it).

Maximizing step :

$$\theta^{(m+1)} = \arg\max Q(\theta \mid \theta^{(m)}).$$

1.5. *Estimation of first-order CHARN latent processes*

We recall that the parameter vector is $\theta = (\beta_1^\top, \cdots, \beta_q^\top, \gamma^\top, \delta^\top)^\top$ and for any $k = 1, \cdots, q$, $c_k \geq 1$, $\beta_k = (\beta_k^{1\top}, \cdots, \beta_k^{c_k\top})^\top$. We apply the E-M step for finding the MLE as follows :

Maximizing with respect to the $\beta_k^s, k = 1, \cdots, q, s = 1, \cdots, c_k$, only the part G_2 yields

$$\begin{aligned} \beta_k^{s(m+1)} &= \arg\max_{\beta_k^s} \sum_{i=1}^{n} \sum_{t=0}^{T} \int \cdots \int \mathbf{u}_i^\top(t) D_{ik}^\top(x_t) \Sigma_{ik}^{-1}(x_t) [Y_{ik}^s(t) - \pi_{ik}^s(t)] \\ &\quad \times p(\mathbf{x}_i \mid \mathbf{Y}_i, \theta^{(m)}) d\mathbf{x}_i, \end{aligned} \tag{9}$$

where $D_{ik}(X_t)$ is a matrix with generic elements

$$D_{ik}^s(X_t) = \pi_{ik}^s(t)[1 - \pi_{ik}^s(t)] \tag{10}$$

and $\Sigma_{ik}(X_t) = Cov(Y_{ik}(t))$ has generic elements $\sigma_{ik}^{sm}(t) = \pi_{ik}^s(t)[1 - \pi_{ik}^s(t)], s = m$
and $\sigma_{ik}^{sm}(t) = -\pi_{ik}^s(t)\pi_{ik}^m(t), s \neq m$.

Maximizing with respect to γ, we apply the chain rule to calculate the score function for γ

$$\frac{\partial l(\gamma)}{\partial \gamma} = \frac{\partial l(\gamma)}{\partial \nu_i(t)} \times \frac{\partial \nu_i(t)}{\partial \mu_i(t)} \times \frac{\partial \mu_i(t)}{\partial \gamma}.$$

One obtains :

$$\begin{aligned} \gamma^{(m+1)} &= \arg\max_{\gamma} \sum_{i=1}^{n} \int \cdots \int \sum_{t=0}^{T} \left[\frac{x_i(t) - b'[\nu_i(t)]}{\phi_i(t)} \right] \frac{\partial \nu_i(t)}{\partial \mu_i(t)} \frac{\partial \mu_i(t)}{\partial \gamma} \\ &\quad \times p(\mathbf{x}_i \mid \mathbf{Y}_i, \theta^{(m)}) d\mathbf{x}_i, \end{aligned} \tag{11}$$

where $b'[\nu_i(t)]$ denotes the first derivative of the function $b[\nu_i(t)]$ with respect to $\nu_i(t)$. The derivative of $\nu_i(t)$ respect to $\mu_i(t)$ depends on the link function of the distribution.

For maximizing with respect to δ, where $\phi_i(t)$ is function of $(X_i(t), \mathbf{u}_i(t), \delta)$, it results again from the chain rule that

$$\frac{\partial l(\delta)}{\partial \delta} = \frac{\partial l(\delta)}{\partial \phi_i(t)} \times \frac{\partial \phi_i(t)}{\partial \delta},$$

from which one has

$$\begin{aligned} \delta^{(m+1)} &= \arg\max_{\delta} \sum_{i=1}^{n} \int \cdots \int \sum_{t=0}^{T} \left[\frac{-\{x_i(t)\nu_i(t) - b[\nu_i(t)]\}}{\phi_i^2(t)} + \frac{\partial c_i[x_i(t), \phi_i(t)]}{\partial \phi_i(t)} \right] \frac{\partial \phi_i(t)}{\partial \delta} \\ &\quad \times p(\mathbf{x}_i \mid \mathbf{Y}_i, \theta^{(m)}) d\mathbf{x}_i. \end{aligned} \tag{12}$$

2. The Posterior distribution

The parameters estimation formulas need the posterior distribution $p(\mathbf{X}_i \mid \mathbf{Y}_i)$. This can be computed by a Bayesian approach. For the traditional linear-Gaussian state-space model, the posterior distribution

is Gaussian and the usual Kalman filter recursions are used for estimating the states. In a general non-Gaussian state-space context, the distribution of the state variable $X_i(t)$ is generally non-Gaussian. In this situation, particle filters approaches can be used to find approximation of the posterior distribution. This is what we do in this paper.

We compute the posterior distribution using the auxiliary iterated extended Kalman particle filter (AIEKPF) method, proposed by Yanhui Xi et al. (2015). We derive equations of posterior mode and posterior covariance of the residuals. The main idea of this algorithm is to use an Auxiliary Particle Filter (APF) technique to generate the importance density function by the Iterated Extended Kalman Particle Filter (IEKF). The general form of AIEKPF algorithm for the individual i is outlined in Algorithm 1. There, the symbol $\mathscr{M}$ refers to multinomial distribution, $\mathscr{N}$ denotes to normal distribution and the symbol $expf$ refers to exponential family distribution.

3. Posterior mode estimation

In this section, the penalized likelihood estimation approach for finding the posterior mode is presented. The following two techniques are used :

1. Gauss-Newton and Fisher-scoring Filtering and smoothing algorithms.
2. Working extended Kalman filter and smoother algorithms.

3.1. *Penalized likelihood estimation*

The posterior mode smoother is defined by $a \equiv \{a^\top(0 \mid T), a^\top(1 \mid T), \cdots, a^\top(T \mid T)\} \in R^m$, where $m = (T+1)n$. The posterior distribution of $\mathbf{X}$ computed by Bayes' theorem is given by

$$p(\mathbf{X} \mid \mathbf{Y}) = \frac{1}{p(\mathbf{Y})} \prod_{i=1}^{n} \prod_{k=1}^{q} \prod_{t=1}^{T} p(Y_{ik}(t) \mid X_i(t)) \times \prod_{i=1}^{n} \prod_{t=1}^{T} g_i(X_i(t)) \prod_{i=1}^{n} g_i(X_i(0)). \quad [14]$$

Since $p(\mathbf{Y})$ does not depend on $\mathbf{X}$, one has

$$p(\mathbf{X} \mid \mathbf{Y}) \propto \prod_{i=1}^{n} \prod_{k=1}^{q} \prod_{t=1}^{T} p(Y_{ik}(t) \mid X_i(t)) \times \prod_{i=1}^{n} \prod_{t=1}^{T} g_i(X_i(t)) \times \prod_{i=1}^{n} g_i(X_i(0)), \quad [15]$$

where for any $t = 0, \ldots, T$,

$$g_i(x_i(t)) = \exp\left\{ \frac{x_i(t)\upsilon_i(t) - b[\upsilon_i(t)]}{\phi_i(t)} + c[x_i(t), \phi_i(t)] \right\}.$$

Taking the logarithm of both sides of (15), the penalized log-likelihood function writes

$$PL(\mathbf{X}) := \sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=1}^{T} \{\log p[Y_{ik}(t) \mid X_i(t)]\} + \sum_{i=1}^{n} \sum_{t=1}^{T} \{\log g_i[X_i(t)]\} + \sum_{i=1}^{n} \log g_i[X_i(0)]. \quad [16]$$

Algorithm 1 : AIEKPF algorithm for finding the posterior distribution

1. Initial step ($t=0$) : generate states (particles) $x_i^m(0)$ via the prior $p(x_i(0)) \sim expf(\nu_i^m(0), \phi_i^m(0); z)$.
2. generate $x_i^m(t) \sim expf(\nu_i^m(t), \phi_i^m(t); z)$.
3. upgrade the particles via the IEKF algorithm
   (a) Calculate
   $$A_i^m(t) = \frac{\partial F(x, \mathbf{u}_i(t), \gamma)}{\partial x} \mid_{x=x_i^m(t-1|t-1)}, \quad C_i^m(t) = H(x_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \delta)$$
   (b) Predict the particle with the IEKF :
   $$\begin{aligned} X_i^m(t \mid t-1) &\approx F(x_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \gamma) \\ P_i^m(t \mid t-1) &= A_i^m(t) P_i^m(t-1 \mid t-1) A_i^{\top m}(t) + C_i^m(t) R C_i^{\top m}(t) \end{aligned}$$
   (c) For $j = 1, \cdots c$
      i. Calculate
      $$B_{ij}^m(t) = \frac{\partial \pi_{it}}{\partial x}(\mathbf{u}_i(t), x) \mid_{x=x_{ij}^m(t|t-1)}$$
      ii. Update the error covariance matrix :
      $$\begin{aligned} P_{ij}(t \mid t) &= (I - K_{ij}(t) B_{ij}(t) P_{ij}(t \mid t-1) \\ K_{ij}(t) &= P_{ij}(t \mid t-1) B_{ij}^{\top}(t) [B_{ij}(t) P_{ij}(t \mid t-1) B_{ij}(t) + \Sigma_i^{-1}(t)]^{-1} \end{aligned}$$
      iii. Update the state estimate : $X_{ij}(t \mid t) = X_{ij}(t \mid t-1) + K_{ij}(t)[Y_i(t) - \widehat{\pi}_i(t)]$
4. For $m = 1, \cdots N$, calculate $w_i^m(t) = q(m \mid \mathbf{Y}_i(t)) \propto \prod_{k=1}^{q} \mathscr{M}[\pi_{ik}(\mathbf{u}_i^m(t), \mu_i^m(t))] w_i^m(t-1)$, and normalize the weights
5. Re-sample to get the indicator ς_i^m of particle m's parent.
6. Generate the importance sampling : for $m = 1, \cdots, N$,
   (a) Create samples $X_i^m(t) \sim q(X_i(t), \varsigma_i^m \mid \mathbf{Y}_i(t)) = \mathscr{N}(\widehat{X}_{ij}^{\varsigma_i^m}(t), P_{ij}^{\varsigma_i^m}(t))$,
   (b) Calculate importance weights of particles by using
   $$w_i^m(t) = \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} = \prod_{k=1}^{q} \frac{\mathscr{M}[\pi_{ik}(x_i^m(t), \mathbf{u}_i^m(t)]}{\mathscr{M}[\pi_{ik}(\mu_i^{\varsigma_i^m}(t), \mathbf{u}_i^m(t)]} \qquad [13]$$
   (c) Normalize the weights $w_i^m(t) = \frac{w_i^m(t)}{\sum_{m=1}^{N} w_i^m(t)}$.
7. Output : sequences of weighted particles (samples) $[\{X_i^m(t), w_i^m(t)\}_{m=1}^N], i = 1, \cdots, n.]$

Since

$$\log p(Y_{ik}(t) \mid X_i(t)) = \log \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{Y_{ik}^s(t)} = \sum_{s=1}^{c_k} Y_{ik}^s(t) \log \pi_{ik}^s(t),$$

the penalized log-likelihood function becomes

$$PL(\mathbf{X}) = \sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=1}^{T} \sum_{s=1}^{c_k} \{Y_{ik}^s(t) \log \pi_{ik}^s(t)\} + G_1 + G_2, \quad [17]$$

where

$$G_1 = \sum_{i=1}^{n} \left\{ \frac{X_i(0)\upsilon_i(0) - b[\upsilon_i(0)]}{\phi_i(0)} + c[X_i(0), \phi_i(0)] \right\}$$
$$G_2 = \sum_{i=1}^{n} \sum_{t=0}^{T} \left\{ \frac{X_i(t)\upsilon_i(t) - b[\upsilon_i(t)]}{\phi_i(t)} + c[X_i(t), \phi_i(t)] \right\}. \quad [18]$$

Several methods can achieve numerical maximization of the penalized log-likelihood. We make use of Gauss-Newton (Fisher-scoring) algorithm and Working Extended Kalman Filter and Smoother (WEKFS) algorithm.

3.2. *Gauss-Newton iteration and Fisher-scoring Filtering*

In a compact matrix notation, the penalized log-likelihood criterion (16) can be described as :

$$PL(\mathbf{X}) = l_1(\mathbf{X}) - l_2(\mathbf{X}), \quad [19]$$

where

$$l_1(\mathbf{X}) = \sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=0}^{T} \sum_{s=1}^{c_k} Y_{ik}^s(t) \log(\pi_{ik}^s(t))$$

and

$$l_2(\mathbf{X}) = \mathbf{X}^\top A\upsilon - \mathbf{1}^\top Ab(\upsilon) + c(\mathbf{X}, \phi)$$

with

- $\mathbf{X}$ is a matrix of latent variables with size $(n \times T)$.
- $A = \text{diag}(1/\phi)$ with size $(n \times n)$.
- υ is the $(n \times T)$ matrix of link function.
- $\mathbf{1}$ is a $(n \times T)$ matrix of ones.
- $b(\upsilon), c(X, \phi)$ are a $(n \times T)$ matrices taking a different forms depending on the distribution of the $X_i(t)$'s.

We define the tables of expectations by $\Pi_i(\mathbf{X}) = (\pi_{i0}^\top(X_0), \pi_{i1}^\top(X_1), \cdots, \pi_{iT}^\top(X_T))^\top$. Fahrmeir and Wagenpfeil (1997) assumed $\pi_i^\top(0) = X_i(0)$ and $\mathbf{Y}_i^\top(0) = a_i(0)$. We recall that the conditional mean and variance of each individual are given by

$$E(Y_{ik}(t) \mid X_i(t)) \quad = \quad \pi_{ik}^s(t),\ s = 1, \cdots, c_k$$

$$\text{Var}(Y_{ik}(t) \mid X_i(t)) \;=\; \Sigma_{ik}(X_t),$$

where $\Sigma_{ik}(X_t)$ has generic elements

$$\sigma_{ik}^{sm}(t) = \begin{cases} \pi_{ik}^{s}(t)[1-\pi_{ik}^{s}(t)], & \text{if } s = m \\ -\pi_{ik}^{s}(t)\pi_{ik}^{m}(t) & \text{if } s \neq m. \end{cases} \quad [20]$$

The diagonal covariance matrix of an individual i at time t is

$$\Sigma_i(\mathbf{X}) = \text{diag}(V_i(0), \Sigma_{i1}(X_1), \cdots, \Sigma_{iT}(X_T)),$$

where $V_i(0) = H^2[X_i(0), \mathbf{u}_i(0), \delta]R_0$.

Define the diagonal matrix

$$D_i(\mathbf{X}) = \text{diag}(1, D_{i1}(X_1), \cdots, D_{iT}(X_T)),$$

where for any $i = 1, \cdots, n$, and $t = 0, \ldots, T$, $D_{it}(X_t)$ stands for the first-order derivative of the conditional probability $\pi_i(t)$ evaluated at $\eta_i(t)$. The score function of $l(\mathbf{X})$ in (19) is given for any $i = 1, \cdots, n$, by $S_i(\mathbf{X}) = (\widehat{S}_{i0}(X_0), \widehat{S}_{i1}(X_1), \cdots, \widehat{S}_{iT}(X_T))^\top$ where

$$S_i(\mathbf{X}) := D_i(\mathbf{X})\Sigma_i^{-1}(\mathbf{X})\left\{\mathbf{Y}_i(t) - \Pi_i(\mathbf{X})\right\}, \quad [21]$$

with components

$$\widehat{S}_i(X_0) \;=\; V_i^{-1}(0)(a_i(0) - X_i(0)) \quad [22]$$

$$\widehat{S}_{it}(X_t) \;=\; D_{it}(X_t)\Sigma_{it}^{-1}(X_t)\left\{\mathbf{Y}_i(t) - \pi_{it}(X_t)\right\}, t = 0, \cdots, T, \quad [23]$$

The first-order derivatives of $PL(\mathbf{X})$ in (19) are

$$M(\mathbf{X}) = \frac{\partial PL}{\partial \mathbf{X}}(\mathbf{X}) = S(\mathbf{X}) - S(\upsilon), \quad [24]$$

where

$$S(\upsilon) = X^\top A - \mathbf{1}Ab''(\upsilon).$$

The expected information matrix is given by $\mathscr{I}_i(\mathbf{X}) = (\mathscr{I}_{i0}(X_0), \mathscr{I}_{i1}(X_1), \cdots, \mathscr{I}_{iT}(X_T))$, where for all $i = 1, \cdots, n$,

$$\mathscr{I}_i(\mathbf{X}) = D_i(\mathbf{X})\Sigma_i^{-1}(\mathbf{X})D_i^\top(\mathbf{X}) \quad [25]$$

with diagonal blocks

$$\mathscr{I}_{i0}(X_0) \;=\; V_i^{-1}(0) \quad [26]$$

$$\mathscr{I}_{it}(X_t) \;=\; D_{it}(X_t)\Sigma_{it}^{-1}(X_t)D_{it}^\top(X_t), t = 0, \cdots, T. \quad [27]$$

The Taylor expansion of the score function around $\mathbf{X}^0$ yields

$$M(\mathbf{X}^1) \approx M(\mathbf{X}^0) - \mathscr{I}(\mathbf{X}^0) \times \left(\mathbf{X}^1 - \mathbf{X}^0\right).$$

Since $M(\mathbf{X}^1) = 0$, a single Fisher scoring to the next iterate $\mathbf{X}^1 \in R^m$, with $m = (T+1)n$ can be obtained from the following equation

$$\left(\mathscr{I}(\mathbf{X}^0) + \mathscr{J}(v^0)\right)\left(\mathbf{X}^1 - \mathbf{X}^0\right) = M(\mathbf{X}^0).$$

This can be rewritten as

$$\mathbf{X}^1 = \left(\mathscr{I}(\mathbf{X}^0) + \mathscr{J}(v^0)\right)^{-1} \mathscr{I}(\mathbf{X}^0)\widetilde{\mathbf{Y}}, \qquad [28]$$

where

$$\mathscr{J}(\upsilon^0) = -\mathbf{1}Ab''(\upsilon^0),$$

with "working" observation $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1^\top, \cdots, \widetilde{Y}_n^\top)^\top$ and $\widetilde{Y}_i = (\widetilde{Y}_i^\top(0), \widetilde{Y}_i^\top(1), \cdots, \widetilde{Y}_i^\top(T))^\top$, which can compute as

$$\widetilde{Y}_i := \left[D_i^{-1}(\mathbf{X})\right]^\top [\mathbf{Y}_i - \Pi_i(\mathbf{X})] + \eta_i(\mathbf{X}), \qquad [29]$$

with components

$$\begin{aligned} \widetilde{Y}_i(0) &= a_i(0) \\ \widetilde{Y}_i(t) &= \left[D_{it}^{-1}(X_t)\right]^\top [\mathbf{Y}_i(t) - \pi_{it}(X_t)] + \eta_{it}(X_t),\ t = 0, \cdots, T, \end{aligned}$$

where $\eta_i(\mathbf{X}) = (\eta_{i1}(X_1), \cdots, \eta_{iT}(X_T))^\top$ is the vector of link function for the ith individual.

3.3. *Working Extended Kalman Filter and Smoother (WEKFS)*

The collection of posterior mode estimated values of $\mathbf{X}$ (predicted, filtered and smoothed) in this algorithm are respectively denoted by $a_{t|t-1}, a_{t|t}$ and $a_{t|T}$, and the corresponding estimated values of the error covariance matrices to the collection of posterior mode (predicted, filtered, smoothed) are respectively denoted by $P_{t|t-1}, P_{t|t}$ and $P_{t|T}$.

Initialization :

$$\begin{aligned} a_i(0 \mid 0) &= a_i(0), \\ V_i(0 \mid 0) &= V_i(0). \end{aligned} \qquad [30]$$

Prediction

For $t = 0, \cdots, T$

$$\begin{aligned} a_i(t \mid t-1) &= F(a_i(t-1 \mid t-1), \mathbf{u}_i(t), \gamma) \\ P_i(t \mid t-1) &= A_i(t)P_i(t-1 \mid t-1)A_i^\top(t) + C_i(t)R_tC_i^\top(t). \end{aligned} \qquad [31]$$

Filtering

For $t = 0, \cdots, T$,

$$a_i(t \mid t) = a_i(t) + K_i(t)(\widetilde{Y}_i(t) - a_i(t \mid t-1))$$

$$
\begin{aligned}
K_i(t) &= P_i(t-1 \mid t-1)B_i^{\top}(t)(B_i(t)P_i(t-1 \mid t-1)B_i^{\top}(t) + \mathscr{I}^{-1}(t))^{-1} \\
P_i(t \mid t) &= (I - K_i(t)B_i(t))P_i(t-1 \mid t-1), \qquad [32]
\end{aligned}
$$

where

$$
A_i(t) = \frac{\partial F}{\partial x}(x, \mathbf{u}_i(t), \gamma) \mid_{x=a_i(t-1|t-1)}
$$

$$
B_i(t) = \frac{\partial \pi_i}{\partial x}(x) \mid_{x=a_i(t|t-1)} .
$$

4. Practical considerations

Our objective in this section is to estimate the latent variables by posterior mode via the working extended Kalman filtering recursions. Two situations are studied. The first is done by simulation where longitudinal multi-categorical data are generated based on latent variables from CHARN models with noises from exponential families distributions. The second situation is an application to a real data set from patients surged for breast cancer. In both cases, R-codes for the methods described in this paper has been written. The numerical results obtained with are presented and discussed.

4.1. *Simulation experiments*

In this part, we produce data from the observation equation described by a multinomial distribution defined in equation (1), and the state equation described by a CHARN model defined in equation (3) with standard Gaussian noise and a standardized exponential distributed noise with parameter 1. There are two scenarios. The first aims at testing the efficiency of the working extended Kalman filter recursions. Here, the parameters of the models are assumed to be known. The second scenario uses the EM algorithm for estimating the parameters of the model, before applying the working extended Kalman filter recursions.

4.2. *Simulation experiments I*

In order to investigate the efficiency of the working extended Kalman filter recursions (WEKF), we consider that an individual fills out a questionnaire constituted of multiple choice questions administered at t occasions. The outline of simulation experiments I is as follows

- **Outline of the simulations**

Algorithm 2 below is designed for **one individual**. So we omit the subscript i in the equations.

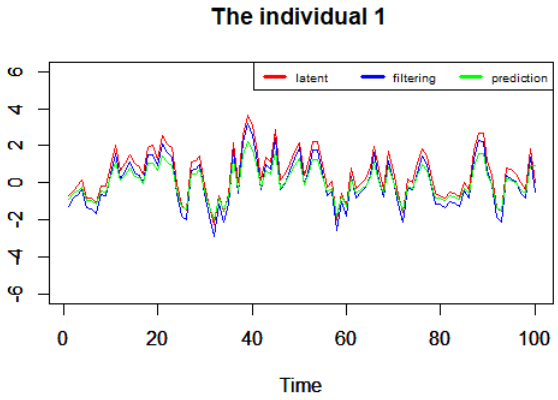
Figure 1 shows the graphs for one individual and for different types of models. There, the red color refers to the latent variable, while the blue refers to the posterior mode via filtering step, and the green color refers to the posterior mode via the prediction step.

It is clear that the results are very good. So, the working Kalman filter recursion succeeds in producing the posterior mode with different types of state-space models. Moreover, the values via two steps are equal to the actual value of the latent variables. There is no curve for the posterior mode via the predictive step in **(e)-(f)** of Figure 1 because the prediction step does not exist for the model considered. Indeed, it depends on the function F of the state equation, which does not exist for the CHARN(0,1) model studied.

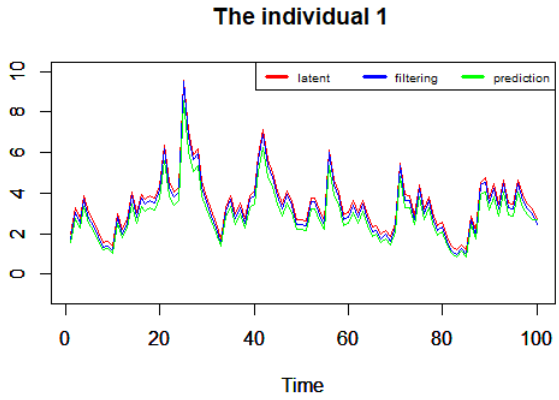
4.3. *Simulation experiments II*

In this scenario, the longitudinal multi-categorical data are generated and considered as given. More precisely, after we generate these data, they are considered as observed, while the state variables and the parameters of models are considered as unknown. We then proceed to their estimation by our results. In Algorithm 3, we outline how we estimate the state (latent) variables.

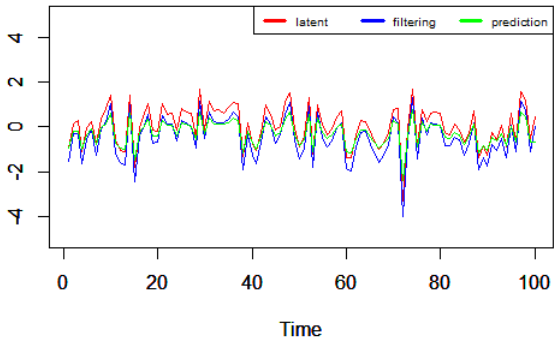
- **Outline of the simulations**



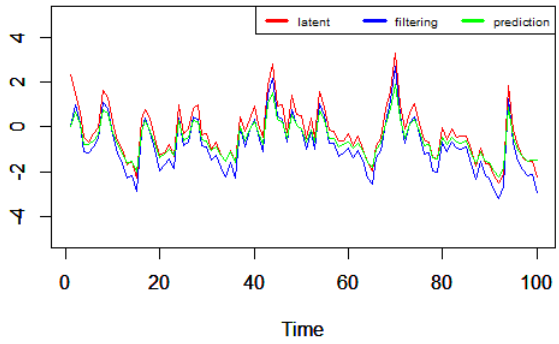
(a)



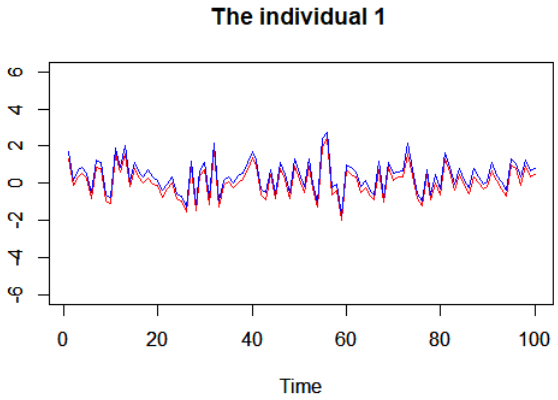
(b)



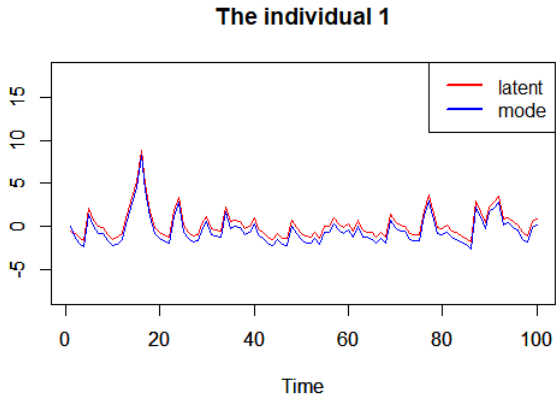
(c)



(d)



(e)



(f)

Figure 1. – AR(1) with **(a)** standard Gaussian and **(b)** standardized exponential noises ; CHARN(1,1) with **(c)** standard Gaussian and **(d)** standardized exponential noises ; CHARN(0,1) with **(e)** standard Gaussian and **(f)** standardized exponential noises.

Algorithm 2 : estimating latent variables in simulation experiment I

1. Generate the multi-categorical longitudinal data as follows :
   (a) Suppose one individual in the longitudinal study to whom is asked a set of $(q = 5)$ items ; for each item k, $c_k = 6$ categories are administered at t occasions.
   (b) Draw the samples of latent variables $X(t)$ by a state equation (CHARN model) with state noise from exponential families distributions (Gaussian or exponential).
   (c) For this individual produce 2 covariates $\mathbf{u}^\top(t)$: Age $u_1(t)$ and Sex $u_2(t)$, where $u_1(t) \sim \mathcal{N}(\mu_{u_1}, \sigma^2_{u_1})$ and $u_2(t) \sim Bin(n, p)$.
   (d) For each item k, assume the β_k^s's are known, and calculate the probabilities

$$\pi_k^s(t) = \frac{\exp[\eta_k^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_k^j(t)]},$$

   where

$$\eta_k^s(t) = \mathbf{u}^\top(t)\beta_k^s + X(t).$$

   (e) For each item k, and each category s at t occasions, generate the responses $Y_k(t) \sim \mathcal{M}(\pi_k^s(t))$.
2. Use the probabilities calculated at step (1-d) to calculate the adjusted observations by using (29).
3. Set values for γ, δ and the variance-covariance matrix of the state noise R_t.
4. Apply the Working Extended Kalman Filtering Recursions (WEKF) to calculate the posterior mode $a(t)$.

The latent variables are used to calculate the observations probabilities $\pi_{ik}(t)$ for generating the individuals' responses $Y_{ik}(t)$. Next, following the steps of Algorithm 3, the latent variables and their estimates are calculated via the working extended Kalman filtering, and compared.

• **Numerical computation**

The EM algorithm has the property of increasing the likelihood at each step, with a low convergence rate. As an alternative to convergence to a local maximum, we use Fisher scoring iteration method with an initial estimate $\hat{\theta}^{(0)}$.

Fisher scoring iteration method is given by

$$\widehat{\theta}^{(m+1)} = \widehat{\theta}^{(m)} + \mathcal{I}(\widehat{\theta}^{(m)})s(\widehat{\theta}^{(m)}), m = 0, 1, \cdots \quad [33]$$

where $\theta = (\beta^\top, \gamma^\top, \delta^\top)^\top$ and $s(\widehat{\theta}^{(m)})$ the Fisher scoring of the parameters and $\mathcal{I}(\widehat{\theta}^{(m)})$ the Fisher information matrix of the parameters. As in glm-R package the convergence occurs if

$$\frac{dev - dev_{old}}{(0.1 + \mid dev \mid)} \leq \varepsilon.$$

where, $dev = -2\log(L)$. We have taken $\varepsilon = 0.001$.

• **A state AR(1) model**

The simulation experiments II is performed with latent variables from an AR(1) model :

$$X_i(t) = \rho X_i(t-1) + \varepsilon_i(t), \quad [34]$$

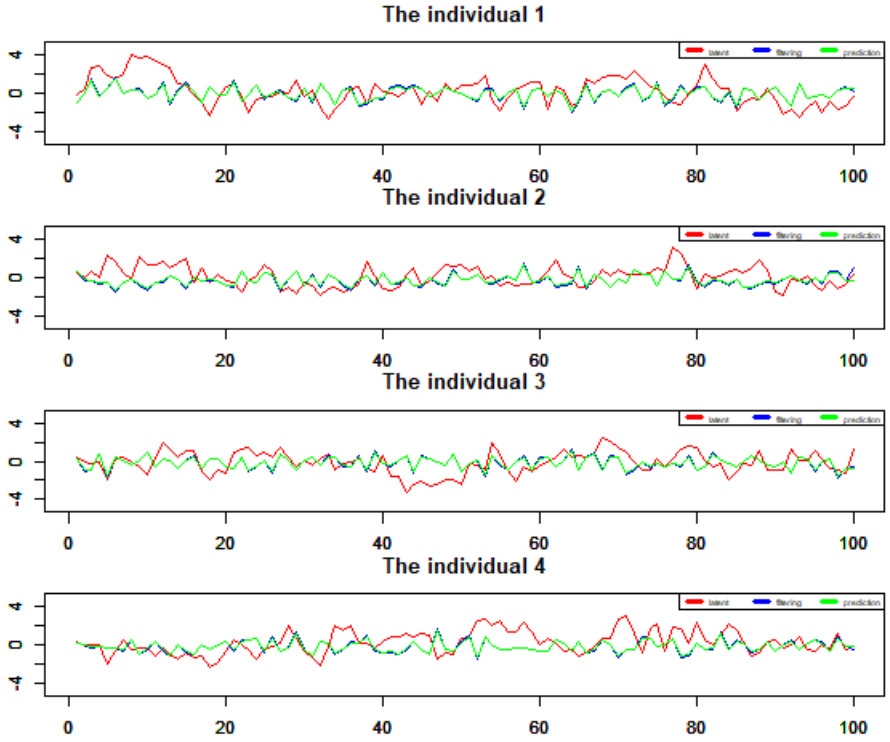


Figure 2. – AR(1) with standard Gaussian noise for $n = 4$ and $T = 100$.

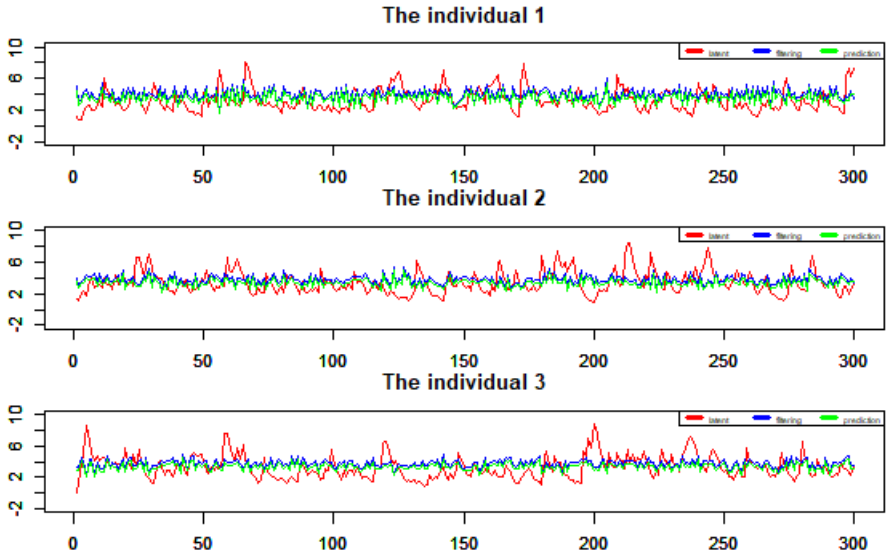


Figure 3. – AR(1) with standardized exponential noise for $n = 3$ and $T = 300$.

Algorithm 3 : estimating latent variables in simulation experiments II

1. Generate the multi-categorical longitudinal data as follows :
   (a) Suppose a sample of n individuals in a longitudinal study to whom is asked a set of $q = 5$ items, for each item k, $c_k = 6$ categories are administered at t occasions.
   (b) Draw the latent variables $X_i(t)$ by a state equation (CHARN model) with state noise from exponential families distributions (Gaussian or exponential).
   (c) For each individual i, generate 2 covariates $\mathbf{u}_i'(t)$: Age $u_1(t)$ and Sex $u_2(t)$.
   (d) For each individual i, item k and category s, set values to β_k^s and calculate the probabilities

$$\pi_{ik}^s(t) = \frac{\exp[\eta_{ik}^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]}$$

   where

$$\eta_{ik}^s(t) = \mathbf{u}_i^\top(t)\beta_k^s + X_i(t).$$

   (e) For each individual i, item k, and category s, at t occasions, generate the responses $Y_{ik}(t) \sim \mathscr{M}(\pi_{ik}^s(t))$.
2. Recall $\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top$ and $\mathbf{X}_i = (\mathbf{X}_i^\top(0), \mathbf{X}_i^\top(1), \cdots, \mathbf{X}_i^\top(T))^\top$. Calculate the posterior distribution $p(\mathbf{X}_i \mid \mathbf{Y}_i)$ via the Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF) algorithm.
3. Set iteration $m = 0$, and apply the classical Kalman Filtering Recursions to calculate the initial value $a_i^0(t)$ of posterior mode.
4. Starting with $a_i^0(t)$, compute the model's parameters $\beta^{m+1}, \gamma^{m+1}, \delta^{m+1}$ via EM algorithm.
5. Perform the Working Extended Kalman Filtering Recursions (WEKF) to calculate the posterior mode $a_i^{m+1}(t)$. If $\mid a_i^{m+1}(t) - a_i^m(t) \mid < 0.001$, STOP, either set $m = m+1$,and proceed to step 4.

We recall that in this model and in the subsequent ones, the noise is either standard Gaussian ($\varepsilon_i(t) \sim \mathscr{N}(0,1)$), or standardized exponential ($\varepsilon_i(t) = (e-1)$ with $e \sim \mathscr{E}(1)$.) The graphs on the figure 2 show the latent variables and their estimates (posterior mode) obtained by the Working Extended Kalman Filter (WEKF). The red color refers to the latent variables, while the blue refers to the posterior mode obtained from filtering step and the green color refers to the posterior mode obtained from the prediction step. As can be seen, the filtering and the prediction match but are sometimes not always close to the latent variable that they estimate. This is probably due to the initialization of the algorithms that are used.

- **A state CHARN(1,1) model**

The state equation considered is

$$X_i(t) = \rho_1 X_i(t-1) + \sqrt{\rho_1 + \rho_2 X_i^2(t-1)}\varepsilon_i(t), \qquad [35]$$

The graphs on Figure 4 display better results than those on Figure 3. As in the previous cases the lack of accuracy may result from the initialization of the algorithms used.

- **A state CHARN(0,1) model**

Here, the state equation is

$$X_i(t) = \sqrt{\rho_1 + \rho_2 X_i^2(t-1)}\varepsilon_i(t). \qquad [36]$$

The graphs in Figure 6 shows more accurate estimators than those in Figure 7. Here, the posterior mode

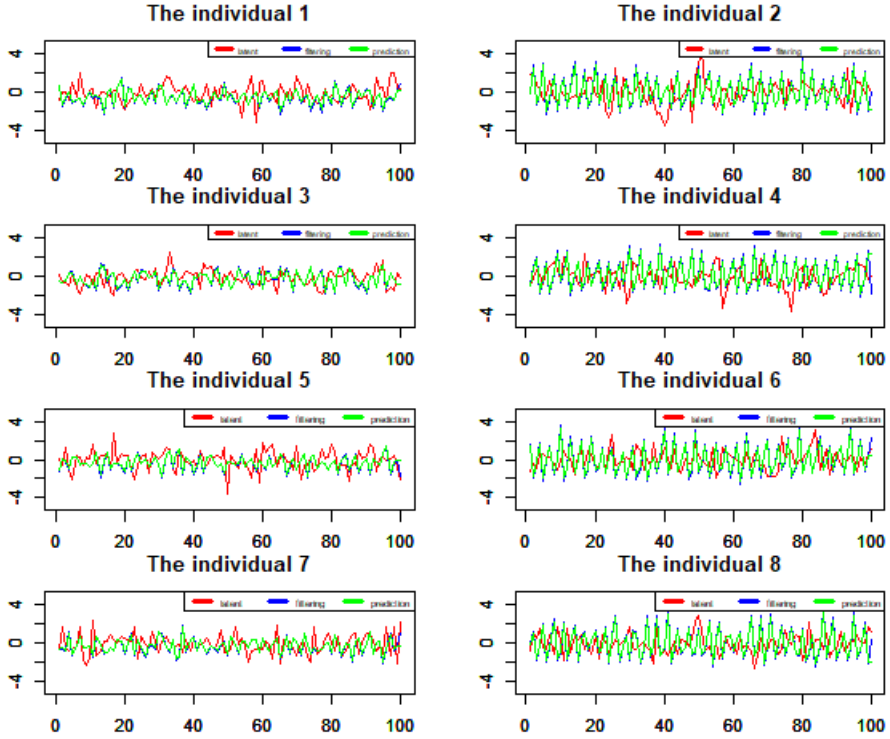


Figure 4. – AR(1) with standard Gaussian noise for $n = 4$ and $T = 100$.

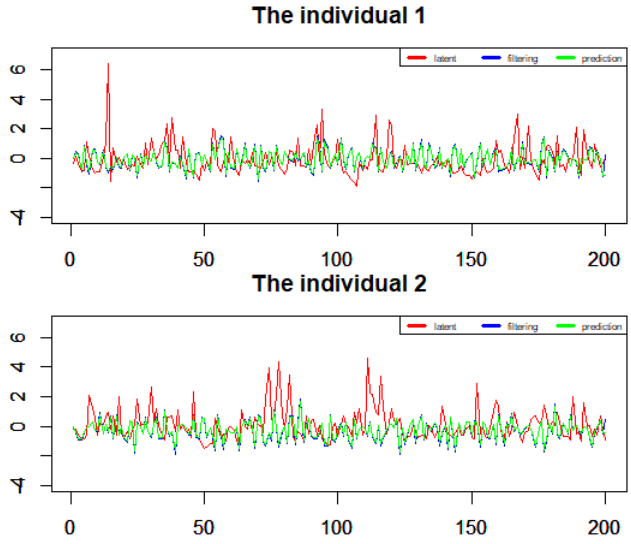


Figure 5. – CHARN(1,1) with standardized exponential noise for $n = 2$ and $T = 200$.

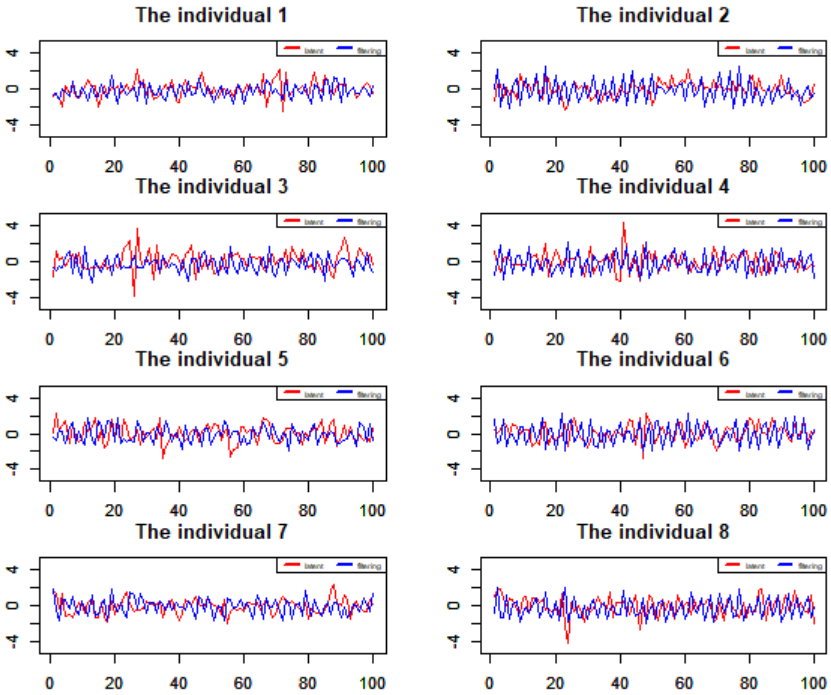


Figure 6. – CHARN(0,1) with standardized exponential noise for $n = 10$ and $T = 100$.

in predictive step is nil. This is due to the fact that we assume a nil function F in the state equation. Hence the posterior mode in WEKF is $a_{t|t-1} = F(a_{t|t}, \mathbf{u}_i(t), \gamma) = 0$.

4.4. *Application to real data*

Rotonda et al. (2011) presented a study on quality of life. They investigated factors correlated with cancer-related fatigue for women surged for breast cancer. In this study, 502 patients were recruited from September 2008 to September 2010. Three French cancer centers in eastern France received the patients : the Alexis Vautrin anti-cancer centre of Lorraine, the Georges-François Leclerc anti-cancer centre of Burgundy and the Paul Strauss anti-cancer centre of Alsaca. The patients filled a questionnaire several times. This was completed at their clinic visits, or a postage-paid envelope was issued to return them. The questionnaires considered personality traits completed before the surgery. The Life Orientation Test (LOT) questionnaire and the trait section of the State-Trait Anxiety Inventory (STAI-B) instrument were used. The patient responses to each item are classified into 4 categories (almost never, sometimes, often, and almost always). Here, the latent variable is the patient fatigue after surgery. This variable is assumed to be quantitative and varying over time around a mean value assumed to be nil in our work. Ten covariates are determined for the study : age, marital status, family situation, number of children and their ages, education, employment status, the chemotherapy group, the step of treatment, and the distance between patient's home and hospital are collected at the baseline assessment.

4.4.1. *Data analysis*

We obtained data from the above mentioned centres. Over the 502 patients, 435 had complete information and the others were not recorded. Each of them filled the twenty items questionnaire at 10 instants. For the analysis, we selected 3 covariates : the marital status, the chemotherapy group and the step of treatment. These are said by the experts to be related to the fatigue of the patient which (taken as our latent variable in this study). The covariates are scored as follows :

- **marital status** : 1= "single" ; 2= "cohabitation" ; 3= "bride" ; 4="widow" ; 5="divorced" ; 6= "bride/cohabitation".
- **treatment step** : 1= "step I" ; 2= "step II" ; 3= "step III".
- **chemotherapy group** : 1= " the group without chemotherapy " 2= "the group with chemotherapy"

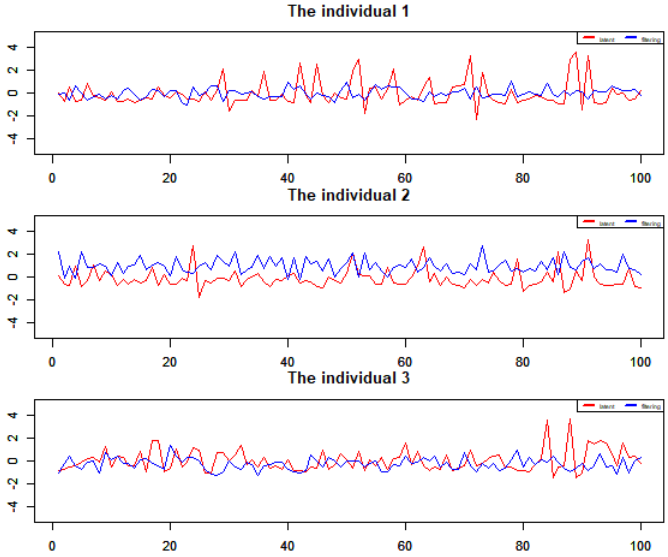


Figure 7. – CHARN(0,1) with standardized exponential noise for $n = 3$ and $T = 100$.

The latent variable of interest, the fatigue of the patient, is assumed to follow an AR(1) (state) model, and is estimated by the posterior mode by using the following algorithm.

Algorithm 4 : estimating the fatigue of the patients

1. Read the data from the Excel file.
2. Compute the posterior distribution $p(\mathbf{X} \mid \mathbf{Y})$ via the Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF) algorithm.
3. Set iteration $m = 0$, apply the classical Kalman Filtering Recursions to calculate the initial value $a_i^0(t)$ to posterior mode.
4. Starting with $a_i^0(t)$, calculate the model's parameters $\beta^{m+1}, \gamma^{m+1}, \rho^{m+1}$ via EM algorithm. Set the initial values for R_t^0, ρ^0 and $\beta_k^{s(0)}$
5. Implement the Working Extended Kalman Filtering Recursions (WEKF) to compute the posterior mode $a_i^{m+1}(t)$. If $\mid a_i^{m+1}(t) - a_i^m(t) \mid < 0.001$, STOP, else set $m = m+1$ and go to step 3.

Figure 8 shows the chronograms of prediction and filter steps for the 10th and 300th individuals. As can be seen, the 10th individual is tired at $t = 0$ and feels rested at $T = 2,3,4$. He is tired at $T = 5$ but feels rested at $T = 6,7$ and tired again at $T = 8$ and, once more, feels rested at $T = 9,10$.

For the 300th individual, he feels rested at $t = 0$, tired at $T = 2$, rested at $T = 3,4,5$, tired again at $T = 6,7,8$, rested at $T = 9$ and tired again at $t = 0$.
It is clear that the fatigue is time-depend. A similar analysis can be done for the other individuals. We believe that a complemetary analysis must be done to evaluate the impact of the covariables considered.

5. Conclusion

Working extended Kalman filtering recursions has been applied in a simulation study of state-space models for (AR(1), CHARN(1,1) and CHARN(0,1) states models each with either standard Gaussian or standardized exponential noises.

The estimated state variables computed with our methods and presented graphically give some approximations of the simulated states observations. Although their seem to depend strongly on the initialization

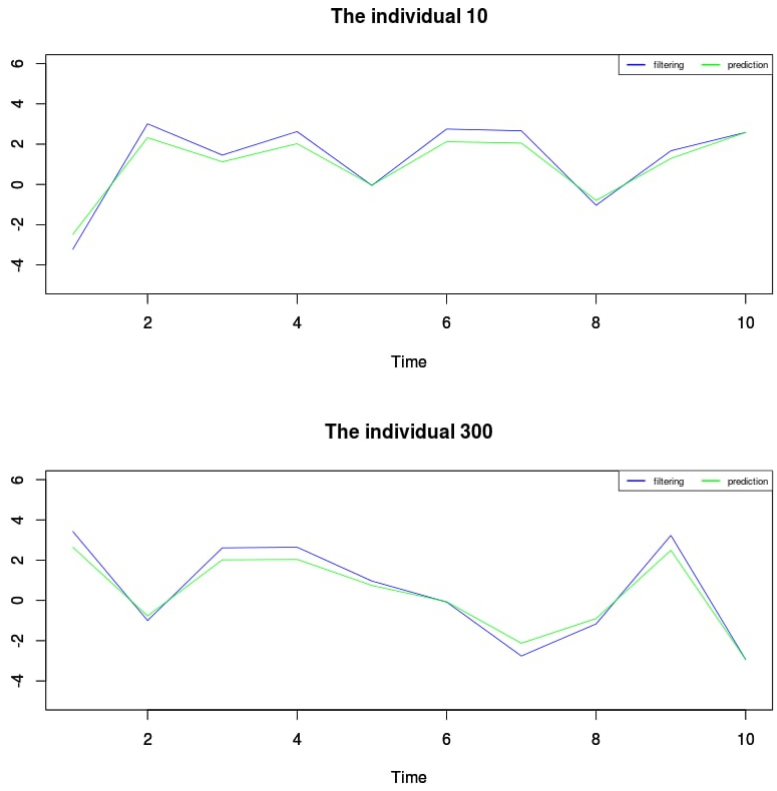


Figure 8. – The graphs for the 10th and 300th individuals.

of the algorithms used, our approach seems to be an alternative to existing tools for estimating latent variables "hidden" by discrete observed random variables, both of the types considered here. Applied to a real data set from patients surged for breast cancer, the two estimation series of the latent variables are very close. Unfortunately, it is not possible to compare them with the true but unobserved latent variable (the fatigue), since they cannot be measured and, therefore, are not available. However, because of the theoretical connection between these estimations and the questionnaire, there is a great hope that they are in a certain way related to the fatigue of the patients, and by this, give credible information about it.

If one admits that the estimators obtained with these real data are those of the fatigue of the patient, even if they only give reliable information about the fatigue, it seems reasonable to try our approach in other fields for estimating latent numerical quantities as for example, the business confidence or morale of customers (in economic), the level of anxiety due to the machines or robots on workers in factories (in industry) etc.

Bibliographie

BARTOLUCCI F., (2014). *Modeling Longitudinal Data by Latent Markov Models with Application to Educational and Psychological Measurement. In Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, pp. 11-19. Springer, Cham.

BARTOLUCCI F., BACCI S., PENNONI F., *Longitudinal analysis of self-reported health status by mixture latent auto-regressive models.* Journal of the Royal Statistical Society : Series C (Applied Statistics), 63(2), pp.267-288.

BARTOLUCCI F., FARCOMENI A., PENNONI F., *A note on the application of the Oakes' identity to obtain the observed information matrix of hidden Markov models.* arXiv preprint arXiv :1201.5990.

BARTOLUCCI F., FARCOMENI A., *Information matrix for hidden Markov models with covariates. Statistics and Computing* 25(3), pp.515-526,2015.

BOUSSEBOUA M., MESBAH M., *Processus de Markov longitudinal latent Rasch observable.* Publ Inst Stat Univ Paris UV fasc, 1-2.

BROCKWELLP.J, DAVISR.A., *Introduction to Time Series and Forecasting.* (1996) Springer-Verlag.

CREAL, D. D. *A Class of Non-Gaussian State Space Models with Exact Likelihood Inference.* Journal of Business & Economic Statistics, 1-13, (2017).

CZADO, C. , SONG, P. X. K. *State space mixed models for longitudinal observations with binary and binomial responses*. Statistical Papers, 49(4),pp. 691-714, (2008).

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*. Journal of the Royal Statistical Society, Series B, pp.39 :1-38, (1977).

DURBIN, J. , KOOPMAN, S. J, *Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives.* Journal of the Royal Statistical Society : Series B (Statistical Methodology), 62(1), pp. 3-56,(2000).

DUNSMUIR, W. T. , SCOTT, D. J. *The glarma package for observation driven time series regression of counts.* Journal of Statistical Software, 67(7), pp.1-36, (2015).

FAHRMEIR, L. , WAGENPFEIL , S. *Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models.* Computational Statistics & Data Analysis, 24(3), pp.295-320, (1997).

FAHRMEIR, L.,TUTZ, G. *Multivariate statistical modelling based on generalized linear models.* Springer Science & Business Media, (2013).

GODSILL , S.,DOUCET, A., WEST, M. *Maximum a posteriori sequence estimation using Monte Carlo particle filters.* Annals of the Institute of Statistical Mathematics, 53(1), pp.82-96, (2001).

HÄRDLE, W., TSYBAKOV, A.,YANG, L. *Nonparametric vector autoregression.* J. Statist. Plann. Infer., 68(2), 221-245, (1998).

JOHANSEN, A. M. DOUCET, A. *A note on auxiliary particle filters.* Stat. Probab. Lett. 78, pp. 1498-1504, 2008).

KITAGAWA, G. *Introduction to time series modeling.* CRC press, (2010).

MESBAH, M. MOUSSEDEK, B. *Processus de Markov Longitudinal Latent Rasch Observable. Pub. Inst. Stat. Univ. Paris. LIV, fasc. 1-2, 2010, 35 à 50 In French.*

MORRELL, D. *Extended Kalman Filter Lecture Notes.* EEE 581-Spring, (1997).

OAKES, D. *Direct calculation of the information matrix via the EM.* Journal of the Royal Statistical Society : Series B (Statistical Methodology), 61(2), pp.479-482, (1999).

PITT, M. K. SHEPHARD, N. *Filtering via Simulation : Auxiliary Particle Filters.* J. Amer. Statist. Asso., 94, pp.590-599, (1999).

POYIADJIS, G., DOUCET, A.,SINGH, S. S. *Maximum likelihood parameter estimation in general state-space models using particle methods.* In Proc of the American Stat. Assoc, (2005).

ROTONDA, C., GUILLEMIN, F., BONNETAIN, F., CONROY, T. *Factors correlated with fatigue in breast cancer patients before, during and after adjuvant chemotherapy : The FATSEIN study.* Contemporary clinical trials, 32(2), pp.244-249, (2011).

SAHA, S.,MANDAL, P. K., BAGCHI, A.,BOERS, Y., DRIESSEN, H. *On the Monte Carlo marginal MAP estimator for general state space models*,(2008).

XI, Y., PENG, H., KITAGAWA, G., CHEN, X. *The auxiliary iterated extended Kalman particle filter.* Optimization and Engineering, 16(2), pp.387-407,(2015).