Kalman recursion generalizations and their applications

Sadeq Kadhim

To cite this version:

Sadeq Kadhim. Kalman recursion generalizations and their applications. Statistics [math.ST]. Université de Lorraine, 2018. English. NNT : 2018LORR0030 . tel-01791912

HAL Id: tel-01791912
https://theses.hal.science/tel-01791912

Submitted on 15 May 2018

Collegium Sciences et Technologies
Université de LORRAINE
École doctorale IAEM
Commission Mathématiques

Les généralisations des récursivités de Kalman et leurs applications

THÈSE

présentée et soutenue publiquement
le 20 Avril 2018

pour l'obtention du

Doctorat de l'Université de Lorraine
(mention mathématique)

Par
Sadeq Awad KADHIM

Composition du jury

Joseph NGATCHOU-WANDJI	Professeur à l'Université de Lorraine	*Directeur*
Mounir MESBAH	Professeur à l'Université Pierre et Marie Curie, Paris 6	*Rapporteur*
Alexandre BERRED	Professeur à l'Université du Havre Normandie	*Rapporteur*
Anne GEGOUT-PETIT	Professeur à l'Université de Lorraine	*Examinateur*
Myriam MAUMY	Maître de Conférences à l'Université de Strasbourg	*Examinateur*

Institut Élie Cartan de Lorraine,
Laboratoire de Mathématiques, BP 70239,
54506 VANDŒUVRE-LÈS-NANCY Cedex

Résumé

Les modèles espace d'état, aussi connus sous le nom de modèles dynamiques, relient des séries chronologiques *observées* à des séries chronologiques *non observées*, par un système de deux équations dont l'une décrit en général une relation entre les termes de la série *observée* et ceux de la série *non observée* (équation d'observation), et l'autre une relation autorégressive entre les valeurs de la série *non observée* (équation d'état). Les valeurs de la série non observée sont appelées *états* et celles de la série observée *observations.*

Nous considérons dans cette thèse, des modèles espace d'état très généraux dans lesquels les fonctions associées peuvent être non-linéaires et les bruits des modèles non-gaussiens. Nous estimons les états de ces modèles en utilisant les récursivités de Kalman généralisées, les filtres particulaires et l'algorithme EM.

Notre travail est motivé par l'envie d'estimer le trait latent en qualité de vie, et d'autres types de variables latentes en économie ou dans le monde industriel. Des exemples concrets de variables latentes sont : la santé des patients, la confiance dans les entreprises, le moral des clients d'une firme, le niveau d'anxiété d'utilisateurs de machines ou de robots dans les usines. Ici, les questionnaires sont considérés comme les observations et les variables latentes comme les états. Notre approche est à la fois une alternative aux travaux existant dans la littérature et leur généralisation.

Plus précisément, dans notre travail, nous nous intéressons aux variables latentes $X_i(t)$ produites par un individu $i, (i = 1, \cdots, n)$, au temps $t, (t = 1, \cdots, T)$. Les $X_i(t)$ peuvent être la santé du patient, un trait latent, etc. Nous observons seulement $Y_i(t)$ au lieu de $X_i(t)$. Les $Y_i(t)$ sont les réponses des individus aux questionnaires. Conformément aux objectifs et à la portée de l'étude, le contenu de cette thèse est structuré en six chapitres, que nous résumons ci-dessous.

Chapitre un : Dans ce chapitre nous faisons un survol de la littérature. Nous y présentons notamment quelques-uns des travaux où l'équation d'état est un modèle linéaire.

Chapitre deux : Ce chapitre présente brièvement les principaux outils mathématiques utilisés. Premièrement, nous rappelons la définition du modèle espace-état donnée par Fahrmeir et Tutz (2013). Aussi connu sous le nom de modèle dynamique, il relie des observations de séries chronologiques ou données longitudinales $\{Y_t\}$ à des "états" non

observés $\{X_t\}$ par un modèle d'observations donnant $\{Y_t\}$ sachant $\{X_t\}$. Les états sont supposés suivre un modèle de transition stochastique. Dans ce chapitre, nous présentons différents modèles espace-état : linéaires avec des bruits gaussiens, non-linéaires avec bruits gaussiens, et non-linéaires avec bruits non-gaussiens.
Les récursivités de Kalman sont des algorithmes qui utilisent une série de mesures (variables) observées au fil du temps pour produire des estimations des variables latentes de trois manières différentes : prédiction pour $t > T$, filtrage pour $t = T$, et lissage pour $t < T$.
Dans les modèles espace-état linéaire ou non-linéaire, si les bruits sont gaussiens, la distribution à posteriori est gaussienne. Par conséquent, les récursivités de Kalman linéaires sont utilisées pour estimer la moyenne à posteriori et la matrice de variance-covariance.
Dans le modèle espace-état non-linéaires avec bruits gaussiens, la distribution à posteriori est encore approximativement gaussienne. Nous utilisons dans ce cas le développement de Taylor pour linéariser les fonctions non-linéaires impliquées dans les modèles étudiés. L'algorithme obtenu s'appelle "récursivités de Kalman étendues". Le modèle espace-état non-linéaire et non-gaussien applique la perspective bayésienne en utilisant les densités conditionnelles. Les filtres particulaires permettent d'approximer les distributions a posteriori permettant d'estimer les valeurs latentes par la prédiction, le filtrage et le lissage. Plusieurs algorithmes de filtres particulaires ont été proposés. Nous présentons les plus connus.

1. **L'algorithme du filtre particulaire** de Kitagawa (1996). L'idée fondamentale est d'utiliser un grand nombre d'échantillons (particules) auxquels nous attribuons des poids, pour approximer la probabilité a posteriori des états.

2. **L'algorithme des filtres particulaires auxiliaires :** L'algorithme précédent présente un problème de dégénérescence. Pitt et Shephard (1999) ont proposé "l'algorithme des filtres particulaires auxiliaires" pour résoudre ce problème. L'idée de base est d'introduire une variable auxiliaire ς dans la distribution d'importance pour générer les particules.

3. **L'algorithme du filtre particulaire auxiliaire de Kalman étendu itéré.** Xi *et al.* (2015) ont proposé un nouvel algorithme du filtre particulaire. Ils ont généré une densité d'importance par l'utilisation de l'algorithme du filtre de Kalman étendu itéré.
   L'estimation du Maximum a Posteriori (MAP) est utilisé pour obtenir une estimation ponctuelle des variables latentes X_t par maximum de vraisemblance :
   $$X_t^{MAP} = \arg\max_{X_t} p(X_t \mid Y_t).$$

Dans ce chapitre, nous présentons deux applications du modèle espace-état non-Gaussien. Nous présentons d'abord les modèles espace-état dont la loi du bruit appartient aux familles exponentielles (modèles linéaires généralisés). Ces modèles ont été étudiés par Fahrmeir et Wagenpfiel (1997) et Fahrmer et Tutz (2013). Ceux-ci ont considéré que la série observée $\{Y_t\}$ est discrète et ont estimé, par maximum de vraisemblance pénalisé, la variable d'état par le mode a posteriori. Leur approche peut également être interprétée comme une méthode nonparamétrique pour le modèle espace-état. Ils ont estimé le mode a posteriori en utilisant la méthode du score de Fisher via le filtrage et le lissage itératif de Kalman.

Deuxièmement, Bousseboua et Mesbah (2010) ont proposé une nouvelle classe de processus longitudinaux multivariés qui sont des résultats dichotomiques ($Y_{ik}(t)$: $i = 1, \cdots, n, k = 1, \cdots, q,\ t = 1, \cdots, T$), où n est le nombre d'individus, q le nombre d'items et t un instant. À chaque instant, nous observons des réponses binaires au lieu de variables latentes $X_i(t)$. Les probabilités conditionnelles sont données par le modèle de Rasch qui est largement utilisé dans divers domaines psychométriques comme la recherche en éducation ou l'analyse de la santé, et en particulier en qualité de vie. Ils ont considéré une étude longitudinale, où les patients répondent aux questions d'un formulaire à des dates régulières de visite, afin que soit mesurée leur santé du moment. Il est reconnu que la santé est un concept multidimensionnel latent. En pratique, chaque dimension est généralement évaluée par une ou plusieurs questions. Les auteurs se sont concentrés sur un cas particulier d'options de réponse dichotomique pour chaque question (oui, non, d'accord, en désaccord, etc.).

Chapitre trois : Dans ce chapitre, nous présentons une nouvelle classe de processus longitudinaux multivariés multicatégorielles. Les données proviennent d'une étude longitudinale, où les patients participent à une entrevue. L'entrevue vise à mesurer la santé des patients à intervalles réguliers, dont les dates sont déterminées avant l'étude. Il s'agit souvent de remplir un questionnaire avec des questions à choix multiples. Ce type questionnaire est utilisé pour des études en qualité de vie pour estimer le bien être des patients, en économie pour estimer la confiance des entreprises ou le moral des clients, et enfin dans le domaine industriel pour estimer le niveau d'anxiété des utilisateurs de machines ou de robots dans les usines. Notre approche est une alternative aux travaux de Bousseboua et Mesbah (2010). C'est aussi une généralisation de ceux-ci, ainsi que de ceux de Bartolucci et Bacci (2014), Bartolucci (2014), Fahrmeir et Wagenpfiel (1997) et Fahrmeir et Tutz (2013). Comme nous l'avons déjà dit, nous nous intéressons aux variables latentes $X_i(t)$ produites par un individu $i, (i = 1, \cdots, n)$, au temps $t, (t = 1, \cdots, T)$. Nous observons seulement $Y_i(t)$ au lieu de $X_i(t)$.

Dans les travaux précédents, les variables latentes $X_i(t)$ sont décrites par le modèle autorégressif du premier ordre AR(1) avec un bruit gaussien. Dans ce chapitre, ils sont décrits par des modèles conditionnellement hétéroscédastiques non-linéaires (CHARN) du premier ordre avec un bruit gaussien. Ici nous faisons l'hypothèse que les observations proviennent d'une distribution multinomiale et que les variables latentes sont gaussiennes. Puisque la distribution a posteriori n'est pas symétrique, le mode a posteriori est une estimation ponctuelle de la variable latente. Dans ce chapitre, nous présentons deux approches pour estimer le mode postérieur. La première est basée sur l'utilisation de la récursivité de Kalman étendue. Dans l'appendice A.2, nous rapellons l'équation du filtre de Kalman étendu. La deuxième approche est basée sur l'utilisation du"maximum a posteriori (MAP)" issu de l'algorithme du filtre particulaire auxiliaire de Kalman étendu itéré. Les paramètres des modèles sont estimés par la méthode du maximum de vraisemblance via l'algorithme EM. Nous rappellons que l'algorithme EM est un algorithme itératif qui génère une séquence d'estimations du paramètre étudié. Chaque itération se décompose en deux étapes. La première étape E "Expectation", c'est-à-dire "Espérance" calcule une vraisemblance à partir de la formule d'un modèle espace-état. La seconde étape M "Maximisation" consiste à rechercher un jeu de paramètres maximisant la vraisemblance estimée à l'étape E. Nous calculons la matrice de variance-covariance des paramètres en exécutant l'"Identité d'Oakes" (1999). L'estimation du vecteur des paramètres de notre modèle est basé sur la distribution a posteriori $p(\mathbf{X}_i \mid \mathbf{Y}_i)$. dont la densité peut être calculée par une approche bayésienne. Dans le modèle état-espace linéaire gaussien, la distribution posteriori est gaussienne. Par conséquent, la moyenne conditionnelle et la matrice de variance-covariance du vecteur d'état estimé sont calculées par les récursivités du filtre de Kalman. Avec le modèle état-espace généralisé non-gaussien comme celui que nous étudions, la distribution de $X_i(t)$ sachant $Y_i(t)$ est généralement non-gaussienne. Par conséquent, il est nécessaire d'utiliser les méthodes des filtres particulaires pour trouver l'approximation de la distribution a posteriori. L'algorithme que nous utilisons ici est celui du filtre particulaire auxiliaire de Kalman étendu itéré proposé par Xi *et al.* (2015). Nous développons cet algorithme avec notre modèle et nous obtenons les équations du mode a posteriori et celle de la covariance a posteriori. La méthode utilisée est présentée à l'appendice A.2.

Chapitre quatre : Dans ce chapitre, nous généralisons notre modèle au cas où le bruit du modèle d'état est supposé issu d'une famille exponentielle. En pratique, les lois parmi les plus courantes de ces familles sont la loi normale, la loi exponentielle, la loi Gamma, la loi du khi-deux, la loi Beta, la loi de Dirichlet, la loi de Bernoulli, la loi multinomiale, la loi de Poisson, la loi de Wishart, la loi inverse Wishart et plusieurs autres. Comme au chapitre 3, nous trouvons aussi le mode a posteriori par deux approches : celle basée sur l'utilisation de la récursivité de Kalman étendue puis celle basée sur le "maximum a posteriori (MAP)" calculé par l'algorithme du filtre particulaire auxiliaire de Kalman étendu itéré. Comme précédemment, les paramètres des modèles sont estimés par la méthode du maximum de vraisemblance par l'algorithme EM. Nous calculons aussi la matrice de variance-covariance des estimateurs en utilisant l'"Identité d'Oakes" (1999). La loi a posteriori $p(\mathbf{X}_i \mid \mathbf{Y}_i)$ est approchée par l'algorithme du filtre particulaire auxiliaire de Kalman étendu itéré.

Chapitre cinq : Dans ce chapitre, nous présentons et commentons les résultats des simulations numériques que nous avons effectuées et l'application de nos résultats aux données réelles de santé. Il y a deux parties. La première, concerne les simulations numériques où les données longitudinales sont engendrées avec des variables latentes décrites par un modèle CHARN dont le bruit est de loi appartenant à une famille exponentielle. Dans la deuxième partie, nous appliquons nos résultats aux données réelles sur la qualité de vie des femmes ayant subi une opération pour cause de cancer du sein. Dans les deux parties, des routines R sont créées à partir des méthodes décrites dans les chapitres 3 et 4. Notre objectif dans ce chapitre est d'estimer les variables latentes par mode a posteriori via les récursivités du filtre de Kalman étendu.

1. **Les simulations :** Dans cette partie, nous produisons des données à partir de l'équation d'observation décrite par une distribution multinomiale définie par l'équation (3.3.1), et l'équation d'état décrite par un modèle CHARN défini par l'équation (3.3.3) avec bruit de loi appartenant à une famille exponentielle. Nous considérons les lois gaussienne et exponentielle. Il y a deux parties. La première partie vise à tester l'efficacité des récursivités du filtre de Kalman étendu, où nous considérons les paramètres du modèle connus. La deuxième partie utilise l'algorithme EM pour estimer les paramètres du modèle, avant d'appliquer des récursivités du filtre de Kalman étendu. Pour simuler, nous tenons compte de la taille des échantillons (nombre d'individus et durée), la forme de la distribution du bruit d'état (Gaussien ou exponentiel) et le type d'équation d'état (modèle AR (1), CHARN (1,1) et CHARN (0,1)). Afin d'évaluer notre approche dans cette étude, nous présentons les figures montrant la série des variables d'état simulées et celles de leurs valeurs obtenues par les algorithmes de prédiction et de filtrage.

2. **Application aux données réelles :** Rotonda *et al.* (2011) ont conduit une étude sur la qualité de vie. Elle porte sur les facteurs corrélés à la fatigue liée au cancer du sein, chez les femmes ayant subi une opération pour cause de cette maladie. Des patients en nombre 502 ont été recrutés de septembre 2008 à septembre 2010 dans trois centres cancéreux français : le centre anticancéreux Alexis Vautrin de Lorraine, le centre anticancéreux Georges-Francois Leclerc de Bourgogne et le centre anticancéreux Paul Strauss d'Alsace, France. Ils ont rempli le questionnaire plusieurs fois, soit lors de leurs visites en clinique, soit à leur domicile après avoir reçu une enveloppe affranchie pour retourner leurs réponses. Le questionnaire utilisé est “State-Trait Anxiety Inventory” dont la version française (STAI-B) comporte vingt items. Pour chaque item, les réponses des patients sont classées en 4 catégories (non, plutôt non, plutôt oui, oui).
   Ici, la variable latente est la fatigue du patient après la chirurgie. Cette variable est supposée être quantitative et varie dans le temps autour d'une valeur moyenne supposée nulle dans notre travail. Nous estimons la fatigue du patient par le mode a posteriori en utilisant deux modèles pour la variable latente : les modèles AR(1) et CHARN (1,1).

Chapitre six : Ce chapitre présente les contributions de l'étude, ses conclusions et quelques recommandations pour les études futures.

1. **Conclusion :** Nous avons appliqué notre approche sur différentes études de simulation, qui ont été conçues avec les différents types de questionnaires et les différentes formes d'équations d'état. Nous l'avons aussi appliquée directement aux données réelles sur la qualité de vie des femmes ayant subi une opération pour le cancer du sein. Notre approche a fourni une estimation de la variable latente à partir de données sur la qualité de vie. Ce qui permet donc d'avoir une idée plus précise de l'intensité, à chaqu'instant, de la fatique de chaqu'une des patientes ayant participé à l'expérience. Ce qui peut aussi permettre une meilleure prise en charge et même dans une certaine mesure de faire de la prévention dans le traitement du cancer du sein.

2. **Perspectives :**

   - Notre approche, qui permet d'estimer une variable latente dans le domaine de la santé, peut aussi s'appliquer dans des domaines tels que le domaine économique, pour estimer la confiance des entreprises ou le moral des clients, le domaine industriel, pour estimer le niveau d'anxiété des utilisateurs de machines ou de robots en usine.

   - Dans le chapitre 5, nous avons utilisé la première approche des récursivités de Kalman pour la création de routines R. Mais celles-ci peuvent aussi être créées pour la méthode du "maximum a posteriori (MAP)" basée sur le filtre particulaire auxiliaire de Kalman étendu itéré. On pourrait comparer les deux approches.

   - Dans l'analyse des données réelles, nous avons constaté que les données étaient incomplètes : il y manquait des covariables ou des réponses des individus. Analyser ces données fut difficile. Nous avons donc choisi d'ignorer les individus pour lesquels il manquait des données. Dans nos travaux à venir, nous allons étudier la possibilité d'étendre notre approche au cas où les données sont incomplètes.

   - Dans notre étude, nous avons supposé que la fonction de lien est un prédicteur linéaire $\eta^s_{ik}(t) = \mathbf{u}_i^\top(t)\beta^s_k + X_i(t)$. Mais Hastie et Tibshirani (1990) ont introduit une classe des modèles additifs généralisés qui remplace la forme linéaire par une somme de fonctions lisses $\sum_j s_j(u_j)$, où les $sj(.)$ sont des fonctions non spécifiées. De la même manière nous pouvons développer notre modèle avec $\eta^s_{ik}(t) = s_i(\mathbf{u}_i(t)) + X_i(t)$.

Mot-clés

- Récursivités de Kalman généralisées

- Modèles à espace d'état non-linéaires

- Données multicatégorielles longitudinales

- Variables latentes

- Filtres particulaires

- Algorithme EM

Abstract

We consider state space models where the observations are multicategorical and longitudinal, and the state is described by CHARN models. We estimate the state by generalized Kalman recursions, which rely on a variety of particle filters and EM algorithm. Our results are applied to estimating the latent trait in quality of life, and this furnishes an alternative and a generalization of existing methods. These results are illustrated by numerical simulations and an application to real data in the quality of life of patients surged for breast cancer.

- Generalized Kalman recursions

- Generalized state space models

- Multicategorical longitudinal data

- Latent variables

- Particle filters

- EM algorithm

Remerciements

Je tiens tout d'abord à remercier le directeur de cette thèse, M. Joseph NGATCHOU-WANDJI, pour m'avoir fait confiance malgré les connaissances plutôt légères que j'avais en octobre 2013 sur le modèle espace-état, puis pour m'avoir guidé, encouragé, conseillé, et pour avoir corrigé erreurs scientifiques et fautes d'orthographe pendant presque quatre ans.
Je remercie Messieurs Didier MAQUIN (directeur de l'école doctorale), Xavier ANTOINE (directeur du laboratoire), Olivier Garet (directeur du département de mathématiques) pour la confiance qu'ils m'ont accordée.
Je voudrais remercier les rapporteurs de cette thèse M. Mounir MESBAH, Professeur des Universités Pierre et Marie Curie, Paris 6, et M. Alexandre Berred Professeur des Universités du Havre-Normandie. Je leur suis très reconnaissant du temps qu'ils ont consacré à l'expertise de ce travail. J'apprécie à sa juste valeur leur présence dans le jury. J'adresse de sincères remerciements à Mme Anne GEGOUT-PETIT et Mme Myriam MAUMY pour l'intérêt qu'elles ont accordé à mon travail et l'honneur qu'elles me font en participant à mon jury.
Bien sûr, atteindre ces objectifs n'aurait pas été possible sans l'aide des membres de l'équipe probabilités et statistiques. Je souhaite notamment remercier M. Antoine Lejay (responsable de l'équipe) pour la confiance qu'il m'a accordée.
J'aimerais remercier Messieurs Francis GUILLEMIN, Jean-François SCHEID, Benoît DANIEL pour avoir fait partie du comité de suivi de ma thèse, pour (ses remarques, ses conseils, ses encouragements, ses notes) ce qui m'a fait continuer pour terminer ma thèse .
J'associe à ces remerciements Madame Sabrina Ferry, Monsieur Georges BILLANT, Madame Laurence QUIROT (responsable administrative de l'IECL), Madame Elodie CUNAT (secrétaire du département) et Madame Nathalie BENITO (assistante de l'équipe probabilités et statistiques) qui m'ont facilité l'accomplissement des démarches administratives auprès de l'université de Lorraine.
Je souhaite remercier les bibliothécaires Madame Valérie DAUBENFELD et Madame Stéphanie JOURDAN, pour m'avoir aidé dans la recherche de références bibliographiques.
Je remercie M. Didier GEMMERLE pour son important travail d'assistance technique et informatique.
J'adresse ma reconnaissance à Monsieur Salem BEN SAID (IECL) pour ses précieux conseils. Je remercie spécialement M.Khalid AL-GHARRAWI (Université de Iowa - USA) et M. Waleed DHHAN SLEABI (Université de Putra-Malaysia) pour leur aide.
Je n'oublie pas mes collègues doctorants : Anis AMRI, Benjamin ALVAREZ, Clèmence KARMANN, Hassan MOHSEN, Mattieu BRACHET, Nassim SAHKI, Simon JACQUES et Tom RIBLET, pour leur soutien permanent.
Mes remerciements vont aussi à mes amis en France et en Irak pour leurs soutiens indéfectibles. Enfin, je remercie ma famille (ma femme, mes filles, ma mère, mes frères et sœurs).

Dédicaces

- A la mémoire de mon père qui me dit, peu avant sa mort: ”mon fils, je veux te voir professeur”.
- A la mémoire de mon frère, décédé le 23/01/2015 alors que j'étais en France. Je me souviens encore de ses paroles lorsqu'il m'a appelé la nuit avant sa mort:”Tu me manques tellement, je voudrais te serrer dans mes bras”.
- A ma respectueuse mère, qui a prié pour moi tout au long de mes études doctorales.
- A ma chère épouse pour sa contribution, sa patience et sa compréhension tout au long de mes études doctorales. Elle m' a incroyablement soutenu et sans elle, rien de tout cela n'aurait été possible pour moi.
- Pour mes filles, Zahraa, Tuqa, Jannah et Safa qui n'ont cessé de m'accompagner au travail par leur amour, qui a toujours été ma plus grande source d'inspiration.

Communications

1. **Avril 2017:** Exopsé sur les “ Longitudinal multicategorical processes : Generalized state space models. ” Rencontres des Jeunes Statisticiens 2017, organisées à Porquerolles, France.

2. **Janvier 2018:** Exopsé sur les “ Estimation du trait latent : application aux donnés sur des femmes opérées du cancer du sein ” Des journées de la Fédération Charles Hermite, LORIA. organisées à Nancy, France .

3. **Mars 2018:** Exopsé sur les “ Estimating latent variable by generalized Kalman recursions ” Groupe de Travail des Doctorants Institut Elie Cartan, Université de Lorraine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

State space models, also known as dynamic models, relate time series observations or longitudinal data $\{Y_t\}$ to unobserved "states" $\{X_t\}$ by an observation model for $\{Y_t\}$ given $\{X_t\}$. The states are assumed to follow a stochastic transition model (Fahrmeir and Tutz (2013)). We consider generalized state space models where the functions involved may be nonlinear and the noises non-gaussian. We aim at estimating the state observations by using generalized Kalman recursions, which involve the use of particle filters and the EM algorithm.

Our work is motivated by the desire to estimate the latent trait in quality of life, where we consider questionnaires as the observations and the latent variables as the states.

Our approach is an alternative to Bousseboua and Mesbah (2010), but derived in a more general context. It is also a generalization of Bartolucci and Bacci (2014), Bartolucci (2014), Fahrmeir and Wagenpfiel (1997) and Fahrmeir and Tutz (2013).

Bousseboua and Mesbah (2010) give a natural extension 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 defined their model and estimated the parameters by maximizing the likelihood function via E-M algorithm. They illustrated their model with two classes of latent processes: i) the general Markov latent processes, ii) the auto-regressive AR(1) latent processes. They assumed the noise in the state equation has a gaussian distribution. They determined the distribution of latent trait and estimated the parameters of their models.

Bartolucci and Bacci (2014) proposed a model for longitudinal categorical data. The latent process was modelled by a mixture of autoregressive AR(1) processes with different means and correlation coefficients, but with equal variance. They considered an application based on a longitudinal data set, concerning self-reported health status (SRHS) derived from the Health and Retirement Study (HRS) and conducted by the University of Michigan and supported by the US National Institute on Aging and the Social Security Administration.

Bartolucci (2014) reviewed a class of models for longitudinal data. In this type of models, the unobserved individual characteristics of interest was represented by a sequence of discrete latent variables, which follows a Markov chain. He used two applications to illustrate his models. The first application is *the evolution of the ability level in mathematics.* There, Bartolucci considered testing the hypothesis of absence of tiring or learning-through-training phenomena during the administration of series of 12 items on Mathematics. The second application is *the evolution of psychological traits in children.* He considered data collected through a mental experiment based on tests administered at different occasions to pre-school children to measure two types of ability: inhibitory control and attentional flexibility.

Fahrmeir and Wagenpfiel (1997) and Fahrmeir and Tutz (2013) studied non-linear time series or discrete longitudinal observations. They created a method of inference based on the posterior mode or, equivalently, maximum penalized likelihood estimation. They acquired efficient smoothing by using the working Kalman filtering and smoothing. They analysed data taken from the IFO Institute for economic research in Munich which collects categorical monthly data of firms in various industrial branches. The monthly questionnaire contained questions on the tendency of realization and expectations of variables like production, orders in hand, and demand. Answers were categorical, most of them trichotomous with categories like “increase”, “decrease”, and “no change”. Thus, for each firm, the data form a categorical time series. Considering all firms within a specific branch, they had a categorical panel or longitudinal data.

Durbin and Koopman (2000) discussed the non-gaussian state space models from both classical and Bayesian perspectives with real non-gaussian time series data. They proposed an approach based entirely on importance sampling and antithetic variables. They gave advantage to their approach: first, they entirely avoided the convergence problems that were associated with MCMC algorithms. Second, they easily computed error variances due to simulation as a routine part of the analysis, where they confirmed that the investigator could attain any predetermined level of simulation accuracy by increasing the sample size, where necessary, by a specific amount.

Czado and Song (2008) proposed a new class of state space models for longitudinal discrete response data where the observation equation specified in an additive form involving both deterministic and random linear predictors. They developed a Markov Chain Monte Carlo (MCMC) algorithm to carry out statistical inference for models with binary and binomial responses, they illustrated the applicability of their model in both simulation studies and data examples.

Creal (2017) calculated the likelihood function for a large class of non-gaussian state space models that includes stochastic intensity, stochastic volatility, and stochastic duration models among others. The state variables in his model followed a non-negative stochastic process that is popular in econometrics for modelling volatility and intensities.

Dunsmuir and Scott (2015) made the **R** Package **glarma**, they considered the generalized state space models for non-gaussian time series (GLARMA) described in Davis, Dunsmuir, and Wang (1999), Brockwell and Davis (2013) and Durbin and Koop-

man(2000).
Almost all these above cited works considered linear models for the latent variable. In **our work**, we considered a non-linear model for this variable. We first assume the noise in state equation from gaussian distribution. We develope the work of Fahrmeir and Wagenpfiel (1997) where the Extended Kalman filter and smoother is used.
Our work provides a new class of multicategorical longitudinal multivariate processes of observed outcomes. Its context is that of a longitudinal study, where participant patients were interviewed about their health at regular intervals. The dates of which established before the study. The interviews typically involved filling out a questionnaire in which they were asked multiple choice questions. These questionnaires were constructed to measure the patient's perceived health at the time of the visit. It is recognized that health is a latent multidimensional concept.
More precisely, we are interested in latent variables $X_i(t)$ produced by an individual $i, (i = 1, \cdots, n)$, at a time $t, (t = 1, \cdots, T)$. The $X_i(t)'$s might be **the patients health**, a **latent trait.** Our work can be applied in health field for studying the patients health. It can be applied in economics field for studying the business confidence or the morale of customers. It can be applied in industrial field for the study of the level of anxiety of workers from machines or robots in factories.
We only observe $Y_i(t)$ instead of $X_i(t)$. The $Y_i(t)'$s are the responses of the individuals to the questionnaire. The latent processes are described by Conditional Heteroscedastic Autoregressive Nonlinear (CHARN) models in two cases: i) gaussian noise processes ii) noise processes from exponential family distributions.
We estimate the parameters of the model by the EM algorithm, after we use the Auxiliary Iterated Extended Kalman Particle Filter(AIEKPF) algorithm for determining the posterior likelihood. Next, the latent variables are estimated by the penalized likelihood and working extended Kalman filtering recursions and the maximum a posteriori (MAP) approach.

1.2 Overview of the dissertation :

In conformity with the objectives and the scope of the study, the contents of this dissertation is structured in six chapters. The dissertation chapters are organized so that the study objectives are apparent and conducted in the sequence outline.

Chapter Two : This chapter briefly presents the main mathematical tools used: State space models: linear, non-linear, gaussian, and non-gaussian. Particle filter methods. The Rasch models. The EM algorithm.

Chapter Three : This chapter presents a new class of longitudinal data through the state space models, where the observations are multicategoricals and the state variables are described by CHARN model with gaussian noise process. We discuss the estimation of the model parameters by the maximum likelihood via the EM algorithm, and the consistency and asymptotic normality to these estimators. Estimation of the parameters needed the posterior distribution which is found by using the Auxiliary Iterated Extended Kalman filter(AIEKPF). The optimality of the particles is presented. In this chapter, the objective is achieved through estimating the latent trait, which is estimated by the posterior mode via two methods. First, the penalized likelihood and working extended Kalman filtering. Second, the maximum a posteriori estimation (MAP).

Chapter Four : The objective of this chapter is to generalize our approach of chapter 3 to the case of noise processes from exponential family distribution. As in chapter 3, we find the estimation of the latent trait and model parameters. The consistency and asymptotic normality of the parameters is discussed. We present the algorithms to some of the methods explained in chapter 3 to avoid the prolongation .

Chapter Five : In this chapter, the numerical results are carried out to estimating latent variables. This chapter contains two parts. In the first, we generate longitudinal data with the latent trait from exponential family distributions. In the second, the real data are from a longitudinal study in women aged 18 years and older, who were surged for breast cancer in France. R-packages were created for the algorithms were mentioned in chapters 3.

Chapter Six : This chapter provides with the summary and detailed discussions of the dissertation conclusions. Areas for future research are also invoked.

Chapter 2

Mathematical tools

2.1 Introduction

This chapter focuses on the theoretical side of the state space models through the linear and nonlinear state space models. Kalman Filter and Smoother recursions are widely used to estimate the state variables and their variances. Therefore, it is presented for linear and nonlinear state space models. Two applications are given. First, the exponential family state space models discussed by Fahrmeir Wagenfiel (1997). Second, the Longitudinal Rasch process studied by Bousseboua and Mesbah (2010).

2.2 State space representations

2.2.1 Linear state space models

A state space model for a (possibly multivariate) time series $\{Y_t, t = 1, 2, \cdots\}$ consists of two equations. The first equation, known as the **observation equation**, expresses the $\ell-$ dimensional observation Y_t as a linear function of a $k-$ dimensional state variable X_t plus noise :

$$Y_t = G_t X_t + \xi_t, \quad t = 1, 2, \cdots, \tag{2.2.1}$$

where $\{\xi_t\}$ for all $t \in \mathbb{N}^*, \xi_t \sim \mathcal{N}(0; \Sigma_t)$ is called *observation noise* and G_t is an $\ell \times k$ matrix called *observation* or *design matrix.*

The second equation, called the **State equation**, determines the state X_t at time t in an expression of the previous state X_{t-1} and a state noise. The state equation as follows

$$X_t = F_t X_{t-1} + H_t \varepsilon_t, \quad t = 1, 2, \cdots \tag{2.2.2}$$

where $\{\varepsilon_t, t = 1, 2, \cdots\}$ is the state noise, that is, a $k-$ dimensional white noise. ε_t for all $t = 1, 2, \cdots$, $\varepsilon_t \sim \mathcal{N}(0; R_t), R_t$ is a $k \times k$ matrix variance-covarince of state noise. F_t is $k \times k$ a matrix called *state matrix*, and H_t is $k \times m$ a matrix called *input matrix.* $\{\varepsilon_t\}$ is uncorrelated with $\{\xi_t\}$ (i.e., $E(\xi_t \varepsilon'_s) = 0$ for all s and t). To complete the specification, it is assumed that the initial state X_0 is uncorrelated with all of the noise terms $\{\varepsilon_t\}$ and $\{\xi_t\}$, and $X_0 \sim \mathcal{N}(x_0; R_0)$. Basically, the system matrices G_t, F_t, Σ_t, R_t and x_0, R_0 (the mean and variance for initial state) are assumed to be deterministic and known.

2.2.2 Non-linear state space models

The observation process $(Y_t), Y_t \in \mathbb{R}^\ell$, is described by an equation of the form

$$Y_t = G(\mathbf{u}_t, X_t, \lambda) + \xi_t. \tag{2.2.3}$$

where $G(.)$ is a non-linear function, ξ_t is the observation noise, λ is a parameter, $(\mathbf{u}_t) \in \mathbb{R}^r$ is the covariate process and (X_t) is the state process described by an equation of the form

$$X_t = F(\mathbf{u}_t, X_{t-1}, \gamma) + H(\mathbf{u}_t, X_{t-1}, \delta) \cdot \varepsilon_t, \tag{2.2.4}$$

where $F(.)$ and $H(.)$ are non-linear functions, (ε_t) is the state noise, and γ is a parameter.

Gaussian State space models make the assumption that:

- $\{\xi_t, t = 1, 2, \cdots\}$ is assumed to have a gaussian distribution: $\xi_t \sim \mathcal{N}(0; \Sigma_t)$, Σ_t the variance-covariance matrix of the observations.

- $\{\varepsilon_t, t = 1, 2, \cdots\}$ has also a gaussian distribution for all, $t = 1, 2, \cdots$, $\varepsilon_t \sim \mathcal{N}(0; R_t)$, $R_t \in \mathbb{R}^k$, where R_t is the matrix variance-covariance of state noise.

- (ξ_t) and (ε_t) are uncorrelated that is, for all t, j

  $\mathrm{E}[\xi_t] = 0$, $\mathrm{E}[\xi_t \xi_t^\top] = \Sigma_t$, $\mathrm{E}[\xi_t \xi_j^\top] = 0$ for $t \neq j$. $\mathrm{E}[\varepsilon_t] = 0, \mathrm{E}[\varepsilon_t \varepsilon_t^\top] = R_t$, $\mathrm{E}[\varepsilon_t \varepsilon_j^\top] = 0$ for $t \neq j$, $\mathrm{E}[\xi_t \varepsilon_j^\top] = 0$ for all t and j.

Since this state space model is nonlinear and gaussian, the posterior distribution $p(X_t \mid Y_t)$ is also approximately gaussian. For that, we use the extended Kalman filter and smoother to compute the *posterior mode* and *covariance matrix.*

Non-gaussian state space models make the assumption that, (ξ_t) and (ε_t) have density functions $r(\xi)$ and $q(\varepsilon)$ respectively both of known forms. We can easily check that

$$\mathrm{E}[X_t \mid X_{t-1}, \mathbf{u}_t] = F[\mathbf{u}_t, X_{t-1}, \gamma]$$

$$\mathrm{Var}[X_t \mid X_{t-1}, \mathbf{u}_t] = H^2[\mathbf{u}_t, X_{t-1}, \delta] R_t.$$

Non-linear state space models are called **Generalized state space models (GSSM)**. There are two important types of these models "*parameter driven*" and "*observation driven.*" Both of which are frequently used in time series analysis. In a parameter-driven model, the state equation is constructed by using the previous state. In an observation-driven model, the state equation is constructed by using the past observation.

2.2.2.1 Parameter-Driven models

The observation and state equations (2.2.3) and (2.2.4) are replaced by the conditional probability densities of the observation and the state variables. The distribution of Y_t given $(X_t, \mathbf{X}_{t-1}, \mathbf{Y}_{t-1})$ is independent of $(\mathbf{X}_{t-1}, \mathbf{Y}_{t-1})$, then the observation equation is:

$$p(Y_t \mid X_t) \coloneqq p(Y_t \mid X_t, \mathbf{X}_{t-1}, \mathbf{Y}_{t-1}), \quad t = 1, 2, \cdots . \tag{2.2.5}$$

where $\mathbf{Y}_t = (Y_1, Y_2, \cdots, Y_t)$ and $\mathbf{X}_t = (X_1, X_2, \cdots, X_t)$. The distribution of X_t given $(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{Y}_{t-1})$ is independent of $(\mathbf{X}_{t-2}, \mathbf{Y}_{t-1})$, then the state equation yields :

$$p(X_t \mid X_{t-1}) \coloneqq p(X_t \mid \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{Y}_{t-1}), \quad t = 1, 2, \cdots . \tag{2.2.6}$$

Note : $x := y, y =: x$ or $x \overset{\text{def}}{=} y$ means x is defined to be another name for y, under certain assumptions taken in context.

Example Assume $Y_t \mid \mu_t \sim$ Poisson (μ_t), this means:

$$\Pr(Y_t = y_t \mid \mu_t) = \frac{\exp(-\mu_t)\mu_t^{Y_t}}{\mu_t!}, t = 1, 2 \cdots$$

and

$$\log(\mu_t) = \mathbf{u}_t^\top \beta + X_t,$$

where β is a parameter vector, $(\mathbf{u}_t)$ is the covariate process and $\{X_t\}$ is a stationary gaussian process, e.g. (AR(1) process)

$$(X_t + \sigma^2/2) = \phi(X_{t-1} + \sigma^2/2) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0; \sigma^2(1 - \phi^2)),$$

with $\sigma > 0, \sigma^2$ an ϕ standing for unknown parameters. This example shows that the state variables are a function of previous states.

2.2.2.2 Observation-Driven models

In an observation-driven model, it is again assumed that Y_t conditional to $(X_t, \mathbf{X}_{t-1}, \mathbf{Y}_{t-1})$ is independent of $(\mathbf{X}_{t-1}, \mathbf{Y}_{t-1})$. The model is specified by the conditional probability densities

$$\begin{aligned} p(Y_t \mid X_t) &= p(Y_t \mid \mathbf{X}_t, \mathbf{Y}_{t-1}), \quad t = 1, 2, \cdots && (2.2.7) \\ p(X_t \mid X_{t-1}, \mathbf{Y}_t) &= p(X_t \mid \mathbf{Y}_t), && (2.2.8) \end{aligned}$$

where $p(X_1 \mid \mathbf{Y}_0) =: p_1(X_1)$ is the initial probability density.

Example Assume $Y_t \mid \mu_t \sim \text{Poisson}(\mu_t)$, this means for y_t:

$$\Pr(Y_t = y_t \mid \mu_t) = \frac{\exp(-\mu_t)\mu_t^{y_t}}{\mu_t!}, t = 1, 2 \cdots$$

and

$$\log(\mu_t) = \mathbf{u}_t^\top \beta + X_t,$$

where β is a parameter vector, $(\mathbf{u}_t)$ is the covariate process and $\{X_t, t = 1, 2, \cdots\}$ is a function of the past observations $Y_s, s < t$.

$$X_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p},$$

with $\phi_i, i = 1, \cdots, p$, standing for unknown parameters. From this example, one can see that the state equation is a function of the past observations $(Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p})$.

2.3 Kalman Filter and smother recursions

2.3.1 Linear Kalman filtering and smoothing recursions

In the linear state space models defined by the equations (2.2.1) and (2.2.2), the observations are $Y_1, \cdots, Y_T$, and the $X_1, \cdots, X_T$ are the state variables. Kalman filtering and smoothing recursions are used to estimate the state variables, as shown in Figure (2.1). It can be done by three algorithms:

- Filtering for $t = T$,
- Smoothing for $t < T$,
- Prediction for $t > T$.

Under the gaussian assumption, the optimal solution to the filtering problem is given by the *conditional* or *posterior mean* of X_t given Y_t

$$x_{t|t} := \mathrm{E}(X_t \mid Y_t).$$

Since the model is linear and gaussian, the *posterior distribution* of X_t is also gaussian,

$$X_t \mid Y_t \sim \mathcal{N}(x_{t|t}; V_{t|t}),$$

the *posterior covariance matrix*

$$V_{t|t} := \mathrm{E}[(X_t - x_{t|t})(X_t - x_{t|t})^\top].$$

The famous linear Kalman filter and smoother calculates the posterior means and covariance matrices in an efficient recursive way. The usual derivations of the Kalman filter and smoother take advantage of the fact that the posterior distributions are gaussian.

In Figure (2.1) the Kalman filter algorithm for $t = 1, \cdots, T$, can be performed as follows: start with initial value x_0. Compute $x_{1|0}$ by using x_0 in prediction step, then using $x_{1|0}$ to find the $x_{1|1}$ via the filter step. Thereafter, compute $x_{2|1}$ by using $x_{1|0}$ in prediction step and in filter step find $x_{2|2}$ by using $x_{2|1}$. Thus, the prediction and filter steps continue until T where each instant t has a prediction value and a filter value. The smoothing step consists of backward recursions for $t = T, \cdots, 1$. Namely for each instant $t = 1, \cdots, T-1$ computing $x_{t|T}$ depend to the value of instant T.
Therefor, in Figure (2.1), we used the lines to illustrate the three algorithms filtering, prediction and smoothing, another paths can be produced by using these algorithms. In Figure (2.1), it is important to appoint that we have one filtering step after the prediction step.

Figure 2.1: **The paths of recursive computation by the Kalman filter and smoothing algorithm.** ⇒: prediction, ⇓: filter, ←: smoothing, →: increasing horizon prediction.

$$\begin{tikzcd}
x_{1|0} \arrow[r] \arrow[d, Rightarrow] & x_{2|0} \arrow[r] & x_{3|0} \arrow[r] & x_{4|0} \arrow[r] & x_{5|0} \arrow[r] & {} \\
x_{1|1} \arrow[r, Rightarrow] & x_{2|1} \arrow[r] \arrow[d, Rightarrow] & x_{3|1} \arrow[r] & x_{4|1} \arrow[r] & x_{5|1} \arrow[r] & {} \\
x_{1|2} & x_{2|2} \arrow[l] \arrow[r, Rightarrow] & x_{3|2} \arrow[r] \arrow[d, Rightarrow] & x_{4|2} \arrow[r] & x_{5|2} \arrow[r] & {} \\
x_{1|3} & x_{2|3} \arrow[l] & x_{3|3} \arrow[l] \arrow[r, Rightarrow] & x_{4|3} \arrow[r] \arrow[d, Rightarrow] & x_{5|3} \arrow[r] & {} \\
x_{1|4} & x_{2|4} \arrow[l] & x_{3|4} \arrow[l] & x_{4|4} \arrow[l] \arrow[r, Rightarrow] & x_{5|4} \arrow[r] \arrow[d, Rightarrow] & {} \\
 & & & & {} &
\end{tikzcd}$$

2.3.1.1 Linear Kalman filter

- The predictive step is $x_{t|t-1}$ and $V_{t|t-1}$ is the covariance matrix of predictive step.
- The filter step is $x_{t|t}$ and $V_{t|t}$ is the covariance matrix of filter step.
- The smoother step is $x_{t|T}$ and $V_{t|T}$ is the covariance matrix of smoother step.

The Kalman algorithm for the linear state space models in equations (2.2.1, 2.2.2) is as follows

Initialization:

$$x_{0|0} = x_0, \quad V_{0|0} = R_0.$$

For $t = 1, \cdots, T$:

Prediction step:

$$\begin{aligned} x_{t|t-1} &= F_t x_{t-1|t-1}, \\ V_{t|t-1} &= F_t V_{t-1|t-1} F_t^\top + H_t R_t H_t^\top, \end{aligned} \tag{2.3.1}$$

Correction or filtering step:

$$\begin{aligned} x_{t|t} &= x_{t|t-1} + K_t(Y_t - G_t x_{t|t-1}), \\ V_{t|t} &= (I - K_t G_t) V_{t|t-1}. \end{aligned} \tag{2.3.2}$$

Kalman gain:

$$K_t = V_{t|t-1} G_t^\top [G_t V_{t|t-1} G_t^\top + \Sigma_t]^{-1}.$$

The optimal Kalman gain gives the formula for the updated estimate variance- covariance matrix $V_{t|t}$. Usage of other gain values leads to a non optimal variance-covariance matrix.

Smoother step: The smoother for X_t given Y_T where $t < T$, is as follows :

$$x_{t|T} \coloneqq \mathrm{E}(X_t \mid Y_T),$$

and again the posterior is normal,

$$X_t \mid Y_T \sim \mathcal{N}(x_{t|T}; V_{t|T}),$$

with

$$V_{t|T} \coloneqq \mathrm{E}[(X_t - x_{t|T})(X_t - x_{t|T})^\top].$$

The ("fixed-interval") smoother consists of backward recursions for $t = T, \cdots, 1$:

$$\begin{aligned} x_{t|T} &= x_{t|t} + A_t(x_{t+1|T} - x_{t+1|t}), \\ V_{t|T} &= V_{t|t} + A_t(V_{t+1|T} - V_{t+1|t}) A_t^\top, \end{aligned} \tag{2.3.3}$$

with

$$A_t = V_{t|t} F_{t+1}^\top V_{t+1|t}^{-1}.$$

The derivatives of linear Kalman equations are presented in Appendix (**A.1**).

2.3.2 Extended Kalman Filter and smoother (EKF) recursions

In many practical applications, especially in engineering, most systems are nonlinear. Therefore, some attempts immediately applied the filtering methods to nonlinear systems. Most of this work was done at NASA Ames by McElhoe (1966) and Smith *et*

al. (1962). The EKF adapted techniques from calculus, namely multivariate *Taylor Series expansions*, to linearize a model about a working point.
If the system model is nonlinear as equations (2.2.3 and 2.2.4), and the observation and state noises from non-gaussian distributions, Monte Carlo methods as particle filters are employed for estimation. The extended Kalman filter on $p(X_t \mid Y_t)$ is approximated by a gaussian density. Sage's and Melsa's (1971) extended an approximate posterior mode in a nonlinear system from conditionally gaussian to exponential family observations. They derived an algorithm that depends on linearizing the non-linear functions by the Taylor expansion about the points recursively. Under the gaussian assumption, the optimal solution to the filtering problem is given by the *conditional* or *posterior mode*

$$x_{t|t} \coloneqq \mathrm{E}(X_t \mid Y_t)$$

of X_t given Y_t. Since the model is non-linear and gaussian, the *posterior distribution* of X_t is also approximately gaussian,

$$X_t \mid Y_t \sim \mathcal{N}(x_{t|t}; V_{t|t}),$$

with *posterior covariance matrix*

$$V_{t|t} \coloneqq \mathrm{E}[(X_t - x_{t|t})(X_t - x_{t|t})^\top].$$

The derivation of Extended Kalman filter and smoother algorithm is given in Appendix (**A.2**).

Initialization:

$$x_{0|0} = x_0, \quad V_{0|0} = R_0.$$

For $t = 1, \cdots, T$:

Prediction step:

$$\begin{aligned} x_{t|t-1} &= F(\mathbf{u}_t, X_{t-1|t-1}, \lambda) \\ V_{t|t-1} &= A_t V_{t-1|t-1} A_t^\top + C_t R_t C_t^\top, \end{aligned} \tag{2.3.4}$$

where

$$A_t = \frac{\partial F(\mathbf{u}_t, x, \gamma)}{\partial x} \mid_{x = x_{t-1|t-1}}$$
$$C_t = H_t(\mathbf{u}_t, X_{t-1|t-1}, \delta).$$

Correction or filtering step:

$$x_{t|t} = x_{t|t-1} + K_t(Y_t - G(\mathbf{u}_t, \widehat{x}_{t|t-1}, \lambda), \tag{2.3.5}$$
$$V_{t|t} = V_{t|t-1} - K_t B_t V_{t|t-1}, \tag{2.3.6}$$

where

$$B_t = \frac{\partial G(\mathbf{u}_t, x, \lambda)}{\partial x} \mid_{x = x_{t-1|t-1}} .$$

Kalman gain:

$$K_t = V_{t|t-1} B_t^\top [B_t V_{t|t-1} B_t^\top + \Sigma_t]^{-1}.$$

Smoothing step : EKF's fundamental idea is that linearizing the non-linear functions by the Taylor expansion about the points recursively, several attempts to linearizing the system non-linear implemented as Backward-Smoothing Extended Kalman Filter by Mark (2005). We found challenging to deriving the smoothing step of Extended Kalman Filter with equations (2.2.3 and 2.2.4).

2.3.3 General Filtering and smoothing

Nonlinear non-gaussian state space model stratifies the Bayesian perspective by using the conditional densities probability. The prediction, filtering, and smoothing distribution can be approached using the relations $p(X_t \mid X_{t-1}, \mathbf{Y}_{t-1}) = p(X_t \mid X_{t-1})$ and $p(Y_t \mid X_t, \mathbf{Y}_{t-1}) = p(Y_t \mid X_t)$, where $\mathbf{Y}_t = (Y_1, \cdots, Y_t)$. The derivation of recursive Bayesian for the one-step-ahead predictive distribution $p(X_t \mid \mathbf{Y}_{t-1})$ and the filtering distribution $p(X_t \mid \mathbf{Y}_t)$ is as follows

One-step-ahead prediction distribution

$$\begin{aligned} p(X_t \mid \mathbf{Y}_{t-1}) &= \int_{-\infty}^{\infty} p(X_t, X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} \\ &= \int_{-\infty}^{\infty} p(X_t \mid X_{t-1}, \mathbf{Y}_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} \\ &= \int_{-\infty}^{\infty} p(X_t \mid X_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1}. \end{aligned} \tag{2.3.7}$$

The formula (2.3.7) is an extension of the one-step-ahead prediction of the Kalman filter. Here, $p(X_t \mid X_{t-1})$ is the density function of the state X_t when the past state X_{t-1} is given, which is specified by the state equation (2.2.4). Consequently, if the filter $p(X_{t-1} \mid \mathbf{Y}_{t-1})$ of X_{t-1} is given, the one-step-ahead predictor $p(X_t \mid \mathbf{Y}_{t-1})$ can be estimated.

Filtering distribution

$$\begin{aligned} p(X_t \mid \mathbf{Y}_t) &= p(X_t \mid Y_t, \mathbf{Y}_{t-1}) \\ &= \frac{p(Y_t \mid X_t, \mathbf{Y}_{t-1}) p(X_t \mid \mathbf{Y}_{t-1})}{p(Y_t \mid \mathbf{Y}_{t-1})} \\ &= \frac{p(Y_t \mid X_t) P(X_t \mid \mathbf{Y}_{t-1})}{p(Y_t \mid \mathbf{Y}_{t-1})} \end{aligned} \tag{2.3.8}$$

where

$$p(Y_t \mid \mathbf{Y}_{t-1}) = \int_{-\infty}^{\infty} p(Y_t \mid X_t) p(X_t \mid \mathbf{Y}_{t-1}) dX_t.$$

In fact, the filter formula (2.3.8) is an extension of the filtering step of the Kalman filter. On other hand, $p(Y_t \mid X_t)$ is the conditional distribution of the observation Y_t, when the state X_t is given. It is established from the observation equation (2.2.3). Consequently, if the predictive distribution $p(X_t \mid \mathbf{Y}_{t-1})$ of X_t is given, then the filter density $p(X_t \mid \mathbf{Y}_t)$ is computable.
Now, the smoothing problem, by using the equation

$$p(X_t \mid X_{t+1}, \mathbf{Y}_T) = p(X_t \mid X_{t+1}, \mathbf{Y}_t),$$

with the state-space models of (2.2.3) and (2.2.4), then holds as follows

$$\begin{aligned} p(X_t, X_{t+1} \mid \mathbf{Y}_T) &= p(X_{t+1} \mid \mathbf{Y}_T) p(X_t \mid X_{t+1}, \mathbf{Y}_T) \\ &= p(X_{t+1} \mid \mathbf{Y}_T) p(X_t \mid X_{t+1}, \mathbf{Y}_t) \\ &= p(X_{t+1} \mid \mathbf{Y}_T) \frac{p(X_t \mid \mathbf{Y}_t) p(X_{t+1} \mid X_t, \mathbf{Y}_t)}{p(X_{t+1} \mid \mathbf{Y}_t)} \\ &= p(X_{t+1} \mid \mathbf{Y}_T) \frac{p(X_t \mid \mathbf{Y}_t) p(X_{t+1} \mid X_t)}{p(X_{t+1} \mid \mathbf{Y}_t)} \quad . \end{aligned} \tag{2.3.9}$$

Integrating both sides of (2.3.9), yields the following sequential formula for the smoothing problem:

The smoothing distribution

$$\begin{aligned} p(X_t \mid \mathbf{Y}_T) &= \int_{-\infty}^{\infty} p(X_t, X_{t+1} \mid \mathbf{Y}_T) dX_{t+1} \\ &= p(X_t \mid \mathbf{Y}_t) \int_{-\infty}^{\infty} \frac{p(X_{t+1} \mid \mathbf{Y}_T) p(X_{t+1} \mid X_t)}{p(X_{t+1} \mid \mathbf{Y}_t)} dX_{t+1}. \end{aligned} \tag{2.3.10}$$

In formula (2.3.10), $p(X_{t+1} \mid X_t)$ is specified by the state equation (2.2.4). On the other hand, $p(X_t \mid \mathbf{Y}_t)$ and $p(X_{t+1} \mid \mathbf{Y}_t)$ are calculated by equation (2.3.7) and (2.3.8), respectively. Thus, by the smoothing formula (2.3.10), if $p(X_{t+1} \mid \mathbf{Y}_T)$ is given, the smoothing density $p(X_t \mid \mathbf{Y}_T)$ can be computed. In contrast, $p(X_T \mid \mathbf{Y}_T)$ can be computed by filtering formula (2.3.8), by repeating the smoothing formula (2.3.10) for $t = T-1, \cdots, 1$, the smoothing distribution $p(X_{T-1} \mid \mathbf{Y}_T), \cdots, p(X_1 \mid \mathbf{Y}_T)$ is obtained successively.

2.3.4 The sequential Monte Carlo filter

The sequential Monte Carlo filter and smoother is also known as a *particle filter* or *bootstrap filter*. Gordon *et al.* (1993) used the Sequential Monte Carlo in advanced

Signal processing and Bayesian inference. Their work was considered as the first application of a Monte Carlo resampling algorithm in Bayesian statistical inference. The authors named their algorithm “the bootstrap filter”, and proved that compared to other filtering methods, their algorithm does not require any assumption about the state-space or the noise of the system. Several articles were published on a related “Monte Carlo filter”, or “ particle filters” by Kitagawa (1996), Del Moral (1996), and Carvalho, Del Moral, Monin and Salut (1997) .
The famous methods of Monte Carlo filter and smoother are:

1. The Particle Filter (PF).

2. The Auxiliary Particle Filters (APF).

3. The Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF).

2.3.4.1 Particle Filter (PF)

Kitagawa (1996), and Kitagawa and Gersch (1996) proposed an algorithm that can be applied to general nonlinear state space model. Their algorithm is based on the approximation of successive prediction and filtering density functions by many of their realizations. The difference between this algorithm and other Monte Carlo-Gibbs sampling methods as Metropolis *et al.* (1953), Hastings (1970) and Tierney (1994) is that the Monte Carlo method is used for the entire filtering and smoothing procedures, whereas the others are used only for numerical integration. The advantage of this algorithm is that it can be applied to an extensive class of nonlinear non-Gaussian of higher dimensional state space models. They succeeded in estimating the posterior density of the state variables given the observations by using their algorithm.
The sequential Monte Carlo method can characterize the true distribution by using many particles. Each particle is considered as a realization drawn from the true distribution.
The predictive, the filter and the smoothing distributions are approximated by m particles. The number of particles m is usually set between 1000 and $100,000$. The particles number is chosen based on the complexity of the distribution and the necessary accuracy. A true cumulative distribution function can be approximated by an empirical distribution function defined using m particles.
The particles $\{p_t^{(1)}, \cdots, p_t^{(m)}\}$ where m is the particles number, that follow the predictive distribution, are generated from the particles $\{f_t^{(1)}, \cdots, f_t^{(m)}\}$ used for the approximation of the filter distribution of the previous state. Then, the realizations $\{f_t^{(1)}, \ldots, f_t^{(m)}\}$ of the filter can be generated by re-sampling the realizations $\{p_t^{(1)}, \cdots, p_t^{(m)}\}$ of the predictive distribution. The table(2.1) shows the distributions and their approximations in the sequential Monte Carlo filter and smoother algorithm.

Table 2.1: ***The distributions and their approximations in the sequential Monte Carlo filter and smoother algorithm.***

Distributions	Density function	Approximations by particles
Predictive distribution	$p(X_t \mid \mathbf{Y}_{t-1})$	$\{p_t^{(1)}, \cdots, p_t^{(m)}\}$
Filter distribution	$p(X_t \mid \mathbf{Y}_t)$	$\{f_t^{(1)}, \cdots, f_t^{(m)}\}$
Smoothing distribution	$p(X_t \mid \mathbf{Y}_T)$	$\{s_{t\|T}^{(1)}, \cdots, s_{t\|T}^{(m)}\}$
Distribution of state noise	$p(\varepsilon_t)$	$\{\varepsilon_t^{(1)}, \cdots, \varepsilon_t^{(m)}\}$

One-step-ahead prediction

The particles $\{f_t^{(1)}, \cdots, f_t^{(m)}\}$ can be considered as realizations of the filter step, which can be generated from the filter distribution $p(X_{t-1} \mid \mathbf{Y}_{t-1})$ of the previous step X_{t-1} by m particles. The particles $\varepsilon_t^{(1)}, \cdots, \varepsilon_t^{(m)}$ can be considered as independent realizations of the state noise ε_t. For $j = 1, \cdots, m$, we can write

$$f_{t-1}^{(j)} \sim p(X_{t-1} \mid \mathbf{Y}_{t-1}), \quad \varepsilon_t^{(j)} \sim q(\varepsilon_t). \tag{2.3.11}$$

Therefore, $p_t^{(j)}$ is the jth particles defined by using the state equation (2.2.4)

$$p_t^{(j)} = F(\mathbf{u}_t^{(j)}, ft - 1^{(j)}, \gamma) + H(\mathbf{u}_t^{(j)}, f_{t-1}^{(j)}, \delta)\varepsilon_t^{(j)}, \tag{2.3.12}$$

where $\{p_t^{(1)}, \cdots, p_t^{(m)}\}$ can be considered as independent realization of the one step ahead predictor distribution of the state X_t.

Filtering

In the filtering step, the Bayes factor (or likelihood) $\alpha_t^{(j)}$ of the particle $p_t^{(j)}$ can be computed with respect to the observation $Y_t(t)$. For $j = 1, \cdots, m$,

$$\alpha_t^{(j)} = r_t(S_t(Y_t, p_t^{(j)})) \times \frac{\partial S_t}{\partial Y_t}, \tag{2.3.13}$$

where S_t the inverse function of G_t in the Non-linear observation equation (2.2.3), and r_t is the probability density function of the observation noise ξ_t. Here $\alpha_t^{(j)}$ can be considered as a weighting factor representing the importance of the particle $p_t^{(j)}$. Then, m particles $f_t^{(1)}, \cdots, f_t^{(m)}$ are obtained by re-sampling $p_t^{(1)}, \cdots, p_t^{(m)}$ with probabilities proportional to the "likelihoods" $\alpha_t^{(1)}, \cdots, \alpha_t^{(m)}$. That is, a new particle $f_t^{(j)}, j = 1, \cdots, m$ is obtained according to

$$f_t^{(j)} = \begin{cases} p_t^{(1)} & \text{probability} \quad \alpha_t^{(1)}/(\alpha_t^{(1)} + \cdots + \alpha_t^{(m)}) \\ \vdots & \\ p_t^{(m)} & \text{probability} \quad \alpha_t^{(m)}/(\alpha_t^{(1)} + \cdots + \alpha_t^{(m)}). \end{cases} \tag{2.3.14}$$

Then, $f_t^{(1)}, \cdots, f_t^{(m)}$ can be considered as the realizations generated from the filter distribution $p(X_t \mid \mathbf{Y}_t)$. The derivation of particle filter algorithm is presented in Appendix (**A.3**).

Algorithm for the Monte Carlo filter

Kitagawa (1996) described the Monte Carlo filter algorithm as follows:

1. A k-dimensional random number Generate from $f_0^{(j)} \sim p_0(X)$ for $j = 1, \cdots, m$. k is the dimensional for state vector X_t
2. For $t = 1, \cdots, T$, repeat the following steps

(a) Generate ℓ-dimensional random numbers $\varepsilon_t^{(j)} \sim q(\varepsilon_t),\ j = 1, \cdots, m$.

(b) Find the new particles: for $j = 1, \cdots, m$.

$$p_t^{(j)} = F(\mathbf{u}_t^{(j)}, f_{t-1}^{(j)}, \gamma) + H(\mathbf{u}_t^{(j)}, f_{t-1}^{(j)}, \delta)\varepsilon_t^{(j)},$$

(c) Calculate the Bayes factor : for $j = 1, \cdots, m$.

$$\alpha_t^{(j)} = r_t(S_t(Y_t, p_t^{(j)})) \times \frac{\partial S_t}{\partial Y_t},$$

(d) Generate $\{f_t^{(1)}, \cdots, f_t^{(m)}\}$ by applying re-sampling method (sampling with replacement) m times from $\{p_t^{(1)}, \cdots, p_t^{(m)}\}$ with probabilities proportional to $\{\alpha_t^{(1)}, \cdots, \alpha_t^{(m)}\}$.

Re-sampling method

The re-sampling method is essentially based on random sampling as described below For $j = 1, \cdots, m$, repeat the following step (a) -(c).

(a) Generate uniform random number $z_t^{(j)} \in U[0; 1]$.

(b) Search for i that achieves

$$C^{-1} \sum_{l=1}^{i-1} \alpha_t^{(l)} < z_t^{(j)} \leq C^{-1} \sum_{l=1}^{i} \alpha_t^{(l)}$$

where $C = \sum_{l=1}^{m} \alpha_t^{(l)}$.

(c) Get a particle which approximates the filter by setting $f_t^{(j)} = p_t^{(j)}$.

The objective of re-sampling is to re-express the distribution function specified by the particles $\{p_t^{(1)}, \cdots, p_t^{(m)}\}$ with weights $\{\alpha_t^{(1)}, \cdots, \alpha_t^{(m)}\}$ through representing the empirical distribution function determined by re-sampled particles with equal weights. As consequence, it is not essential to perform exact random sampling.

Monte Carlo smoothing method

The Monte Carlo smoothing method is created by extending the Monte Carlo filtering method through preserving past particles. In the following, the vector of the particles $(s^{(j)}_{1|t}, \cdots, s^{(j)}_{t|t})$ indicates the $j-$th realization from the t-dimensional joint distribution function $p(X_1, \cdots, X_t \mid \mathbf{Y}_t)$.
It should be only modifying step 2(d) of the Monte Carlo filtering algorithm to achieve the objective of smoothing as follows.

(d-S) For $j = 1, \cdots, m$ by re-sampling the t-dimensional vector $(s^{(j)}_{1|t-1}, \cdots, s^{(j)}_{t|t-1}, p^{(j)}_t)^\top$, generate $(s^{(j)}_{1|t}, \cdots, s^{(j)}_{t|t}, s^{(j)}_{t|t})^\top$.

In this modification, by re-sampling $\{(s^{(j)}_{1|t-1}, \cdots, s^{(j)}_{t-1|t-1}, p^{(j)}_t)^\top,\ j = 1, \cdots, m\}$ with the same weights as in step $2-d$, fixed interval smoothing for non-linear non-Gaussian state space model can be achieved (Kitagawa(1996)). In practical computation, since the re-sampling methods repeated a finite number of particles (m particles), the number of different particles progressively decreases and then the weights become concentrated on a small number of particles. As result, the shape of distribution finally collapses. Accordingly, in the smoothing algorithm, it is necessarily the step $(d-S)$ performs as follows:

(d-L) For $j = 1, \cdots, m$, generate $(s^{(j)}_{t-L|t}, \cdots, s^{(j)}_{t-1|t}, s^{(j)}_{t|t})$. Here, L is a fixed integer, usually less than 30, and assume $f^{(j)}_t = s^{(j)}_{t|t}$ by re-sampling $(s^{(j)}_{t-L|t-1}, \cdots, s^{(j)}_{t-1|t-1}, p^{(j)}_t)$. Then, the modification to algorithm turns out to conform to the L-lag fixed-lag smoothing algorithm. If L is fixed and large, the fixed interval smoother $p(X_t \mid \mathbf{Y}_T)$ is approximated by the fixed-lag smoother $p(X_t \mid \mathbf{Y}_{t+L})$. Furthermore, the distribution is determined by $X^1_{t|t+L}, \cdots, X^m_{t|t+L}$ not distant from $p(X_t \mid \mathbf{Y}_{t+L})$. Subsequently, L should be taken not very large, i.e., $L = 20$ or 30.

2.3.4.2 The Auxiliary Particle Filters (APF)

Pitt and Shephard (1999) proposed the Auxiliary Particle Filters (APF). They extended standard particle filtering methods by adding an auxiliary variable which permits the particle filter to be adapted more efficiently. The work of the auxiliary variable ς, is only to improve a simulation performance. The essential idea of the APF is to insert an auxiliary variable ς. It plays an important role of index of the mixture component. The augmented joint distribution $p(X_t, \varsigma \mid \mathbf{Y}_t)$ with this additionally auxiliary variable is updated as :

$$\begin{aligned} p(X_t, \varsigma = m \mid \mathbf{Y}_t) &\propto p(\mathbf{Y}_t \mid X_t) p(X_t, \varsigma = m \mid \mathbf{Y}_t) \\ &= p(\mathbf{Y}_t \mid X_t) p(X_t \mid \varsigma = m, \mathbf{Y}_t) \\ &\quad \times p(m \mid \mathbf{Y}_t). \\ &= p(\mathbf{Y}_t \mid X_t) p(X_t \mid X^m_{t-1}) w^m_{t-1}, \end{aligned} \tag{2.3.15}$$

where $y \propto x$ means that $y = kx$ for some constant k. Thus, the APF generates a sample $\{X_t^m, \varsigma^m\}_{m=1}^N$ from the joint distribution $p(X_t, \varsigma \mid \mathbf{Y}_t)$, where $\varsigma^m(\varsigma^m \equiv \{\varsigma = m\})$ refers to the index of particle $m's$ parent. The importance density is used in APF as follows

$$q(X_t, \varsigma \mid \mathbf{Y}_t) \quad \propto \quad p(\mathbf{Y}_t \mid \mu_t^m) p(X_t \mid X_{t-1}^m) w_{t-1}^m, \tag{2.3.16}$$

where μ_t^m is some description of X_t given X_{t-1}^m, such as the mean, mode or another statistics. If it is the mean, then $\mu_t^m = \mathrm{E}(X_t \mid X_{t-1}^m)$ or a sample $\mu_t^m \sim p(X_t \mid X_{t-1}^m)$. Then, by forgetting the auxiliary variable, the sample $\{X_t^m\}_{m=1}^N$ is simply obtained. The importance density is written as

$$q(X_t, \varsigma \mid \mathbf{Y}_t) = q(\varsigma \mid \mathbf{Y}_t) q(X_t \mid \varsigma, \mathbf{Y}_t). \tag{2.3.17}$$

Then, defining

$$q(X_t \mid \varsigma^m, \mathbf{Y}_t) \quad := \quad p(X_t \mid X_{t-1}^m), \tag{2.3.18}$$

substituting Eqs. (2.3.17) and (2.3.18) into Eq. (2.3.16), the formula is :

$$q(\varsigma^m \mid \mathbf{Y}_t) \quad \propto \quad p(Y_t \mid \mu_t^m) w_{t-1}^m. \tag{2.3.19}$$

Recursively, the importance weights are updated as

$$w_t^m \propto \frac{p[X_t^m, \varsigma^m \mid \mathbf{Y}_t]}{q[X_t^m, \varsigma^m \mid \mathbf{Y}_t]}. \tag{2.3.20}$$

Substituting Eqs.(2.3.15) and (2.3.16) into Eq.(2.3.20) yields

$$\begin{aligned} w_t^m \quad &\propto \quad \frac{p(\mathbf{Y}_t \mid X_t^m) p(X_t \mid X_{t-1}^m) w_{t-1}^m}{p(\mathbf{Y}_t \mid \mu_t^{\varsigma^m}) p(X_t \mid X_{t-1}^m) w_{t-1}^m} \\ &= \quad \frac{p[\mathbf{Y}_t \mid X_t^m]}{p[\mathbf{Y}_t \mid \mu_t^{\varsigma^m}]}. \end{aligned} \tag{2.3.21}$$

Importance density: Importance sampling is favourite among the variance reduction techniques that can be used with complex distributions when Monte Carlo sampling is used. The necessary procedure in importance sampling is to choose a density function which "encourages" the important values. If the importance density is applied primary in the simulation, the result is a biased estimator. However, the new importance sampling estimator is unbiased due to the fact that the outputs of a simulation are weighted to correct for the use of the biased distribution. On other words, the importance weight is given by $w_t = \frac{p(x)}{q(x)}$, which is called the likelihood ratio. $q(x), p(x)$ are known as the importance density and the nominal density respectively. The good choice of importance density leads to minimize the variance of importance weights (Kroese and Rubinstein (2012)).

Auxiliary Particle Filter (APF) Algorithm

The APF algorithm can be performed as follows:

1. Initialization $(t = 0):$ for $m = 1, \cdots, N$, generate the states (particles) X_0^m from the prior $p(X_0)$, set
$$\widehat{x}_0^m = E[X_0^m] = F(\mathbf{u}_0^m, x_0^m, \delta^m)$$
and
$$\begin{aligned} V_0^m &= \mathrm{E}[(X_0^m(0) - \widehat{x}_0^m)(X_0^m - \widehat{x}_0^m)^\top] = \mathrm{Var}(X_0^m) \\ &= H^2(\mathbf{u}_0^m, x_0^m, \delta^m) R_0. \end{aligned}$$
For $t = 1, \cdots, T$, repeat the following steps:

2. For $m = 1, \cdots, N$, generate $\mu_t^m \sim p(X_t \mid X_{t-1})$.

3. For $m = 1, \cdots N$, compute
$$\begin{aligned} w_t^m &= q(m \mid \mathbf{Y}_t) \\ &\propto p(Y_t \mid \mu_t^m) w_{t-1}^m. \end{aligned}$$

4. Resample to obtain the index ς^m of particle m's parent.

5. For $m = 1, \cdots, N$, generate
$$X_t^m \sim p(X_t^m \mid X_{t-1}^m, \varsigma^m).$$

6. Update the second-stage weights
$$w_t^m = \frac{p[\mathbf{Y}_t \mid X_t^m]}{p[\mathbf{Y}_t \mid \mu_t^{\varsigma^m}]}.$$

- Normalize the weights to avoid the Degeneracy problem is as follows
$$w_t^m = \frac{w_t^m}{\sum_{m=1}^N w_t^m}.$$

7. Output: a set of weighted particles (samples) $[\{X_t^m, w_t^m\}_{m=1}^N]$.

2.3.4.3 The Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF)

Xi *et al.* (2015) proposed a new particle filter called the auxiliary iterated extended Kalman particle filter (AIEKPF). The basic idea of the AIEKPF algorithm is the use of the iterated extended Kalman filter (IEKF) algorithm within an APF algorithm to generate the importance density function. The proposed approach uses the iterated extended Kalman filter (IEKF) to merge the last measurements into state distribution. The robustness of the APF and the importance density generated by the IEKF leads to the new filter that can match the posterior distribution well.
We adapted this algorithm to the non-linear non-gaussian models defined by (2.2.3,2.2.4) by deriving the equations of extended Kalman filtering. The derivation of equations for extended Kalman filtering recursions is presented in the Appendix (**A.2**).
Let $\{\mathbf{X}_t^m, m = 1, 2, \cdots, N\}$ be a set of support points with corresponding weights $\{w_t^m, m = 1, 2, \cdots, N\}$.

Remark: In the state space representation, the posterior densities $\{p(x_{1:t} \mid y_{1:t}), t \geq 1\}$ are sequentially approximated by the Sequential Monte Carlo methods (SMC). The posterior distributions that are approximated by a set of N weighted random sample called particles, for $t \geq 1$ leads to the approximation equation,

$$p(X_t \mid Y_{t-1}) \approx \sum_{m=1}^{N} w_{t|t-1}^{(m)} \delta(X_t - X_t^{(i)}),$$

where $\delta(.)$ indicates the Dirac delta function. The posterior mean are approximated by

$$\widehat{X}_{t|t-1} \approx \sum_{m=1}^{N} w_{t|t-1}^{(m)} X_t^{(i)}.$$

Then, the posterior distribution can be approximated as

$$\widehat{p}(\mathbf{X}_t \mid \mathbf{Y}_t) \approx \sum_{m=1}^{N} w_t^m \delta(\mathbf{X}_t - \mathbf{x}_t) \tag{2.3.22}$$

where $\mathbf{x}_t = (x_1^m, \cdots, x_t^m)$ is a set of particles, the weights are normalized such that $\sum_{m=1}^{N} w_t^m = 1$. The weight w_t^m is chosen using the precept of importance sampling, which is given by

$$w_t^m(t) \propto \frac{p[\mathbf{X}_t^m \mid \mathbf{Y}_t]}{q[\mathbf{X}_t^m \mid \mathbf{Y}_t]}, \tag{2.3.23}$$

where $q[\mathbf{X}_t^m \mid \mathbf{Y}_t]$ is known as the importance density. As we mentioned earlier in APF algorithm the auxiliary variable, ς, played a role of index of the mixture component.

The augmented joint distribution $p(X_t, \varsigma \mid \mathbf{Y}_t)$ is updated as:

$$\begin{aligned} p(X_t, \varsigma = m \mid \mathbf{Y}_t) &\propto p(Y_t \mid X_t) p(X_t, \varsigma = m \mid \mathbf{Y}_t) \\ &= p(Y_t \mid X_t) p(X_t \mid \varsigma = m, \mathbf{Y}_t) \\ &\quad \times p(m \mid \mathbf{Y}_t). \\ &= p(Y_t \mid X_t) p(X_t \mid X_{t-1}^m) w_{t-1}^m. \end{aligned} \tag{2.3.24}$$

Thus, the importance density is

$$q(X_t, \varsigma \mid \mathbf{Y}_t) \propto p(Y_t \mid \mu_t^m) p(X_t \mid X_{t-1}^m) w_{t-1}^m, \tag{2.3.25}$$

where μ_t^m is mean, mode or another statistic. Then, by deleting the auxiliary variable, the sample $\{X_t^m\}_{m=1}^N$ is obtained. The importance density yields

$$q(X_t, \varsigma \mid \mathbf{Y}_t) = q(\varsigma \mid \mathbf{Y}_t) q(X_t \mid \varsigma, \mathbf{Y}_t). \tag{2.3.26}$$

Then, defining

$$q(X_t \mid \varsigma^m, \mathbf{Y}_t) = p(X_t \mid X_{t-1}^m), \tag{2.3.27}$$

substituting Eqs. (2.3.26) and (2.3.27) into Eq. (2.3.25), the formula is :

$$q(\varsigma^m \mid \mathbf{Y}_t) \propto p(Y_t \mid \mu_t^m) w_{t-1}^m. \tag{2.3.28}$$

Recursively, the importance weights are updated as

$$w_t^m \propto \frac{p[X_t^m, \varsigma^m \mid \mathbf{Y}_t]}{q[X_t^m, \varsigma^m \mid \mathbf{Y}_t]}. \tag{2.3.29}$$

Substituting Eqs.(2.3.24) and (2.3.25) into Eq.(2.3.29), the importance weights are defined as

$$\begin{aligned} w_t^m &\propto \frac{p(Y_t \mid X_t^m) p(X_t \mid X_{t-1}^m) w_{t-1}^m}{p(Y_t \mid \mu_t^{\varsigma^m}) p(X_t \mid X_{t-1}^m) w_{t-1}^m} \\ &= \frac{p[Y_t \mid X_t^m]}{p[Y_t \mid \mu_t^{\varsigma^m}]}. \end{aligned} \tag{2.3.30}$$

Occasionally, the likelihood function is also narrow compared to the prior distribution, or it happens to lie in the tails of the prior distribution, then the posterior distribution is poorly approximated by the importance density. To devise the optimal importance density, the IEKF is used to update particles by blending the most current observation with the optimal gaussian approximation of the state. In other words, the IEKF computes the following recursive approximation to the real posterior filtering density

$$p(X_t \mid \mathbf{Y}_t) = \mathcal{N}\left(\widehat{X}_{tj}; V_{tj}\right) \tag{2.3.31}$$

where $t = 0, \cdots, T, j = 1, \cdots, c$ and c is the number of the Iterated Extended Kalman Filter (IEKF), $\widehat{X}_{tj}$ is the iterative value of the X_t on the jth iteration, and V_{tj} is the covariance matrix of $\widehat{X}_{tj}$. The derivation of equations to iterated extended Kalman filter for the non-linear non-gaussian as equations (2.2.3, 2.2.4) in Appendix **A.2**. According to the IEKF update equation, $\widehat{X}_{tj}$ and V_{tj} can be updated as follows :

$$\begin{aligned} \widehat{X}_{t|t,j} &= \widehat{X}_{t|t-1,j} + K_{tj}[Y_t - G(\mathbf{u}_t, X_t, \lambda)] \qquad (2.3.32) \\ K_{tj} &= V_{t|t-1,j}B_{tj}^{\top}[B_{tj}V_{t|t-1,j}B_{tj} + \Sigma_t]^{-1} \\ V_{t|t,j} &= (I - K_{tj}B_{tj})V_{t|t-1,j} \qquad (2.3.33) \end{aligned}$$

where B_{tj} is the Jacobian of G_t, the function appearing in the observation equation (2.2.3), and K_{tj} is a Kalman gain matrix. Within the APF algorithm, the importance density is generated for each particle by the IEKF algorithm as follows

$$q(X_t^m, \varsigma \mid \mathbf{Y}_t) = \mathcal{N}\left(\widehat{X}_{tj}^{\varsigma m}; V_{tj}^{\varsigma m}\right) \qquad (2.3.34)$$

where $m = 1, 2, \cdots N$, and N is the number of the particles.
In short, the AIEKPF uses the IEKF to update the equations with the new observations to compute the mean and covariance of the importance density for each particle at time $t-1$. Next, the mth particle is sampled from the distribution.

Auxiliary Iterated Extended Kalman Filter (AIEKPF) Algorithm

The proposed AIEKPF algorithm can be performed by the following steps:

1. Initialization ($t = 0$) : For $m = 1, \cdots, N$, generate the states (particles) X_0^m from the prior $p(X_0)$, and set

   $$\widehat{x}_0^m = \mathrm{E}(\mathbf{u}_0^m, X_0^m) = F(x_0^m, \delta^m)$$

   and

   $$\begin{aligned} V_0^m &= \mathrm{E}[(X_0^m(0) - \widehat{x}_0^m)(X_0^m - \widehat{x}_0^m)^{\top}] = \mathrm{Var}(X_0^m) \\ &= H^2(\mathbf{u}_0^m, x0^m, \delta^m)R_0 \end{aligned}$$

   For $t = 1, \cdots, T$, repeat the following steps :

2. For $m = 1, \cdots, N$, generate $\mu_t^m \sim p(X_t \mid X_{t-1})$.

3. For $m = 1, \cdots, N$, using the IEKF algorithm to update the particles

   - Calculate the Jacobians A_t^m, C_t^m of the process model

   $$A_t^m = \frac{\partial F(\mathbf{u}_t, x, \gamma)}{\partial x} |_{x=x_{t-1|t-1}^m}$$
   $$C_t^m = H(\mathbf{u}_t, x_{t-1|t-1}^m, \delta)$$

- Using the IEKF to predict the particle :

$$x^m_{t|t-1} \approx F(\mathbf{u}_t, x^m_{t-1|t-1}, \gamma)$$

$$V^m_{t|t-1} = A^m_t V^m_{t-1|t-1} A^{m\top}_t + C^m_t R_t C^{m\top}_t$$

- For $j = 1, \cdots, c$ (c is the iteration number of the IEKF)
  a - Compute the Jacobian B^m_{tj} of G_t function in observation equation

$$B^m_{tj} = \frac{\partial G(\mathbf{u}_t, x, \lambda)}{\partial x} |_{x=x^m_{t|t-1,j}}$$

  b - Update the state estimation error covariance V^m_{tj} by Eq.(2.3.33)
  c - Update the state estimate X^m_{tj} by Eq.(2.3.32).

4. For $m = 1, \cdots N$, Compute

$$\begin{aligned} w^m_t &= q(m \mid \mathbf{Y}_t) \\ &\propto p(Y_t \mid X_t) \times w^m_{t-1}, \end{aligned} \tag{2.3.35}$$

5. Resample to find the index ς^m of particle m's parent.

6. Find the importance sampling : for $m = 1, \cdots, N$,

   - Draw samples

$$\begin{aligned} X^m_t &\sim q(X_t, \varsigma^m \mid \mathbf{Y}_t) \\ &= \mathcal{N}(\widehat{X}^{\varsigma^m}_{tj}; V^{\varsigma^m}_{tj}),\ j = c, \end{aligned} \tag{2.3.36}$$

   we recall c is the iteration number of the IEKF.

   - Compute the importance weights of particles as follows

$$w^m_t = \frac{p(\mathbf{Y}_t \mid X^m_t)}{p(\mathbf{Y}_t \mid \mu^{\varsigma^m}_t)} \tag{2.3.37}$$

   - Normalize the weights to avoid the Degeneracy problem is as follows

$$w^m_t = \frac{w^m_t}{\sum_{m=1}^N w^m_t}.$$

7. Output: a set of weighted particles (samples) $[\{X^m_t, w^m_t\}^N_{m=1}]$.

Remark: practically, the iteration to perform the equations (2.3.36, 2.3.37) leads to *Degeneracy problem.* On other words, all the particles will have negligible weights but a few of the particles will have a significant weight. The term effective sample size is used to measure the degeneracy problem as follows:

$$N_{eff} = \frac{1}{\sum_{m=1}^{N} (w_t^m)^2}$$

where a smaller N_{eff} means a larger variance for the weights. One of the techniques to solve this problem is *Resampling* with replacement N particles $\{\widetilde{X}_t^m, m = 1, \cdots, N\}$ from the set $\{X_t^m, m = 1, \cdots, N\}$. The importance weights can be normalized as follows :

$$w_t^m = \frac{w_t^m}{\sum_{m=1}^{N} w_t^m}.$$

2.3.4.4 Optimality of the Auxiliary Iterated Extended Kalman Particle Filter(AIEKPF) Algorithm

Johansen and Doucet (2008) proposed a novel interpretation of the APF, which allows them to present the first convergence results for APF algorithm. They defined the *consistency* and *asymptotic normality* of a weighted particles (samples) $[\{X_t^m, w_t^m\}_{m=1}^N]$.

Definition 3.1 (Asymptotic normality). Under the regularity condition given in [Chopin (2004), Theorm 1] or [Del Moral, 2004, Section 9.4, pp.300-306], one has the convergence in distribution

$$\sqrt{N}(\widehat{\varphi}_t - \bar{\varphi}_t) \Rightarrow \mathcal{N}(0; \sigma_{APF}^2(\varphi_t)), \tag{2.3.38}$$

where φ_t is the function of trajectories X_t and $\bar{\varphi}_t$ is the expectation of $\widehat{\varphi}_t$ with respect to the filtering distribution.

$$\bar{\varphi}_t = \int \varphi_t p(X_t \mid \mathbf{Y}_t) dX_t$$

and

$$\widehat{\varphi}_t = \sum_{m=1}^{N} W_t^m \varphi_t^m, \tag{2.3.39}$$

where $W_t^m = w_t^m [\sum_{m=1}^{N} w_t^m]^{-1}$, and

$$w_t^m = \frac{p[\mathbf{Y}_t \mid X_t^m]}{p[\mathbf{Y}_t \mid \mu_t^{\varsigma_t^m}]}. \tag{2.3.40}$$

At time $t = 1$, obtain

$$\sigma^2_{APF}(\varphi_1) = \int \frac{(p(X_1 \mid \mathbf{Y}_1)^2}{q_1(X_1)}(\varphi_1 - \bar{\varphi}_1)^2 dX_1, \tag{2.3.41}$$

whereas for $t > 1$,

$$\begin{aligned}\sigma^2_{APF}(\varphi_t) &= \int \frac{[p(X_1 \mid \mathbf{Y}_t)]^2}{q_1(X_1)} \vartriangle \varphi^2_{1,t} dX_1 \\ &+ \sum_{r=2}^{t-1} \int \frac{[p(X_{1:r} \mid \mathbf{Y}_t)]^2}{\widehat{p}(X_{r-1} \mid \mathbf{Y}_r) q_r(X_r \mid X_{r-1})} \\ &\quad \times \vartriangle \varphi^2_{r,t} dX_r \\ &+ \int \frac{[p(X_t \mid \mathbf{Y}_t]^2}{\widehat{p}(X_{t-1} \mid \mathbf{Y}_t) q_t(X_t \mid X_{t-1})} \\ &\quad \times (\varphi_t - \bar{\varphi}_t)^2 dX_t. \end{aligned} \tag{2.3.42}$$

where

$$\vartriangle \varphi_{r,t} = \int \varphi_t p(X_{r+1:t} \mid \mathbf{Y}_{r+1:t}) dX_{r+1:t} - \bar{\varphi}_t$$

2.3.5 Maximum a posteriori estimation (MAP)

The Bayesian approach certainly leads to compute the posterior distribution by using the particle filtering methods. If the posterior distribution is gaussian, the point that maximizes the posterior distribution is posterior mean. In contrast, the mode maximize the non-gaussian posterior distribution which is called maximum a posteriori probability (MAP).

$$X_t^{MAP} = \arg\max_{X_t} p(X_t \mid \mathbf{Y}_{t-1}), t = 1, \cdots, T. \tag{2.3.43}$$

The MAP can be used to compute an estimation of state variable X_t. We present the MAP via three methods:

1. Viterbi algorithm.
2. Particle filter algorithm.
3. Auxiliary Iterated Extended Kalman Filter (AIEKPF) algorithm.

2.3.5.1 Particle filters MAP via the Viterbi algorithm

Godsill *et al.*(2000) developed a technique which was implemented by using maximum a posteriori (MAP) sequence estimation in non-linear non-gaussian dynamic models. They used Monte Carlo filtering and Viterbi algorithm, and focused on estimating the MAP sequence as follows $\mathbf{X}_t^{MAP} = (X_1^{MAP}, \cdots, X_t^{MAP})^\top$

$$\begin{aligned}\mathbf{X}_t^{MAP} &\cong \arg\max_{\mathbf{X}_t} \left(p(\mathbf{X}_t \mid \mathbf{Y}_t)\right) \\ &= \arg\max_{\mathbf{X}_t} \left(\frac{p(\mathbf{Y}_t \mid \mathbf{X}_t).p(\mathbf{X}_t \mid \mathbf{Y}_{t-1})}{p(\mathbf{Y}_t \mid \mathbf{Y}_{t-1})}\right) \qquad (2.3.44)\end{aligned}$$

The dependency upon $\mathbf{Y}_t$ is indicated by t. Note that $p(\mathbf{Y}_t \mid \mathbf{Y}_{t-1})$ does not depend on the value of $\mathbf{X}_t$. Consequently, it can be equivalently written

$$\mathbf{X}_t^{MAP} \cong \arg\max_{\mathbf{X}_t} \left[p(\mathbf{Y}_t \mid \mathbf{X}_t).p(\mathbf{X}_t \mid \mathbf{Y}_{t-1})\right] \qquad (2.3.45)$$

practically, computing $p(X_t \mid \mathbf{Y}_{t-1})$ can only be executed in closed form for linear gaussian models by using the Kalman filter-smoother and for finite state space hidden Markov models. In contrast, the numerical techniques must be employed to approximate the posterior distribution, such as the extended Kalman filter, gaussian sum methods and general numerical integration procedures Monte- Carlo filtering (particle filtering) as the methods proposed by Kitagawa (1987), Kitagawa (1996), Doucet *et al.* (2000a) and Doucet *et al.* (2000b). Then, by Bayes' rule the predictive distribution is as follows

$$\begin{aligned} p(X_t \mid \mathbf{Y}_{t-1}) &= \int p(X_t, X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} \\ &= \int p(X_t \mid X_{t-1}, \mathbf{Y}_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} \\ &= \int p(X_t \mid X_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1}, \qquad (2.3.46)\end{aligned}$$

the approximation can be made at time t as follows

$$\widehat{p}(\mathbf{X}_t \mid \mathbf{Y}_t) \approx \sum_{m=1}^{N} w_t^{(m)} \delta(\mathbf{X}_t - \mathbf{x}_t^m), \qquad (2.3.47)$$

where δ indicates the Dirac delta function and $w_t^{(m)}$ is the weight joined to particle $\mathbf{x}_t^{(m)}, w_t^{(m)} \geq 0$ and

$$\sum_{m=1}^{N} w_t^{(m)} = 1.$$

The approximation of $p(\mathbf{X}_t \mid \mathbf{Y}_{t-1})$ is as follows

$$\widehat{p}(\mathbf{X}_t \mid \mathbf{Y}_{t-1}) \approx \sum_{j}^{N} p(X_t \mid X_{t-1}^j) w_{t-1}^j. \qquad (2.3.48)$$

The Viterbi algorithm calculate the MAP as

$$\mathbf{X}_t^{MAP} \cong \arg\max_{X_t} \sum_{k=1}^{t} [\log p(Y_k \mid X_k) + \log p(X_k \mid X_{k-1})]. \tag{2.3.49}$$

The Viterbi algorithm is a famous technique to estimating discrete state-space hidden Markov models. Godsill *et al.* (2000) developed this algorithm to estimating continuous state-space Markov model via a discrete approximation of the state space as particles. The Viterbi algorithm is summarized as follows:

Viterbi algorithm

Initialization : For $1 \leq m \leq N$

$$\delta_1(m) = \log p(X_1^{(m)} \mid X_0^{(m)}) + \log p(Y_1 \mid X_1^{(m)}) \tag{2.3.50}$$

Recursion: For $2 \leq k \leq t$ and $1 \leq j \leq N$

$$\begin{aligned} \delta_k(j) &= \log p(Y_k \mid X_k^{(j)}) + \max_m \left[\delta_{k-1}(m) + \log p(X_k^{(j)} \mid X_{k-1}^{(m)}) \right] \\ \psi_k(j) &= \arg\max_m \left[\delta_{k-1}(m) + \log p(X_k^{(j)} \mid X_{k-1}^{(m)}) \right] \end{aligned} \tag{2.3.51}$$

Termination:

$$\begin{aligned} m_t &= \arg\max_m \delta_t(m) \\ \widehat{X}_t^{MAP} &= X_t^{m_t} \end{aligned} \tag{2.3.52}$$

Backtracking: For $k = t-1, t-2, \cdots, 1$

$$\begin{aligned} m_k &= \psi_{k+1}(m_{k+1}) \\ \widehat{X}_k^{MAP} &= X_k^{(m_k)} \end{aligned} \tag{2.3.53}$$

2.3.5.2 The maximum a posteriori (MAP) with particle filtering and smoothing algorithm

Saha *et al.* (2008) proposed a method to estimating MAP to a general state space model, in which they utilised gradient based optimization method. In this method, the MAP locate can be approximated by computing the posterior (filtering) density at the predicted particle $\{x_t^{(m)}\}_{m=1}^N$, and choosing the particle with the highest density. One of the advantages of this method is that the posterior density is approximate to any specific point, in addition to be approximate to particles forming the clouds.

They compared the particle filters using the Viterbi algorithm and their method (pf-MAP) via numerical simulations. They observed that their method was better, as an estimate of current state space. They concluded that their results were not certainly similar to the Viterbi algorithm results. Furthermore, the Viterbi algorithm goal is maximization of $p(\mathbf{x}_t \mid \mathbf{y}_t)$, whereas the pf-MAP goal is maximization of $p(x_t \mid \mathbf{y}_t)$. They noted that the two estimates would be principally similar in the linear-gaussian state space models. In contrast, the difference would certainly appear for the above two estimators in nonlinear and/or non-gaussian state space models. They investigated the behaviours of the two algorithms through numerical simulations. First, they studied the linear gaussian state space model and compared these two estimators. The criteria is used a root mean square error estimate (RMSE) to a true MAP. The RMSE's for the two methods behaved similarly. Second, they took the nonlinear-gaussian state space model, three different initial values for estimating the RMSE are used. They observed through RMSE estimate that the pf-MAP implements better as an estimate of current state whereas it is also computationally much easy. The filtering and smoothing steps to find the MAP as follows:

Filtering step

The MAP estimate of the filtering density at time t is defined as

$$X_t^{MAP} = \arg\max_{X_t} p(X_t \mid \mathbf{Y}_t), \tag{2.3.54}$$

where $\mathbf{Y}_t = (Y_1, \cdots, Y_t)^\top$ and $\mathbf{X}_t = (X_1, \cdots, X_t)^\top$. By using the Baye's rule, the posterior (filtering) density can be written as

$$\begin{aligned} p(X_t \mid \mathbf{Y}_t) &= p(X_t \mid Y_t, \mathbf{Y}_{t-1}) \\ &= \frac{p(Y_t \mid X_t, \mathbf{Y}_{t-1}) p(X_t \mid \mathbf{Y}_{t-1})}{p(Y_t \mid \mathbf{Y}_{t-1})} \\ &= \frac{p(Y_t \mid X_t) p(X_t \mid \mathbf{Y}_{t-1})}{p(Y_t \mid \mathbf{Y}_{t-1})} \end{aligned} \tag{2.3.55}$$

Therefore, the MAP estimate of the posterior distribution is

$$X_t^{MAP} = \arg\max_{X_t} \left[\frac{p(Y_t \mid X_t) p(X_t \mid \mathbf{Y}_{t-1})}{p(Y_t \mid \mathbf{Y}_{t-1})} \right], \tag{2.3.56}$$

the denominator $p(Y_t \mid \mathbf{Y}_{t-1})$ is independent of X_t, the MAP estimate yields

$$X_t^{MAP} \simeq \arg\max_{X_t} \left[p(Y_t \mid X_t) p(X_t \mid \mathbf{Y}_{t-1}) \right]. \tag{2.3.57}$$

Since the conditional likelihood $p(Y_t \mid X_t)$ is known for each X_t. The main problem for computing MAP is the computation of the predictive density $p(X_t \mid \mathbf{Y}_{t-1})$ (the second terms in equation (2.3.57)), whereas it not available in closed form. Nonetheless, using the formula of prediction density in Bayesian approach yields

$$\begin{aligned} p(X_t \mid \mathbf{Y}_{t-1}) &= \int p(X_t, X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} \\ &= \int p(X_t \mid X_{t-1}, \mathbf{Y}_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} \\ &= \int p(X_t \mid X_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1}, \qquad (2.3.58) \end{aligned}$$

Then, using the equation (2.3.58) one obtains

$$p(X_t \mid \mathbf{Y}_{t-1}) = \int p(X_t \mid X_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1}. \qquad (2.3.59)$$

Using particle filtering technique, this posterior distribution can be approximated by a set of N weighted particles as

$$\widehat{p}(\mathbf{X}_t \mid \mathbf{Y}_t) \approx \sum_{m=1}^{N} w_t^m \delta(\mathbf{X}_t - \mathbf{x}_t^m) \qquad (2.3.60)$$

where $\mathbf{x}_t^m = (x_1^{m\top}, \cdots, x_t^{m\top})^\top$. Now, use equation (2.3.60) to approximate $p(X_t \mid \mathbf{Y}_{t-1})$ as

$$\widehat{p}(X_t \mid \mathbf{Y}_{t-1}) \approx \sum_{j}^{N} p(X_t \mid X_{t-1}^j) w_{t-1}^{(j)}. \qquad (2.3.61)$$

Substituting (2.3.61) into (2.3.57) the MAP estimation is obtained by computing the global maxima of the posterior distribution by particles as

$$X_t^{MAP} = \arg\max_{X_t^{(m)}} p(\mathbf{Y}_t \mid X_t^{(m)}) \times \sum_{j}^{N} p(X_t^{(m)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}. \qquad (2.3.62)$$

It should be noted that for each time step, the memory necessity of this MAP estimator is $O(N)$ and the computational complexity is $O(N^2)$.

Smoothing step: Saha *et al.* (2008) extended the MAP estimation concept to the marginal smoothing

$$\begin{aligned} X_{t|T}^{MAP} &= \arg\max_{X_t} p(X_t \mid \mathbf{Y}_T) \\ &= \arg\max_{X_t^{(m)}} p(X_t^{(m)} \mid \mathbf{Y}_t) \frac{w_{t|T}^{(m)}}{w_t^{(m)}}, \qquad (2.3.63) \end{aligned}$$

where for the filtering density $p(X_t \mid \mathbf{Y}_t)$ at the particle cloud $\{X_t^{(m)}\}_{m=1}^N$ the evaluation during the forward filtering step can be defined as

$$p(X_t \mid \mathbf{Y}_t) \approx \frac{p(\mathbf{Y}_t \mid X_t^{(m)}) \sum_j^N p(X_t^{(m)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}}{p(Y_t \mid \mathbf{Y}_{t-1})}, \tag{2.3.64}$$

since $p(Y_t \mid \mathbf{Y}_{t-1})$ in equation (2.3.64) is independent of $X_t^{(m)}$. Consequently, obtaining $x_{t|T}^{MAP}$ requires to replace $p(X_t^{(m)} \mid \mathbf{Y}_t)$ by the filtered density

$$q(X_t^{(m)} \mid \mathbf{Y}_t) \quad = \quad p(\mathbf{Y}_t \mid X_t^{(m)}) \sum_j^N p(X_t^{(m)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}. \tag{2.3.65}$$

The derivation of the Forward-Backward smoothing is given in Appendix **A.4**.

The MAP via Particle filter Algorithm :

- Given observations $\mathbf{Y}_t = (Y_1, \cdots, Y_t)^\top$,
- For $m = 1, \cdots, N$ where N is the number of particles
  Forward filtering step:
- Assume $p(X_0)$, draw $X_0^{(m)}$ from $p(X_0)$, set $w_0^{(m)} = \frac{1}{N}$.
- Apply particle filter to generate and store $\{X_t^{(m)}, w_t^{(m)}\}$ for $t = 0, \cdots, T$.
- Evaluate (un-normalized) filtering distribution for $t = 1, \cdots, T$, at cloud points m

  $$q(X_t^{(m)} \mid \mathbf{Y}_t) = p(\mathbf{Y}_t \mid X_t^{(m)}) \sum_j^N p(X_t^{(m)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}.$$

  starting with $q(X_0^{(m)}) = p(X_0^{(m)})$ and store
  Backward smoothing step:
- Set $w_{T|T}^{(m)} = w_T^{(m)}$
- For $t = T-1, \cdots, 1$ the smoother importance weights is computed as

  $$w_{t|T}^{(m)} = w_t^{(m)} \sum_{j=1}^N \left[w_{t+1|T}^{(m)} \frac{p(X_{t+1}^{(j)} \mid X_t^{(m)})}{\sum_{k=1}^N p(X_{t+1}^{(j)} \mid X_t^{(k)}) w_t^{(k)}} \right].$$

- Compute the approximate smoother MAP as

  $$X_{t|T}^{MAP} = \arg\max_{X_t^{(m)}} q(X_t^{(m)} \mid \mathbf{Y}_t) \frac{w_{t|T}^{(m)}}{w_t^{(m)}}.$$

2.3.5.3 Auxiliary Iterated Extended Kalman Particle Filter-MAP(AIEKPF-MAP) Algorithm

We can extend the MAP estimation algorithm by using the weights calculated by Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF) algorithm :

$$w_t^m \quad = \quad \frac{p[Y_t \mid X_t^m]}{p[Y_t \mid \mu_t^{\varsigma_i^m}]}.$$

The AIEKPF algorithm can be performed in the following steps:

1. Initialization ($t = 0$) : For $m = 1, \cdots, N$, generate the states (particles) X_0^m from the prior $p(X_0)$, and set

   $$\widehat{X}_0^m = \mathrm{E}(X_0^m) = F(\mathbf{u}_0^m, X_0^m, \delta^m)$$

   and

   $$\begin{aligned} V_0^m &= \mathrm{E}[(X_0^m - \widehat{X}_0^m)(X_0^m - \widehat{X}_0^m)^\top] = \mathrm{Var}(X_0^m) \\ &= H^2(\mathbf{u}_0^m, X_0^m, \delta^m)R_0 \end{aligned}$$

   For $t = 1, \cdots, T$, repeat the following steps:

2. For $m = 1, \cdots, N$, generate $\mu_t^m \sim p(X_t \mid X_{t-1})$.

3. For $m = 1, \cdots, N$, using the IEKF algorithm to update the particles

   - Compute the Jacobians A_t^m, C_t^m of the process model

     $$A_t^m = \frac{\partial F(\mathbf{u}_t, x, \gamma)}{\partial x} \mid_{x = X_{t-1|t-1}^m}$$

     $$C_t^m = H(\mathbf{u}_t, X_{t-1|t-1}^m, \delta)$$

   - Predict the particle with the IEKF:

     $$x_{t|t-1}^m \approx F(\mathbf{u}_t, x_{t-1|t-1}^m, \gamma)$$

     $$V_{t|t-1}^m = A_t^m V_{t-1|t-1}^m A_t^{m\top} + C_t^m R_t C_t^{m\top}$$

   - For $j = 1, \cdots c$ (c is the iteration number of the IEKF)
     a - Compute the Jacobian , B_{tj}^m

     $$B_{tj}^m = \frac{\partial G(\mathbf{u}_t, x, \lambda)}{\partial x} \mid_{x = x_{t|t-1,j}^m}$$

     b -Update the state estimation error covariance V_{tj}^m with Eq.(2.3.33)
     c - Update the state estimate X_{tj}^m with Eq.(2.3.32).

4. For $m = 1, \cdots N$, calculate

$$\begin{aligned} w_t^m &= q(m \mid \mathbf{Y}_t) \\ &\propto p(Y_t \mid X_t) \times w_{t-1}^m, \end{aligned} \tag{2.3.66}$$

5. Perform the Re-sampling methods to obtain the index ς^m of particle m's parent.

6. Find the importance sampling : for $m = 1, \cdots, N$,

   - Draw samples

$$\begin{aligned} X_t^m &\sim q(X_t, \varsigma^m \mid \mathbf{Y}_t) \\ &= \mathcal{N}(\widehat{X}_{tj}^{\varsigma^m}; V_{tj}^{\varsigma^m}), j = c \end{aligned} \tag{2.3.67}$$

   - Compute the importance weights of particles by using

$$w_t^m = \frac{p(Y_t \mid X_t^m)}{p(Y_t \mid \mu_t^{\varsigma^m})} \tag{2.3.68}$$

   - Normalize the weights

$$w_t^m = \frac{w_t^m}{\sum_{m=1}^N w_t^m}.$$

Forward Filtering step :

$$X_t^{MAP} = \arg\max_{X_t^{(m)}} p(Y_t \mid X_t^{(m)}) \sum_j p(X_t^{(j)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}$$

Backward smoothing step :

- Set $w_{T|T}^{(m)} = w_t^{(m)}$
- For $t = T-1, \cdots, 1$, compute the smoother importance weights as

$$w_{t|T}^{(m)} = w_t^{(m)} \sum_{j=1}^N \left[w_{t+1|T}^{(j)} \frac{p(X_{t+1}^{(j)} \mid X_t^{(m)})}{\sum_{r=1}^N p(X_{t+1}^{(j)} \mid X_t^{(r)}) w_t^{(r)}} \right]$$

- Compute the approximate smoother MAP as

$$X_{t|T}^{MAP} = \arg\max_{X_t^{(m)}} \left\{ q(X_t^{(m)} \mid \mathbf{Y}_t) \frac{w_{t|T}^{(m)}}{w_t^{(m)}} \right\}.$$

2.4 The state space representation and latent variable models

State space models methodology provides a good technique for analyzing non-gaussian time series and longitudinal data, as a model for discrete longitudinal observations. In this section, we present two applications of non-gaussian state space model. First, the exponential family state space models (generalized linear models) was studied by Fahrmeir and Wagenpfiel (1997) and Fahrmer and Tutz (2013). They considered the discrete time series observation $\{Y_t\}$ and estimated the state variable by posterior mode. Second, Bousseboua and Mesbah (2010) proposed a new class of longitudinal multivariate processes which are dichotomous outcomes $(Y_{ik}(t) : i = 1, \cdots, n,\ k = 1, \cdots, q,\ t = 1, \cdots, T)$, where n is the number of individual and q is the number of item and t the instant.

2.4.1 The exponential family state space models

Consider the time series observations $\{Y_t\}$ and the state variables $\{X_t\}$ with dimensions ℓ and k, respectively. The exponential family observation equation is

$$Y_t \mid X_t \sim p(Y_t \mid X_t) = \exp\left\{\frac{\theta_t^\top Y_t - b_t(\theta_t)}{\phi}\right\} + c_t(Y_t, \phi), \tag{2.4.1}$$

where θ_t the natural parameter is a function of $\eta_t = Z_t X_t$, Z_t is a $\ell \times k$ matrix. ϕ is a dispersion parameter, and $c_t(.)$ and $b_t(.)$ are known functions. By the properties of exponential families, mean and variance functions can be written as

$$\begin{aligned} \mathrm{E}(Y_t \mid X_t) &= \mu_t = \frac{\partial b_t(\theta_t)}{\partial \theta_t} && (2.4.2)\\ \mathrm{Var}(Y_t \mid X_t) &= \Sigma_t = \frac{\partial^2 b_t(\theta_t)}{\partial \theta_t \partial \theta_t^\top}. && (2.4.3) \end{aligned}$$

As in generalized linear models, the mean μ_t is related to the linear predictor $\eta_t = Z_t X_t$ by

$$\mu_t = h(Z_t X_t), \tag{2.4.4}$$

where $h : \mathbb{R}^\ell \to \mathbb{R}^\ell$ is a response function which is two-times continuously differentiable. Z_t is a $\ell \times k$ matrix, which may be expressed by covariates $\mathbf{u}_t$ or also by past responses $\mathbf{Y}_{t-1}$.

For example in binomial distribution $h(\pi_t) = \exp(\eta_t)/\{1 + \exp(\eta_t)\}$ where $\eta_t = \mathbf{u}_i^\top \beta$, with $(\mathbf{u}_i)$ standing for the covariate variables and β the regression parameters. This function h is known as the logit model. In contrast, $h(\pi_t) = \Phi(\pi_t)$ is called the probit model.

The state process $\{X_t, t = 1, 2, \cdots\}$, $X_t \in \mathbb{R}^k$, is the solution of the following equation, called state equation:

$$X_t = F_t X_{t-1} + \varepsilon_t, \tag{2.4.5}$$

where F_t is the transition matrix and $\{\varepsilon_t\}$ is a Gaussian noise with $\varepsilon_t \sim \mathcal{N}(0; R_t)$. The initial state $X_0 \sim \mathcal{N}(x_0; R_0)$.
Fahrmeir and Wagenpfiel (1997) assumed the following conditional independence assumptions:

(A_1) Conditional on X_t, the current responses Y_t are independent of past states $X_{t-1}, \cdots, X_0$, i.e.

$$p(Y_t \mid X_t, X_{t-1}, \cdots, X_0) = p(Y_t \mid X_t), t = 1, \cdots, T.$$

(A_2) The sequence $\{X_t, t = 1, 2, \cdots\}$ is Markovian, i.e.

$$p(X_t \mid X_{t-1}, \cdots, X_0) = p(X_t \mid X_{t-1}).$$

According to (2.4.5) one obtains

$$p(X_t \mid X_{t-1}) \sim \mathcal{N}(F_t X_{t-1}, R_t).$$

2.4.1.1 Penalized likelihood estimation

Fahrmeir and Wagenpfeil (1997) proposed a simple method of inference founded on posterior mode, or equivalently maximum penalized likelihood estimation. Their approach can also be interpreted as a nonparametric method for smoothing time-varying coefficients. They estimated the posterior mode by using Fisher scoring method via the iterative Kalman filtering and smoothing. They used the numerical method to maximum penalized likelihood called *"Working Kalman filtering and smoother (WKFS)"* with the exponential family distribution. Define $\mathbf{Y}_t = (Y_0^\top, Y_1^\top, \cdots, Y_t^\top)^\top$, and $\mathbf{X}_t = (X_0^\top, X_1^\top, \cdots, X_t^\top)^\top$, the posterior mode smoother is defined as

$$a \equiv \left\{a_{0|T}^\top, a_{1|T}^\top, \cdots, a_{T|T}^\top\right\} \in \mathbb{R}^m,\ t = 1, \cdots, T.$$

with $m = (T+1)k$,
The posterior distribution of X_T given Y_T is obtained by Bayes' theorem :

$$\begin{aligned} p(\mathbf{X}_T \mid \mathbf{Y}_T) &= \frac{1}{p(\mathbf{Y}_T)} \left\{ \prod_{t=1}^{T} p(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{Y}_{t-1}) \right\} \times \\ & \left\{ \prod_{t=1}^{T} p(\mathbf{X}_t \mid \mathbf{X}_{t-1}, \mathbf{Y}_{t-1}) \times p(X_0) \right\} \end{aligned} \tag{2.4.6}$$

$p(\mathbf{Y}_T)$ does not depend on $\mathbf{X}_T$

$$p(\mathbf{X}_T \mid \mathbf{Y}_T) \propto \prod_{t=1}^{T} p(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{Y}_{t-1}) \prod_{t=1}^{T} p(\mathbf{X}_t \mid \mathbf{X}_{t-1}).p(X_0). \tag{2.4.7}$$

Taking logarithms and substituting the densities of $\mathbf{X}_t$ of the state equation (2.4.5) and X_0, one can obtain the penalized log-likelihood function

$$PL : \mathbb{R}^m \longrightarrow \mathbb{R},\ m = (T+1)k$$

$$\begin{aligned} PL(\mathbf{X}_t) &= \sum_{t=1}^{T} \log p(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{Y}_{t-1}) \\ &\quad + \sum_{t=1}^{T} \log p(\mathbf{X}_t \mid \mathbf{X}_{t-1}) + \log p(X_0), \end{aligned} \tag{2.4.8}$$

which can be written as:

$$PL(\mathbf{X}_t) = \sum_{t=1}^{T} p(\mathbf{Y}_t \mid X_t, \mathbf{Y}_{t-1}) + G_1 + G_2, \tag{2.4.9}$$

where

$$\begin{aligned} G_1 = &- \frac{1}{2}(X_0 - a_0)R_0^{-1}(X_0 - a_0) \\ G_2 = &-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{X}_t - F_t\mathbf{X}_{t-1})^\top R_t^{-1}(\mathbf{X}_t - F_t\mathbf{X}_{t-1}). \end{aligned}$$

The densities $P(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{Y}_{t-1})$ are defined by the exponential family observation model equation (5.17). Thus, the posterior mode smoother is given by

$$a \equiv (a_{0|T}^\top, a_{1|T}^\top, \cdots, a_{T|T}^\top)^\top = \arg\max_{\mathbf{X}}\{PL(\mathbf{X})\}, \tag{2.4.10}$$

The Maximization of $p(\mathbf{X}_T \mid \mathbf{Y}_T)$ is equivalent to the maximization of the penalized log-likelihood (2.4.9).
Numerical maximization of the penalized log-likelihood can be achieved by various algorithms. Fahrmeir(1992) suggested the generalized extended Kalman filter and smoother as an approximative posterior mode estimator in dynamic generalized linear models. Fahrmeir and Kaufmann (1991) developed iterative forward-backwards Gauss-Newton (Fisher-scoring) algorithms. The iterative application of linear Kalman filtering and smoothing to a ”working” model give the same results by using Gauss-Newton smoothers.

2.4.1.2 Gauss-Newton and Fisher-scoring Filtering and smoothing

A maximization of the penalized log-likelihood $PL(\mathbf{X}_t)$ with the best performance to approximation can be computed by Gauss-Newton or Fisher-scoring iterations. In other words, this can also be achieved by applying the linear Kalman filtering and smoothing to a "working model" in each Fisher-scoring iteration. Then, the penalized log-likelihood criterion (2.4.8) can be written in a matrix notation as:

$$PL(\mathbf{X}_t) = l(\mathbf{X}_T) - \frac{1}{2}\mathbf{X}_t^\top \mathcal{K}\mathbf{X}_T, \tag{2.4.11}$$

where

$$l(\mathbf{X}_t) = \sum_{t=0}^{T} l_t(\mathbf{X}_t),$$

with

$$l_t(\mathbf{X}_t) = \ln p(\mathbf{Y}_t \mid \mathbf{X}_t), \quad t = 0, \cdots, T,$$

and

$$l_0 = -(X_0 - a_0)^\top R_0^{-1}(X_0 - a_0))/2,$$

the penalty matrix $\mathcal{K}$ is symmetric and block-tridiagonal, with blocks that can easily be computed from (2.4.8):

$$\mathcal{K} = \begin{bmatrix} \mathcal{K}_{00} & \mathcal{K}_{01} & \cdots & \cdots & 0 \\ \mathcal{K}_{10} & \mathcal{K}_{11} & \mathcal{K}_{12} & \cdots & \vdots \\ \vdots & \mathcal{K}_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & & \ddots & \mathcal{K}_{T-1,T} \\ 0 & \cdots & \cdots & \mathcal{K}_{T,T-1} & \mathcal{K}_{T,T} \end{bmatrix}$$

with

$$\begin{aligned} \mathcal{K}_{t-1,t} &= \mathcal{K}_{t,t-1}^\top, \quad t = 1, \cdots, T \\ \mathcal{K}_{00} &= F_1^\top R_1^{-1} F_1, \\ \mathcal{K}_{tt} &= R_t^{-1} + F_{t+1}^\top R_t^{-1} F_{t+1}, \quad t = 1, \cdots, T \\ F_{T+1} &= 0, \\ \mathcal{K}_{t-1,t} &= -F_t^\top R_t^{-1}, \quad t = 1, \cdots, T. \end{aligned}$$

R_t is the variance-covariance matrix of the noise process ε_t of the state equation (2.4.5). The description of Fisher scoring steps in matrix notation is as follows:
rewrite the observations matrix with Y_0

$$\mathbf{Y}^\top = (Y_0^\top, Y_1^\top, \cdots, Y_T^\top)^\top.$$

Fahrmeir and Wagenpfeil (1997) assumed $Y_0^\top = a_0$. In contrast, the matrices of expectations are defined by adding μ_0

$$\mu = (\mu_0^\top, \mu_1^\top, \cdots, \mu_T^\top)^\top.$$

Fahrmeir and Wagenpfeil (1997) assumed $\mu_0 = X_0$, where $\mu_t = Z_t X_t$, the block-diagonal covariance matrix

$$\Sigma = \text{diag}(R_0, \Sigma_1, \cdots, \Sigma_T),$$

the block-diagonal design matrix

$$\mathbf{Z} = \text{diag}(I, Z_1, \cdots, Z_T),$$

with $I \in \mathbb{R}^{k\times k}$ is the unit matrix and the block-diagonal matrix

$$\mathbf{D} = \text{diag}(I, D_1, \cdots, D_T),$$

where for all $t = 1, \ldots, T$, $D_t = \partial h(\eta_t)/\partial\eta_t$ first- order derivative of the response function $h(\eta)$ is evaluated at $\eta_t = Z_t X_t$. The score function of $l(\mathbf{X}_T)$ in (2.4.11) can be written as

$$\mathbf{S} = (S_1, \cdots, S_T) = \mathbf{Z}^\top \mathbf{D}\Sigma^{-1}\{\mathbf{Y} - \mu\}, \tag{2.4.12}$$

with components

$$\begin{aligned} S_0 &= R_0^{-1}(a_0 - X_0) & (2.4.13)\\ S_t &= Z_t D_t \Sigma_t^{-1}\{\mathbf{Y}_t - \mu_t\} \quad t = 1, \cdots, T, & (2.4.14)\end{aligned}$$

the weight matrix

$$\mathbf{W} = \text{diag}(W_0, W_1, \cdots, W_T) := \mathbf{D}\Sigma^{-1}\mathbf{D}^\top, \tag{2.4.15}$$

with diagonal blocks

$$\begin{aligned} W_0 &= R_0^{-1} \\ W_t &= D_t \Sigma_t^{-1} D_t^\top, \quad t = 1, \cdots, T. & (2.4.16)\end{aligned}$$

Denote the first-order derivative of $PL(\mathbf{X})$ in (2.4.11) by

$$\mathbf{M} = \partial PL(\mathbf{X})/\partial\mathbf{X} = \mathbf{S} - \mathcal{K}\mathbf{X}, \tag{2.4.17}$$

and the (expected) information matrix of $l(\mathbf{X})$ by

$$\mathcal{I} = -\mathbf{E}\left\{\frac{\partial^2 PL(\mathbf{X})}{\partial \mathbf{X}\mathbf{X}^\top}\right\} = \mathbf{S} + \mathcal{K} = \mathbf{Z}^\top\mathbf{W}\mathbf{Z} + \mathcal{K}, \tag{2.4.18}$$

with diagonal blocks

$$\begin{aligned} \mathcal{I}_0 &= R_0^{-1} & (2.4.19)\\ \mathcal{I}_t &= Z_t^\top W_t Z_t, \quad t = 1, \cdots, T. & (2.4.20)\end{aligned}$$

A single Fisher scoring to the next iterate $\mathbf{X}^1 \in \mathbb{R}^m$, with $m = (T+1)k$ is given by

$$\mathcal{I}^0\left\{\mathbf{X}^1 - \mathbf{X}^0\right\} = \mathbf{M}^0.$$

where $\mathbf{X}^0 \in \mathbb{R}^m$, is the current iterate, and $\mathcal{I}^0, \mathbf{M}^0$ the first and second derivatives of the current iterate. This can be rewritten as

$$\mathbf{X}^1 = \left\{\mathbf{Z}^\top \mathbf{W}(\mathbf{X}^0)\mathbf{Z} + \mathcal{K}\right\}^{-1} \mathbf{Z}^\top \mathbf{W}(\mathbf{X}^0)\widetilde{Y}(\mathbf{X}^0), \tag{2.4.21}$$

the "working" observation $\widetilde{\mathbf{Y}}(\mathbf{X}^0) = (\widetilde{Y}_1^\top(X_1^0), \cdots, \widetilde{Y}_T^\top(X_T^0))^\top$ can be computed as

$$\widetilde{\mathbf{Y}}(\mathbf{X}^0) := \left\{\mathbf{D}^{-1}(\mathbf{X}^0)\right\}^\top \left\{\mathbf{Y} - \mu(\mathbf{X}^0))\right\} + \mathbf{Z}\mathbf{X}^0, \tag{2.4.22}$$

with components

$$\begin{aligned} \widetilde{Y}_0 &= a_0 \\ \widetilde{Y}_t &= \left\{D_t^{-1}(X_t^0)\right\}^\top \left\{\mathbf{Y}_t - \mu(X_t^0)\right\} + Z_t X_t^0, \end{aligned} \tag{2.4.23}$$

where $\mu(\mathbf{X}^0) = h(\mathbf{Z}\mathbf{X}^0)$ and $\mathbf{X}^0$ is the value of the state variable of current iterate. Fahrmeir and Wagenpfeil (2013) mentioned one can apply their approach to a linear Gaussian state space model. Then, one can achieve above by assume $\mathbf{D}$ is the identity matrix and the score function yields

$$\mathbf{S} = \mathbf{Z}\Sigma^{-1}(\mathbf{Y} - \mathbf{Z}\mathbf{X}), \tag{2.4.24}$$

with $\Sigma = diag(R_0, \Sigma_1, \cdots, \Sigma_T)$, where $\Sigma_t = cov(\mathbf{Y}_t \mid \mathbf{X}_t)$. The weight matrix $\mathbf{W}_t$ reduces to Σ_t^{-1}, and one uses the actual observations $\mathbf{Y}$ instead of the "working" observations $\widetilde{\mathbf{Y}}$, since $\mathbf{D} = I, \mu(\mathbf{X}^0) = \mathbf{Z}\mathbf{X}^0$, then

$$a = (\mathbf{Z}^\top \Sigma^{-1}\mathbf{Z} + \mathcal{K})^{-1}\mathbf{Z}^\top \Sigma^{-1}\mathbf{Y}, \tag{2.4.25}$$

where $a = (a_{0|t}^\top, a_{1|t}^\top, \cdots, a_{T|T}^\top)^\top$ is the vector of smoother estimates. As noted earlier, the classical linear Kalman filter and smoother solves (2.4.25) efficiently, without the expression of the block-tridiagonal matrix $(\mathbf{Z}^\top\Sigma^{-1}\mathbf{Z}+\mathcal{K})$, for more details see Fahrmeir and Tutz (2013), chapter (8).

2.4.1.3 Working Kalman Filter and Smoother (WKFS)

In the following algorithm, the predictive, filter and smoother values $a_{t|t-1}, a_{t|t}, a_{t|T}$ respectively are numerical approximations of X_t, and $V_{t|t-1}, V_{t|t}$ and $V_{t|T}$ are numerical approximations of error covariance matrices of predicted, filtered and smoothed values respectively.

Initialization:

$$\begin{aligned} a_{0|0} &= a_0, \\ V_{0|0} &= R_0. \end{aligned} \tag{2.4.26}$$

For $t = 1, \cdots, T$:

Prediction:

$$\begin{aligned} a_{t|t-1} &= F_t a_{t-1|t-1} \\ V_{t|t-1} &= F_t V_{t-1|t-1} F_t^\top + R_t \end{aligned} \tag{2.4.27}$$

Filtering:

$$\begin{aligned} a_{t|t} &= a_{t|t-1} + K_t(\widetilde{Y}_t - Z_t a_{t|t-1}) \\ K_t &= V_{t|t-1} Z_t^\top \{Z_t V_{t|t-1} Z_t^\top + (W_t^{-1})^\top\}^{-1} \\ V_{t|t} &= V_{t|t-1} - K_t Z_t V_{t|t-1} \end{aligned} \tag{2.4.28}$$

Smoothing : For smoothing one uses the classical fixed interval smoother for $t = T, \cdots, 1$:

$$\begin{aligned} a_{t-1|T} &= a_{t-1|t-1} + B_t(a_{t|T} - a_{t|t-1}) \\ V_{t-1|T} &= V_{t-1|t-1} + B_t(V_{t|T} - V_{t|t-1}) B_t^\top \end{aligned}$$

where

$$B_t = V_{t-1|t-1} F_t^\top V_{t|t-1}.$$

2.4.2 Longitudinal Rasch process : Binary data

Bousseboua and Mesbah (2010) introduced a class of longitudinal latent processes where, at any time, a set of categorical binary observations are observed instead of the latent variables. The conditional probabilities are given by Rasch model which is widely used in various psychometric scopes such as educational research and the health analysis, especially in the quality of life.

They considered a longitudinal study, where participant patients are interviewed at regular dates of visit. In this interviews, the patients are asked to answer questions in a questionnaire. This questionnaire was constructed to measure their health at the time. It is recognized that the health is a latent multidimensional concept. In practice, each dimension is usually assessed by one or more questions. They focused on the measurement of a single dimension, and in a particular case of dichotomous response options for each item (yes-no, agree-disagree, etc.).

A binary state space model consists of two processes. First, the observed process $Y_i(t) = (Y_{i1}(t), \cdots, Y_{iq}(t))^\top$, where the conditional distribution of $Y_i(t)$ given the $n-$ dimensional state variable $X_i(t)$ is Bernoulli. That is $Y_i(t) \sim$ Bernoulli $(\pi_i(t))$. Second, the state process $\{X_i(t)\}$ is assumed to follow a $n-$ dimensional Markov process.

2.4.2.1 Description of the Rasch latent process

The model defined in Bousseboua and Mesbah (2010) consists in a finite trajectory of a multivariate process:

$$\{(Y_{ik}(t), X_i(t)) : 1 \leq i \leq n, 1 \leq k \leq q, 1 \leq t \leq T\}$$

where $Y_{ik}(t)$ represents the longitudinal process of observations. For all $i, X_i(t)$ is an unobservable (latent) process. The variable $Y_{ik}(t)$ is the response of the person i at instant t to the item k. This questionnaire is administered on different occasions to the same individuals. These models are characterized by the fact that each these response variable Y_{it} depends only on the corresponding latent trait variable $X_i(t)$. The path of the model can be clarified at any time t in Figure (2.2).

Figure 2.2: **The Path of longitudinal Rasch processes**

$$\begin{tikzcd}
\underbrace{\{Y_{11}(t)\cdots\cdots Y_{1q}(t)\}}_{Y_1(t)} & \cdots\cdots & \cdots\cdots & \underbrace{\{Y_{n1}(t)\cdots\cdots Y_{nq}(t)\}}_{Y_n(t)} \\
\mathrm{X}_1(t) \arrow[u] & \cdots\cdots & \cdots\cdots & X_n(t) \arrow[u]
\end{tikzcd}$$

The process of observation $Y_i(t) = (Y_{i1}(t), \cdots, Y_{iq}(t))^\top$ has values in $\{0, 1\}$, where $i = 1, \cdots, n, t = 1, \cdots, T, k = 1, \cdots, q$. The value 1 denotes a correct response of the i individual to the item k . Practically, in the education studies, the latent variable $X_i(t)$ can be interpreted as a measure of the ability of an individual. Whereas, in the quality of life, in the health field, the latent variable can represent a level of the health of an individual. In general, it is a measure of the individual position in scalar axis corresponding to the latent unidimensional trait measured by the questionnaire. Consequently, the latent variable $X_i(t)$ depends only on the individual i.
The observation and state equations of the model are written as follows.

2.4.2.2 The observation equation

They assumed that:

1. The conditional probability of $Y_{ik}(t)$ given $X_i(t)$ is

$$\Pr(Y_{ik}(t) = y_{ik}(t) \mid X_i(t)) = \pi_{ik}(t)^{y_{ik}(t)}(1 - \pi_{ik}(t))^{1-y_{ik}(t)}, \tag{2.4.29}$$

   where

$$\Pr(Y_{ik}(t) = 1 \mid X_i(t)) = \pi_{ik}(t) = \frac{\exp\{y_{ik}(t)(X_i(t) - \beta_k)\}}{1 + \exp\{X_i(t) - \beta_k\}}, \tag{2.4.30}$$

   where $X_i(t)$ denotes the scalar latent trait or the ability of the individual i, and β_k denote a real parameter related to item k is known as difficulty parameter in the educational context. The equation (2.4.30) is called the Rasch model.

2. The response variables $Y_{ik}(1), \cdots, Y_{ik}(T)$ are assumed to be conditionally independent of the vector of latent variable $(X_i(1) = x_i(1), \cdots X_i(T) = x_i(T))^\top$:

$$\begin{aligned} \Pr\left[Y_{ik}(1)\right. &= y_{ik}(1), \cdots, Y_{ik}(T) = y_{ik}(T) \mid X_i(1) = x_i(1), \cdots, X_i(T) = x_i(T)\left.\right] \\ &= \prod_{t=1}^{T} \Pr\left[Y_{ik}(t) = y_{ik}(t) \mid X_i(t) = x_i(t)\right], \end{aligned}$$

that means the response variables at an instant t is based only on the corresponding latent variable.

3. The response variables $Y_{i1}, \cdots, Y_{ik}$, are assumed to be conditionally independent to the latent variable $X_i(t)$, where for all $k = 1, \cdots, q$, $Y_{ik} = (Y_{ik}(1), \cdots, Y_{ik}(T))^\top$:

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

This means that the q items are conditionally independent relatively to the corresponding latent variable.

2.4.2.3 State equation

Bousseboua and Mesbah (2010) proposed two classes of latent process: a general first-order Markov latent process and a first-order autoregressive latent process (AR(1) latent process).

First-order Markov latent process

This model considers the latent process $(X_i(t) : 1 \leq t \leq T)$ as being a Markov chain of order one. They assumed the chain $(X_i(t) : 1 \leq t \leq T)$ has a real support and a Gaussian distribution with variance σ_i^2. Then:

$$p(X_i(t) \mid X_i(t-1), X_i(t-2), ..., X_i(1)) = p(X_i(t) \mid X_i(t-1)) \sim \mathcal{N}(\mu_i; \sigma_i^2). \quad (2.4.31)$$

The vector $\mathbf{X}_i = (X_i(1), \cdots, X_i(T))$ is gaussian and has probability density g_i defined by :

$$g_i(\mathbf{X}_i) = \frac{1}{\sigma_i^T \sqrt{(2\pi)^T}} \exp\left\{-\frac{1}{2\sigma_i^2}\left[X_i^2(1) + \sum_{t=2}^{T}(X_i(t) - X_i(t-1))^2\right]\right\} \quad (2.4.32)$$

Autoregressive latent process (AR(1) Process)

$\{X_i(t)\}$ is a stationary gaussian process satisfying a first-order autoregressive AR(1) equation

$$X_i(t) = \rho_i X_i(t-1) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0; \sigma_i^2) \tag{2.4.33}$$

$$p(X_i(t) \mid X_i(t-1), X_i(t-2), ..., X_i(1)) = p(X_i(t) \mid X_i(t-1)) \sim \mathcal{N}(\mu_i; \frac{\sigma_i^2}{1-\rho_i^2}). \tag{2.4.34}$$

The joint distribution of the variables $(X_i(1), \cdots, X_i(t))$ can be written as :

$$g_i(\mathbf{X}_i) = \frac{(\sqrt{1-\rho_i^2})^T}{\sqrt{(2\pi)^T \sigma_i^T}} \exp\left\{ -\frac{(1-\rho_i^2)}{2\sigma_i^2} \left[X_i^2(1) + \sum_{t=2}^{T} (X_i(t) - X_i(t-1))^2 \right] \right\} \tag{2.4.35}$$

where $\mathbf{X}_i = (X_i(1), \cdots, X_i(T))$ and the distribution of $X(0)$ is normal, namely $g_i(X(0)) \sim \mathcal{N}\left(\mu_i; \frac{\sigma_i^2}{1-\rho_i^2}\right)$.

2.4.2.4 The Marginal Likelihood

The state space models can be considered as a particular case of incomplete data models. The observations $Y_i(t)$ can be interpreted as incomplete due to missing state variable $X_i(t)$. The likelihood function known as Marginal likelihood can be written :

$$\begin{aligned}
p(\theta_i; \mathbf{Y}) &= \prod_{i=1}^{n} \int \cdots \int p(Y_i \mid X_i; \theta) g_i(\mathbf{X}_i) d(\mathbf{X}_i). && (2.4.36) \\
&= \prod_{i=1}^{n} \int \cdots \int \prod_{t=1}^{T} \prod_{k=1}^{q} \left[\pi_{ik}(t)^{y_{ik}(t)} (1-\pi_{ik}(t))^{1-y_{ik}(t)} \right] g_i(\mathbf{X}_i) d(\mathbf{X}_i). \\
&= \prod_{i=1}^{n} \int \cdots \int \prod_{t=1}^{T} \prod_{k=1}^{q} [1-\pi_{ik}(t)] \left[\frac{\pi_{ik}(t)}{1-\pi_{ik}(t)} \right]^{y_{ik}(t)} g_i(\mathbf{X}_i) d(\mathbf{X}_i). \\
&= \prod_{i=1}^{n} \int \cdots \int \prod_{t=1}^{T} [1-\pi_{ik}(t)]^q \left[\frac{\pi_{ik}(t)}{1-\pi_{ik}(t)} \right]^{r_i(t)} g_i(\mathbf{X}_i) d(\mathbf{X}_i). \\
&= \prod_{i=1}^{n} \int \cdots \int \frac{\exp\{\sum_{t=1}^{T} r_i(t).X_t - \sum_{k=1}^{q} \beta_k.r_k\}}{\prod_{t=1}^{T} \prod_{k=1}^{q} [1+\exp\{X_t - \beta_k\}]} .g_i(\mathbf{X}_i) d(\mathbf{X}_i), && (2.4.37)
\end{aligned}$$

where $\theta_i = (\beta, \sigma_i^2)$ are the parameters for first- order Markov process, or $\theta_i = (\beta, \rho_i, \sigma_i^2)$ for the AR(1) latent process model. Since $r_i(t) = \sum_{k=1}^{q} y_{ik}(t)$ is the score for the $i-$ individual at the $t-$ th occasion, and $g_i(X)$ is the joint distribution for first- order Markov chain latent process is described in equation (2.4.32) or joint distribution for AR(1) latent process is described in (2.4.35).

2.4.2.5 The EM algorithm

The Expectation-Maximization (EM) algorithm was proposed by Dempster, Laird, and Rubin (1977). In fact, EM algorithm is a good approach to the iterative computation of maximum likelihood estimation (MLE). In a variety of incomplete-data problems, the EM algorithm is better than algorithms such as the Newton-Raphson method which may turn out to be more complicated. On each iteration of the EM algorithm, there are two steps-called the Expectation step or the E-step and the Maximization step or the M-step. The cases where the EM algorithm can be performed include not only evidently incomplete-data situations, where there are missing data, truncated distributions, or censored or grouped observations. The term "incomplete data" means there are two samples Y the observed data and X the state variables are not observed, or there are missing data.

The EM algorithm aims at computing the MLE of the marginal likelihood by iteratively applying two steps. More precisely, at the step $p+1$, to compute $\theta^{(p+1)}$ using the value $\theta^{(p)}$ which is obtained at the p step. The (p + 1)-th cycle of the EM algorithm consists of the following two steps for $p = 0, 1, \cdots$:

Expectation -step: This step computes the conditional mean of the complete log-likelihood knowing the current values of the estimators

$$\begin{aligned} Q(\theta \mid \theta^{(p)}) &= \mathrm{E}\left\{\log\left[f(\mathbf{Y}, \mathbf{X}; \theta)\right] \mid \mathbf{Y}, \theta^{(p)}\right\}, \\ &= \sum_{i=1}^{n} \int \cdots \int \left[\log\{g_i(\mathbf{X}, \theta)\} + \log\{p(\mathbf{Y}_i, \mathbf{X}_i; \theta \mid \mathbf{Y}, \theta^{(p)})\}\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d\mathbf{X}_i, \end{aligned} \tag{2.4.38}$$

Bousseboua and Mesbah (2010) calculated this step for Rasch latent Markov processes as follows

$$Q(\theta \mid \theta^{(p)}) = -\frac{T}{2} \sum_{i=1}^{n} \log(2\pi\sigma_i^2) + H_1 + H_2, \tag{2.4.39}$$

where

$$\begin{aligned} H_1 &= -\sum_{i=1}^{n} \frac{1}{2\sigma_i^2} \int \cdots \int \left[X_i(1)^2 + \sum_{t=2}^{T} (X_i(t) - X_i(t-1))^2\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i). \\ H_2 &= +\sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=1}^{T} \int \cdots \int \log\left[\frac{\exp\{Y_{ik}(t)(X_i(t) - \beta_k\}}{1 + \exp\{X_i(t) - \beta_k\}}\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i). \end{aligned}$$

Here, $P(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})$ is the conditional density of the latent vector $\mathbf{X}_i = (X_i(1), \cdots, X_i(T))$ given the observation vector $\mathbf{Y}_i$.

Bousseboua and Mesbah (2010) approximated this density by:

$$p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) \propto \frac{\exp\left\{\sum_{t=1}^{T} X_i(t).Y_{ik}(t) - \sum_{k=1}^{q} \beta_k^p.Y_{ik}(t) - C\right\}}{\prod_{t=1}^{T}\prod_{k=1}^{q}\left[1+\exp(X_i(t)-\beta_k^p\right]}, \tag{2.4.40}$$

where

$$C = \frac{1}{2\sigma_i^{2(p)}}\left[X_i^2(1) + \sum_{t=2}^{T}(X_i(t)-X_i(t-1))^2\right].$$

The maximization with respect to $\sigma_i^2{}'$s yields :

$$\begin{aligned}\sigma_i^{2(p+1)} &= \frac{1}{T}\int\cdots\int\left[X_i^2(1) + \sum_{t=2}^{T}(X_i(t)-X_i(t-1))^2\right]\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i).\end{aligned} \tag{2.4.41}$$

Maximizing with respect to β includes only the double sum over i and t of the last terms of (2.4.39). That requires solving the q following equations, for $i = 1, \cdots, n$

$$\begin{aligned}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{ik}(t) &= \sum_{i=1}^{n}\sum_{t=1}^{T}\int\cdots\int\left[1+\exp(X_i(t)-\beta_k)\right]^{-1}\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i),\ k = 1, \cdots, q.\end{aligned} \tag{2.4.42}$$

On other hand, the estimation of Rasch latent AR(1) processes is as follows. Denote the parameters of the model by $\theta = (\beta, \rho, \sigma^2)$ with $\beta = (\beta_1, \cdots, \beta_k)^\top$, $\rho = (\rho_1, \cdots, \rho_n)^\top$ and $\sigma^2 = (\sigma_1^2, \cdots, \sigma_n^2)^\top, \forall i, \sigma_i^2 \neq 0$. Then, E-step is given by:

$$Q(\theta \mid \theta^{(p)}) = -\frac{T}{2}\sum_{i=1}^{n}\log(2\pi\sigma_i^2) + \frac{1}{2}\sum_{i=1}^{n}\log(1-\rho_i^2) + G_1 + G_2, \tag{2.4.43}$$

where

$$\begin{aligned}G_1 &= -\sum_{i=1}^{n}\frac{1}{2\sigma_i^2}\int\cdots\int\left[(1-\rho_i^2)X_1^2 + \sum_{t=1}^{T}(X_t-\rho_i X_i(t-1))^2\right]\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i).\\ G_2 &= +\sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\int\cdots\int\log\left[\frac{\exp\{Y_{ik}(t)(X_i(t)-\beta_k\}}{1+\exp\{X_i(t)-\beta_k\}}\right]\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i).\end{aligned}$$

and $p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})$ is the conditional density of latent vector $\mathbf{X}_i = (X_i(1), \cdots, X_i(T))^\top$ given $\mathbf{Y}_i$.
Bousseboua and Mesbah (2010) approximated this density by:

$$p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) \propto \frac{\exp\left\{\sum_{t=1}^{T} X_i(t).Y_{ik}(t) - \sum_{k=1}^{q} \beta_k^p.Y_{ik(t)-C}\right\}}{\prod_{t=1}^{T}\prod_{k=1}^{q}\left[1+\exp(X_i(t)-\beta_k^p)\right]} \tag{2.4.44}$$

$$C = \frac{1}{2\sigma_i^{2(p)}}\left[(1-\rho_i^{2(p)})X_i^2(1) + \sum_{t=2}^{T}(X_i(t)-\rho_i^{(p)}X_i(t-1))^2\right].$$

Here, the maximization with respect to σ_i^2 also includes only the first and third terms in (2.4.43), and maximizing with respect to ρ_i includes the second and third term of this expression. One obtains :

$$\begin{aligned}\sigma_i^{2(p+1)} &= \frac{1}{T}\sum_{i=1}^{n}\int\cdots\int\left[(1-\rho_i^2)X_i^2(1) + \sum_{t=2}^{T}(X_i(t)-\rho_i X_i(t-1)^2\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i).\ \ i = 1, \cdots, n, \end{aligned} \tag{2.4.45}$$

$$\begin{aligned}\frac{\rho_i}{1-\rho_i^2} &= \frac{1}{\sigma_i^2}\int\cdots\int\left[\rho_i X_i^2(1) + \sum_{t=2}^{T} X_i(t-1)(X_i(t)-\rho_i X_i(t-1))\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i).\end{aligned} \tag{2.4.46}$$

and maximizing with respect to β includes only the double sum over i and t of the last term of the expression (2.4.43). For $i = 1, \cdots, n$, that leads to solving the q following equations :

$$\begin{aligned}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{ik}(t) &= \sum_{i=1}^{n}\sum_{t=1}^{T}\int\cdots\int\left[1+\exp(X_i(t)-\beta_k)\right]^{-1} \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i).\ \ k = 1, \cdots, q.\end{aligned} \tag{2.4.47}$$

Maximization-step: The second step aims to find the value $\theta^{(p+1)}$ that achieves the maximum of the quantity $Q(\theta \mid \theta^{(p)})$:

$$Q(\theta \mid \theta^{(p+1)}) = \arg\max_{\theta} Q(\theta \mid \theta^{(p)}) \tag{2.4.48}$$

The iterative of the E and M stage are concluded if

$$\mid Q(\theta^{(p+1)}) - Q(\theta^{(p)}) \mid < \epsilon,$$

where ϵ is a prior fixed quantity. Bousseboua and Mesbah (2010) applied the Newton Raphson method to calculate the value $\theta^{(p+1)}$, and the integral are approximated numerically using the Gauss Hermite quadrature formulas.

2.4.3 Summary of the convergence properties of the EM algorithm

The convergence properties of the EM algorithm was studied by Wu (1983). He proved the EM sequence converges to the unique maximum likelihood estimate. He specified the following two special cases:

(a) The unobserved complete-data density function can be described by a curved exponential family distribution with parameters having compact space. The likelihood is unimodal when it has a unique mode.

(b) The likelihood function is unimodal and can undoubtedly satisfy a differentiability condition.

These properties of the EM algorithm help us to prove the asymptotic normality of the parameters. The properties of the EM algorithm can be briefly presented as follows:

(i) The likelihood and $L(\theta^{(p)})$ increases with any EM sequence $\{\theta^{(p)}\}$. If it is bounded, converges to some L^*, where L^* is a stationary value of L.

(ii) The continuity of Q achieved for every density function has a curved exponential family if $Q(\theta \mid \theta^{(p)})$ is continuous in both $\theta^{(p)}$ and θ. If $\theta^{(p)}$ converges to some point θ^*, θ^* is a stationary point under the continuity condition of $DQ(\theta \mid \theta^{(p)})$ in $\theta^{(p)}$ and θ. $DQ(\theta \mid \theta^{(p)})$ is first- order derivative with respect to the $\theta^{(p)}$, that is

$$DQ(\theta \mid \theta^{(p)}) = \frac{\partial}{\partial\theta} Q(\theta \mid \theta^{(p)}) \mid_{\theta=\theta^{(p)}} .$$

(iii) If, in addition to (ii), Q is not trapped at any a stationary point θ_0, but not a local maximum of L, i.e.

$$\sup_{\theta\in\Theta} Q(\theta \mid \theta_0) > Q(\theta_0 \mid \theta_0),$$

consequently, L^* is also a local maximum of L.

(iv) If, in addition to (ii) or (iii), $\parallel \theta^{(p+1)} - \theta^{(p)} \parallel\rightarrow 0$, [1] as $p \rightarrow \infty$ and the stationary point (local maxima) set with a L took a discrete value, thereafter $\theta^{(p)}$ converges to a stationary point (local maximum).

(v) If, in addition to (ii) or (iii), unacceptable, there exist two different stationary points (local maxima) relates to the same L value, then $\theta^{(p)}$ converges to a stationary point (local maximum).

[1] $\parallel . \parallel$ denote the absolute value norm, if p, q are vectors, then

$$\parallel p - q \parallel_1 = \sum_{i=1}^{n} \mid p_i - q_i \mid .$$

(vi) $\theta^{(p)}$ converges to the unique maximizer θ^* of $L(\theta)$, if $L(\theta)$ achieves that it is unimodal in Θ and has only stationary point and $D(\theta \mid \theta^{(p)})$ is continuous in θ and $\theta^{(p)}$,

2.4.4 Consistency and Asymptotic normality of the maximum likelihood estimator (MLE)

The inferential methods for any model must achieve the asymptotic properties of ML estimators. Under regularity assumptions which are determined by the proposed to the model, the following asymptotic properties of the estimators θ for the generalized state space model

Asymptotic existence and uniqueness: The estimators exist and are (locally) unique as a sample size $n \to +\infty$.

Consistency: If θ refer to the "true" parameter value, then as $n \to \infty$, then $\hat{\theta} \to^{P} \theta$ in probability (weak consistency) or with probability $p > 0.5$ (strong consistency).

Asymptotic normality: The distribution of the MLE is normal for $n \to \infty$, namely,

$$\hat{\theta} \sim^{a.s} \mathcal{N}(\theta, \mathcal{I}(\theta))$$

and

$$n^{1/2}(\hat{\theta} - \theta) \sim \mathcal{N}(0, \mathcal{I}^{-1}(\theta)).$$

i.e., $\hat{\theta}$ approximately is normal with approximate (or "asymptotic") covariance matrix

$$cov(\hat{\theta}) = \mathcal{I}(\theta),$$

where $\mathcal{I}^{-1}(\theta)$ is the inverse of the Fisher matrix and it can be calculated.
Furthermore, the MLE is asymptotically efficient according to a broad class of other estimators.

2.5 Oakes' identity to find the information matrix via the EM algorithm

Oakes (1999) assumed that the observed data Y with likelihood $L(\phi, Y)$ depending on the parameter vector ϕ may be specified as a many-to-one parameter-free map [2] of full

[2] **many-to-one function** : The parameters are a functions of full data (X, Y). A function f which may (but does not necessarily) companion a given member of the range of f with more than one member of the domain of f. For example, trigonometric functions such as $\sin x$ are many-to-one where $sinx = sin(2\pi + x) = sin(4\pi + x) = ...,$.

data X with log-likelihood $L(\phi, Y)$. The objective is to maximize $L(\phi, Y)$ in ϕ.

$$Q(\phi^p \mid \phi) = E_{X|Y;\phi} L_0(\phi^p, x).$$

The fundamental identity

$$L(\phi^p, X) = Q(\phi^p \mid \phi) - E_{X|Y;\phi} \log\{p(X \mid Y; \phi^p)\}, \tag{2.5.1}$$

ϕ^p denoted the result from the iteration p of the algorithm. The EM algorithm yields iteratively, at each step selecting ϕ^p to maximize $(\phi^p \mid \phi)$ in its first argument ϕ^p with its second argument ϕ held fixed. The procedure will be useful when as is often the case, the functional form of $L(\phi^p, x)$ is appreciably simpler than that of $L(\phi, y)$. If $L(\phi^p, X)$ is of exponential family form the maximization can be divided into a separate "E-step" $-$ calculate the conditional expectation of the sufficient statistics in X given the observed data Y, using the current parameter estimate $\phi-$ and M-step $-$ maximize the log-likelihood $L(\phi^p, x)$ with these computed values for the sufficient statistics and choose the new parameter value ϕ^p.

2.5.1 Derivative of $Q(\phi^p \mid \phi)$ and $L(\phi)$

Oakes (1999) assumed that the procedures of expectation usually interchange with respect to X and differentiation in ϕ hold for $\log\{p(X \mid Y; \phi^p)\}$. Thus

$$E_{X|Y;\phi} \frac{\partial[\log\{p(X \mid Y; \phi^p)\}]}{\partial \phi} = 0 \tag{2.5.2}$$

and

$$E_{X|Y;\phi} \frac{\partial^2[\log\{p(X \mid Y; \phi^p)\}}{\partial \phi^2} = -E_{X|Y;\phi} \frac{\partial[\log\{p(X \mid Y; \phi^p)\}]}{\partial \phi} \frac{\partial[\log\{p(X \mid Y; \phi^p)\}]^T}{\partial \phi}. \tag{2.5.3}$$

Differentiation of equation (2.5.1) in ϕ^p gives

$$\frac{\partial L}{\partial \phi^p} = \frac{\partial Q(\phi^p \mid \phi)}{\partial \phi^p} - E_{X|Y;\phi} \frac{\partial[\log\{p(X \mid Y; \phi^p)\}]}{\partial \phi^p} \tag{2.5.4}$$

The substitution $\phi = \phi^p$ makes the last term disappear by equation (2.5.2). So the score statistic for the observed data yields

$$\frac{\partial L}{\partial \phi} = \left\{ \frac{\partial Q(\phi^p \mid \phi)}{\partial \phi^p} \right\}_{\phi^p = \phi}. \tag{2.5.5}$$

Differentiation of equation (2.5.4) in ϕ^p and ϕ gives respectively

$$\begin{aligned}
\frac{\partial^2 L}{\partial \phi^{p^2}} &= \frac{\partial^2 Q(\phi^p \mid \phi)}{\partial \phi^{p^2}} - E_{X|Y;\phi} \frac{\partial^2[\log\{p(X \mid Y; \phi^p)\}]}{\partial \phi^{p^2}}, \\
0 &= \frac{\partial^2 Q(\phi^p \mid \phi)}{\partial \phi \partial \phi^p} - E_{X|Y;\phi} \frac{\partial[\log\{p(X \mid Y; \phi^p)\}]}{\partial \phi^p} \frac{\partial[\log\{p(X \mid Y; \phi^p)\}]^\top}{\partial \phi}.
\end{aligned}$$

Substituting $\phi = \phi^p$, adding the two equations and using equation (2.5.3) yields

$$\frac{\partial^2 L}{\partial \phi^2} = \left\{ \frac{\partial^2 Q(\phi^p \mid \phi)}{\partial {\phi^p}^2} + \frac{\partial^2 Q(\phi^p \mid \phi)}{\partial \phi^p \partial \phi} \right\}_{\phi^p = \phi}, \tag{2.5.6}$$

which is valid for all ϕ. The second term in equation (2.5.6) is naturally called the “missing information”,i.e. due to the fact that only Y and not X is observed.

Chapter 3

Longitudinal multicategorical processes : Generalized state space models with Gaussian noises

3.1 Introduction

In this chapter, we introduce a new class of longitudinal multivariate processes which are multicategorical. The data are from a longitudinal study, where the patients participate in an interview. The interview aims at measuring the patients' health at regular intervals, the dates of which are determined before the study. It typically involves filling out a questionnaire in which they are asked multiple choice questions. These questionnaires are used in the quality of life studies in health scope, constructed in order to measure the patient's perceived health at the time of the visit.

It is recognized that health is a latent multidimensional concept. We are interested to latent variables $X_i(t)$ produced by an individual $i, (i = 1, \cdots, n)$, at time $t, (t = 1, \cdots, T)$.

The $X_i(t)'$s may be **the patients health**, a **latent trait**, etc.. We only observe $Y_i(t)$ instead of $X_i(t)$. The $Y_i(t)'$s are the responses of the individuals to the questionnaire where each item in the questionnaire has multiple options, the individual selects one of them.

In the past studies, the latent variables $X_i(t)'s$ were described by first-order autoregressive model AR(1) with gaussian noise. In this chapter they are described by a first-order conditional heteroscedastic non-linear (CHARN) models with gaussian noise. In the state space models whether linear or non-linear if the observations and the latent variable have a gaussian distribution, the posterior distribution also has a gaussian distribution (Arulampalam (2002), Kitagawa(2010)). Consequently, the Kalman filtering is used to estimate the posterior mean and variance-covariance matrix. In the present context, the observations are from a multinomial distribution and the latent variables from a gaussian distribution. Since the posterior distribution is not symmetric we con-

sider the posterior mode as an estimator of the latent variable.
In this chapter, two approaches are presented for estimating the latent variable. First, the Working Extended Kalman Filtering recursions (WEKF). Second, Maximum A Posteriori via the Auxiliary Iterated Extended Kalman Particle Filtering (AIEKPF-MAP). The models parameters are estimated through the Maximum Likelihood Estimation method (MLE) via the Expectation-Maximization (EM) algorithm. The consistency and asymptotic normality for MLE's are established. The posterior distribution $p[X_i(t) \mid \mathbf{Y}_i(t)]$ is computed through the Bayesian approach, where we develop the Auxiliary Iterated Extended Kalman Particle filter(AIEKPF) with our model by deriving the equations of extended Kalman filter recursions.

3.2 The path of the process

In the applications of life, like biometrics, economics, finance, psychometrics, public health, social sciences, the data are obtainable as longitudinal data or categorical time series.
The categorical time series concept can be explained with an example. Let $\{U_t, t = 1, 2, \cdots\}$ be a time series that has c categories. The values of U_t are $0, 1, 2, \cdots, c-1$. To clarify more, in health study for the children the sleep state score virtually is observed for an individual (newborn infants) the responses are (*quiet sleep, indeterminate sleep, active sleep* and *awake*), these scores can be recorded as :
1 (quiet sleep), 2 (indeterminate sleep), 3 (Active sleep), 4 (awake).
An alternative recording might be:
1 (awake) , 2 (active sleep), 3 (indeterminate sleep), 4 (quiet sleep)
However, the diverse records leads to different results. In order to reduce the risk of multiple interpretations, the categorical time series is rewritten (regardless of the measurement scale) as the vector $U_t = (U_{t1}, \cdots, U_{ts})^\top$ of length $s = c-1$, with elements

$$U_{tj} = \begin{cases} 1, & \text{if the category } j \text{ is observed at time } t \\ 0, & \text{with otherwise.} \end{cases}$$

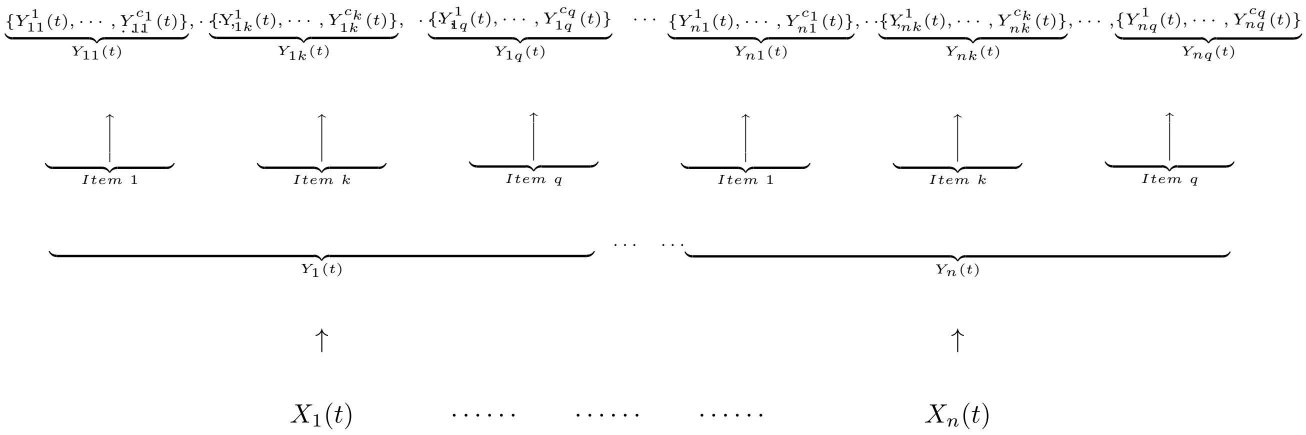
The first recording of the example is written as follows:

- $U_t = (1, 0, 0)^\top$ indicate "quiet sleep"
- $U_t = (0, 1, 0)^\top$ indicate "indeterminate sleep "
- $U_t = (0, 0, 1)^\top$ indicate " active sleep"
- $U_t = (0, 0, 0)^\top$ indicate " awake".

After this explanation, we present our model which assumes that n individuals are investigated. The model consists of a finite trajectory of a multivariate process:

$$\{(Y_{ik}(t), X_i(t)) : 1 \leq i \leq n, 1 \leq k \leq q, 1 \leq t \leq T\}$$

Figure 3.1: The Path of longitudinal multicategorical processes


$\{Y_{11}^1(t), \cdots, Y_{11}^{c_1}(t)\}$
$Y_{11}(t)$
$\{Y_{1k}^1(t), \cdots, Y_{1k}^{c_k}(t)\}$
$Y_{1k}(t)$
$\{Y_{1q}^1(t), \cdots, Y_{1q}^{c_q}(t)\}$
$Y_{1q}(t)$
$\{Y_{n1}^1(t), \cdots, Y_{n1}^{c_1}(t)\}$
$Y_{n1}(t)$
$\{Y_{nk}^1(t), \cdots, Y_{nk}^{c_k}(t)\}$
$Y_{nk}(t)$
$\{Y_{nq}^1(t), \cdots, Y_{nq}^{c_q}(t)\}$
$Y_{nq}(t)$
Item 1
Item k
Item q
Item 1
Item k
Item q
$Y_1(t)$
$Y_n(t)$
$X_1(t)$
$X_n(t)$

where for all $i = 1, \cdots, n, k = 1, \cdots, q$ $(Y_{ik}(t) = (Y_{ik}^1(t), \cdots, Y_{ik}^{c_k}(t))^\top)$ is the vector of longitudinal observations and, for all $i, (X_i(t))$ is an unobservable (latent) process. The variable $Y_{ik}^s(t), s = 1, \cdots, c_k$, denotes the response to category s at instant t, of the person i to the item k. These model designs the $q \times (c_1 + c_2 + \cdots + c_q)$ vector responses $(Y_{i1}(t), Y_{i2}(t), \cdots, Y_{iq}(t))$ at every occasion t, the responses of individual i are categorical for q-items. In other words, every item has a vector of categories responses. This questionnaire is administered on different occasions to the same individuals. A characteristic of the models considered is that each response vector $Y_{ik}(t)$ depends only on the corresponding latent trait $X_i(t)$, as shown in figure (3.1). In this context , the model can be represented at any time t by the processes of observations $\{Y_{ik}^s(t)\}$ which have values 0 or 1. The value 1 indicates that the category s has been observed, and 0 if it is not, $s = 1, \cdots, c_k$, with c_k denoting the categories number of item k. For each individual at instant t one has $q(c_1 + \cdots + c_q)$ vectors of responses with values 0 and 1.

3.3 Modeling

In this section, we construct the observation and state equations for our model. The marginal likelihood is written. We compute the parameters estimation of the model, and study their consistency and asymptotic normality.
One denotes the conditional probability π by

$$\Pr[Y_{ik}(t) = (y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t)) \mid X_i(t) = x_i(t)] = \pi((y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t)) \mid X_i(t)),$$

$i = 1, \cdots, n,\ k = 1, \cdots, q,\ t = 1, \cdots, T,$

3.3.1 The observation equation

We make the following assumptions:

1. The conditional probability of $Y_{ik}(t)$ given $X_i(t)$ is a multinomial distribution. More explicitly, for all $i = 1, \cdots, n,\ k = 1, \cdots, q,\ t = 1, \cdots, T$,

$$\Pr[Y_{ik}(t) = (y^1_{ik}(t), \cdots y^{c_k}_{ik}(t)) \mid X_i(t) = x_i(t)] = \prod_{s=1}^{c_k} [\pi^s_{ik}(t)]^{y^s_{ik}(t)}, \tag{3.3.1}$$

where

$$\begin{aligned} \pi^s_{ik}(t) &= \frac{\exp[\eta^s_{ik}(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta^j_{ik}(t)]} \quad \text{if } s < c_k, \\ \pi^{c_k}_{ik}(t) &= \frac{1}{1 + \sum_{j=1}^{c_k} \exp[\eta^j_{ik}(t)]}, \end{aligned} \tag{3.3.2}$$

and $\sum_{s=1}^{c_k} y^s_{ik}(t) = 1, \sum_{s=1}^{c_k} \pi^s_{ik}(t) = 1$.

The link function $\eta^s_{ik}(t)$ is defined with the logit function as follows

$$\begin{aligned} \eta^s_{ik}(t) = logit(\pi^s_{ik}(t)) &= \log\left[\frac{\pi^s_{ik}(t)}{\pi^{c_k}_{ik}(t)}\right] \\ &= \log\left[\frac{\pi^s_{ik}(t)}{1 - \sum_{j=1}^{c_k-1} \pi^j_{ik}(t)}\right] = \mathbf{u}_i^\top(t)\beta^s_k + X_i(t), \end{aligned}$$

and

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

$\mathbf{u}_i(t) = (u_{i1}(t), \cdots, u_{ir}(t))^\top$ is the vector of independent covariate variables and r is their number. For $k = 1, \cdots, q$, the $\beta^s_k = (\beta^s_{k1}, \cdots, \beta^s_{kr})^\top$ are vectors of unknown regression parameters.

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

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

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

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

3.3.2 The state equation

We first give some examples of usual nonlinear time series models.

3.3.2.1 The nonlinear time series models

Tjøstheim (1994) and Turkman et al.(2014) defined nonlinear time series models that broadly are classified into the following categories:

1. **Parametric models for the conditional mean**
   In these models appears the conditional mean function of the process X_t as a nonlinear function of the past observations. In other hand, the conditional variance is constant. The general model is written as :

   $$X_t = F(X_{t-1}, \theta) + \varepsilon_t,$$

   where F is a known nonlinear function, (ε_t) is the noise process and θ is the unknown parameter vector.

   $$\mathrm{E}[X_i(t) \mid X_i(t-1) = x] = F(x, \theta).$$

   Here are some examples of these models :

   (a) Exponential AR:

   $$F(x, \theta) = \{\theta_1 + \theta_2) \exp(-\theta_3 x^2)\}x, \quad \theta = (\theta_1, \theta_2, \theta_3)^\top$$

   (b) Logistic AR:

   $$F(x, \theta) = \theta_1 x + \theta_2 x\{[1 + \exp(-\theta_3(x - \theta_4))]^{-1} - 1/2\}, \ \theta_3 > 0, \ \theta = (\theta_1, \theta_2, \theta_3, \theta_4)^\top$$

2. **Parametric models for the conditional variance**
   In these models, the conditional variance function of the process X_t is a nonlinear function of the past observations. Whereas the conditional mean is constant, the general model is written as :

   $$X_t = H(X_{t-1}, \theta)\varepsilon_t$$

   This model is widely used in financial applications as ARCH and GARCH models. ARCH(1) model has the form

   $$X_t = (\theta_1 + \theta_2 X_{t-1}^2)^{1/2} \varepsilon_t,$$

   where $\mathrm{Var}[X_t \mid X_{t-1} = x] = (\theta_1 + \theta_2 x_{t-1}^2)\sigma_\varepsilon^2$, with $\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2$.

3. **Mixed parametric models for the conditional mean and variance**
   In these models both the conditional mean and variance functions of the process X_t appear as nonlinear functions of the past observations. A general model containing both a conditional mean and conditional variance component is the conditional heteroscedastic autoregressive nonlinear (CHARN) model, written on the form
   $$X_t = F(X_{t-1}, \theta_1) + H(X_{t-1}, \theta_2)\varepsilon_t.$$
   This model contains special bilinear models of the form:
   $$X_t = \theta_1 + \theta_2 X_{t-1} + \theta_3 X_{t-1}\varepsilon_t + \varepsilon_t.$$
   The above example includes nonlinear dynamics both in the mean and the variance, since
   $$\mathrm{E}(X_t \mid X_{t-1}) = \theta_1 + \theta_2 X_{t-1}$$
   and
   $$\mathrm{Var}[X_t \mid X_{t-1}] = (1 + \theta_3 X_{t-1})^2 \sigma_\varepsilon^2.$$

Some properties of such models are studied in Tjøstheim (1986), Tjøstheim (1994), Brockwell and Davis (1991), Brockwell and Davis(2013), Shumway and Stoffer(2001), and Tong(1990). The parameter estimation is discussed in Ngatchou-Wandji (2008) among others.

3.3.2.2 First- order CHARN latent process

We describe the state equation by first- order CHARN model :

$$X_i(t) = F[X_i(t-1), \mathbf{u}_i(t), \gamma] + H[X_i(t-1), \mathbf{u}_i(t), \delta]\varepsilon_i(t), \tag{3.3.3}$$

where:

- γ, δ are the model parameters.
- $(\varepsilon_i(t))$ is the gaussian noise process for the state process :
  $$\varepsilon_i(t) \sim \mathcal{N}(0; R_t), R_t > 0.$$
- $(\mathbf{u}_i(t))$ is the covariate variable process, $\mathbf{u}_i(t) \in \mathbb{R}^r$, we recall r is the number of covariate variable process.
- $F(.,.,.) : \mathbb{R} \times \mathbb{R}^r \times \mathbb{R}^l \longrightarrow \mathbb{R}$, is a non-linear function.
- $H(.,.,.) : \mathbb{R} \times \mathbb{R}^r \times \mathbb{R}^l \longrightarrow \mathbb{R}$, is a non-linear function.

Remark: From simple computations, if one assumes that the ε_i 's have common density function f, then the conditional distribution of $X_i(t)$ given $X_i(t-1) = x$ is given by

$$\begin{aligned} f^{(i)}_{X_i(t-1)=x}(z) &= \frac{1}{H(x, \mathbf{u}_i(t), \delta)} f\left[\frac{z - F(x, \mathbf{u}_i(t), \gamma)}{H(x, \mathbf{u}_i(t), \delta)}\right] \\ &= \frac{1}{\sqrt{2\pi R_t^2 H(x, \mathbf{u}_i(t), \delta)}} \exp\left[\frac{-(z - F(x, \mathbf{u}_i(t), \gamma))^2}{2R_t H^2(x, \mathbf{u}_i(t), \delta)}\right], \end{aligned} \tag{3.3.4}$$

From the above remark, it's clear that $(X_i(t) : 1 \leq t \leq T)$ satisfies

$$p(X_i(t) \mid X_i(t-1), X_i(t-2), \cdots, X_i(1)) = p(X_i(t) \mid X_i(t-1)) \sim \mathcal{N}(\mu_i(t), V_i(t)),$$

where for all $i = 1, \cdots, n,\ t = 1, \cdots, T$, the conditional mean $\mu_i(t)$ and the conditional variance $V_i(t)$ are defined by

$$\mu_i(t) = \mathrm{E}[X_i(t) \mid X_i(t-1), \mathbf{u}_i(t)] = F[X_i(t-1), \mathbf{u}_i(t), \gamma]$$

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

Then, the joint law g_i of $\mathbf{X}_i = (X_i(0), X_i(1), \cdots, X_i(T))^\top$ is obtained easily by successive conditionings:

$$\begin{aligned} g_i(\mathbf{X}_i) &= \prod_{t=1}^{T} p(X_i(t) \mid X_i(t-1)) \times p(X_i(0)) = \prod_{t=0}^{T} \mathcal{N}(\mu_i(t), V_i(t)), \\ &= \frac{1}{\sqrt{(2\pi)^T} \prod_{t=0}^{T} \sqrt{V_i(t)}} \exp\left\{-\sum_{t=0}^{T} \frac{[X_i(t) - \mu_i(t)]^2}{2V_i(t)}\right\}. \end{aligned} \tag{3.3.5}$$

3.3.3 The marginal likelihood

For all $i = 1, \cdots, n$, let

$$\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top, \quad \mathbf{Y}_i(t) = (Y_{i1}^\top(t) \cdots, Y_{iq}^\top(t))^\top, t = 0, 1, 2, \cdots, T.$$

with

$$Y_{ik}(t) = (Y_{ik}^{(1)}(t), \cdots, Y_{ik}^{(c_k)}(t))^\top \quad, t = 0, 1, 2, \cdots, T, k = 1, \cdots, q,$$

and

$$\mathbf{X}_i = (X_i(0), X_i(1), \cdots, X_i(T))^\top, \quad d(\mathbf{X}_i) = (d(X_i(0)), d(X_i(1)), \cdots, d(X_i(T)))^\top.$$

Let $\theta = (\beta, \gamma, \delta)$ with β a $q-$ dimensional vectors $\beta = (\beta_1^\top, \cdots, \beta_q^\top)^\top$, with $\beta_k^\top$s, being

$c_k \times r$ matrices, where c_k is categories number of 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 $\mathbf{y}_i(t) = (y_{i1}^\top(t) \cdots, y_{iq}^\top(t))^\top$, define

$$\begin{aligned}
\Pr(\mathbf{Y}_1^\top &= \mathbf{y}_1^\top, \mathbf{Y}_2^\top = \mathbf{y}_2^\top, \cdots, \mathbf{Y}_n^\top = \mathbf{y}_n^\top) = \prod_{i=1}^{n} \int \cdots \int p(\mathbf{y}_i \mid \mathbf{X}_i; \theta) g_i(\mathbf{X}_i) d(\mathbf{X}_i) \\
&= \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \Pr(Y_{ik}(t) = (y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t)) \mid X_i(t)) g_i(\mathbf{X}_i) d(\mathbf{X}_i) \\
&= \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{y_{ik}^s(t)} \times g_i(\mathbf{X}_i) d(\mathbf{X}_i). \\
&= \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \prod_{s=1}^{c_k} \left[\frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \right]^{y_{ik}^s(t)} \times g_i(\mathbf{X}_i) d(\mathbf{X}_i). \quad (3.3.6)
\end{aligned}$$

where $g_i(\mathbf{X}_i)$ is the joint density function of the latent variables $\mathbf{X}_i$ defined by equation (3.3.5). Then the likelihood is given by

$$\begin{aligned}
p(\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \cdots, \mathbf{Y}_n^\top) &= \prod_{i=1}^{n} \int \cdots \int \prod_{t=0}^{T} \prod_{k=1}^{q} \prod_{s=1}^{c_k} \left[\frac{\exp[\mathbf{u}_i^\top(t) \beta_k^s + X_i(t))]}{1 + \sum_{j=1}^{c_k} \exp[\mathbf{u}_i^\top(t) \beta_k^j + X_i(t)]} \right]^{Y_{ik}^s(t)} \\
&\quad \times g_i(\mathbf{X}_i) d(\mathbf{X}_i), \quad (3.3.7)
\end{aligned}$$

3.3.4 The EM algorithm

The EM algorithm is a general iterative method to obtain maximum likelihood estimators in incomplete data cases. If $\theta^{(0)}$ denotes a starting value for θ, the $(p+1)$-th cycle of the EM algorithm consists of the following two steps for $p = 0, 1, \cdots$:

Expectation -step: compute the expectation $Q(\theta \mid \theta^{(p)})$ as follows

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

The E-step for our model yields :

$$Q(\theta \mid \theta^{(p)}) = -\frac{nT}{2} \log(2\pi) + G_1 + G_2 \quad (3.3.8)$$

where

$$\begin{aligned}
G_1 &= -\frac{1}{2} \sum_{i=1}^{n} \int \cdots \int \sum_{t=0}^{T} \left\{ \log V_i(t) + \frac{[X_i(t) - \mu_i(t)]^2}{V_i(t)} \right\} \\
&\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i). \quad (3.3.9)
\end{aligned}$$

$$\begin{aligned} G_2 &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=0}^{T}\sum_{s=1}^{c_k} Y_{ik}^s(t)\int\cdots\int \log[\pi_{ik}^s(t)] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i), \end{aligned} \tag{3.3.10}$$

and $p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})$ is the conditional density of the latent vectors $\mathbf{X}_i$ given the observation vectors $\mathbf{Y}_i$ which is later computed by the particle filtering algorithm.

Maximizing step:

$$\theta^{(p+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(p)})$$

3.3.5 Estimation of first- order CHARN latent processes

We recall that the parameter vector is $\theta = (\beta_1^\top, \cdots, \beta_q^\top, \gamma^\top, \delta^\top)^\top, k = 1, \cdots, q$, and $\beta_k = (\beta_k^{1\top}, \cdots, \beta_k^{c_k\top})^\top$. By applying the E-M algorithm MLE can be found as follows : maximize with respect to the β, only the part G_2, $\beta_k^{s(p+1)}$ is the solution of the following equation:

$$\sum_{i=1}^{n}\sum_{t=0}^{T}\int\cdots\int \left\{\mathbf{u}_i^\top(t)D_{ik}^\top(X_t)\Sigma_{ik}^{-1}(X_t)[Y_{ik}^s(t) - \pi_{ik}^s(t)]\right\} \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i) = 0,$$

where

$$\pi_{ik}^s(t) = \frac{\exp[\mathbf{u}_i^\top(t)\beta_k^s + X_i(t)]}{1 + \sum_{j=1}^{c_k}\exp[\mathbf{u}_i^\top(t)\beta_k^j + X_i(t)]}.$$

The derivation of β_k^s is given in Appendix **(B.1)**. Where

$$\begin{aligned} D_{ik}^s(X_t) &= \frac{\partial \pi_{ik}^s(t)}{\partial \eta_{ik}^s} \\ &= \frac{\partial\{\exp[\eta_{ik}^s(t)]/[1 + \sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]\}}{\partial \eta_{ik}^s(t)} \\ &= \frac{(1 + \sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]).(\exp[\eta_{ik}^s(t)]) - (\exp[\eta_{ik}^s(t)])^2}{(1 + \sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)])^2} \\ &= \frac{\exp[\eta_{ik}^s(t)]\{(1 + \sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]) - (\exp[\eta_{ik}^s(t)])\}}{(1 + \sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)])^2} \end{aligned}$$

$$\begin{aligned}
&= \frac{\exp[\eta_{ik}^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]} \times \frac{(1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]) - \exp[\eta_{ik}^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]} \\
&= \pi_{ik}^s(t) \times \left\{ \frac{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]} - \frac{\exp[\eta_{ik}^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]} \right\} \\
&= \pi_{ik}^s(t) \times \left\{ 1 - \frac{\exp[\eta_{ik}^s(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^j(t)]} \right\} \\
&= \pi_{ik}^s(t)[1-\pi_{ik}^s(t)]
\end{aligned}$$

$\Sigma_{ik}(X_t) = cov(Y_{ik}(t))$ has generic elements

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

Maximizing the term G_1 with respect to $\gamma, \gamma^{(p+1)}$ is the solution of the following equation:

$$\sum_{i=1}^{n} \int \cdots \int \left\{ \sum_{t=0}^{T} \frac{[X_i(t)-\mu_i(t)]}{V_i(t)} \frac{\partial \mu_i(t)}{\partial \gamma} \right\} \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i) = 0.$$

Maximizing the term G_1 with respect to $\delta, \delta^{(p+1)}$ is the solution of the following equation:

$$\frac{1}{2}\sum_{i=1}^{n} \int \cdots \int \left\{ \sum_{t=0}^{T} \left[\frac{-1}{V_i(t)} + \frac{[X_i(t)-\mu_i(t)]^2}{V_i^2(t)} \right] \frac{\partial V_i(t)}{\partial \delta} \right\} \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i) = 0. \qquad (3.3.11)$$

3.3.6 Information matrix via the EM algorithm

In this subsection, we compute the observed information matrix of our model by implementing Oakes' identity (1999) presented in chapter 2.
Oakes (1999) derived the following identity that involves two components:

$$\mathcal{I}(\theta) = -\left\{ \frac{\partial^2 Q(\theta \mid \theta^{(p)})}{\partial \theta^2} \mid_{\theta^{(p)}=\theta} + \frac{\partial^2 Q(\theta \mid \theta^{(p)})}{\partial \theta^{(p)} \partial \theta} \mid_{\theta^{(p)}=\theta} \right\}. \qquad (3.3.12)$$

The first component is directly computed by the EM algorithm. This component is the second-order derivative of the conditional expected value of the complete data log-likelihood given the observed data.
The first component in (3.3.12) the block-diagonal matrix [1] is defined as follows:

$$\frac{\partial^2 Q(\theta \mid \theta^{(p)})}{\partial \theta^2} = diag\left(\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial \beta^2}, \frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial \gamma^2}, \frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial \delta^2} \right)$$

[1] **Block matrix**: In mathematics, a block matrix is a matrix that is interpreted as having been broken into sections called blocks or submatrices.

The second component in (3.3.12) is the first-order derivative of the score, for the same expected log-likelihood, with respect to the current value of the parameters. This component known as the "missing information", is defined as :

$$\frac{\partial^2 Q(\theta \mid \theta^{(p)})}{\partial\theta^{(p)}\partial\theta} = \left(\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\theta^{(p)}\partial\beta}, \frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\theta^{(p)}\partial\gamma}, \frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial\theta^{(p)}\partial\delta}\right)$$

First component : Second-order derivative with respect to β_k^s yields :

$$\begin{aligned}\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\beta_k^{s2}} &= -\sum_{i=1}^{n}\sum_{t=0}^{T}\int\cdots\int\left[\mathbf{u}_i^\top(t)D_{ik}^\top(X_t)\Sigma_{ik}^{-1}(X_t)D_{ik}(X_t)\mathbf{u}_i(t)\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i). \qquad (3.3.13)\end{aligned}$$

second-order derivative with respect to γ and δ as follows:

$$\begin{aligned}\frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\gamma^2} &= \sum_{i=1}^{n}\int\cdots\int\left[\frac{\partial^2\mu_i(t)/\partial\gamma^2}{V_i(t)}\right] \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i). \\ &= \sum_{i=1}^{n}\int\cdots\int\left[\frac{\frac{\partial^2}{\partial\gamma^2}(F[X_i(t-1), \mathbf{u}_i(t), \gamma])}{H^2(x_{t-1}, \mathbf{u}_i(t), \delta)R^2}\right] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i). \qquad (3.3.14)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial\delta^2} &= \frac{1}{2}\sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\left[\frac{1}{V_i^2(t)} - 2\frac{[X_i(t)-\mu_i(t)]^2}{V_i^3(t)}\right]\frac{\partial^2 V_i(t)}{\partial\delta^2}\right\} \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i). \qquad (3.3.15)\end{aligned}$$

Second component : In order to compute the second component, first-order derivatives of the expected values in (3.3.11), (3.3.11) and (3.3.11) are used to obtain the following formulas

$$\begin{aligned}\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\theta^{(p)}\partial\beta_k^s} &= \sum_{i=1}^{n}\sum_{t=0}^{T}\int\cdots\int\left\{\mathbf{u}_i^\top(t)D_{ik}^\top(X_t)\Sigma_{ik}^{-1}(X_t)[Y_{ik}^s(t) - \pi_{ik}^s(t)]\right\} \\ &\quad \times\frac{\partial p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})\theta^{(p)})}{\partial\theta^{(p)}}d(\mathbf{X}_i). \qquad (3.3.16)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\theta^{(p)}\partial\gamma} &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\frac{[X_i(t)-\mu_i(t)]}{V_i(t)}\frac{\partial\mu_i(t)}{\partial\gamma}\right\} \\ &\quad \times\frac{\partial p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})\theta^{(p)})}{\partial\theta^{(p)}}d(\mathbf{X}_i). \qquad (3.3.17)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial \theta^{(p)} \partial \delta} &= \frac{1}{2}\sum_{i=1}^{n}\int \cdots \int \left\{ \sum_{t=0}^{T} \left[\frac{-1}{V_i(t)} + \frac{[X_i(t)-\mu_i(t)]^2}{V_i^2(t)} \right] \frac{\partial V_i(t)}{\partial \delta} \right\} \\ &\quad \times \frac{\partial p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})\theta^{(p)})}{\partial \theta^{(p)}} d(\mathbf{X}_i). \qquad (3.3.18)\end{aligned}$$

The derivation of the posterior density is obtained by using a numerical methods to approximate the derivative via the particle filter proposed by Poyiadjis et al. (2005). The formula of derivation of the posterior density is presented later in the section of the posterior distribution.

3.3.6.1 Fisher information :

The Fisher information matrix for each parameter can be written as follows :

- The Fisher information matrix of parameter β as follows :

$$\mathcal{I}(\beta) = -\left\{ \frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial \beta_k^{s2}} + \frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial \theta^{(p)} \partial \beta_k^s} \right\}, \qquad (3.3.19)$$

  where $\frac{\partial^2 Q(\beta|\beta^{(p)})}{\partial \beta_k^{s2}}$ and $\frac{\partial^2 Q(\beta|\beta^{(p)})}{\partial \theta^{(p)} \partial \beta_k^s}$ are obtained by equations (3.3.13) and (3.3.16), respectively.

- The Fisher information matrix of parameter γ is :

$$\mathcal{I}(\gamma) = -\left\{ \frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial \gamma^2} + \frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial \theta^{(p)} \partial \gamma} \right\}, \qquad (3.3.20)$$

  where $\frac{\partial^2 Q(\gamma|\gamma^{(p)})}{\partial \gamma^2}$ and $\frac{\partial^2 Q(\gamma|\gamma^{(p)})}{\partial \theta^{(p)} \partial \gamma}$ are computed by equations (3.3.14) and (3.3.17), respectively.

- The Fisher information matrix of parameter δ is given by :

$$\mathcal{I}(\delta) = -\left\{ \frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial \delta^2} + \frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial \theta^{(p)} \partial \delta} \right\}, \qquad (3.3.21)$$

  where $\frac{\partial^2 Q(\delta|\delta^{(p)})}{\partial \delta^2}$ and $\frac{\partial^2 Q(\delta|\delta^{(p)})}{\partial \theta^{(p)} \partial \delta}$ are given by equations (3.3.15) and (3.3.18), respectively.

3.3.6.2 Assumptions

In the literature to have the consistency of MLE's, the underlying density (likelihood function) must satisfy certain "regularity conditions" about the sample $\mathbf{Y}_i$ and the parameters θ. For more details, see Lehmann and Casella (1998, section 6.3). Under the following assumptions, the asymptotic properties of MLE's of our model can be established:

(A.1) For all $i = 1, \cdots, n$, the observations $\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top$, are such that for all $t = 1, \cdots, T$ and $k = 1, \cdots, q$, the conditional probability $\Pr(Y_{ik}(t) \mid \mathbf{X}_i)$ is described by a multinomial distribution.

(A.2) For all $i = 1, \cdots, n$ and $k = 1, \cdots, q$, the vectors $Y_{ik}(1), \cdots, Y_{ik}(T)$ are conditionally independent given the vector of latent variable $X_i(1) = x_i(1), \cdots X_i(T) = x_i(T)$:

$$\begin{aligned} \Pr(Y_{ik}(1) &= y_{ik}(1), \cdots, Y_{ik}(T) = y_{ik}(T) \mid X_i(1) = x_i(1), \cdots, X_i(T) = x_i(T)) \\ &= \prod_{t=1}^{T} \Pr(Y_{ik}(t) = y_{ik}(t) \mid X_i(t) = x_i(t)) \end{aligned}$$

where $\Pr(Y_{ik}(1) = y_{ik}(1), \cdots, Y_{ik}(T) = y_{ik}(T) \mid X_i(1) = x_i(1), \cdots, X_i(T) = x_i(T))$ is the joint mass function of $\mathbf{Y}_i$ conditional to $\mathbf{X}_i$.

(A.3) For all $i = 1, \cdots, n$ and $t = 1, \cdots, T$, the vectors $Y_{i1}(t), \cdots, Y_{iq}(t)$ are conditionally independent knowing the latent variable $X_i(t)$:

$$\begin{aligned} \Pr(Y_{i1}(t) &= y_{i1}(t), \cdots, Y_{ik}(t) = y_{ik}(t) \mid X_i(t) = x_i(t)) \\ &= \prod_{k=1}^{q} \Pr(Y_{ik}(t) = y_{ik}(t) \mid X_i(t) = x_i(t)) \end{aligned}$$

where $\Pr(Y_{i1}(t) = y_{i1}(t), \cdots, Y_{ik}(t) = y_{ik}(t) \mid X_i(t) = x_i(t))$ is the joint mass function conditional to $X_i(t)$.

(A.4) The process $X_i(t)$ follows equation (4.2.3) with noise process having Gaussian density function.

(A.5) The distribution of the observation has common support. This assumption holds since $Y_{ik}(t) = (y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t))$ where $y_{ik}^s(t)$ has a value 0 or $1, i = 1, \cdots, n, k = 1, \cdots, q, t = 1, \cdots, T, s = 1, \cdots, c_k$.

(A.6) The parameters are *identifiable;* that is, if $\theta \neq \theta'$, then the marginal likelihood satisfies

$$p(Y_1, \cdots, Y_n, \theta) \neq p(Y_1, \cdots, Y_n, \theta').$$

(A.7) The parameter space Θ contains an open ball $\mathcal{H}$ such that the true parameter value θ_0 is an interior point.

For the vector of parameters $\theta = (\beta, \gamma, \delta)$, as $n \to +\infty$, and $T \to +\infty$, we give a short outline of standard $(N)^{1/2}$-asymptotic where $N = (n \times T)$. For this case typical "regularity assumptions " (A.1-A.7) are weak conditions, in particular convergence of $\mathcal{I}(\theta_N)/N = cov(\hat{\theta}_N)/N$, say as $N \to +\infty$,

$$\mathcal{I}(\theta_N)/N \to \mathcal{I}(\theta). \tag{3.3.22}$$

We have under regularity assumptions (A.1-A.7)

$$N^{-1/2}L'(\theta) \to \mathcal{N}(0, \mathcal{I}(\theta)), \tag{3.3.23}$$

where $L'(\theta) = s(\theta)$ is first-order derivative of the likelihood function with respect to θ, and is called the score function. Then the following Theorem hold under regularity assumptions (A.1-A.7) and the convergence properties to the EM algorithm

Theorem Let $\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top, i = 1, \cdots, n$, and $N = T \times n$. Let $\widehat{\theta}$ be the MLE's of θ. Under the regularity assumptions (A.1-A.7) and convergence properties of EM algorithm $\widehat{\theta}$ is asymptotically normal :

$$N^{1/2}(\widehat{\theta} - \theta) \sim \mathcal{N}(0; \mathcal{I}^{-1}(\theta)). \tag{3.3.24}$$

Proof: The likelihood function $L(\theta; Y)$ depends on θ. Denote first-order derivative of the log likelihood (with respect to $\widehat{\theta}$) by L', and second-order, and third-order derivative by L'' and L'''. Now, by using the Taylor expansion for $L'(\widehat{\theta})$ in a neighbourhood of θ. For some $\bar{\theta}$ between $\widehat{\theta}$ and θ, one has

$$\begin{aligned} L'(\widehat{\theta}) &= L'(\theta) + (\widehat{\theta} - \theta)L''(\theta) + \frac{1}{2}(\widehat{\theta} - \theta)^2 L'''(\bar{\theta}) \\ &= L'(\theta) + (\widehat{\theta} - \theta)L''(\theta) + O(\mid \widehat{\theta} - \theta \mid^2). \end{aligned} \tag{3.3.25}$$

O is used for describing the limiting behaviour of sequences for which the term $O(\mid \widehat{\theta} - \theta \mid^2)$ is asymptotically negligible. As we know $L'(\widehat{\theta})$ is 0, rearranging and multiply through by $\sqrt{N}$, one obtains:

$$\sqrt{N}(\widehat{\theta} - \theta) = \sqrt{N}L'(\theta)(-L''(\theta))^{-1}. \tag{3.3.26}$$

Therefore, the first factor in (3.3.26) can be written as

$$\begin{aligned} \sqrt{N}L'(\theta) &= \sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T} L'(\theta) - 0\right] \\ &= \sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T} L'(\theta) - E[L'(\theta)]\right] \to \mathcal{N}[0; \mathcal{I}(\theta)], \end{aligned}$$

where the score or efficient score[2] is the gradient (the vector of partial derivatives) with respect to some parameter θ. Namely $L'(\theta) = s(\theta)$ is the score function with $E[L'(\theta)] = 0$ and the Fisher information for θ is $\mathcal{I}(\theta)$ with blok-diagonal matrix as follows $\mathcal{I}(\theta) = diag(\mathcal{I}(\beta), \mathcal{I}(\gamma), \mathcal{I}(\delta))^\top$. By the law of large numbers

[2] Cox and Hinkley. (1979), p.107

$$-L''(\theta) \to \mathcal{I}(\theta)$$

Then

$$\begin{aligned}\sqrt{N}(\widehat{\theta}-\theta) = \sqrt{N}L'(\theta)(-L''(\theta))^{-1} &\to \mathcal{N}\left(0; \mathcal{I}(\theta)[\mathcal{I}(\theta)\mathcal{I}'(\theta)]^{-1}\right) \\ &\to \mathcal{N}(0; \mathcal{I}^{-1}(\theta))\end{aligned}$$

3.4 The Posterior distribution

The estimation formulas for parameters uses the conditional distribution for the state space given the observations (posterior distribution) $p(\mathbf{X}_i \mid \mathbf{Y}_i)$. This density can be computed by a Bayesian approach. In the linear-gaussian state-space model the posterior distribution certainly is gaussian. Consequently, the conditional mean and variance-covariance matrix of the state vector are computed by Kalman filter and smoother recursions. With the general non-gaussian state-space model, the distribution of the state space $X_i(t)$ is generally non-gaussian. Actually, the model in equations (3.3.1)and (3.3.3) is a *non-gaussian state space model.* Therefore, it is necessary to use the particles methods to find the approximation to the posterior distribution.
Here, we compute this distribution by the auxiliary iterated extended Kalman particle filter (AIEKPF) method, proposed by Xi *et al.* (2015). We develop this algorithm with our model where we find the derivation of the equations of posterior mode and posterior covariance to the extended Kalman filtering method, presented in Appendix **A.2**.

3.4.1 The Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF)

As we mentioned in chapter 2, the main idea of the (AIEKPF) algorithm is that the importance density function is generated by the Iterated Extended Kalman Particle Filter (IEKF) method within an Auxiliary Particle filter (APF) method. The support of state variable $(X_i(t))$ for each individual of state and observation equations (3.3.1, 3.3.3) is given as particle form $\{\mathbf{X}_i^m(t), m = 1, 2, \cdots, N\}$ with associated weights $\{w_i^m(t), m = 1, 2, \cdots, N\}$ and $\mathbf{X}_i(t) = (X_i(0), X_i(1), \cdots, X_i(T))^\top$ is the set of the state variables for one individual at time $t = 1, 2, \cdots, T$. Recall the observation to an individual $\mathbf{Y}_i(t) = (Y_i(1), \cdots, Y_i(T))^\top$ where $i = 1, \cdots, n, t = 1, \cdots, T$.
In chapter 2, we noted that one can approximate the posterior distribution by a set of N particles as follows

$$p(X_t \mid Y_{t-1}) \approx \sum_{m=1}^{N} w_{t|t-1}^{(m)} \delta(X_t - X_t^{(m)}).$$

The posterior mean is approximated as

$$\widehat{X}_{t|t-1} \approx \sum_{m=1}^{N} w_{t|t-1}^{(m)} X_t^{(m)}.$$

Then, the posterior probability density function can be approximated for an individual as

$$\widehat{p}(\mathbf{X}_i(t) \mid \mathbf{Y}_i(t)) \approx \sum_{m=1}^{N} w_i^m(t)\delta(\mathbf{X}_i(t) - \mathbf{X}_i^m(t)), \tag{3.4.1}$$

where $\delta(.)$ refers to the Dirac delta function, and the weights are normalized as $\sum_{m=1}^{N} w_i^m(t) = 1$. The weight $w_i^m(t)$ of an individual is selected using the formula of importance sampling, which is written as

$$w_i^m(t) \propto \frac{p[\mathbf{X}_i^m(t) \mid \mathbf{Y}_i(t)]}{q[\mathbf{X}_i^m(t) \mid \mathbf{Y}_i(t)]}, \tag{3.4.2}$$

where $q[\mathbf{X}_i^m(t) \mid \mathbf{Y}_i(t)]$ defines the importance density.
Actually, APF methodology depends on the idea that introduces an auxiliary variable, $\varsigma_i, i = 1, \cdots, n$ which plays an important role of index of the mixture component, the augmented joint distribution $p(X_i(t), \varsigma_i \mid \mathbf{Y}_i(t))$ with this extra auxiliary variable is updated for an individual i as follows :

$$\begin{aligned} p(X_i(t), \varsigma_i = m \mid \mathbf{Y}_i(t)) &\propto p(\mathbf{Y}_i(t) \mid X_i(t))p(X_i(t), \varsigma_i = m \mid \mathbf{Y}_i(t)) \\ &= p(\mathbf{Y}_i(t) \mid X_i(t))p(X_i(t) \mid \varsigma_i = m, \mathbf{Y}_i(t)) \\ &\quad \times p(m \mid \mathbf{Y}_i(t)). \\ &= p(\mathbf{Y}_i(t) \mid X_i(t))p(X_i(t) \mid X_i^m(t-1))w_i^m(t-1) \\ &= \prod_{k=1}^{q} p(Y_{ik}(t) \mid X_i(t)) \times p(X_i(t) \mid X_i^m(t-1))w_i^m(t-1) \\ &= \prod_{k=1}^{q} \mathcal{M}(\pi_{ik}(t)) \times \mathcal{N}(\mathrm{E}[X_i^m(t)], \mathrm{Var}[X_i^m(t)]) \\ &\quad \times w_i^m(t-1), \end{aligned} \tag{3.4.3}$$

Previously, we assume that $\mathbf{Y}_i(t) = (Y_{i1}^\top(t), \cdots, Y_{iq}^\top(t))^\top$, and

$$p(\mathbf{Y}_i(t) \mid X_i(t)) = \prod_{k=1}^{q} p(Y_{ik}(t) \mid X_i(t)),$$

where the symbol $\mathcal{M}$ refers to multinomial distribution and symbol $\mathcal{N}$ refers to Normal distribution with mean and variance given respectively by

$$\begin{aligned} \mathrm{E}[X_i^m(t)] &= F(x_i^m(t-1), \mathbf{u}_i^m(t), \gamma^m), \\ \mathrm{Var}[X_i^m(t)] &= H^2(x_i^m(t-1), \mathbf{u}_i^m(t), \delta^m)R_t. \end{aligned} \tag{3.4.4}$$

The APF generates a sample $\{X_i^m(t), \varsigma_i^m\}_{m=1}^N$, from the joint distribution $p(X_i(t), \varsigma_i \mid \mathbf{Y}_i(t))$, where $\varsigma_i^m(\varsigma_i^m \equiv \{\varsigma_i = m\})$ refers to the index of particle m's parent to an individual. The importance density used in APF has the property that

$$\begin{aligned} q(X_i(t), \varsigma_i \mid \mathbf{Y}_i(t)) &\propto p(\mathbf{Y}_i(t) \mid \mu_i^m(t))p(X_i(t) \mid X_i^m(t-1))w_i^m(t-1) \\ &\propto \prod_{k=1}^{q} p(Y_{ik}(t) \mid X_i(t)) \times p(X_i(t) \mid X_i^m(t-1))w_i^m(t-1) \\ &\propto \prod_{k=1}^{q} \mathcal{M}\left[\pi_{ik}(\mathbf{u}_i^m(t), \mu_i^m(t))\right] \times \mathcal{N}(\mathrm{E}[X_i^m(t)]; \mathrm{Var}[X_i^m(t)]) \\ &\times \; w_i^m(t-1) \end{aligned} \tag{3.4.5}$$

and $\mu_i^m(t)$ is some description of $X_i(t)$ given $x_i^m(t-1)$ like the mean, mode or another statistics. If it is the mean then $\mu_i^m(t) = \mathrm{E}[X_i(t) \mid X_i^m(t-1)]$ or a sample $\mu_i^m(t) \sim P(X_i(t) \mid X_i^m(t-1))$. Then, the auxiliary variable is omitted, the sample $\{X_i^m(t)\}_{m=1}^N$ is obtained. The importance density can be rewritten as

$$q(X_i(t), \varsigma_i \mid \mathbf{Y}_i(t)) = q(\varsigma_i \mid \mathbf{Y}_i(t))q(X_i(t) \mid \varsigma_i, \mathbf{Y}_i(t)). \tag{3.4.6}$$

Then

$$\begin{aligned} q(X_i(t) \mid \varsigma_i^m, \mathbf{Y}_i(t)) &\simeq p(X_i(t) \mid X_i^m(t-1)) \\ &= \mathcal{N}(\mathrm{E}[X_i^m(t)]; \mathrm{Var}[X_i^m(t)]), \end{aligned} \tag{3.4.7}$$

where $\mathrm{E}[X_i^m(t)]$ and $\mathrm{Var}[X_i^m(t)]$ are as in equations (3.4.4). By substituting Eqs. (3.4.6) and (3.4.7) into Eq. (3.4.5), one obtains for an individual i :

$$\begin{aligned} q(\varsigma_i^m \mid \mathbf{Y}_i(t)) &\simeq p(\mathbf{Y}_i(t) \mid \mu_i^m(t))w_i^m(t-1) \\ &= \prod_{k=1}^{q} \mathcal{M}\left[\pi_{ik}(\mathbf{u}_i^{m\top}(t), \mu_i^m(t))\right] w_i^m(t-1). \end{aligned} \tag{3.4.8}$$

The importance weights of an individual i are recursively updated as

$$w_i^m(t) \propto \frac{p[X_i^m(t), \varsigma_i^m \mid \mathbf{Y}_i(t)]}{q[X_i^m(t), \varsigma_i^m \mid \mathbf{Y}_i(t)]}. \tag{3.4.9}$$

Substituting Eqs.(3.4.3) and (3.4.5) into Eq.(3.4.9), one obtains

$$\begin{aligned} w_i^m(t) &\propto \frac{p(\mathbf{Y}_i(t) \mid X_i^m(t))p(X_i(t) \mid X_i^m(t-1))w_i^m(t-1)}{p(\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t))p(X_i(t) \mid X_i^m(t-1))w_i^m(t-1)} \\ &= \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} = \frac{\prod_{k=1}^{q} p[Y_{ik}(t) \mid X_i^m(t)]}{\prod_{k=1}^{q} p[Y_{ik}(t) \mid \mu_i^{\varsigma_i^m}(t)]} \\ &= \prod_{k=1}^{q} \frac{\mathcal{M}\left[\pi_{ik}(X_i^m(t), \mathbf{u}_i^m(t))\right]}{\mathcal{M}\left[\pi_{ik}(\mu_i^{\varsigma_i^m}(t), \mathbf{u}_i^m(t))\right]}. \end{aligned} \tag{3.4.10}$$

The IEKF recursions computes the following recursive approximation to the true posterior filtering density of an individual

$$p(X_i(t) \mid \mathbf{Y}_i(t)) = \mathcal{N}\left(\widehat{X}_{ij}(t); P_{ij}(t)\right) \tag{3.4.11}$$

where $i = 1, \cdots, n$, $t = 0, \cdots, T, j = 1, \cdots, c$ and c is the number of the Iterated Extended Kalman Filter (IEKF), $\widehat{X}_{ij}(t)$ is the iterative value of the $X_i(t)$ on the jth iteration, and $P_{ij}(t)$ is the covariance matrix of $\widehat{X}_{ij}(t)$.
According to the IEKF update equation , $\widehat{X}_{ij}(t)$ and $P_{ij}(t)$ can be updated for an individual as follows :

$$X_{ij}(t \mid t) = X_{ij}(t \mid t-1) + K_{ij}(t)[\widetilde{Y}_i(t) - \widehat{\pi}_i(t)] \tag{3.4.12}$$

$1 \le i \le n, 1 \le k \le q, 1 \le t \le T, 1 \le s \le c_k$, where

$$\widetilde{Y}_i(t) = D_i^{-1}(t)\{Y_i(t) - \widehat{\pi}_i(t)\} + \eta_i(t)$$

and $\widehat{\pi}_i(t) = (\widehat{\pi}_{i1}^\top(t), \cdots, \widehat{\pi}_{ik}^\top(t), \cdots \widehat{\pi}_{iq}^\top(t))^\top$, $\widehat{\pi}_{ik}(t) = (\widehat{\pi}_{ik}^1(t), \cdots, \widehat{\pi}_{ik}^{ck}(t))^\top$ since

$$\widehat{\pi}_{ik}^s(t) = \frac{\exp[\mathbf{u}_i^\top(t)\beta_k^s + X_i(t \mid t-1)]}{1 + \sum_{r=1}^{c_k} \exp[\mathbf{u}_i^\top(t)\beta_k^s + X_i(t \mid t-1)]}, \ s = 1, \cdots, c_k.$$

$$\begin{aligned} K_{ij}(t) &= P_{ij}(t \mid t-1)B_{ij}^\top(t)[B_{ij}(t)P_{ij}(t \mid t-1)B_{ij}(t) + \Sigma_{it}^{-1}]^{-1} \\ P_{ij}(t \mid t) &= (I - K_{ij}(t)B_{ij}(t)P_{ij}(t \mid t-1), \end{aligned} \tag{3.4.13}$$

where $K_{ij}(t)$ the Kalman gain matrix for individual i at instant t in the jth iteration. The importance density for each particle to one individual i is generated by using IEKF method within the APF method as follows

$$q(X_i^m(t), \varsigma_i \mid \mathbf{Y}_i(t)) = \mathcal{N}\left(\widehat{X}_{ij}^{\varsigma_i^m}(t); P_{ij}^{\varsigma_i^m}(t)\right), \tag{3.4.14}$$

where $m = 1, 2, \cdots N$, and N is the number of the particles.
In brief, AIEKPF method uses IEKF method to update the equations with the new observations to compute the mean and covariance of the importance density for each particle at time $t-1$ to one individual. Subsequently, the mth particle is sampled from the distribution.

3.4.1.1 Auxiliary Iterated Extended Kalman Filter (AIEKPF) Algorithm

The AIEKPF algorithm can be performed for an individual by the following steps:

1. Initialization $(t = 0)$: For $m = 1, \cdots, N$, generate the states (particles) $X_i^m(0)$ from the prior $p(X_i(0)) = \mathcal{N}(\widehat{X}_i^m(0); P_i^m(0))$, and set

$$\widehat{X}_i^m(0) = \mathrm{E}[X_i^m(0)] = F(x_i^m(0), \mathbf{u}_i^m(0), \delta^m)$$

and

$$\begin{aligned}P_i^m(0) &= \mathrm{E}[(X_i^m(0) - \widehat{X}_i^m(0))(X_i^m(0) - \widehat{X}_i^m(0))^\top] = \mathrm{Var}(X_i^m(0)) \\ &= H^2(x_i^m(0), \mathbf{u}_i^m(0), \delta^m)R_0\end{aligned}$$

For $t = 1, \cdots, T$, repeat the following steps:

2. For $m = 1, \cdots, N$, generate $\mu_i^m(t) \sim \mathcal{N}(\mathrm{E}[X_i^m(t)]; \mathrm{Var}[X_i^m(t)])$.

3. For $m = 1, \cdots, N$, using the IEKF algorithm to update the particles

   3-1. Calculate the Jacobians $A_i^m(t), C_i^m(t)$ of the process model

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

   $$C_i^m(t) = H(x_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \delta)$$

   3-2. Predict the particle with the IEKF:

   $$X_i^m(t \mid t-1) \approx F(x_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \gamma)$$

   $$P_i^m(t \mid t-1) = A_i^m(t)P_i^m(t-1 \mid t-1)A_i^{m\top}(t) + C_i^m(t)R_tC_i^{m\top}(t)$$

   3-3. For $j = 1, \cdots c$ (c is the iteration number of the IEKF)

   a- Calculate the Jacobians $, B_{ij}^m(t)$

   $$B_{ij}^m(t) = \frac{\partial \pi_{it}(\mathbf{u}_i(t), x)}{\partial x} \mid_{x=x_{ij}^m(t|t-1)}$$

   b- The state estimation error covariance $P_{ij}^m(t)$ is updated with equation 3.4.13

   c- The state estimate is updated $X_{ij}^m(t)$ with equation 3.4.12.

4. For $m = 1, \cdots N$, compute

$$\begin{aligned}w_i^m(t) &= q(m \mid \mathbf{Y}_i(t)) \\ &\propto \prod_{k=1}^{q} \mathcal{M}\left[\pi_{ik}(\mathbf{u}_i^m(t), \mu_i^m(t))\right] w_i^m(t-1),\end{aligned}$$

5. Apply the Resample method to obtain the index ς_i^m of particle m's parent.

6. Importance sampling : for $m = 1, \cdots, N$,

6-1. Draw samples

$$\begin{aligned} X_i^m(t) &\sim q(X_i(t), \varsigma_i^m \mid \mathbf{Y}_i(t)) \\ &= \mathcal{N}(\widehat{X}_{ij}^{\varsigma_i^m}(t), P_{ij}^{\varsigma_i^m}(t)), \end{aligned} \tag{3.4.15}$$

where $j = c$, where the $j = 1, \cdots, c$ the last iteration of IEKF before this step is equal to c.

6-2. Compute the importance weights of particles by using

$$\begin{aligned} w_i^m(t) &= \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} = \frac{\prod_{k=1}^q p[Y_{ik}(t) \mid X_i^m(t)]}{\prod_{k=1}^q p[Y_{ik}(t) \mid \mu_i^{\varsigma_i^m}(t)]} \\ &= \prod_{k=1}^q \frac{\mathcal{M}\left[\pi_{ik}(X_i^m(t), \mathbf{u}_i^m(t))\right]}{\mathcal{M}\left[\pi_{ik}(\mu_i^{\varsigma_i^m}(t), \mathbf{u}_i^m(t))\right]}. \end{aligned} \tag{3.4.16}$$

6-3. Normalize the weights

$$w_i^m(t) = \frac{w_i^m(t)}{\sum_{m=1}^N w_i^m(t)}.$$

7. Output: a set of weighted particles (samples) to an individual $[\{X_i^m(t), w_i^m(t)\}_{m=1}^N], i = 1, \cdots, n$.

3.4.2 Optimality of the Auxiliary Iterated Extended Kalman Particle Filter(AIEKPF) Algorithm

Under the following regularity condition $C.1, C.2$ and $C.3$ given in [Chopin (2004) , Douc and Moulines (2008)] one can prove the *consistency* and *asymptotic normality* of a weighted particles (samples) $[\{x_t^m, w_t^m\}_{m=1}^N]$.

C.1 The initial sample to an individual $[\{X_i^m(0), w_i^m(0)\}_{m=1}^N]$ is *consistent* and *asymptotically normal.*

C.2 The selection step consists in multinomial resampling.

C.3 $\bar{\varphi}$ and $\sigma_{APF}^2(\varphi)$ are finite quantities, where $\bar{\varphi}$ is the expectation of a function φ and $\sigma_{APF}^2(\varphi)$ the variance .

Definition 3.1 (Asymptotic normality). Under the conditions (C.1-C.3) a sample for one individual $[\{X_i^m(t), w_i^m(t)\}_{m=1}^N]$ as $N \to \infty$, for all $i = 1, \cdots, n$ and $t = 1, \cdots, T$,

$$\sqrt{N}(\widehat{\varphi}_i(t) - \bar{\varphi}_i(t)) \Rightarrow \mathcal{N}[0; \sigma_{APF}^2(\varphi_i(t))], \tag{3.4.17}$$

where $\varphi_i(t)$ is the function of trajectories $X_i(t)$ and $\bar{\varphi}_i(t)$ is the expectation of $\widehat{\varphi}_i(t)$ with respect to the filtering distribution.

$$\bar{\varphi}_i(t) = \int \varphi_i(t) p(X_i(t) \mid \mathbf{Y}_i(t)) dX_i(t)$$

and

$$\widehat{\varphi}_i(t) = \sum_{m=1}^{N} W_i^m(t) \varphi_i^m(t), \tag{3.4.18}$$

where $W_i^m(t) = w_i^m(t)[\sum_{m=1}^N w_i^m(t)]^{-1}$, and

$$\begin{aligned} w_i^m(t) &= \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} = \frac{\prod_{k=1}^q p[Y_{ik}(t) \mid X_i^m(t)]}{\prod_{k=1}^q p[Y_{ik}(t) \mid \mu_i^{\varsigma_i^m}(t)]} \\ &= \prod_{k=1}^{q} \frac{\mathcal{M}\left[\pi_{ik}(X_i^m(t), \mathbf{u}_i^m(t))\right]}{\mathcal{M}\left[\pi_{ik}(\mu_i^{\varsigma_i^m}(t), \mathbf{u}_i^m(t))\right]}. \end{aligned} \tag{3.4.19}$$

At time $t = 1$ for an individual, the formula of the variance yields

$$\sigma^2_{APF}(\varphi_i(1)) = \int \frac{(p(X_i(1) \mid \mathbf{Y}_i(1))^2}{q_1(X_i(1))} (\varphi_i(1) - \bar{\varphi}_i(1))^2 dX_i(1), \tag{3.4.20}$$

and for $t > 1$,

$$\begin{aligned} \sigma^2_{APF}(\varphi_i(t)) &= \int \frac{[p(X_i(1) \mid \mathbf{Y}_i(t))]^2}{q_1(X_i(1))} \triangle\, \varphi_i^{(1,t)}(X_i(1))^2 dX_i(1) \\ &+ \sum_{r=2}^{t-1} \int \frac{[p(X_i(1:r) \mid \mathbf{Y}_i(t))]^2}{\widehat{p}(X_i(r-1) \mid \mathbf{Y}_i(r)) q_r(X_i(r) \mid X_i(r-1))} \\ &\quad \times \triangle\, \varphi_i^{(r,t)}(X_i(r))^2 dX_i(r) \\ &+ \int \frac{[p(X_i(t) \mid \mathbf{Y}_i(t))]^2}{\widehat{p}(X_i(t-1) \mid \mathbf{Y}_i(t)) q_t(X_i(t) \mid X_i(t-1))} \\ &\quad \times (\varphi_i(t) - \bar{\varphi}_i(t))^2 dX_i(t). \end{aligned} \tag{3.4.21}$$

where

$$\triangle\, \varphi_i^{(r,t)}(X_i(r)) = \int \varphi_i(X_i(t)) p(X_i(r+1:t) \mid \mathbf{Y}_i(r+1:t)) dX_i(r+1:t) - \bar{\varphi}_i(t)$$

and

$$q_1(X_i(1)) = \mathcal{N}(\mathrm{E}(X_i^m(1)), \mathrm{Var}(X_i^m(1)))$$

$$q_t(X_i(t) \mid X_i(t-1)) = \mathcal{N}(\mathrm{E}(X_i^m(t)), \mathrm{Var}(X_i^m(t)))$$

$$\widehat{p}(X_i(t) \mid \mathbf{Y}_i(t)) \approx \sum_{m=1}^{N} w_i^m(t)\delta(X_i(t) - X_i^m(t)).$$

where $\delta(.)$ denotes the Dirac delta function.

3.4.3 Derivation of the posterior density function

We calculated the information matrix via the EM algorithm, in the second component the derivative of posterior density with respect to $\theta^{(p)}$ is used, we recall $\theta^{(p)}$ is the vector of parameters at iterative p. By using the approach proposed by Poyiadjis et al. (2005). This approach approximates the filter derivative based on the sequence of marginal distributions $p(X_i(t) \mid \mathbf{Y}_i(t))$. Let

$$p(X_i(t) \mid \mathbf{Y}_i(t), \theta^{(p)}) = \frac{\xi[X_i(t), \mathbf{Y}_i(t), \theta^{(p)}]}{\int \xi[X_i(t), \mathbf{Y}_i(t), \theta^{(p)}]dX_i(t)}, \tag{3.4.22}$$

where

$$\begin{aligned}\xi(X_i(t), \mathbf{Y}_i(t), \theta^{(p)}) &= p(\mathbf{Y}_i(t) \mid X_i(t), \theta^{(p)})p(X_i(t) \mid \mathbf{Y}_i(t-1), \theta^{(p)}) \\ &= \left[\prod_{k=1}^{q} \mathcal{M}[\pi_{ik}(t)]\right] \times p(X_i(t) \mid \mathbf{Y}_i(t-1), \theta^{(p)}),\end{aligned}$$

with $\mathcal{M}$ denoting to the multinomial distribution. Then the gradient of (3.4.22) is given by

$$\begin{aligned}\nabla p(X_i(t) \mid \mathbf{Y}_i(t), \theta^{(p)}) &= \frac{\nabla\xi[X_i(t), \mathbf{Y}_i(t), \theta^{(p)}]}{\int \xi[X_i(t), \mathbf{Y}_i(t), \theta^{(p)}]dX_i(t)} \\ &\quad -p(X_i(t) \mid \mathbf{Y}_i(t))\frac{\int \nabla\xi[X_i(t), \mathbf{Y}_i(t), \theta^{(p)}]dX_i(t)}{\int \xi[X_i(t), \mathbf{Y}_i(t), \theta^{(p)}]dX_i(t)},\end{aligned} \tag{3.4.23}$$

where ∇ is the first-order derivatives of the density with respect to $\theta^{(p)}$, and

$$\begin{aligned}\nabla\xi(X_i(t), \mathbf{Y}_i(t), \theta^{(p)}] &= p(\mathbf{Y}_i(t) \mid X_i(t)) \int p(X_i(t) \mid X_i(t-1))p(X_i(t-1) \mid \mathbf{Y}_i(t-1) \\ &\quad \times[\nabla \log p(\mathbf{Y}_i(t) \mid X_i(t)) + \nabla \log p(X_i(t) \mid X_i(t-1))]dX_i(t-1) \\ &+ p(\mathbf{Y}_i(t) \mid X_i(t)) \int p(X_i(t) \mid X_i(t-1)) \\ &\quad \times\nabla p(X_i(t-1) \mid \mathbf{Y}_i(t-1))dX_i(t-1).\end{aligned}$$

The filter derivative approximations by the weighted particles can be obtained as follows:

$$\widehat{\nabla} p(X_i(t) \mid \mathbf{Y}_i(t), \theta^{(p)}) = \sum_{m=1}^{N} \widetilde{w}_i^m(t)\alpha_i^m(t)\delta(X_i(t) - X_i^m(t)), \tag{3.4.24}$$

where $\delta(.)$ denotes the Dirac delta function, and for all $i = 1, \cdots, n, t \geq 1$,

$$\widetilde{w}_i^m(t) = \frac{w_i^m(t)}{\sum_{m=1}^{N} w_i^m(t)},$$

with

$$w_i^m(t) = \prod_{k=1}^{q} \frac{\mathcal{M}\left[\pi_{ik}(X_i^m(t), \mathbf{u}_i^m(t))\right]}{\mathcal{M}\left[\pi_{ik}(\mu_i^{s_i^m}(t), \mathbf{u}_i^m(t))\right]}. \tag{3.4.25}$$

Recall that $w_i^m(t)$ is the weighted particles calculated by the Auxiliary Particle Filter (APF) algorithm or Auxiliary Iterated Extended Kalman Particle Filter(AIEKPF) Algorithm, and

$$\widetilde{w}_i^m(t)\alpha_i^m(t) = \frac{\rho_i^m(t)}{\sum_{m=1}^{N} w_i^m(t)} - \widetilde{w}_i^m(t)\frac{\sum_{m=1}^{N} \rho_i^m(t)}{\sum_{m=1}^{N} w_i^m(t)}, \tag{3.4.26}$$

where $\alpha_i^m(t)$ corresponds to an approximation of the so-called score $\widehat{\nabla} p(X_i^m(t) \mid Y_i(t), \theta^{(p)})$:

$$\begin{aligned} \alpha_i^m(t) &= [w_i^m(t)]^{-1} \frac{\widetilde{\nabla} p(X_i^m(t) \mid \mathbf{Y}_i(t), \theta^{(p)})}{q[X_i^m(t) \mid \mathbf{Y}_i(t), \theta^{(p)}]} \\ &= [w_i^m(t)]^{-1} \prod_{k=1}^{q} \frac{\widetilde{\nabla} \mathcal{M}\left[\pi_{ik}(X_i^m(t), \mathbf{u}_i^m(t))\right]}{\mathcal{M}\left[\pi_{ik}(\mu_i^{s_i^m}(t), \mathbf{u}_i^m(t))\right]}, \end{aligned} \tag{3.4.27}$$

and

$$\rho_i^m(t) = \frac{\widetilde{\nabla} \xi[X_i^m(t), \mathbf{Y}_i(t), \theta^{(p)}]}{p[X_i^m(t) \mid \mathbf{Y}_i(t)]} \tag{3.4.28}$$

since

$$\begin{aligned} \widetilde{\nabla} \xi[X_i^m(t), \mathbf{Y}_i(t), \theta^{(p)}] &= \sum_{m=1}^{N} \widetilde{w}_i^m(t) p(\mathbf{Y}_i(t) \mid X_i(t)) p(X_i(t) \mid X_i^m(t-1)) \\ &\quad [\nabla \log p(\mathbf{Y}_i(t) \mid X_i(t)) + \nabla \log p(X_i(t) \mid X_i^m(t-1)) + \alpha_i^m(t-1)] \\ &= \sum_{m=1}^{N} \widetilde{w}_i^m [\prod_{k=1}^{q} \mathcal{M}(\pi_{ik}(t)] \times \mathcal{N}(\mu_i^m(t), V_i^m(t)) \\ &\quad \left[\frac{\partial l(\beta)}{\partial \beta_k^s} + \frac{\partial l(\gamma)}{\partial \gamma} + \frac{\partial l(\delta)}{\partial \delta} + \alpha_i^m(t-1)\right] \end{aligned}$$

where

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_k^s} &= \sum_{i=1}^{n}\sum_{t=1}^{T}[y_{ik}^s(t) - \pi_{ik}^s(t)]\mathbf{u}_i^\top(t)D_{ik}^\top(X_t)\Sigma_{ik}^{-1}(X_t) \\
\frac{\partial l(\gamma)}{\partial \gamma} &= \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{[X_i(t) - \mu_i(t)]}{V_i(t)} \times \frac{\partial \mu_i(t)}{\partial \gamma} \\
\frac{\partial l(\delta)}{\partial \delta} &= \sum_{i=1}^{n}\sum_{t=1}^{T}\left[\frac{1}{V_i(t)} - \frac{[X_i(t) - \mu_i(t)]^2}{V_i^2(t)}\right] \times \frac{\partial V_i(t)}{\partial \delta}.
\end{aligned}$$

3.5 Posterior mode estimation

In nonlinear non gaussian state space model, there are two approaches to estimate the state variable. First, the Bayesian approach which uses A maximum a posteriori (MAP). Regrettably, this approach leads to numerical integration problems, often difficult for more complicated models. Several methods are proposed to solving this problem. One of these Recent Markov Chain Monte Carlo methods solving this problem by repeated sampling from approximative posterior distributions. But there are still open questions about achieving the convergence of sampling schemes.
Second, penalized likelihood estimation approach was proposed by Fahrmeir and Kaufmann (1991), Fahrmeir (1992) and Fahrmeir and Wagenpfeil (1997). This approach can also be interpreted as nonparametric method for the state space models. They estimated the posterior mode by using Fisher scoring method by iterative Kalman filtering and smoothing. They used the numerical method to maximize the penalized likelihood called *"Working Kalman filtering and smoother (WKFS)*" with the exponential family distribution.
In this chapter, we used the two approaches to find the estimation of state variable. We develop the two approaches with our model to find the posterior mode as follows :

1. Penalized likelihood estimation :

   a - Gauss-Newton and Fisher-scoring Filtering and smoothing algorithms.

   b - Working extended Kalman filter and smoother algorithms.

2. Maximum a posteriori sequence estimation (MAP):

   a - The auxiliary iterated extended Kalman filter particle filter-MAP.

3.5.1 Penalized likelihood estimation

The posterior mode smoother is defined as follows

$$a \equiv \left\{a^\top(0 \mid T), a^\top(1 \mid T), \cdots, a^\top(T \mid T)\right\} \in \mathbb{R}^m,$$

with $m = (T+1)n$, where

$$\begin{array}{rcl} a(0 \mid T) & = & ((a_1(0 \mid T), \cdots, a_n(0 \mid T))^\top, \\ \vdots & \vdots & \vdots \qquad\qquad \vdots \\ a(T \mid T) & = & (a_1(T \mid T), \cdots, a_n(T \mid T))^\top. \end{array}$$

The posterior likelihood of $\mathbf{X}$ by Bayes' theorem

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{Y}) &= \frac{1}{p(\mathbf{Y})} \left\{ \prod_{i=1}^{n} \prod_{k=1}^{q} \prod_{t=1}^{T} p(Y_{ik}(t) \mid X_i(t)) \right\} \\ &\quad \times \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} g_i(X_i(t)) \right\} \left\{ \prod_{i=1}^{n} g_i(X_i(0)) \right\} \end{aligned} \tag{3.5.1}$$

$p(\mathbf{Y})$ does not depend on $\mathbf{X}$

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{Y}) &\propto \left\{ \prod_{i=1}^{n} \prod_{k=1}^{q} \prod_{t=1}^{T} p(Y_{ik}(t) \mid X_i(t)) \right\} \\ &\quad \times \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} g_i(X_i(t)) \right\} \left\{ \prod_{i=1}^{n} g_i(X_i(0)) \right\} \end{aligned} \tag{3.5.2}$$

where the prior likelihood $g_i(X_i(0))$

$$g_i(X_i(0)) = \frac{1}{\sqrt{2\pi V_i(0)}} \exp\left\{ \frac{-[X_i(0) - \mu_i(0)]^2}{2V_i(0)} \right\}. \tag{3.5.3}$$

Taking logarithms and inserting the densities of $\mathbf{X}$ in equation (3.3.5) and $X_i(0)$ in equation (3.5.3), the panelized log-likelihood function yields

$$PL : \mathbb{R}^m \longrightarrow \mathbb{R},\ m = (T+1)n$$

$$\begin{aligned} PL(\mathbf{X}) &= \sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=1}^{T} \{\log p(Y_{ik}(t) \mid X_i(t))\} \\ &\quad + \sum_{i=1}^{n} \sum_{t=1}^{T} \log g_i(X_i(t)) + \sum_{i=1}^{n} \log g_i(X_i(0)) \end{aligned} \tag{3.5.4}$$

where

$$\begin{aligned}\log p(Y_{ik}(t) \mid X_i(t)) &= \log \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{Y_{ik}^s(t)} \\ &= \sum_{s=1}^{c_k} Y_{ik}^s(t) \log \pi_{ik}^s(t).\end{aligned}$$

Then

$$PL(\mathbf{X}) = \sum_{i=1}^{n} \sum_{k=1}^{q} \sum_{t=1}^{T} \sum_{s=1}^{c_k} \{Y_{ik}^s(t) \log \pi_{ik}^s(t)\} + G_1 + G_2 \tag{3.5.5}$$

where

$$\begin{aligned}G_1 = &- \frac{1}{2}\sum_{i=1}^{n}(X_i(0) - \mu_i(0))V_i^{-1}(0)(X_i(0) - \mu_i(0)) \\ G_2 &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_i(t) - \mu_i(t))'V_i^{-1}(t)(X_i(0) - \mu_i(t)),\end{aligned}$$

The posterior mode smoother yields

$$a \equiv (a^\top(0 \mid T), a^\top(1 \mid T), \cdots, a^\top(T \mid T))^\top = \arg\max_X \{PL(\mathbf{X})\}, \tag{3.5.6}$$

The maximization of $p(\mathbf{X} \mid \mathbf{Y})$ is equivalent to maximization of the penalized log-likelihood (3.5.4).
Numerical maximization of the penalized log-likelihood can be performed by various algorithms. In this chapter we use two algorithms to find the posterior mode. First, the iterative forward-backward Gauss-Newton (Fisher-scoring) algorithms. Second, the Working Extended Kalman Filter and Smoother (WEKFS) which gives the same numerical results computed by Gauss-Newton (Fisher-scoring) algorithms.

3.5.2 Gauss-Newton and Fisher-scoring Filtering and smoothing

A maximization of the penalized log-likelihood $PL(\mathbf{X})$ with generally best performance to approximation quality can be computed by Gauss-Newton or Fisher-scoring iterations. In other words, this can be also achieved by applying extended Kalman filtering and smoothing to a "working model" in each Fisher-scoring iteration. Therefore, the penalized log-likelihood criterion (3.5.4) can be presented in compact matrix notation as:

$$PL(\mathbf{X}) = l(\mathbf{X}) - \frac{1}{2}\mathbf{X}^\top \mathcal{K}\mathbf{X}, \tag{3.5.7}$$

where

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

The penalty matrix for one individual $\mathcal{K}_i$ is symmetric and block-tridiagonal, with blocks that can be easily computed from (3.5.4):

$$\mathcal{K}_i = \begin{bmatrix} \mathcal{K}_{i,00} & \mathcal{K}_{i,01} & \cdots & \cdots & 0 \\ \mathcal{K}_{i,10} & \mathcal{K}_{i,11} & \mathcal{K}_{i,12} & \cdots & \vdots \\ \vdots & \mathcal{K}_{i,21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & & \ddots & \mathcal{K}_{i,T-1,T} \\ 0 & \cdots & \cdots & \mathcal{K}_{i,T,T-1} & \mathcal{K}_{i,T,T} \end{bmatrix}$$

with $\mu_i(t) = F(X_i(t-1), \mathbf{u}_i(t), \gamma)$, the conditional mean is rewritten by matrix notation as follows

$$F(X_i(t-1), \mathbf{u}_i(t), \gamma) = \Psi_{it}X_i(t),$$

then

$$\begin{aligned} \mathcal{K}_{i,t-1,t} &= \mathcal{K}_{i,t,t-1}^{\top}, \quad i = 1, \cdots, n, t = 1, \cdots, T \\ \mathcal{K}_{i,00} &= \Psi_{i,1}^{\top}R_t^{-1}\Psi_{i,1}, \\ \mathcal{K}_{i,tt} &= R_t^{-1} + \Psi_{i,t+1}^{\top}R_t^{-1}\Psi_{i,t+1}, \quad i = 1, \cdots, n, t = 1, \cdots, T \\ \Psi_{i,T+1} &= 0, \\ \mathcal{K}_{i,t-1,t} &= -\Psi_{i,t}^{\top}R_t^{-1}, \quad i = 1, \cdots, n, t = 1, \cdots, T, \end{aligned}$$

We now describe the steps of Fisher scoring in vector notation. Recall that for all $i = 1, \cdots, n$,

$$\mathbf{Y}_i = (\mathbf{Y}_i^{\top}(0), \mathbf{Y}_i^{\top}(1), \cdots, \mathbf{Y}_i^{\top}(T))^{\top}, \;\; \mathbf{Y}_i(t) = (Y_{i1}^{\top}(t) \cdots, Y_{iq}^{\top}(t))^{\top}, t = 0, 1, 2, \cdots, T.$$

Fahrmeir and Wagenpfeil (1997) assumed $\mathbf{Y}_i^{\top}(0) = a_i(0)$ and the vectors of expectations are defined 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 $\pi_i^{\top}(0) = X_i(0)$. We recall the conditional mean and variance of one individual

$$\begin{aligned} \mathrm{E}(Y_{ik}(t) \mid X_i(t)) &= \pi_{ik}^{s}(t), \;\; s = 1, \cdots, c_k \\ \mathrm{Var}(Y_{ik}(t) \mid X_i(t)) &= \Sigma_{ik}(X_t), \end{aligned}$$

where $Y_{ik}(t) = (Y_{ik}^{1}(t), \cdots, Y_{ik}^{c_k}(t))^{\top}$, and $\Sigma_{ik}(X_t)$ has generic elements

$$\sigma_{ik}^{sm}(t) = \begin{cases} \pi_{ik}^{s}(t)[1-\pi_{ik}^{s}(t)], & \text{if } s = m \\ -\pi_{ik}^{s}(t)\pi_{ik}^{m}(t) & \text{if } s \neq m \end{cases} \qquad (3.5.8)$$

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

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

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

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

Since $D_{it}(X_t)$ is first-order derivative of the conditional probability $\pi_i(t)$ evaluated at $\eta_i(t)$, the score function of $l(\mathbf{X})$ in (3.5.7) is given by $S_i(\mathbf{X}) = (\widehat{S}_{i0}(X_0), \widehat{S}_{i1}(X_1), \cdots, \widehat{S}_{iT}(X_T))^\top$, where for all $i = 1, \cdots, n$,

$$S_i(\mathbf{X}) := D_i(\mathbf{X})\Sigma_i^{-1}(\mathbf{X})\left\{\mathbf{Y}_i(t) - \Pi_i(\mathbf{X})\right\}, \qquad (3.5.9)$$

with components

$$\begin{aligned} \widehat{S}_i(X_0) &= V_i^{-1}(0)(a_i(0) - X_i(0)) && (3.5.10) \\ \widehat{S}_{it}(X_t) &= D_{it}(X_t)\Sigma_{it}^{-1}(X_t)\left\{\mathbf{Y}_i(t) - \pi_{it}(X_t)\right\}, t = 1, \cdots, T, && (3.5.11) \end{aligned}$$

The first-order derivatives of $PL(\mathbf{X})$ in (3.5.7) are

$$M(\mathbf{X}) = \partial PL(\mathbf{X})/\partial \mathbf{X} = S(\mathbf{X}) - \mathcal{K}\mathbf{X}, \qquad (3.5.12)$$

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

$$\mathcal{I}_i(\mathbf{X}) = D_i(\mathbf{X})\Sigma_i^{-1}(\mathbf{X})D_i'(\mathbf{X}) \qquad (3.5.13)$$

with diagonal blocks

$$\begin{aligned} \mathcal{I}_{i0}(X_0) &= V_i^{-1}(0) && (3.5.14) \\ \mathcal{I}_{it}(X_t) &= D_{it}(X_t)\Sigma_{it}^{-1}(X_t)D_{it}^\top(X_t), t = 1, \cdots, T. && (3.5.15) \end{aligned}$$

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

$$M(\widehat{\mathbf{X}} \mid \mathbf{Y}) \approx M(\mathbf{X}^0) - \mathcal{I}(\mathbf{X}^0) \times \left\{\mathbf{X}^1 - \mathbf{X}^0\right\}$$

since $M(\widehat{\mathbf{X}} \mid \mathbf{Y}) = 0$, a single Fisher scoring to the next iterate $\mathbf{X}^1 \in R^m$, with $m = (T+1)n$ is as follows

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

This can be rewritten as

$$\mathbf{X}^1 = \left\{\mathcal{I}(\mathbf{X}^0) + \mathcal{K}\right\}^{-1}\mathcal{I}(\mathbf{X}^0)\widetilde{\mathbf{Y}}, \qquad (3.5.16)$$

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

$$\widetilde{Y}_i := \left\{D_i^{-1}((\mathbf{X}))\right\}^\top \{\mathbf{Y}_i - \Pi_i(\mathbf{X}))\} + \eta_i(\mathbf{X}), \tag{3.5.17}$$

with components

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

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

3.5.3 Working Extended Kalman Filter and Smoother (WEKFS)

In the following algorithm, $a_{t|t-1}, , a_{t|t}, a_{t|T}$ are numerical approximations to predicted, filtered, smoothed values of posterior mode to $\mathbf{X}$ and the corresponding $P_{t|t-1}, P_{t|t}, P_{t|T}$ are numerical approximations to predicted, filtered, smoothed values of error covariance matrices. We expand the non-linear functions F and H in the state equation and the conditional probability $\pi_{ik}^s(t)$ in Taylor series expansion about the estimation value of $\mathbf{X}$. The derivation of Extended Kalman filter and smoother is presented in Appendix **(A.2)**

Initialization:

$$\begin{aligned}
a_i(0 \mid 0) &= a_i(0), \\
V_i(0 \mid 0) &= V_i(0).
\end{aligned} \tag{3.5.18}$$

Prediction: for $t = 1, \cdots, T$,

$$\begin{aligned}
a_i(t \mid t-1) &= F(a_i(t-1 \mid t-1), \mathbf{u}_i(t), \gamma) \\
P_i(t \mid t-1) &= A_i(t)P_i(t-1 \mid t-1)A_i^\top(t) + C_i(t)R_tC_i^\top(t),
\end{aligned} \tag{3.5.19}$$

where

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

Filtering : for $t = 1, \cdots, T$,

$$\begin{aligned}
a_i(t \mid t) &= a_i(t) + K_i(t)(\widetilde{Y}_i(t) - a_i(t \mid t-1)) \\
K_i(t) &= P_i(t-1 \mid t-1)B_i^\top(t)(B_i(t)P_i(t-1 \mid t-1)B_i^\top(t) + \mathcal{I}_i^{-1}(t))^{-1} \\
P_i(t \mid t) &= (I - K_i(t)B_i(t))P_i(t-1 \mid t-1),
\end{aligned} \tag{3.5.20}$$

where

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

3.5.3.1 Iteratively extended Kalman filter and smoother (IWEKFS)

Initialization: Calculate $\mathbf{X}^0 = (a^0_{0|t}, a^0_{1|t}, \cdots, a^0_{T|t})^\top$ with General Kalman filter and smoother (GKFS) recursions. Set iteration index $k = 0$.

STEP 1: Starting with $\mathbf{X}^k$, compute $\mathbf{X}^{k+1}$ by performing Working extended Kalman filter and smoother (WEKFS) recursions.

STEP 2: If the convergence criterion is achieved : STOP, else set $k = k + 1$ and go to STEP 1.

3.5.4 The maximum a posteriori (MAP) with particle filtering and smoothing algorithm

The second approach to finding the posterior mode is the MAP. Firstly, the extension of the MAP via the particle filtering is proposed by Saha *et al.* (2008) with our model. Secondly, we developed the MAP via the Auxiliary Iterated Extended Kalman Filter (AIEKPF) Algorithm. The filtering and smoothing step with our model is as follows:

3.5.4.1 Filtering step

As we mentioned in chapter 2, the particle based filter MAP estimator (pf-MAP) proposed by Saha *et al.* (2008) has the advantage that the posterior density can be approximated not only by particles shape at clouds but at any point.
The MAP estimate of the filtering density at time t to an individual is defined as follows

$$X_i^{MAP}(t) = \arg\max_{X_i(t)} p(X_i(t) \mid \mathbf{Y}_i(t)). \tag{3.5.21}$$

By using the Baye's rule , the filtering density yields

$$\begin{aligned} p(X_i(t) \mid \mathbf{Y}_i(t)) &= p(X_i(t) \mid \mathbf{Y}_i(t), \mathbf{Y}_i(t-1)) \\ &= \frac{p(\mathbf{Y}_i(t) \mid X_i(t), \mathbf{Y}_i(t-1))p(X_i(t) \mid \mathbf{Y}_i(t-1))}{p(\mathbf{Y}_i(t) \mid \mathbf{Y}_i(t-1))} \\ &= \frac{p(\mathbf{Y}_i(t) \mid X_i(t))p(X_i(t) \mid \mathbf{Y}_i(t-1))}{p(\mathbf{Y}_i(t) \mid \mathbf{Y}_i(t-1))} \end{aligned} \tag{3.5.22}$$

Therefore, the MAP estimate is as follows

$$X_i^{MAP}(t) = \arg\max_{X_i(t)} \left[\frac{p(\mathbf{Y}_i(t) \mid X_i(t))p(X_i(t) \mid \mathbf{Y}_i(t-1))}{p(\mathbf{Y}_i(t) \mid \mathbf{Y}_i(t-1))}\right]. \tag{3.5.23}$$

Since the denominator $p(\mathbf{Y}_i(t) \mid \mathbf{Y}_i(t-1))$ is independent of $X_i(t)$, the MAP estimate can be expressed as

$$\begin{aligned} X_i^{MAP}(t) &= \arg\max_{X_i(t)} \left[p(\mathbf{Y}_i(t) \mid X_i(t)) p(X_i(t) \mid \mathbf{Y}_i(t-1)) \right] \\ &= \arg\max_{X_i(t)} \left[\prod_{k=1}^{q} p(Y_{ik}(t) \mid X_i(t)) p(X_i(t) \mid \mathbf{Y}_i(t-1) \right] \\ &= \arg\max_{X_i(t)} \left[\prod_{k=1}^{q} \mathcal{M}(\pi_{ik}(t)) p(X_i(t) \mid \mathbf{Y}_i(t-1) \right]. \qquad (3.5.24) \end{aligned}$$

The main principle for evaluating MAP is the computation of the predictive density $P(X_i(t) \mid \mathbf{Y}_i(t-1))$ (the second terms in equation (3.5.24)) which is not available in known formula. The prediction density formula in Bayesian approach is as follows

$$\begin{aligned} p(X_i(t) \mid \mathbf{Y}_i(t-1)) &= \int p(X_i(t), X_i(t-1) \mid \mathbf{Y}_i(t-1)) dX_{t-1} \\ &= \int p(X_i(t) \mid X_i(t-1), \mathbf{Y}_i(t-1)) p(X_i(t-1) \mid \mathbf{Y}_i(t-1)) dX_i(t-1) \\ &= \int p(X_i(t) \mid X_i(t-1)) \times p(X_i(t-1) \mid \mathbf{Y}_i(t-1)) dX_i(t-1) \\ &= \int \mathcal{N}(\mathrm{E}[X_i^m(t)]; \mathrm{Var}[X_i^m(t)]) \times p(X_i(t-1) \mid \mathbf{Y}_i(t-1)) dX_i(t-1). \quad (3.5.25) \end{aligned}$$

The posterior distribution can be approximated in particle form by a set of N weighted particles as:

$$\widehat{p}(X_i(t) \mid \mathbf{Y}_i(T)) \approx \sum_{m=1}^{N} w_i^m(t) \delta(X_i(t) - \mathbf{x}_i^m(t)) \qquad (3.5.26)$$

Now, use equation (3.5.26) to approximate $p(X_i(t) \mid \mathbf{Y}_i(t-1))$ as

$$\begin{aligned} \widehat{p}(X_i(t) \mid \mathbf{Y}_i(t-1)) &\approx \sum_{j}^{N} p(X_i(t) \mid X_i^j(t-1)) w_i^{(j)}(t-1). \\ &\approx \sum_{j}^{N} \mathcal{N}(\mu_i^j(t); V_i^j(t)) w_i^{(j)}(t-1). \qquad (3.5.27) \end{aligned}$$

Substituting (3.5.27) into (3.5.24), the MAP estimation is obtained by finding the global maxima of the posterior distribution which is approximated by particle as

$$\begin{aligned} X_i^{MAP}(t) &= \arg \max_{X_i^{(m)}(t)} p(\mathbf{Y}_i(t) \mid X_i^{(m)}(t)) \\ &\quad \times \sum_j p(X_i^{(m)}(t) \mid X_i^{(j)}(t-1)) w_i^{(j)}(t-1) \\ &= \arg \max_{X_i^{(m)}(t)} \left[\prod_{k=1}^{q} \mathcal{M}(\pi_{ik}(t)) \times \sum_j \mathcal{N}(\mu_i^j(t), V_i^j(t)) w_i^{(j)}(t-1) \right], \end{aligned}$$

where $i = 1, \cdots, n, t = 1, \cdots, T, m = 1, \cdots, N$.

3.5.4.2 Smoothing step

Saha *et al.* (2008) extended the MAP estimation concept to the marginal smoothing

$$\begin{aligned} X_i^{MAP}(t \mid T) &= \arg \max_{X_i(t)} p(X_i(t) \mid \mathbf{Y}_i(T)) \\ &= \arg \max_{X_i^{(m)}(t)} p(X_i^{(m)}(t) \mid \mathbf{Y}_i(t)) \frac{w_i^{(m)}(t \mid T)}{w_i^{(m)}(t)}. \end{aligned} \qquad (3.5.28)$$

The filtered density $p(X_i(t) \mid \mathbf{Y}_i(t))$ at the particle cloud $\{X_i^{(m)}(t)\}_{m=1}^N$ can be approximated during the forward filtering step as

$$p(X_i(t) \mid \mathbf{Y}_i(t)) \approx \frac{p(\mathbf{Y}_i(t) \mid X_i^{(m)}(t)) \sum_j p(X_i^{(m)}(t) \mid X_i^{(j)}(t-1)) w_i^{(j)}(t-1)}{p(\mathbf{Y}_i(t) \mid \mathbf{Y}_i(t-1))} \qquad (3.5.29)$$

Since $p(\mathbf{Y}_i(t) \mid \mathbf{Y}_i(t-1))$ in equation (3.5.29) is independent of $X_i^{(m)}(t)$, $X_i^{MAP}(t \mid T)$ can be obtained by replacing $p(X_i^{(m)}(t) \mid \mathbf{Y}_i(t))$ by the filtered density

$$\begin{aligned} q(X_i^{(m)}(t) \mid \mathbf{Y}_i(t)) &= p(\mathbf{Y}_i(t) \mid X_i^{(m)}(t)) \\ &\quad \sum_j p(X_i^{(m)}(t) \mid X_i^{(j)}(t-1)) w_i^{(j)}(t-1). \\ &= \prod_{k=1}^{q} \mathcal{M}(\pi_{ik}^m(t)) \sum_j \mathcal{N}(\mu_i^j(t), V_i^j(t)) w_i^{(j)}(t-1) \end{aligned} \qquad (3.5.30)$$

The derivation of the Forward-Backward smoothing is given in Appendix **(A.4)**.

3.5.4.3 Auxiliary Iterated Extended Kalman Particle Filter -MAP(AIEKPF-MAP) Algorithm

the MAP estimation can be found via the Auxiliary Iterated Kalman Particle Filter (AIEKPF) with the weights to individual i at time t as follows :

$$
\begin{aligned}
w_i^m(t) &= \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} \\
&= \prod_{k=1}^{q} \frac{\mathcal{M}[\pi_{ik}(X_i^m(t), \mathbf{u}_i(t))]}{\mathcal{M}[\pi_{ik}(\mu_i^{\varsigma_i^m}(t), \mathbf{u}_i(t))]}
\end{aligned}
$$

The AIEKPF algorithm can be performed with the following steps:

1. Initialization $(t = 0)$: For $m = 1, \cdots, N$, draw the states (particles) $X_i^m(0)$ from the prior $p(X_i(0)) = \mathcal{N}(\widehat{X}_i^m(0), P_i(0)^m)$, and set

   $$\widehat{X}_i^m(0) = \mathrm{E}[X_i^m(0)] = F(X_i^m(0), \mathbf{u}_i^m(0), \delta^m)$$

   and

   $$
   \begin{aligned}
   P_i^m(0) &= \mathrm{E}[(X_i^m(0) - \widehat{X}_i^m(0))(X_i^m(0) - \widehat{X}_i^m(0))^\top] \\
   &= \mathrm{Var}(X_i^m(0)) \\
   &= H^2(X_i^m(0), \mathbf{u}_i^m(0), \delta^m)R_0
   \end{aligned}
   $$

   For $t = 1, \cdots, T$, repeat the following steps:

2. For $m = 1, \cdots, N$, generate $\mu_i^m(t) \sim \mathcal{N}(\mathrm{E}(X_i^m(t)); \mathrm{Var}(X_i^m(t)))$.

3. For $m = 1, \cdots, N$, update the particles via the IEKF algorithm

   3-1. Compute the Jacobians $A_i^m(t), C_i^m(t)$ of the model as follows

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

   $$C_i^m(t) = H(X_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \delta)$$

   3-2. Predict the particle with the IEKF:

   $$X_i^m(t \mid t-1) \approx F(X_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \gamma)$$

   $$P_i^m(t \mid t-1) = A_i^m(t) P_i^m(t-1 \mid t-1) A_i^{m\top}(t) + C_i^m(t) R_t C_i^{m\top}(t)$$

   3-3 For $j = 1, \cdots c$ (c is the iteration number of the IEKF)

   a- Compute the Jacobians , $B_{ij}^m(t)$

   $$B_{ij}^m(t) = \frac{\partial \pi_{it}(\mathbf{u}_i(t), x)}{\partial x} \mid_{x = x_{ij}^m(t|t-1)}$$

   b- Update the state estimation error covariance $P_{ij}^m(t)$ with Eq.(3.4.13)

   c- Update the state estimate $X_{ij}^m(t)$ with Eq.(3.4.12).

4. For $m = 1, \cdots N$, calculate

$$\begin{aligned} w_i^m(t) &= q(m \mid \mathbf{Y}_i(t)) \\ &\propto \prod_{k=1}^{q} \mathcal{M}(\pi_{ik}(\mathbf{u}_i(t), \mu_i^m(t)) w_i^m(t-1), \end{aligned}$$

5. Resample to obtain the index ς_i^m of particle m's parent.

6. Importance sampling : for $m = 1, \cdots, N$,

   6-1. Draw samples $X_i^m(t) \sim q(X_i(t), \varsigma_i^m \mid \mathbf{Y}_i(t)) = \mathcal{N}(\widehat{X}_{ij}^{\varsigma_i^m}(t); P_{ij}^{\varsigma_i^m}(t))$, where $j = c$

   6-2. Calculate importance weights of particles by using

$$w_i^m(t) = \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} = \prod_{k=1}^{q} \frac{\mathcal{M}[\pi_{ik}(x_i^m(t), \mathbf{u}_i(t))]}{\mathcal{M}[\pi_{ik}(\mu_t^{\varsigma_i^m}(t), \mathbf{u}_i(t))]}$$

   6-3. Normalize the weights

$$w_i^m(t) = \frac{w_i^m(t)}{\sum_{m=1}^{N} w_i^m(t)}.$$

Forward Filtering step

$$\begin{aligned} X_i^{MAP}(t) &= \arg\max_{X_i^{(m)}(t)} p(\mathbf{Y}_i(t)(t) \mid X_i^{(m)}(t)) \sum_j p(X_i^{(j)}(t) \mid X_i^{(j)}(t-1)) w_i^{(j)}(t-1) \\ &= \arg\max_{X_i^{(m)}(t)} \left[\prod_{k=1}^{q} \mathcal{M}(\pi_{ik}^m(t)) \sum_j \mathcal{N}(\mu_i^j(t); V_i^j(t)).w_i^{(j)}(t-1) \right]. \end{aligned}$$

Backward smoothing step

- Set $w_i^{(m)}(T \mid T) = w_i^{(m)}(t)$
- For $t = T-1, \cdots, 1$ compute the smoother importance weights as

$$\begin{aligned} w_i^{(m)}(t \mid T) &= w_i^{(m)}(t) \sum_{j=1}^{N} \left[w_i^{(j)}(t+1 \mid T) \frac{p(X_i^{(j)}(t+1) \mid X_i^{(m)}(t))}{\sum_{r=1}^{N} p(X_i^{(j)}(t+1) \mid X_i^{(r)}(t)) w_i^{(r)}(t)} \right] \\ &= w_i^{(m)(t)} \sum_{j=1}^{N} \left[w_i^{(j)}(t+1 \mid T) \frac{\mathcal{N}(\mu_i^j(t); V_i^j(t))}{\sum_{r=1}^{N} \mathcal{N}(\mu_i^r(t); V_i^r(t)) w_i^{(r)}(t)} \right]. \end{aligned}$$

- Compute the approximate smoother MAP as

$$
\begin{aligned}
X_i^{MAP}(t \mid T) &= \arg \max_{X_i^{(m)}(t)} \left\{ q(X_i^{(m)}(t) \mid \mathbf{Y}_i(t)) \frac{w_i^{(m)}(t \mid T)}{w_i^{(m)}(t)} \right\}. \\
&= \arg \max_{X_i^{(m)}(t)} \left\{ \prod_{k=1}^{q} \mathcal{M}(\pi_{ik}^{m}(t)) \sum_j \mathcal{N}(\mu_i^j(t); V_i^j(t)) w_i^{(j)}(t-1) \frac{w_i^{(m)}(t \mid T)}{w_i^{(m)}(t)} \right\}.
\end{aligned}
$$

Chapter 4

Longitudinal multicategorical processes : Generalized state space models with exponential family state noise

4.1 Introduction

In this chapter, we present longitudinal multivariate processes which are multicategorical $Y_{ik}(t) = (Y_{ik}^1(t), \cdots, Y_{ik}^{c_k}(t))^\top, i = 1, \cdots, n, t = 1, \cdots, T$ and $k = 1, \cdots, q,$. As in chapter 3, the latent processes $X_i(t)$ are described by a CHARN model but with the state noise from exponential family distribution. Practically, many of the most common distributions belong to the exponential families, as Normal, Exponential, Gamma, Chi-squared, Beta, Dirichlet, Bernoulli, Multinomial, Poisson, Wishart, Inverse Wishart and many others distributions.

In this chapter, we generalize our model with the state noises assumed to be an exponential family distribution. In the state space models if the observations and the latent variable have gaussian distribution, the posterior distribution is also a gaussian distribution. Therefore, the classical Kalman filtering can be used to estimate the posterior mean and variance matrix (Arulampalam (2002), Kitagawa(2010)). Here, the observations are from a multinomial distribution and the latent variables are from a CHARN model with noise from an exponential family distribution. Consequently, the posterior distribution is non symmetric. That is why we estimate the latent variables by the posterior mode instead of the posterior mean.

As in chapter 3, we find the posterior mode by two approaches. First, via the Working Extended Kalman Filtering recursions (WEKF). Second, by using the Auxiliary Iterated Extended Kalman Particle Filtering (AIEKPF-MAP). The models parameters are estimated through the Maximum Likelihood (MLE) method via the Expectation-Maximization (EM) algorithm. Their consistency and asymptotic normality are established. The posterior distribution $p(\mathbf{X}_i \mid \mathbf{Y}_i)$ is approximated by the Auxiliary Iterated Extended Kalman Particle filter (AIEKPF).

4.2 The state space models

In this section we define as in the preceding chapter the observation and state equations. The observation equation does not change. In other words, the observations have multinomial distribution. The state equation CHARN model but the noise process has an exponential family distribution.

4.2.1 The observation equation

1. We assume that the conditional probability of $Y_{ik}(t)$ given $X_i(t)$ is a multinomial distribution. for all $i = 1, \cdots, n,\ k = 1, \cdots, q,\ t = 1, \cdots, T$, the conditional probability can clearly written as follows

$$\Pr[Y_{ik}(t) = (y_{ik}^1(t), \cdots y_{ik}^{c_k}(t)) \mid X_i(t) = x_i(t)] = \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{y_{ik}^s(t)}, \tag{4.2.1}$$

where

$$\begin{aligned} \pi_{ik}^s(t) &= \frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \quad \text{if } s < c_k, \\ \pi_{ik}^{c_k}(t) &= \frac{1}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]}, \end{aligned} \tag{4.2.2}$$

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

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

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 vector of independent covariate variables and r is their number. For $k = 1, \cdots, q$, the $\beta_k^s = (\beta_{k1}^s, \cdots, \beta_{kr}^s)^\top$ are vectors of unknown regression parameters.

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

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

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

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

4.2.2 The state equation

The state equation is the following

$$X_i(t) = F[X_i(t-1), \mathbf{u}_i(t), \gamma] + H[X_i(t-1), \mathbf{u}_i(t), \delta]\varepsilon_i(t), \qquad (4.2.3)$$

where $\varepsilon_i(t) \sim expf(\nu_i(t), \phi_i(t))$, that is the density functions of the $\varepsilon_i(t)'s$ are :

$$f_{\varepsilon_i(t)}(z) = \exp\left\{\frac{z\nu_i(t) - b[\nu_i(t)]}{\phi_i(t)} + c[z, \phi_i(t)]\right\}. \qquad (4.2.4)$$

with γ, δ the model parameters.
It results that the conditional distribution of $x_i(t)$ given $x_i(t-1) = x$ has density function :

$$f^i_{X_i(t-1)=x}(z) = \frac{1}{H[x, \mathbf{u}_i(t), \delta]} f_z \left\{\frac{z - F[x, \mathbf{u}_i(t), \gamma]}{H[x, \mathbf{u}_i(t), \delta]}\right\} \qquad (4.2.5)$$

- $(\varepsilon_i(t))$ is the noise process for the state process.
- $F(.,.,.) : \mathbb{R} \times \mathbb{R}^r \times \mathbb{R}^l \longrightarrow \mathbb{R}$ is a non-linear function.
- $H(.,.,.) : \mathbb{R} \times \mathbb{R}^r \times \mathbb{R}^l \longrightarrow \mathbb{R}$ is a non-linear function.
- $\nu_i(t)$ is a canonical parameter.
- $\phi_i(t)$ is the dispersion or a scale parameter.
- $b[\nu_i(t)], c[z, \phi_i(t)]$ are functions taking different forms depending on the distribution of the $\varepsilon_i(t)'s$.
The common distributions from exponential family with their functions and parameters are presented in Appendix **(C.1)**).

The latent process $(X_i(t) : 1 \leq t \leq T)$ satisfies:

$$p(X_i(t) \mid X_i(t-1), X_i(t-2), \cdots, X_i(1)) = p(X_i(t) \mid X_i(t-1)).$$

Recall that

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

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

$$\begin{aligned} g_i(\mathbf{X}_i) &= \prod_{t=1}^{T} p(X_i(t) \mid X_i(t-1)) \times p(X_i(0)) \\ &= \frac{1}{\prod_{t=0}^{T} H[X_i(t-1), \mathbf{u}_i(t), \delta]} \exp\left\{ \sum_{t=0}^{T} \left[\frac{Z_i(t)\nu_i(t) - b[\nu_i(t)]}{\phi_i(t)} + c[Z_i(t), \phi_i(t)] \right] \right\}. \quad (4.2.6) \end{aligned}$$

where

$$Z_i(t) = \frac{X_i(t) - F[X_i(t-1), \mathbf{u}_i(t), \gamma]}{H[X_i(t-1), \mathbf{u}_i(t), \delta]} \quad (4.2.7)$$

4.2.3 The marginal likelihood

In this section, as chapter 3, we recall that for all $i = 1, \cdots, n$,

$$\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top, \quad \mathbf{Y}_i(t) = (Y_{i1}^\top(t) \cdots, Y_{iq}^\top(t))^\top, t = 0, 1, 2, \cdots, T.$$

with

$$Y_{ik}(t) = (Y_{ik}^{(1)}(t), \cdots, Y_{ik}^{(c_k)}(t))^\top \quad , t = 0, 1, 2, \cdots, T, k = 1, \cdots, q,$$

and

$$\mathbf{X}_i = (X_i(0), X_i(1), \cdots, X_i(T))^\top, \quad d(\mathbf{X}_i) = (d(X_i(0)), d(X_i(1)), \cdots, d(X_i(T)))^\top.$$

The parameters of the model are $\theta = (\beta, \gamma, \delta)$. β $q-$ dimensional vectors $\beta = (\beta_1^\top, \cdots, \beta_q^\top)^\top$, with $\beta_k^\top$s, are $c_k \times r$ matrices, where c_k is categories of item $k, k = 1, \cdots, q$. r denotes the number of covariates.

We recall $\mathbf{y}_i = (\mathbf{y}_i^\top(0), \mathbf{y}_i^\top(1), \cdots, \mathbf{y}_i^\top(T))^\top$ with $\mathbf{y}_i(t) = (y_{i1}^\top(t) \cdots, y_{iq}^\top(t))^\top$, the density function can be written as follows

$$\begin{aligned}
\Pr(\mathbf{Y}_1^\top &= \mathbf{y}_1^\top, \mathbf{Y}_2^\top = \mathbf{y}_2^\top, \cdots, \mathbf{Y}_n^\top = \mathbf{y}_n^\top) = \prod_{i=1}^n \int \cdots \int p(\mathbf{y}_i \mid \mathbf{X}_i; \theta) g_i(\mathbf{X}_i) d(\mathbf{X}_i) \\
&= \prod_{i=1}^n \int \cdots \int \prod_{t=0}^T \prod_{k=1}^q \Pr(Y_{ik}(t) = y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t) \mid X_i(t)) g_i(\mathbf{X}_i) d(\mathbf{X}_i) \\
&= \prod_{i=1}^n \int \cdots \int \prod_{t=0}^T \prod_{k=1}^q \prod_{s=1}^{c_k} [\pi_{ik}^s(t)]^{y_{ik}^s(t)} \times g_i(\mathbf{X}_i) d(\mathbf{X}_i). \\
&= \prod_{i=1}^n \int \cdots \int \prod_{t=0}^T \prod_{k=1}^q \prod_{s=1}^{c_k} \left[\frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \right]^{y_{ik}^s(t)} \times g_i(\mathbf{X}_i) d(\mathbf{X}_i). \qquad (4.2.8)
\end{aligned}$$

where $g_i(\mathbf{X}_i)$ is the joint density function of the latent variables $\mathbf{X}_i$ defined by equation (4.2.6). Thus, the likelihood is

$$\begin{aligned}
p(\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \cdots, \mathbf{Y}_n^\top) &= \prod_{i=1}^n \int \cdots \int \prod_{t=0}^T \prod_{k=1}^q \prod_{s=1}^{c_k} \left[\frac{\exp[\mathbf{u}_i^\top(t)\beta_k^s + X_i(t))]}{1 + \sum_{j=1}^{c_k} \exp[\mathbf{u}_i^\top(t)\beta_k^j + X_i(t)]} \right]^{Y_{ik}^s(t)} \\
&\quad \times g_i(\mathbf{X}_i) d(\mathbf{X}_i), \qquad (4.2.9)
\end{aligned}$$

4.2.4 The EM algorithm

As in chapter 3 the Expectation and Maximizing step are:

Expectation -step: compute the expectation $Q(\theta \mid \theta^{(p)})$ given by

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

The E-step for first-order Markov latent process from Exponential family distribution is :

$$Q(\theta \mid \theta^{(p)}) = G_1 + G_2 + G3 \qquad (4.2.10)$$

where

$$\begin{aligned} G_1 &= \sum_{i=1}^{n} \int \cdots \int \left\{ \sum_{t=0}^{T} \left[\frac{Z_i(t)\nu_i(t) - b[\nu_i(t)]}{\phi_i(t)} + c[Z_i(t), \phi_i(t)] \right] \right\} \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i). \qquad (4.2.11) \\ G_2 &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=0}^{T}\sum_{s=1}^{c_k} y_{ik}^{s}(t) \int \cdots \int \log[\pi_{ik}^{s}(t)] \\ &\quad \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i), \qquad (4.2.12) \\ G3 &= -\log(H[X_i(t-1), \mathbf{u}_i(t), \delta]) \qquad (4.2.13) \end{aligned}$$

and $p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})$ 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:

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

4.2.5 Estimation of first-order CHARN latent processes

We recall that the parameter vector is $\theta = (\beta_1^\top, \cdots, \beta_q^\top, \gamma^\top, \delta^\top)^\top, k = 1, \cdots, q$, and $\beta_k = (\beta_k^{1\top}, \cdots, \beta_k^{c_k\top})^\top$. We apply the E-M step for finding the MLE as follows maximize with respect to the $\beta_k^s, k = 1, \cdots, q, s = 1, \cdots, c_k$, only the part G_2, $\beta_k^{s(p+1)}$ is the solution of the following equation:

$$\sum_{i=1}^{n}\sum_{t=0}^{T} \int \cdots \int \left\{ \mathbf{u}_i^\top(t) D_{ik}^\top(X_t) \Sigma_{ik}^{-1}(X_t)[Y_{ik}^s(t) - \pi_{ik}^s(t)] \right\} \times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)}) d(\mathbf{X}_i) = 0. (4.2.14)$$

The derivation of β_k^s is given in Appendix **(B.1)**.
$D_{ik}(X_t)$ is a matrix with generic elements

$$D_{ik}^s(X_t) = \pi_{ik}^s(t)[1 - \pi_{ik}^s(t)] \qquad (4.2.15)$$

and $\Sigma_{ik}(X_t) = cov(Y_{ik}(t))$ has generic elements

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

$\gamma^{(p+1)}$ is the value for which the derivative of G_1 with respect to γ is nil. That is the solution of $\frac{\partial}{\partial\gamma}(G_1) = 0$. Since

$$\begin{aligned}\frac{\partial}{\partial\gamma}(G_1) &= \frac{\partial}{\partial\gamma}\sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\left[\frac{Z_i(t)\nu_i(t)-b[\nu_i(t)]}{\phi_i(t)}+c[Z_i(t),\phi_i(t)]\right]\right\}\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i),\\ &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\left[\frac{\nu_i(t)}{\phi_i(t)}+c'[Z_i(t),\phi_i(t)]\right]\right\}\times\frac{\partial Z_i(t)}{\partial\gamma}\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i) = 0, \qquad (4.2.16)\end{aligned}$$

where $Z_i(t)$ defined by equation (4.2.7), then, $\gamma^{(p+1)}$ is the solution of equation (4.2.16)= 0.
Also, $\delta^{(p+1)}$ is the value for which the derivative of $G_1 + G_3$ with respect to δ is nil. That is the solution of:

$$\begin{aligned}\frac{\partial}{\partial\delta}\{G_2+G_3\} &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\left[\frac{\nu_i(t)}{\phi_i(t)}+c'[Z_i(t),\phi_i(t)]\right]\right\}\frac{\partial Z_i(t)}{\partial\delta}\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i)\\ &\quad-\frac{H'[X_i(t-1),\mathbf{u}_i(t),\delta]}{H[X_i(t-1),\mathbf{u}_i(t),\delta]} = 0.\end{aligned}$$

4.2.6 Information matrix via the EM algorithm

In this section, as in chapter 3 we compute the Fisher information matrix via the Oakes' identity

First component: The first component is the second-order derivative with respect to β_k^s :

$$\begin{aligned}\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\beta_k^{s2}} &= -\sum_{i=1}^{n}\sum_{t=1}^{T}\int\cdots\int\left[\mathbf{u}_i^\top(t)D_{ik}^{\top(p)}(X_t)\Sigma_{ik}^{-1(p)}(X_t)D_{ik}^{(p)}(X_t)\mathbf{u}_i(t)\right]\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i). \qquad (4.2.17)\end{aligned}$$

The second-order derivatives with respect to γ and δ are given by :

$$\begin{aligned}\frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\gamma^2} &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}c''[Z_i(t),\phi_i(t)]\right\}\times\frac{\partial^2 Z_i(t)}{\partial\gamma^2}\\ &\quad\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i), \qquad (4.2.18)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial\delta^2} &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}c''[Z_i(t),\phi_i(t)]\right\}\frac{\partial^2 Z_i(t)}{\partial\delta^2}\times p(\mathbf{X}_i \mid \mathbf{Y}_i, \theta^{(p)})d(\mathbf{X}_i)\\ &\quad-\frac{H''[X_i(t-1),\mathbf{u}_i(t),\delta]\times H[X_i(t-1),\mathbf{u}_i(t),\delta]-2H'[X_i(t-1),\mathbf{u}_i(t),\delta]}{H^2[X_i(t-1),\mathbf{u}_i(t),\delta]}.(4.2.19)\end{aligned}$$

Second component: The second component is computed as follows

$$\begin{aligned}\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\theta^{(p)}\partial\beta_k^s} &= \sum_{i=1}^{n}\sum_{t=1}^{T}\int\cdots\int\left\{\mathbf{u}_i^\top(t)D_{ik}(X_t)\Sigma_{ik}^{-1}(X_t)[Y_{ik}^s(t)-\pi_{ik}^s(t)]\right\}\\ &\quad\times\frac{\partial p(\mathbf{X}_i \mid \mathbf{Y}_i,\theta^{(p)})}{\partial\theta^{(p)}}d(\mathbf{X}_i). \qquad (4.2.20)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\theta^{(p)}\partial\gamma} &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\left[\frac{\nu_i(t)}{\phi_i(t)}+c'[Z_i(t),\phi_i(t)]\right]\right\}\times\frac{\partial Z_i(t)}{\partial\gamma}\\ &\quad\times\frac{\partial p(\mathbf{X}_i \mid \mathbf{Y}_i,\theta^{(p)})}{\partial\theta^{(p)}}d(\mathbf{X}_i). \qquad (4.2.21)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial\theta^{(p)}\partial\delta} &= \sum_{i=1}^{n}\int\cdots\int\left\{\sum_{t=0}^{T}\left[\frac{\nu_i(t)}{\phi_i(t)}+c'[Z_i(t),\phi_i(t)]\right]\right\}\frac{\partial Z_i(t)}{\partial\delta}\frac{\partial p(\mathbf{X}_i \mid \mathbf{Y}_i,\theta^{(p)})}{\partial\theta^{(p)}}d(\mathbf{X}_i).\\ &\quad-\frac{H'[X_i(t-1),\mathbf{u}_i(t),\delta]}{H[X_i(t-1),\mathbf{u}_i(t),\delta]}. \qquad (4.2.22)\end{aligned}$$

The derivative of the posterior density is done as in chapter 3.

4.2.6.1 Fisher information :

The information matrix (Fisher information) of parameter β can be written as

$$\mathcal{I}(\beta) = -\left\{\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\beta_k^{s2}}+\frac{\partial^2 Q(\beta \mid \beta^{(p)})}{\partial\theta^{(p)}\partial\beta_k^s}\right\}, \qquad (4.2.23)$$

where $\frac{\partial^2 Q(\beta|\beta^{(p)})}{\partial\beta_k^{s2}}$ and $\frac{\partial^2 Q(\beta|\beta^{(p)})}{\partial\theta^{(p)}\partial\beta_k^s}$ are given by equations (4.2.17) and (4.2.20), respectively.

The information matrix of parameter γ is :

$$\mathcal{I}(\gamma) = -\left\{\frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\gamma^2}+\frac{\partial^2 Q(\gamma \mid \gamma^{(p)})}{\partial\theta^{(p)}\partial\gamma}\right\}, \qquad (4.2.24)$$

where $\frac{\partial^2 Q(\gamma|\gamma^{(p)})}{\partial\gamma^2}$ and $\frac{\partial^2 Q(\gamma|\gamma^{(p)})}{\partial\theta^{(p)}\partial\gamma}$ are given by equations (4.2.18) and (4.2.21), respectively.

The information matrix of parameter δ is given by :

$$\mathcal{I}(\delta) = -\left\{\frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial\delta^2}+\frac{\partial^2 Q(\delta \mid \delta^{(p)})}{\partial\theta^{(p)}\partial\delta}\right\}, \qquad (4.2.25)$$

where $\frac{\partial^2 Q(\delta|\delta^{(p)})}{\partial\delta^2}$ and $\frac{\partial^2 Q(\delta|\delta^{(p)})}{\partial\theta^{(p)}\partial\delta}$ are given by equations (4.2.19) and (4.2.22), respectively.

4.2.7 Consistency and Asymptotic normality of the maximum likelihood estimator(MLE)

We recall the parameters of our model is $\theta = (\beta^\top, \gamma^\top, \delta^\top)^\top$, where $\beta = (\beta_1^\top, \cdots, \beta_q^\top)^\top$ and $\beta_k = (\beta_k^{1\top}, \cdots, \beta_k^{c_k\top})^\top, k = 1, \cdots, q$. As in Chapter 3, we discuss the consistency and asymptotic normality to the model parameters.

4.2.7.1 Assumptions

Our assumptions remain the same as in chapter 3. Only (A4) changes to

(A.4)′ The state space $X_i(t)$ follows equation (4.2.3) with state noise from exponential family distribution.

As in chapter 3, under the regularity assumptions (A.1-A.7) and the convergence properties to the EM algorithm (which had been explained in chapter 3) we conclude that

$$N^{1/2}(\widehat{\theta} - \theta) \sim^d \mathcal{N}(0, \mathcal{I}^{-1}(\theta)). \tag{4.2.26}$$

4.3 The Posterior distribution

As in chapter 3, the estimation formulas for parameters uses the the posterior distribution $p(\mathbf{X}_i \mid \mathbf{Y}_i)$. This density can be computed by the Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF) method previously is presented in chapter 3. Therefore, its algorithm is recalled here.

4.3.1 Auxiliary Iterated Extended Kalman Filter (AIEKPF) Algorithm

The AIEKPF algorithm can be performed for the individual i in the following steps:

1. Initialization $(t = 0)$: For $m = 1, \cdots, N$, draw the states (particles) $X_i^m(0)$ from the prior $p(X_i(0)) \sim f^i_{X_i(0)}$, and set

   $$\widehat{X}_i^m(0) = \mathrm{E}[X_i^m(0)] = F(x_i^m(0), \mathbf{u}_i^m(0), \delta^m)$$

   and

   $$\begin{aligned} P_i^m(0) &= \mathrm{E}[(X_i^m(0) - \widehat{X}_i^m(0))(X_i^m(0) - \widehat{X}_i^m(0))^\top] \\ &= \mathrm{Var}(X_i^m(0)) \\ &= H^2(x_i^m(0), \mathbf{u}_i^m(0), \delta^m)R_0 \end{aligned}$$

   For $t = 1, \cdots, T$, repeat the following steps:

2. For $m = 1, \cdots, N$, generate $\mu_i^m(t) \sim f^i_{X_i(t-1)}$.

3. For $m = 1, \cdots, N$, update the particles using the IEKF algorithm

   3-1. Compute the Jacobians $A_i^m(t), C_i^m(t)$ of the process model

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

   $$C_i^m(t) = H(x_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \delta)$$

   3-2. Predict the particle with the IEKF:

   $$X_i^m(t \mid t-1) \approx F(x_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \gamma)$$

   $$P_i^m(t \mid t-1) = A_i^m(t) P_i^m(t-1 \mid t-1) A_i^{\top m}(t) + C_i^m(t) R C_i^{\top m}(t)$$

   3-3. For $j = 1, \cdots c$ (c is the iteration number of the IEKF)

   a- Compute the Jacobians , $B_{ij}^m(t)$

   $$B_{ij}^m(t) = \frac{\partial \pi_{it}(\mathbf{u}_i(t), x)}{\partial x} \mid_{x = x_{ij}^m(t|t-1)}$$

   b- Update the state estimation error covariance $P_{ij}^m(t)$:

   $$\begin{aligned} P_{ij}(t \mid t) &= (I - K_{ij}(t) B_{ij}(t) P_{ij}(t \mid t-1) \qquad (4.3.1) \\ K_{ij}(t) &= P_{ij}(t \mid t-1) B_{ij}^{\top}(t) [B_{ij}(t) P_{ij}(t \mid t-1) B_{ij}(t) + \Sigma_i^{-1}(t)]^{-1} \end{aligned}$$

   c- Update the state estimate $X_{ij}^m(t)$:

   $$X_{ij}(t \mid t) = X_{ij}(t \mid t-1) + K_{ij}(t)[Y_i(t) - \widehat{\pi}_i(t)] \tag{4.3.2}$$

4. For $m = 1, \cdots N$, calculate

   $$\begin{aligned} w_i^m(t) &= q(m \mid \mathbf{Y}_i(t)) \\ &\propto \prod_{k=1}^{q} \mathcal{M}\left[\pi_{ik}(\mathbf{u}_i^m(t), \mu_i^m(t))\right] w_i^m(t-1). \end{aligned}$$

5. Resample to obtain the index ς_i^m of particle m's parent.

6. Importance sampling : for $m = 1, \cdots, N$,

   6-1. Draw samples

   $$\begin{aligned} X_i^m(t) &\sim q(X_i(t), \varsigma_i^m \mid \mathbf{Y}_i(t)) \\ &= \mathcal{N}(\widehat{X}_{ij}^{\varsigma_i^m}(t); P_{ij}^{\varsigma_i^m}(t)), \qquad (4.3.3) \end{aligned}$$

6-2. Calculate importance weights of particles by using

$$\begin{aligned} w_i^m(t) &= \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{S_i^m}(t)]} && (4.3.4) \\ &= \frac{\prod_{k=1}^q p[Y_{ik}(t) \mid X_i^m(t)]}{\prod_{k=1}^q p[Y_{ik}(t) \mid \mu_i^{S_i^m}(t)]} \\ &= \prod_{k=1}^{q} \frac{\mathcal{M}[\pi_{ik}(x_i^m(t), \mathbf{u}_i^m(t)]}{\mathcal{M}[\pi_{ik}(\mu_i^{S_i^m}(t), \mathbf{u}_i^m(t)]} && (4.3.5) \end{aligned}$$

6-3. Normalize the weights

$$w_i^m(t) = \frac{w_i^m(t)}{\sum_{m=1}^N w_i^m(t)}.$$

7. Output: a set of weighted particles (samples) to an individual $[\{X_i^m(t), w_i^m(t)\}_{m=1}^N], i = 1, \cdots, n.$

4.3.2 Optimality of the Auxiliary Iterated Extended Kalman Particle Filter(AIEKPF) Algorithm

Under the following regularity condition $C.1, C.2$ and $C.3$ given in Chapter 3 we can establish the *consistency* and *asymptotic normality* of a weighted particles (samples) $[\{x_t^m, w_t^m\}_{m=1}^N]$. :

$$\sqrt{N}(\widehat{\varphi}_i(t) - \bar{\varphi}_i(t)) \Rightarrow \mathcal{N}[0; \sigma_{APF}^2(\varphi_i(t))], \tag{4.3.6}$$

where $\varphi_i(t)$ is the function of trajectories $X_i(t)$ and $\bar{\varphi}_i(t)$ is the expectation of $\varphi_i(t)$ with respect to the filtering distribution

$$\bar{\varphi}_i(t) = \int \varphi_i(t) p(X_i(t) \mid \mathbf{Y}_i(t) dX_i(t)$$

and

$$\varphi_i(t) = \sum_{m=1}^{N} W_i^m(t) \varphi_i^m(t), \tag{4.3.7}$$

where $W_i^m(t) = w_i^m(t)[\sum_{m=1}^N w_i^m(t)]^{-1}$, and

$$w_i^m(t) = \frac{p(\mathbf{Y}_i(t) \mid X_i^m(t))}{p(\mathbf{Y}_i(t) \mid \mu_i^{S_i^m})} = \prod_{k=1}^{q} \frac{\mathcal{M}(\pi_{ik}(u_i^m(t), X_i^m(t)))}{\mathcal{M}(\pi_{ik}(u_i^m(t), \mu_i^{S_i^m}))}.$$

$\sigma_{APF}^2(\varphi_i(t))$ as equations (3.4.20),(3.4.21) in chapter 3, but with $q_t(X_i(t) \mid X_i(t-1)) = f_{X_i(t-1)}^i$

4.4 Posterior mode estimation

In this chapter, the observations are from a multinomial distribution and the latent variables are described by a CHARN model with noise from an exponential family distributions. In this chapter we extend the numerical solutions used in chapter 3 to find the posterior mode. We used the following two approaches :

1. Penalized likelihood estimation :

   a - Gauss-Newton and Fisher-scoring Filtering and smoothing algorithms.

   b - Working extended Kalman filter and smoother algorithms.

2. Maximum a posteriori sequence estimation (MAP) via the auxiliary iterated extended Kalman filter particle filter-MAP.

4.4.1 Penalized likelihood estimation

The posterior mode smoother is given by

$$a \equiv \left\{a^\top(0 \mid T), a^\top(1 \mid T), \cdots, a^\top(T \mid T)\right\} \in \mathbb{R}^m,$$

with $m = (T+1)n$, where

$$\begin{array}{rcl} a(0 \mid T) & = & ((a_1(0 \mid T), \cdots, a_n(0 \mid T))^\top, \\ \vdots & \vdots & \vdots \qquad\qquad \vdots \\ a(T \mid T) & = & (a_1(T \mid T), \cdots, a_n(T \mid T))^\top. \end{array}$$

The posterior distribution of $\mathbf{X}$ by Bayes' theorem

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{Y}) \quad = \quad & \frac{1}{p(\mathbf{Y})}\left\{\prod_{i=1}^{n}\prod_{k=1}^{q}\prod_{t=1}^{T} p(Y_{ik}(t) \mid X_i(t))\right\} \\ & \times \left\{\prod_{i=1}^{n}\prod_{t=1}^{T} g_i(X_i(t))\right\}\left\{\prod_{i=1}^{n} g_i(X_i(0))\right\} \end{aligned} \qquad (4.4.1)$$

$p(\mathbf{Y})$ does not depend on $\mathbf{X}$

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{Y}) \quad \propto \quad & \left\{\prod_{i=1}^{n}\prod_{k=1}^{q}\prod_{t=1}^{T} p(Y_{ik}(t) \mid X_i(t))\right\} \\ & \times \left\{\prod_{i=1}^{n}\prod_{t=1}^{T} g_i(X_i(t))\right\}\left\{\prod_{i=1}^{n} g_i(X_i(0))\right\} \end{aligned} \qquad (4.4.2)$$

where

$$g_i(X_i(0)) = \frac{1}{H[X_i(0), \mathbf{u}_i(0), \delta]} \exp\left\{\frac{Z_i(0)\upsilon_i(0) - b[\upsilon_i(0)]}{\phi_i(0)} + c[Z_i(0), \phi_i(0)]\right\}. \quad (4.4.3)$$

Taking logarithms and inserting the densities of $X_i(t)$ in equation (4.2.6) and $X_i(0)$ in equation (4.4.3), the penalized log-likelihood function is given by

$$PL : \mathbb{R}^m \longrightarrow \mathbb{R},\ m = (T+1)n$$

$$\begin{aligned} PL(\mathbf{X}) &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\{\log p(Y_{ik}(t) \mid X_i(t))\} \\ &\quad + \sum_{i=1}^{n}\sum_{t=1}^{T}\log g_i(X_i(t)) + \sum_{i=1}^{n}\log g_i(X_i(0)) \end{aligned} \quad (4.4.4)$$

where

$$\begin{aligned} \log p(Y_{ik}(t) \mid X_i(t)) &= \log\prod_{s=1}^{c_k}[\pi_{ik}^s(t)]^{Y_{ik}^s(t)} \\ &= \sum_{s=1}^{c_k} Y_{ik}^s(t)\log\pi_{ik}^s(t). \end{aligned}$$

Then

$$PL(\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\sum_{s=1}^{c_k}\{Y_{ik}^s(t)\log\pi_{ik}^s(t)\} + G_1 + G_2 \quad (4.4.5)$$

where

$$\begin{aligned} G_1 &= \sum_{i=1}^{n}\left\{\frac{Z_i(0)\upsilon_i(0) - b[\upsilon_i(0)]}{\phi_i(0)} + c[Z_i(0), \phi_i(0)]\right\} - \log(H[X_i(0), \mathbf{u}_i(0), \delta]) \\ G_2 &= \sum_{i=1}^{n}\sum_{t=1}^{T}\left\{\left[\frac{Z_i(t)\upsilon_i(t) - b[\upsilon_i(t)]}{\phi_i(t)} + c[Z_i(t), \phi_i(t)]\right]\right\} - \log(H[X_i(t-1), \mathbf{u}_i(t), \delta]). \end{aligned}$$

Numerical maximization of the penalized log-likelihood can be achieved by various algorithms. As in chapter 3 we use two methods. First, iterative forward-backward Gauss-Newton (Fisher-scoring) algorithms. Second, Working Extended Kalman Filter and Smoother(WEKFS).

4.4.2 Gauss-Newton and Fisher-scoring Filtering and smoothing

The penalized log-likelihood criterion (4.4.4) can be presented in compact matrix notation as:

$$PL(\mathbf{X}) = l_1(\mathbf{X}) - l_2(\mathbf{X}), \tag{4.4.6}$$

where

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

and

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

where

- $\mathbf{X}$ is a $(n \times T)$ matrix of latent variables.
- $A = diag(\frac{1}{\phi})$ with size $(n \times n)$, with $\phi \neq 0$.
- 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 $X_i(t)$.

For the description of Fisher scoring steps in matrix notation, the tables of observations for all $i = 1, \cdots, n$, can be written as follows

$$\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top, \quad \mathbf{Y}_i(t) = (Y_{i1}^\top(t)\cdots, Y_{iq}^\top(t))^\top, t = 1, 2, \cdots, T.$$

Fahrmeir and Wagenpfeil (1997) assumed $\mathbf{Y}_i^\top(0) = a_i(0)$. Correspondingly we define the tables of expectations by

$$\Pi_i(\mathbf{X}) = (\pi_{i0}^\top(X_0), \pi_{i1}^\top(X_1), \cdots, \pi_{iT}^\top(X_T))^\top.$$

They assumed $\pi_i^\top(0) = X_i(0)$. We recall the conditional mean and variance of individual i

$$\begin{aligned}\mathrm{E}(Y_{ik}(t) \mid X_i(t)) &= \pi_{ik}^{s}(t),\ s = 1, \cdots, c_k\\ \mathrm{Var}(Y_{ik}(t) \mid X_i(t)) &= \Sigma_{ik}(X_t),\end{aligned}$$

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

$$\sigma_{ik}^{sm}(t) = \begin{cases} \pi_{ik}^{s}(t)[1-\pi_{ik}^{s}(t)], & \text{if } s = m \\ -\pi_{ik}^{s}(t)\pi_{ik}^{m}(t) & \text{if } s \neq m \end{cases} \qquad (4.4.7)$$

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

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

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

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

Since $D_{it}(X_t)$ is first-order derivative of the conditional probability $\pi_i(t)$ evaluated at $\eta_i(t)$.

The score function of $l(\mathbf{X})$ in (4.4.6) is given by $S_i(\mathbf{X}) = (\widehat{S}_{i0}(X_0), \widehat{S}_{i1}(X_1), \cdots, \widehat{S}_{iT}(X_T))^\top$, where for all $i = 1, \cdots, n$,

$$S_i(\mathbf{X}) := D_i(\mathbf{X})\Sigma_i^{-1}(\mathbf{X})\left\{\mathbf{Y}_i(t) - \Pi_i(\mathbf{X})\right\}, \qquad (4.4.8)$$

with components

$$\begin{aligned} \widehat{S}_i(X_0) &= V_i^{-1}(0)(a_i(0) - X_i(0)) && (4.4.9) \\ \widehat{S}_{it}(X_t) &= D_{it}(X_t)\Sigma_{it}^{-1}(X_t)\left\{\mathbf{Y}_i(t) - \pi_{it}(X_t)\right\}, t = 1, \cdots, T, && (4.4.10) \end{aligned}$$

the first-order derivatives of $PL(\mathbf{X})$ in (4.4.6) are

$$M(\mathbf{X}) = \partial PL(\mathbf{X})/\partial \mathbf{X} = S(\mathbf{X}) - S(\upsilon), \qquad (4.4.11)$$

where

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

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

$$\mathcal{I}_i(\mathbf{X}) = D_i(\mathbf{X})\Sigma_i^{-1}(\mathbf{X})D_i^\top(\mathbf{X}) \qquad (4.4.12)$$

with diagonal blocks

$$\begin{aligned} \mathcal{I}_{i0}(X_0) &= V_i^{-1}(0) && (4.4.13) \\ \mathcal{I}_{it}(X_t) &= D_{it}(X_t)\Sigma_{it}^{-1}(X_t)D_{it}^\top(X_t), t = 1, \cdots, T. && (4.4.14) \end{aligned}$$

As in chapter 3, the Taylor expansion of the score function about $\mathbf{X}^0$ yields

$$M(\widehat{\mathbf{X}} \mid \mathbf{Y}) \approx M(\mathbf{X}^0) - \mathcal{I}(\mathbf{X}^0) \times \left\{\mathbf{X}^1 - \mathbf{X}^0\right\}$$

since $M(\widehat{\mathbf{X}} \mid \mathbf{Y}) = 0$, a single Fisher scoring to the next iterate $\mathbf{X}^1 \in \mathbb{R}^m$, with $m = (T+1)n$ is as follows

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

This can be rewritten as

$$\mathbf{X}^1 = \left\{\mathcal{I}(\mathbf{X}^0) + \mathcal{J}(\nu^0)\right\}^{-1} \mathcal{I}(\mathbf{X}^0)\widetilde{\mathbf{Y}}, \tag{4.4.15}$$

where

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

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

$$\widetilde{Y}_i := \left\{D_i^{-1}((\mathbf{X}))\right\}^\top \left\{\mathbf{Y}_i - \Pi_i(\mathbf{X}))\right\} + \eta_i(\mathbf{X}), \tag{4.4.16}$$

with components

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

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

4.4.3 Working Extended Kalman Filter and Smoother (WEKFS)

In the following algorithms, $a_{t|t-1}, , a_{t|t}, a_{t|T}$ are numerical approximations to predicted, filtered, smoothed values of posterior mode to $\mathbf{X}$ and the corresponding $P_{t|t-1}, P_{t|t}, P_{t|T}$ are numerical approximations to predicted, filtered, smoothed values of error covariance matrices.

Initialization:

$$\begin{aligned} a_i(0 \mid 0) &= a_i(0), \\ V_i(0 \mid 0) &= V_i(0). \end{aligned} \tag{4.4.17}$$

Prediction for $t = 1, \cdots, T$

$$\begin{aligned} a_i(t \mid t-1) &= F(a_i(t-1 \mid t-1), \mathbf{u}_i(t), \gamma) \\ P_i(t \mid t-1) &= A_i(t)P_i(t-1 \mid t-1)A_i^\top(t) + C_i(t)R_tC_i^\top(t). \end{aligned} \tag{4.4.18}$$

where

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

Filtering for $t = 1, \cdots, T$,

$$\begin{aligned} a_i(t \mid t) &= a_i(t) + K_i(t)(\widetilde{Y}_i(t) - a_i(t \mid t-1)) \\ K_i(t) &= P_i(t-1 \mid t-1)B_i^\top(t)(B_i(t)P_i(t-1 \mid t-1)B_i^\top(t) + \mathcal{I}^{-1}(t))^{-1} \\ P_i(t \mid t) &= (I - K_i(t)B_i(t))P_i(t-1 \mid t-1) \end{aligned} \tag{4.4.19}$$

where

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

4.4.4 The maximum a posteriori (MAP) with particle filtering and smoothing algorithm

As in chapter 3, we find the maximum a posteriori (MAP) via the Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF) method. We present the algorithm with the state noise processes from exponential family distribution.

4.4.5 Auxiliary Iterated Extended Kalman Particle Filter - MAP(AIEKPF-MAP) Algorithm

We find the MAP estimation by using the Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF), where:

$$\begin{aligned} w_i^m(t) &= \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} \\ &= \frac{\prod_{k=1}^q p[Y_{ik}(t) \mid X_i^m(t)]}{\prod_{k=1}^q p[Y_{ik}(t) \mid \mu_i^{\varsigma_i^m}(t)]} \\ &= \prod_{k=1}^q \frac{\mathcal{M}[\pi_{ik}(X_i^m(t), \mathbf{u}_i(t))]}{\mathcal{M}[\pi_{ik}(\mu_i^{\varsigma_i^m}(t), \mathbf{u}_i(t))]} \end{aligned} \tag{4.4.20}$$

The AIEKPF algorithm can be implemented by the following steps:

1. Initialization $(t = 0)$: For $m = 1, \cdots, N$, draw the states (particles) $X_i^m(0)$ from the prior $p(X_i(0)) = f^i_{X_i(0)}$, and set

   $$\widehat{X}_i^m(0) = \mathrm{E}[X_i^m(0)] = F(X_i^m(0), \mathbf{u}_i^m(0), \delta^m)$$

   and

   $$\begin{aligned} P_i^m(0) &= \mathrm{E}[(X_i^m(0) - \widehat{X}_i^m(0))(X_i^m(0) - \widehat{X}_i^m(0))^\top] \\ &= \mathrm{Var}(X_i^m(0)) \\ &= H^2(X_i^m(0), \mathbf{u}_i^m(0), \delta^m)R_0 \end{aligned}$$

   For $t = 1, \cdots, T$, repeat the following steps:

2. For $m = 1, \cdots, N$, generate $\mu_i^m(t) \sim f^i_{X_i(t-1)}$.

3. For $m = 1, \cdots, N$, Update the particles using the IEKF algorithm

   3-1. Compute the Jacobians $A_i^m(t), C_i^m(t)$ of the process model

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

$$C_i^m(t) = H(X_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \delta)$$

   3-2. Predict the particle with the IEKF:

$$X_i^m(t \mid t-1) \approx F(X_i^m(t-1 \mid t-1), \mathbf{u}_i(t), \gamma)$$

$$P_i^m(t \mid t-1) = A_i^m(t) P_i^m(t-1 \mid t-1) A_i^{m\top}(t) + C_i^m(t) R_t C_i^{m\top}(t)$$

   3-3. For $j = 1, \cdots c$ (c is the iteration number of the IEKF)

      a- Compute the Jacobians , $B_{ij}^m(t)$

$$B_{ij}^m(t) = \frac{\partial \pi_{it}(\mathbf{u}_i(t), x)}{\partial x} \mid_{x = x_{ij}^m(t|t-1)}$$

      b- Update the state estimation error covariance $P_{ij}^m(t)$:

$$\begin{aligned} P_{ij}(t \mid t) &= (I - K_{ij}(t) B_{ij}(t) P_{ij}(t \mid t-1) \qquad (4.4.21) \\ K_{ij}(t) &= P_{ij}(t \mid t-1) B_{ij}^\top(t) [B_{ij}(t) P_{ij}(t \mid t-1) B_{ij}(t) + \Sigma_i^{-1}(t)]^{-1} \end{aligned}$$

      c- Update the state estimate $X_{ij}^m(t)$:

$$X_{ij}(t \mid t) = X_{ij}(t \mid t-1) + K_{ij}(t)[Y_i(t) - \widehat{\pi}_i(t)] \tag{4.4.22}$$

4. For $m = 1, \cdots N$, calculate

$$\begin{aligned} w_i^m(t) &= q(m \mid \mathbf{Y}_i(t)) \\ &\propto \prod_{k=1}^{q} \mathcal{M}(\pi_{ik}(\mathbf{u}_i(t), \mu_i^m(t)) w_i^m(t-1), \end{aligned}$$

5. Resample to obtain the index ς_i^m of particle m's parent.

6. Importance sampling : for $m = 1, \cdots, N$,

   6-1. Draw samples $X_i^m(t) \sim q(X_i(t), \varsigma_i^m \mid \mathbf{Y}_i(t)) = \mathcal{N}(\widehat{X}_{ij}^{\varsigma_i^m}(t), P_{ij}^{\varsigma_i^m}(t))$, where $j = c$

6-2. Calculate importance weights of particles by using

$$w_i^m(t) = \frac{p[\mathbf{Y}_i(t) \mid X_i^m(t)]}{p[\mathbf{Y}_i(t) \mid \mu_i^{\varsigma_i^m}(t)]} = \prod_{k=1}^{q} \frac{\mathcal{M}[\pi_{ik}(x_i^m(t), \mathbf{u}_i(t))]}{\mathcal{M}[\pi_{ik}(\mu_t^{\varsigma_i^m}(t), \mathbf{u}_i(t))]}$$

6-3. Normalize the weights $w_i^m(t) = \frac{w_i^m(t)}{\sum_{m=1}^N w_i^m(t)}$.

Forward Filtering step

$$\begin{aligned}
X_i^{MAP}(t) &= \arg\max_{X_i^{(m)}(t)} \left[p(\mathbf{Y}_i(t) \mid X_i^{(m)}(t)) \sum_j p(X_i^{(j)}(t) \mid X_i^{(j)}(t-1)) w_i^{(j)}(t-1) \right] \\
&= \arg\max_{X_i^{(m)}(t)} \left[\prod_{k=1}^{q} p(Y_{ik}(t) \mid X_i^{(m)}(t)) \sum_j p(X_i^{(j)}(t) \mid X_i^{(j)}(t-1)) w_i^{(j)}(t-1) \right] \\
&= \arg\max_{X_i^{(m)}(t)} \left[\prod_{k=1}^{q} \mathcal{M}(\pi_{ik}^m(t)) \sum_j f_{X_i(t-1)}^i . w_i^{(j)}(t-1) \right].
\end{aligned}$$

Backward smoothing step

- Set $w_i^{(m)}(T \mid T) = w_i^{(m)}(t)$
- For $t = T-1, \cdots, 1$, evaluate the smoother importance weights as

$$\begin{aligned}
w_i^{(m)}(t \mid T) &= w_i^{(m)}(t) \sum_{j=1}^{N} \left[w_i^{(j)}(t+1 \mid T) \frac{p(X_i^{(j)}(t+1) \mid X_i^{(m)}(t))}{\sum_{r=1}^N p(X_i^{(j)}(t+1) \mid X_i^{(r)}(t)) w_i^{(r)}(t)} \right] \\
&= w_i^{(m)(t)} \sum_{j=1}^{N} \left[w_i^{(j)}(t+1 \mid T) \frac{f_{X_i(t-1)}^j}{\sum_{r=1}^N f_{X_i(t-1)}^r) w_i^{(r)}(t)} \right].
\end{aligned}$$

- Evaluate the approximate smoother MAP as

$$\begin{aligned}
x_i^{MAP}(t \mid T) &= \arg\max_{X_i^{(m)}(t)} \left\{ q(X_i^{(m)}(t) \mid \mathbf{Y}_i(t)) \frac{w_i^{(m)}(t \mid T)}{w_i^{(m)}(t)} \right\}. \\
&= \arg\max_{X_i^{(m)}(t)} \left\{ \prod_{k=1}^{q} \mathcal{M}(\pi_{ik}^m(t)) \sum_j f_{X_i(t-1)}^j . w_i^{(j)}(t-1) \frac{w_i^{(m)}(t \mid T)}{w_i^{(m)}(t)} \right\}.
\end{aligned}$$

Chapter 5

Practical considerations

5.1 Introduction

In this chapter, the numerical results and their discussion are presented. There are two parts. In the first part, the numerical simulation where the longitudinal multicategorical data are generated with latent variables described by a CHARN model and the state noise is generated from exponential family distributions. In the second part, we apply our results to real data in the quality of life of patients surged for cancer.
In the two parts, we created R codes from the methods described in chapters 3 and 4. Our objective in this chapter is to estimate the latent variables by posterior mode via the working extended Kalman filtering recursions.

5.2 Simulation experiments

In this part, we produce data from the observation equation described by a multinomial distribution defined in equation (3.3.1), and the state equation described by a CHARN model defined in equation (3.3.3) with state noise from exponential families. We consider gaussian and exponential distributions. There are two parts. 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 part uses the EM algorithm for estimates the parameters of the model, before applying the working extended Kalman filter recursions.

5.2.1 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

5.2.1.1 The outline of simulation I

In this outline we have **one individual**, so we omit the subscript i in the equations.

1. Generate the multicategorical 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 administered at t occasions.

   (b) Draw the samples of latent variables $X(t)$ by a state equation (CHARN model) with state noise from exponential family distributions (gaussian or exponential distribution).

   (c) For this individual produce 2 covariate variables $\mathbf{u}^\top(t)$: 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 $\beta_k^{s\prime}$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 in step (1-d) to calculate the adjusted observation by using equation (4.4.16).

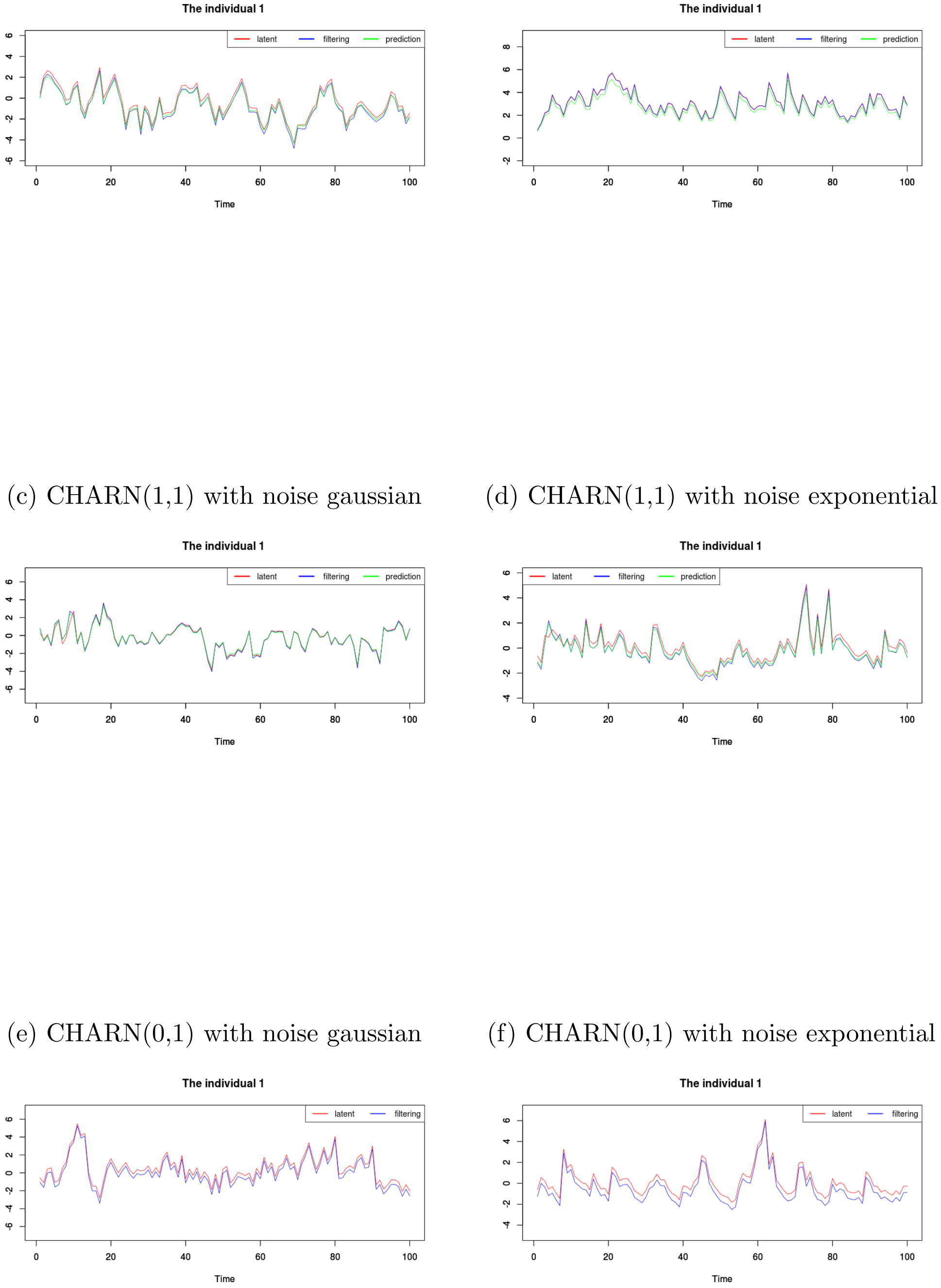
3. Set the values for γ, δ and the matrix of variance-covariance of state noise R_t.

4. Apply the Working Extended Kalman Filtering Recursions (WEKF) to calculate the posterior mode $a(t)$.

Figure 5.1 shows one individual with different type of models and the state noise distribution. There the red colour refers to the latent variable, while the blue colour refers to the postrior mode via filtering step and green colour refers to the posterior mode via the prediction step.

It is clear that the results are very good. The working Kalman filter recursion succeeds in producing the posterior mode with different type of state space model. 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 Figures 5.1(e) and (f) because the prediction step does not exist for the model considered. Indeed, it depends on the function F of the state equation, this function does not exist for the CHARN(0,1) model studied.

Figure 5.1: The graphs for $n = 1$individual and $T = 100$



(a) AR(1) with noise gaussian

(b) AR(1) with noise exponential

(c) CHARN(1,1) with noise gaussian

(d) CHARN(1,1) with noise exponential

(e) CHARN(0,1) with noise gaussian

(f) CHARN(0,1) with noise exponential

5.2.2 Simulation experiments II

In this part, the longitudinal data are generate, and it is considered as real data. In other words, after we generate the data, the state variables and the parameters of models are considered as unknown. In the following, we outline how we estimate the latent variables.

5.2.2.1 The outline of simulation experiments II

1. Generate the multicategorical 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 administered at t occasions.

   (b) Draw the latent variables $X_i(t)$ by a state equation (CHARN model) with state noise from exponential family distributions (gaussian or exponential distribution).

   (c) For each individual i, generate 2 covariate variables $\mathbf{u}_i'(t)$: Age $u_1(t)$ and Sex $u_2(t)$.

   (d) For each individual i, each item k and category s, gives value to parameters β_k^s, and calculate the probabilities

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

   where

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

   (e) For each individual i, each item k, and each category s at t occasions, generate the responses $Y_{ik}(t) \sim \mathcal{M}(\pi_{ik}^s(t))$.

2. Recall that $\mathbf{Y}_i = (\mathbf{Y}_i^\top(0), \mathbf{Y}_i^\top(1), \cdots, \mathbf{Y}_i^\top(T))^\top, \mathbf{X}_i = (\mathbf{X}_i^\top(0), \mathbf{X}_i^\top(1), \cdots, \mathbf{X}_i^\top(T))^\top$. 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 $p = 0$, 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 $\beta^{p+1}, \gamma^{p+1}, \delta^{p+1}$ via EM algorithm.

5. Perform the Working Extended Kalman Filtering Recursions (WEKF) to calculate the posterior mode $a_i^{p+1}(t)$. If $\mid a_i^{p+1}(t) - a_i^p(t) \mid < 0.001$, STOP, else set $p = p + 1$ and go to step 4.

   The latent variables are used to calculate the observations probabilities $\pi_{ik}(t)$ to generate the individuals' responses $Y_{ik}(t)$, after that, the latent variables are considered as unknown variables in the real data. Therefore, we apply the steps of outline to find the estimators for the model's parameters and the latent variables. Then the latent variables and their estimates are calculated via the working extended Kalman filtering are compared.

5.2.2.2 Numerical computation

The EM algorithm has the property of increasing the likelihood at each step, but the convergence is slow. 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

$$\widehat{\theta}^{(p+1)} = \widehat{\theta}^{(p)} + \mathcal{I}(\widehat{\theta}^{(p)})s(\widehat{\theta}^{(p)}), p = 0, 1, \cdots \tag{5.2.1}$$

where $\theta = (\beta^\top, \gamma^\top, \delta^\top)^\top$ and $s(\widehat{\theta}^{(p)})$ the Fisher scoring of the parameters and $\mathcal{I}(\widehat{\theta}^{(p)})$ the Fisher information matrix of the parameters described in Chapters 3 and 4. As glm-R package the convergence occurs if

$$\frac{dev - dev_{old}}{(0.1+ \mid dev \mid)} \leq \epsilon \tag{5.2.2}$$

where, $dev = -2\log(L)$, we take $\epsilon = 0.01$

5.2.2.3 The gaussian state noise with AR(1) model

The outline of simulation experiments II is performed, where the latent variable considered follows an AR(1) model

$$X_i(t) = \rho X_i(t-1) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0, R_t). \tag{5.2.3}$$

To generate the individuals' responses, suppose $\rho = 0.5$, and $R_t = 1$. The steps (2-5) of outline are implemented with iterations $p = 0, 1, \cdots, 10$, and in the initial iteration $p = 0$, suppose $\rho^{(0)} = 0.4$, and $R_t^{(0)} = 0.5$. Recalling that , here

$$\eta_{ik}^s(t) = \beta_{0k}^s + \beta_{1k}^s u_1(t) + \beta_{21k}^s u_{21}(t) + \beta_{22k}^s u_{22}(t) \tag{5.2.4}$$

where $u_1(t)$ is the age variable and $u_2(t)$ is the sex variable. We divide it into two groups according to its value $u_{21}(t)$ and $u_{22}(t)$. We suppose the initial vectors of $\beta_{rk}^{s(0)}$ as follows

$$\beta_{01}^{s(0)} = \beta_{02}^{s(0)} = \cdots = \beta_{0q}^{s(0)} = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

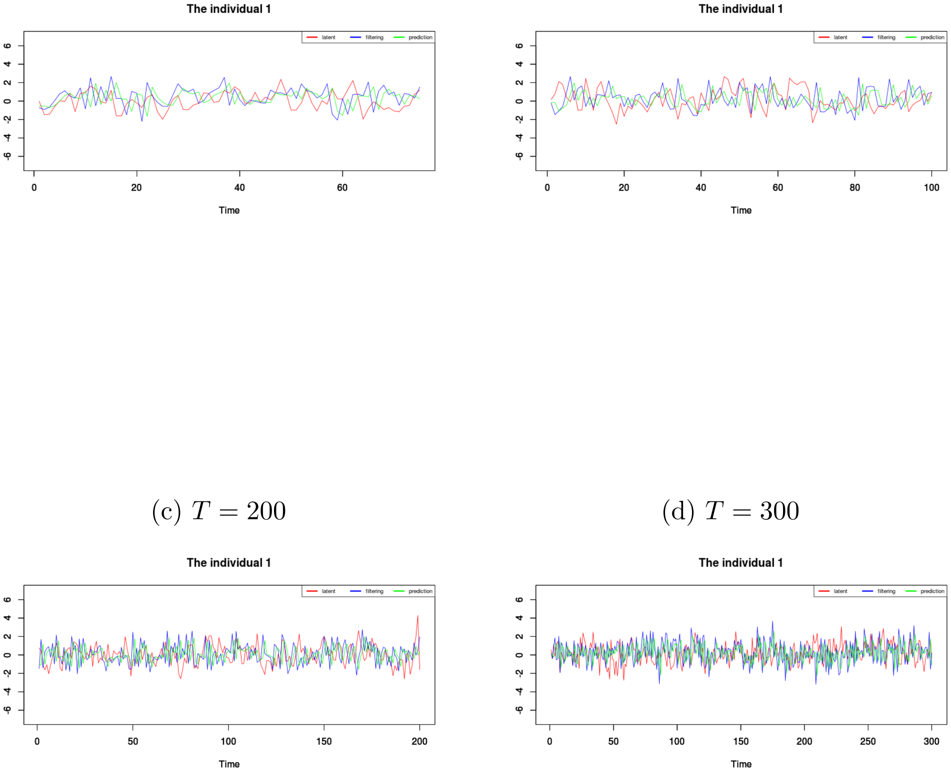
$$\beta_{11}^{s(0)} = \beta_{12}^{s(0)} = \cdots = \beta_{1q}^{s(0)} = \begin{pmatrix} 0.7 & 0.7 & 0.7 & 0.7 & 0.9 & 0.9 \end{pmatrix}$$
$$\beta_{211}^{s(0)} = \beta_{212}^{s(0)} = \cdots = \beta_{21q}^{s(0)} = \begin{pmatrix} 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$
$$\beta_{221}^{s(0)} = \beta_{222}^{s(0)} = \cdots = \beta_{22q}^{s(0)} = \begin{pmatrix} 0.4 & 0.4 & 0.4 & 0.7 & 0.9 & 0.9 \end{pmatrix}$$

Two cases are considered. First, the number of individuals is small ($n \leq 15$) with instant ($T > 75$). Second the number of individuals is large ($n > 75$) with instant ($T \leq 10$).

- **n small with T large**

The graphs (5.2-5.6) show the latent variables with state noise from the gaussian distribution and their estimators (posterior mode) via the Working Extended Kalman Filter (WEKF). On the graphs, the red colour refers to the latent variables, while the blue colour refers to the posterior mode via filtering step and the green colour refers to the posterior mode via the prediction step. We took $n = 1, 2, 3, 4, 10$, and $T = 75, 100, 200, 300$. Figures 5.2-5.6 show the results.
In Figure 5.2 the results are acceptable, but in Figures 5.2 (a) and (b), there are some jumps in the latent curve. Our approach has not adapted to it. In Figure 5.3 The results are good except Figure 5.3 (a) show the posterior mode in two steps (prediction and filtering) smaller than the latent variables. As seen in Figure 5.4, the results are also good and the series stable. Figure 5.5 shows that there are oscillations in the latent variables, while the posterior mod is stationary. Figures 5.6 show the results are good as the previous individuals Figures.

Figure 5.2: The graphs for $n = 1$ individual



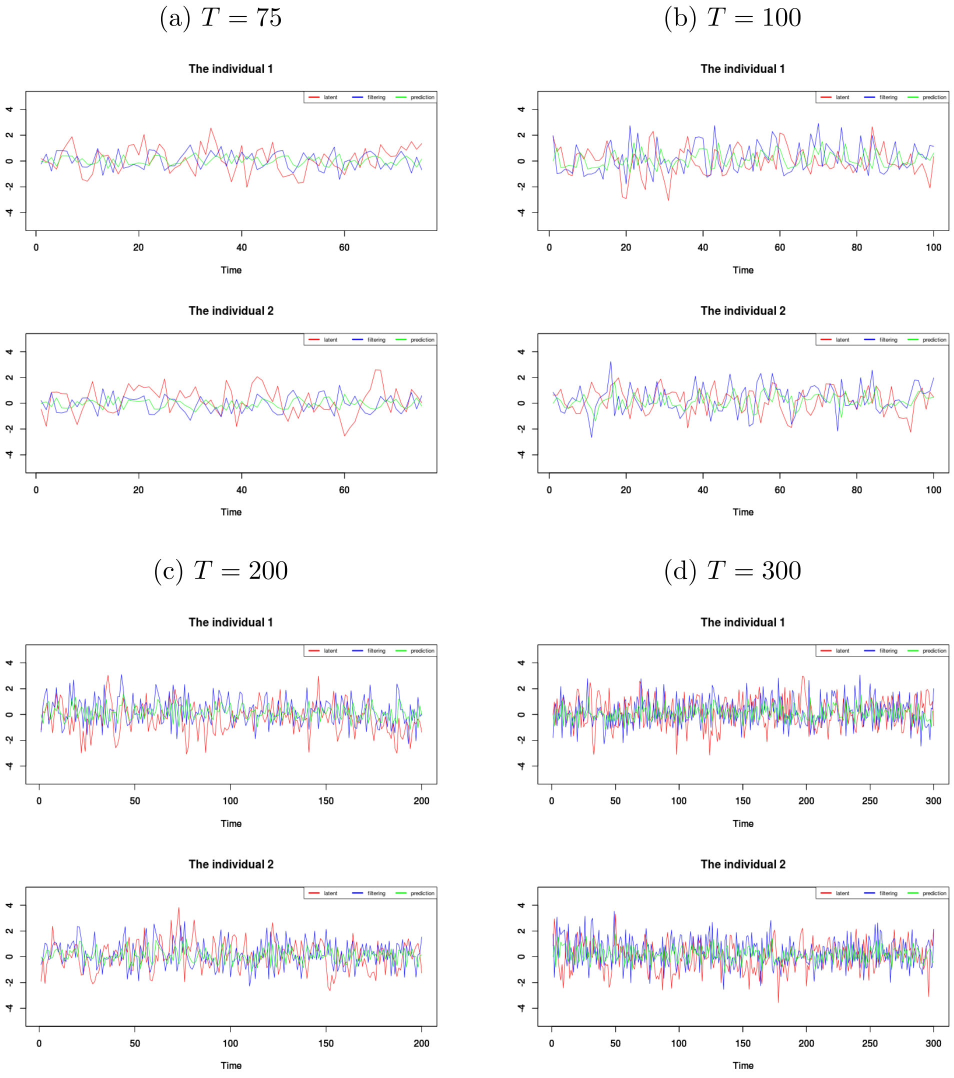
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.3: The graphs for $n = 2$ individuals



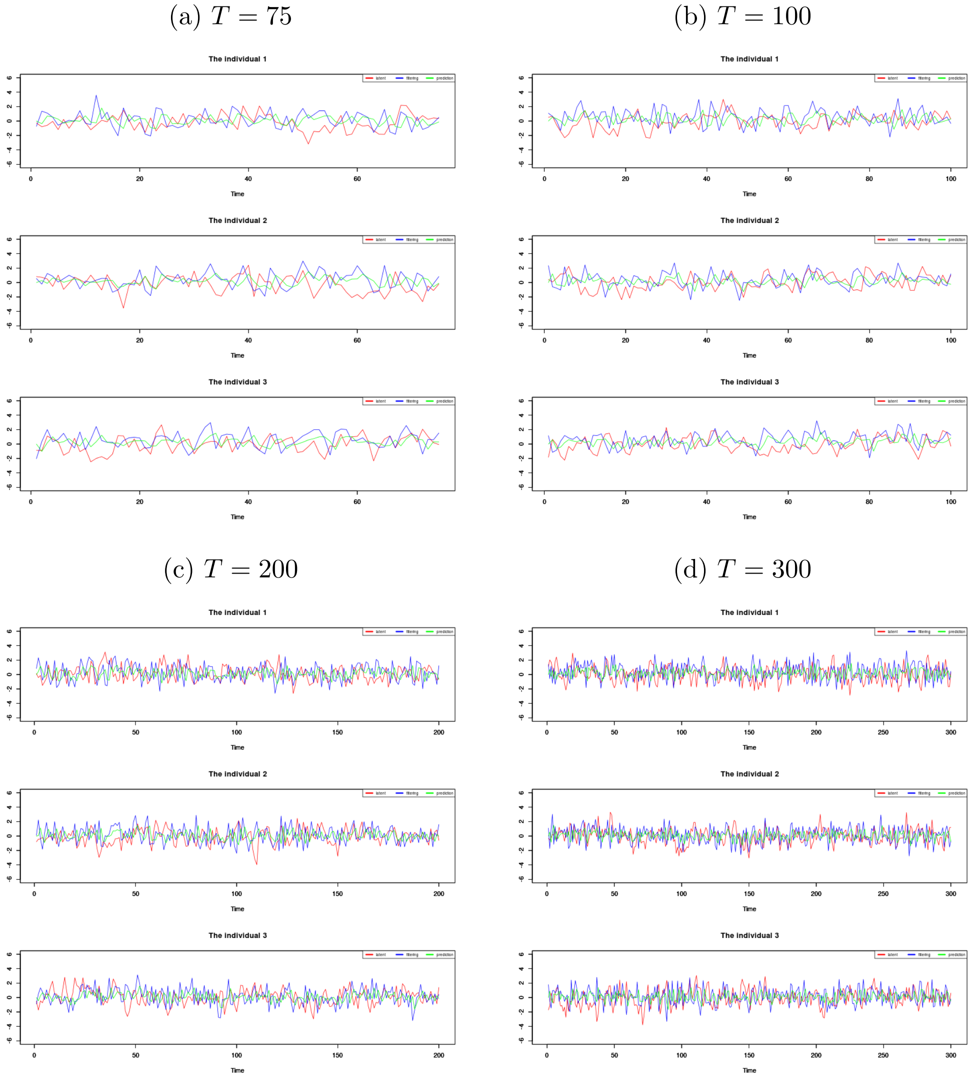
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.4: The graphs for $n = 3$ individuals



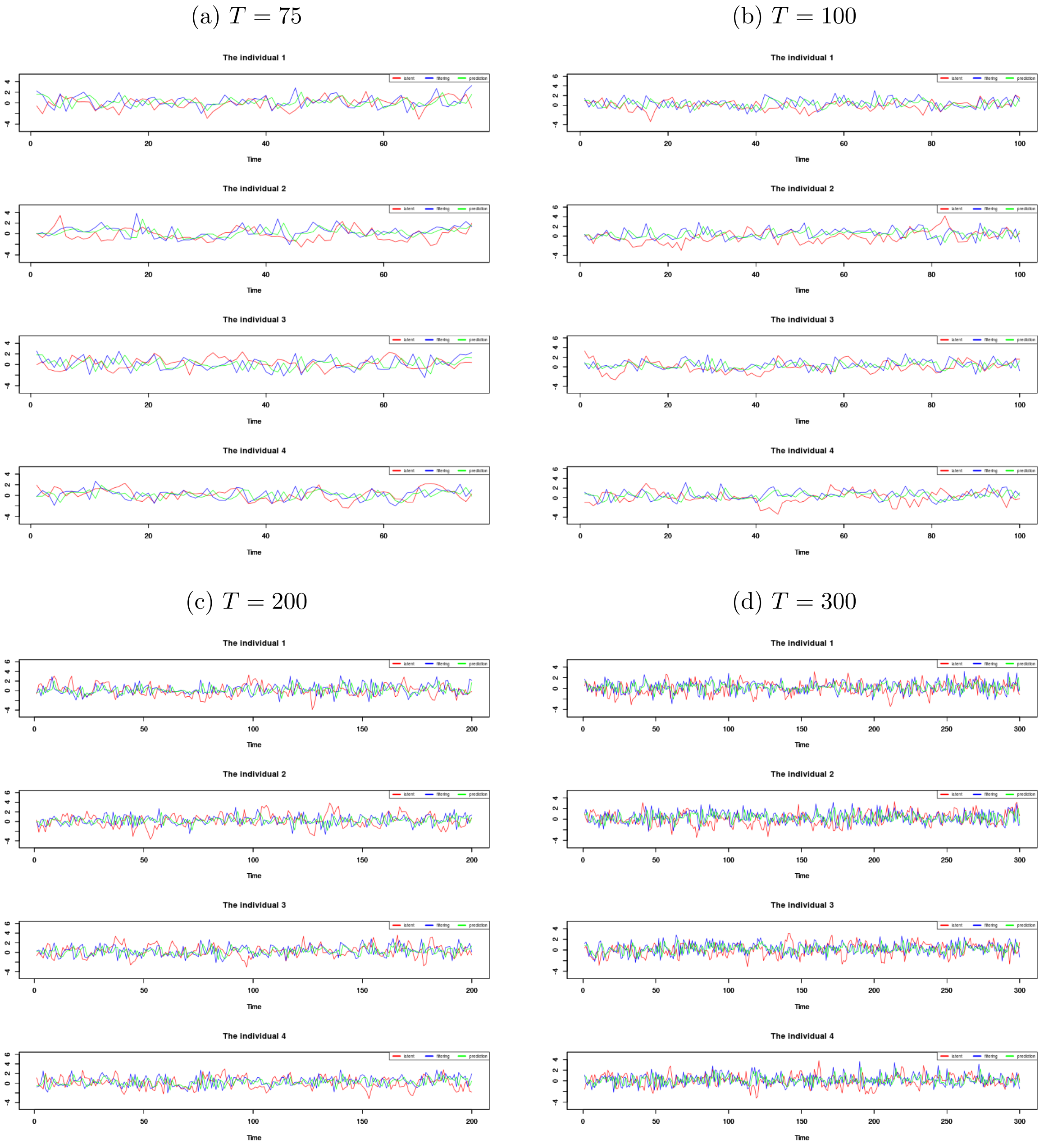
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.5: The graphs for $n = 4$ individuals



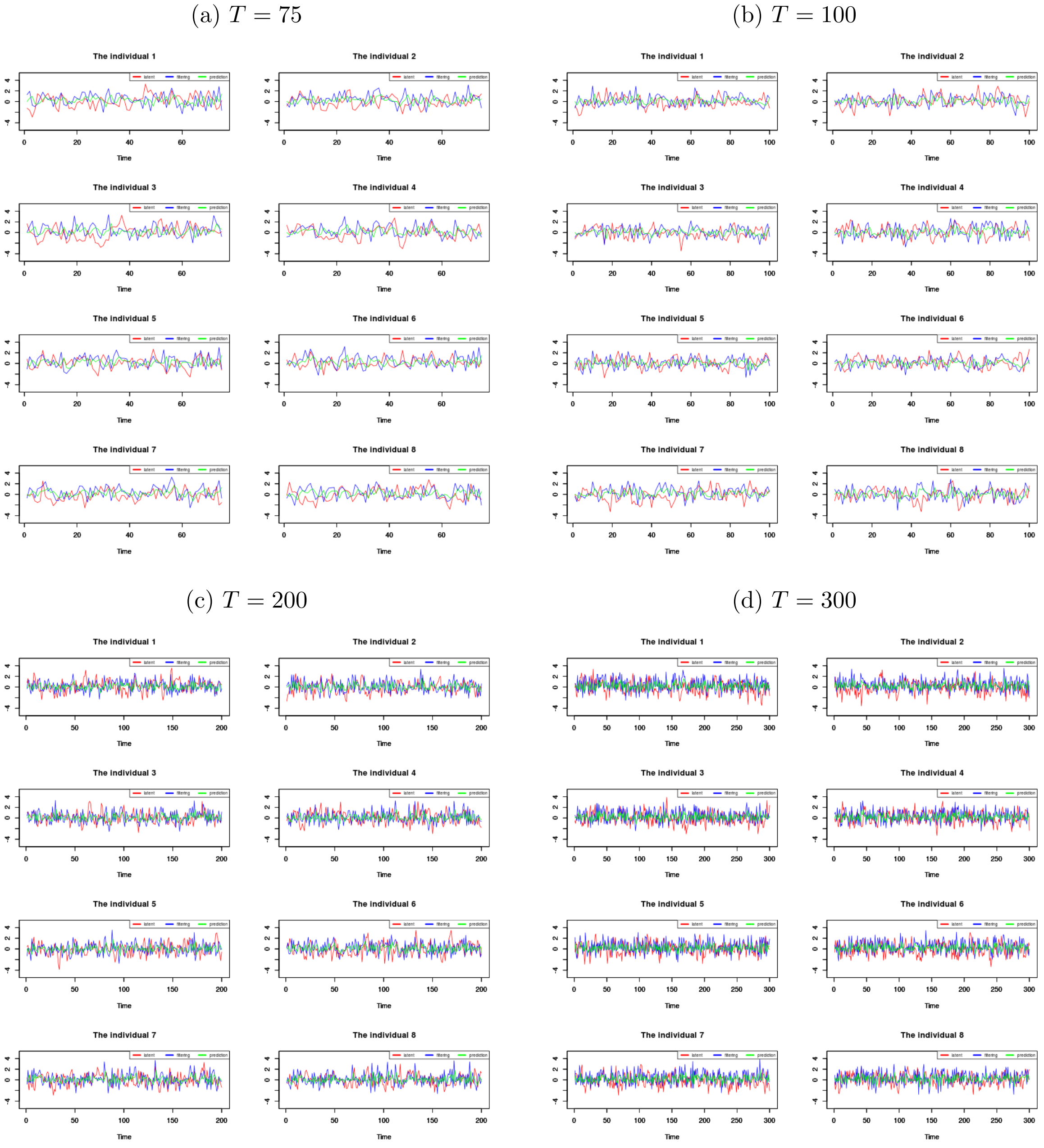
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.6: The graphs for $n = 10$ individuals



(a) $T = 75$

(b) $T = 100$

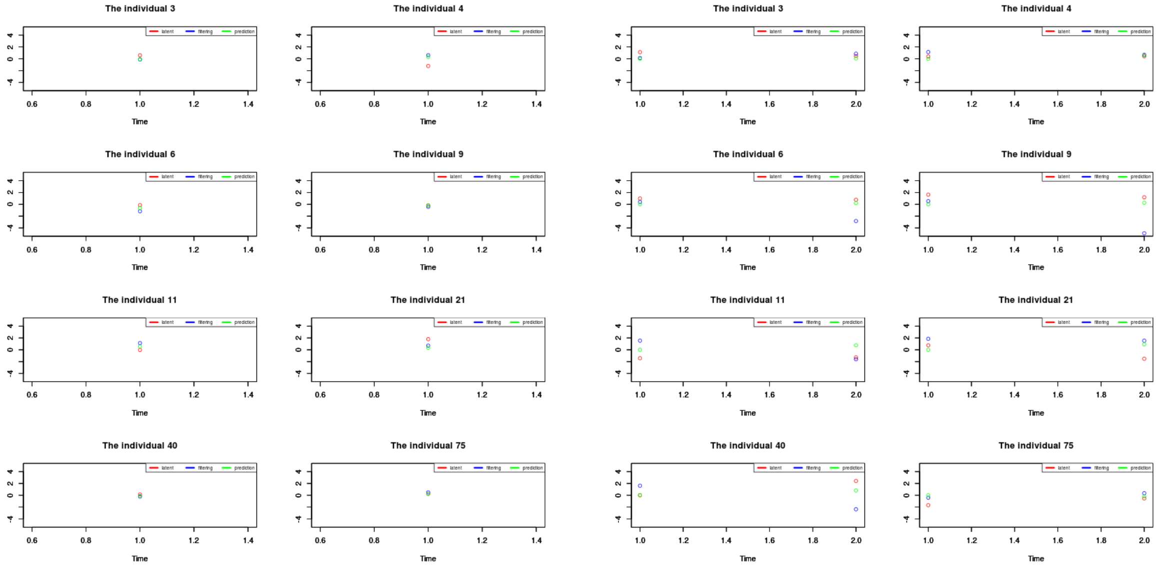
(c) $T = 200$

(d) $T = 300$

- **n large with T small**

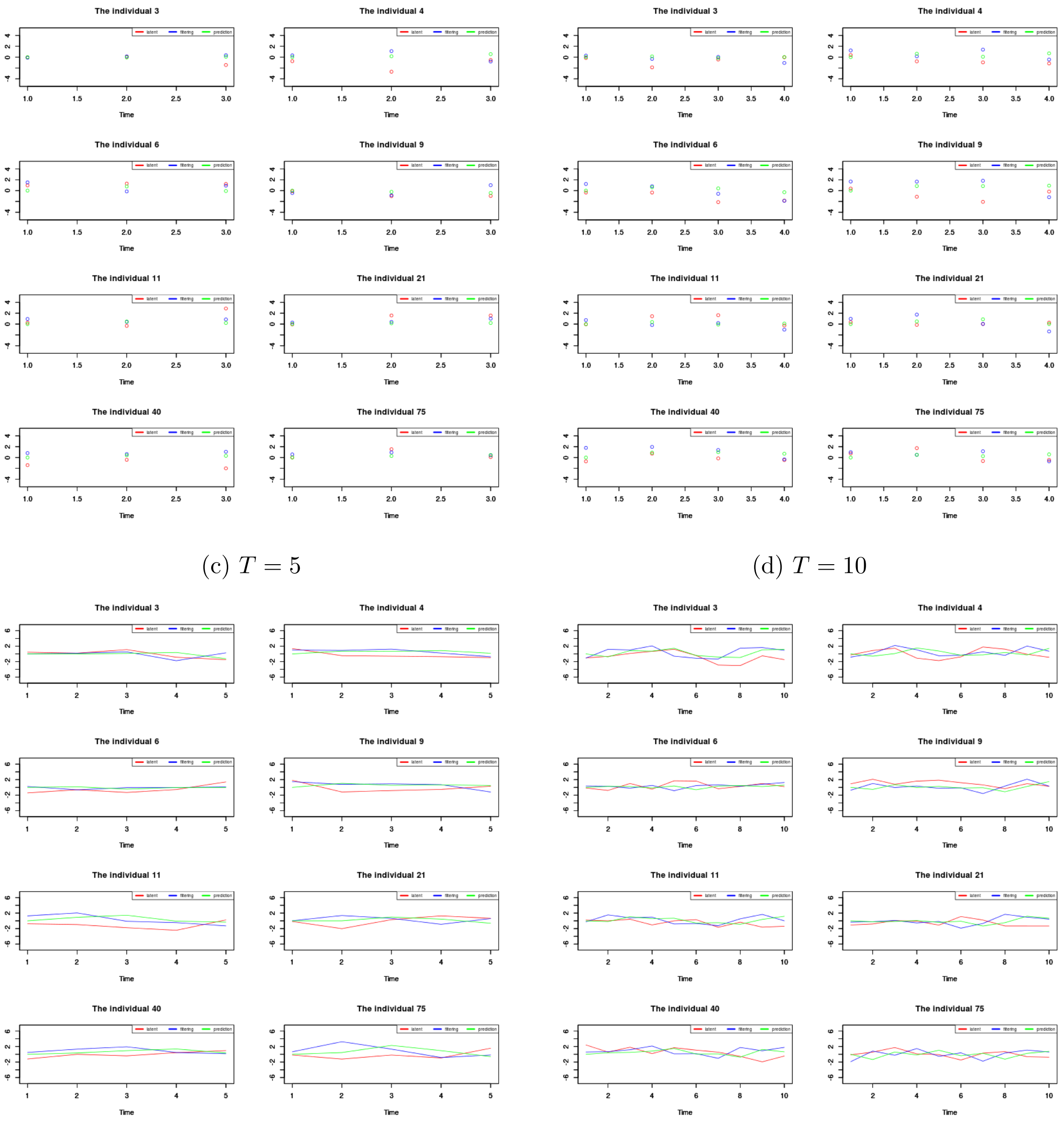
Here, the number of individuals is $n = 75, 100, 200, 300$. The length of time is $T = 1, 2, 3, 4, 5$ and 10. For $T = 1, 2, 3, 4$, the latent variable is the red circle, while the posterior mode via the filtering step is the blue circle, and the posterior mode via the prediction step is the green circle. The results are illustrated in Figures 5.7-5.14 .
As seen in Figures 5.7-5.14, at occasions $T = 1, 2, 3, 4$, the results are excellent, but at $T = 5, 10$, although there are some spacing in the curves but the results are acceptable. An issue that we have neglected up to this point is the problem of the initial value of posterior mode $a_i^{(0)}(t)$ and the initial values of parameters $\beta^{(0)}, \gamma^{(0)}, \delta^{(0)}$.

Figure 5.7: The graphs for $n = 75$ individuals with $T = 1, 2$


(a) $T = 1$

(b) $T = 2$

Figure 5.8: The graphs for $n = 75$ individuals with $T = 3, 4, 5$ and 10



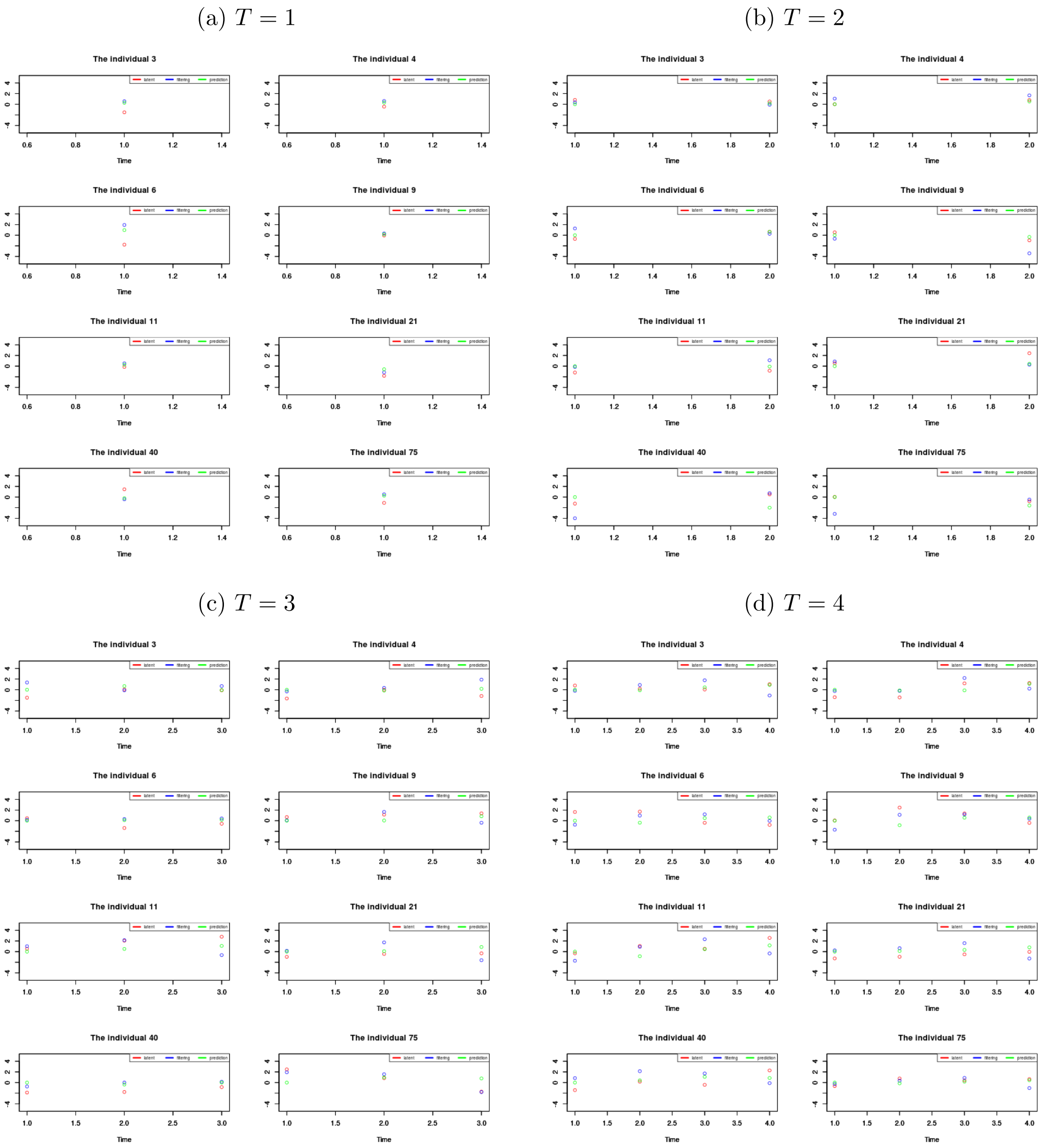
(a) $T = 3$

(b) $T = 4$

(c) $T = 5$

(d) $T = 10$

Figure 5.9: The graphs for $n = 100$ individuals with $T = 1, 2, 3$ and 4


(a) $T = 1$

(b) $T = 2$

(c) $T = 3$

(d) $T = 4$

Figure 5.10: The graphs for $n = 100$ individuals with $T = 5$ and 10

(a) $T = 5$ (b) $T = 10$

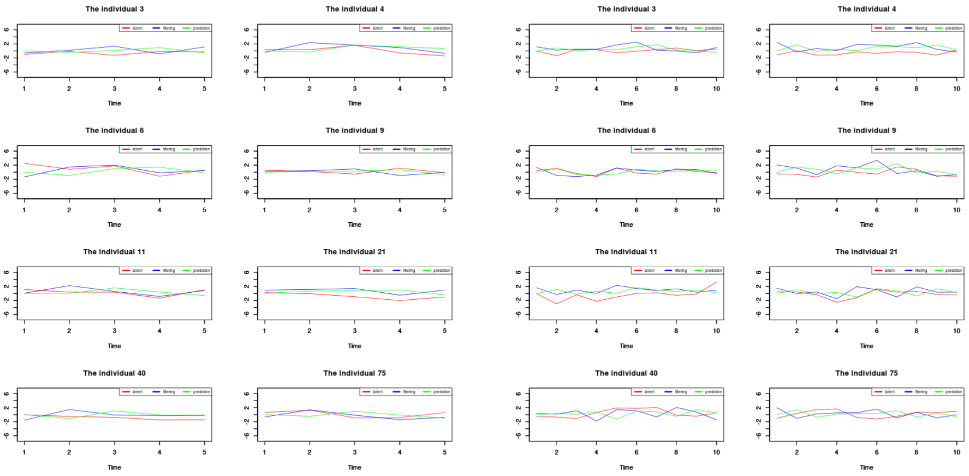


Figure 5.11: The graphs for $n = 200$ individuals with $T = 1, 2$

(a) $T = 1$ (b) $T = 2$

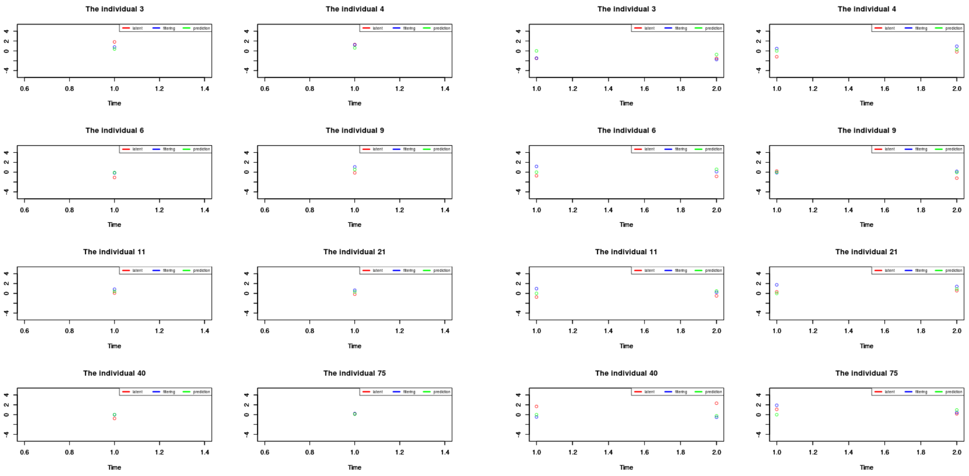
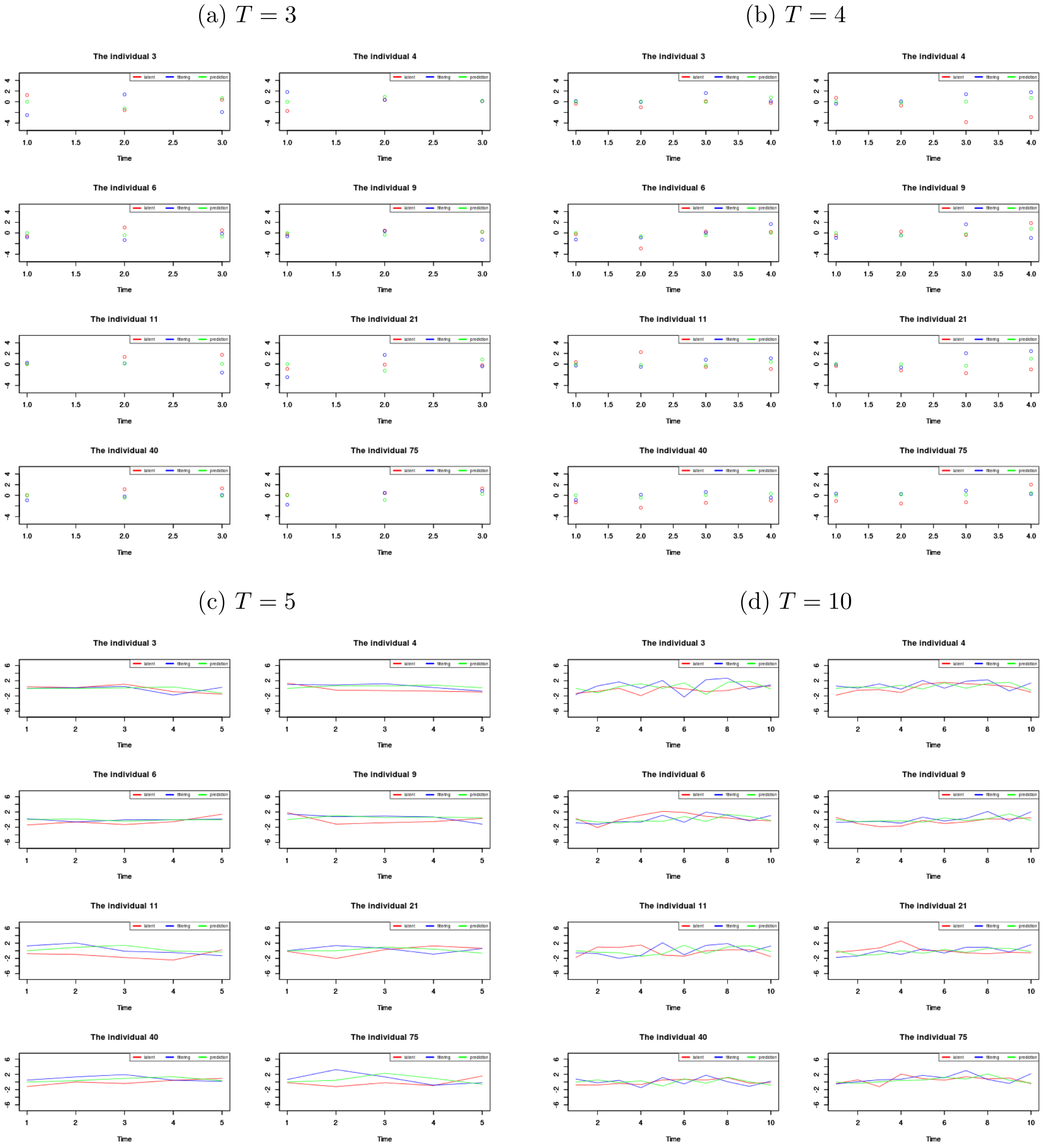


Figure 5.12: The graphs for $n = 200$ individuals with $T = 3, 4, 5$ and 10



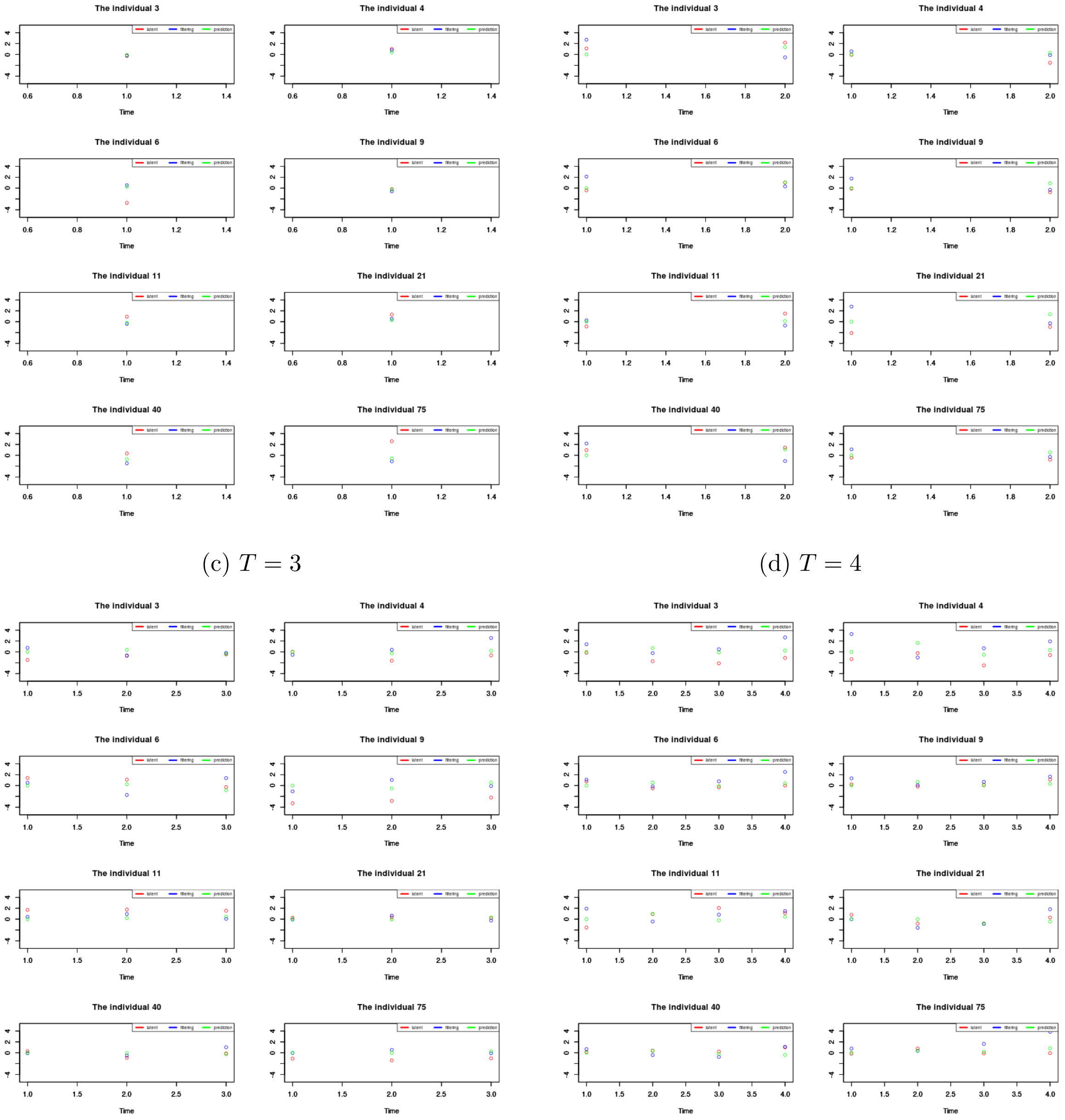
(a) $T = 3$

(b) $T = 4$

(c) $T = 5$

(d) $T = 10$

Figure 5.13: The graphs for $n = 300$ individuals with $T = 1, 2, 3$ and 4



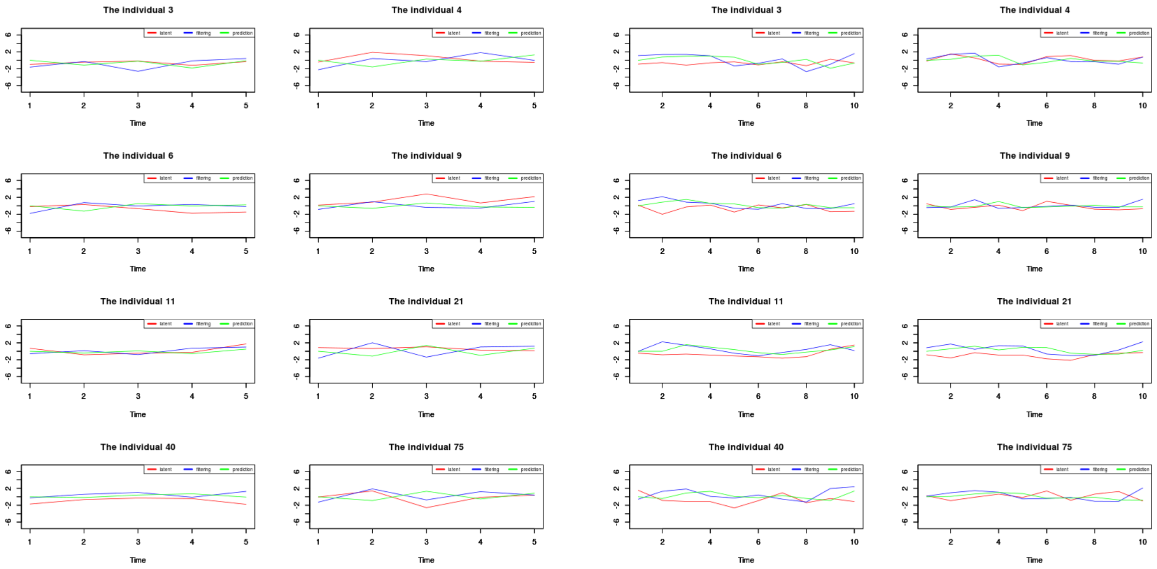
(a) $T = 1$

(b) $T = 2$

(c) $T = 3$

(d) $T = 4$

Figure 5.14: The graphs for $n = 300$ individuals with $T = 5$ and 10



(a) $T = 5$

(b) $T = 10$

5.2.2.4 The gaussian state noise with CHARN(1,1)

The state space equation is described as follows

$$X_i(t) = \rho_1 X_i(t-1) + \sqrt{\rho_1 + \rho_2 X_i^2(t-1)}\varepsilon_i(t), \varepsilon_i(t) \sim \mathcal{N}(0, R_t), \tag{5.2.5}$$

the simulation II outline is implemented. In step (1) set $\rho_1 = 0.5, \rho_1 = 0.2$ and $R_t = 1$. In steps (2-5), set $\rho_1^{(0)} = \rho_2^{(0)} = 0.4$, and $R_t^{(0)} = 0.5$. Figure 5.15 shows the results, where $n = 10$ at instants $T = 75, 100, 200, 300$. The working extended Kalman filter (WEKF) recursion succeeds in producing the best estimators for the state variables described by nonlinear state equation as the linear state equation with gaussian noise processes.

5.2.2.5 The gaussian state noise with CHARN(0,1)

The state space equation considered is as follows

$$X_i(t) = \sqrt{\rho_1 + \rho_2 X_i^2(t-1)}\varepsilon_i(t), \varepsilon_i(t) \sim \mathcal{N}(0, R_t), \tag{5.2.6}$$

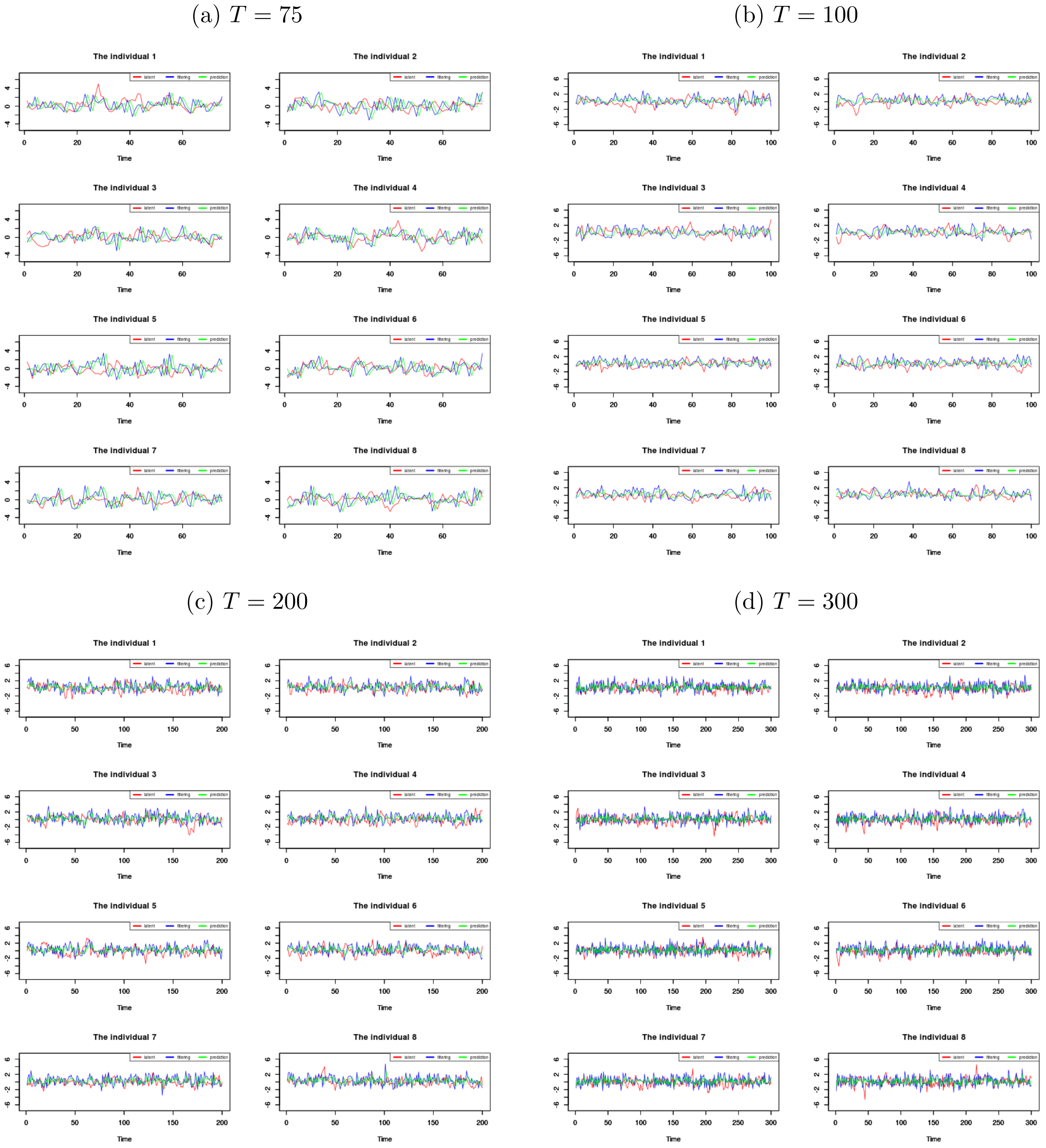
as the previous subsection, set $\rho_1 = 0.5, \rho_1 = 0.2$ and $R_t = 1$. Perform the steps (2-5) of outline with $\rho_1^{(0)} = \rho_2^{(0)} = 0.4$, and $R_t^{(0)} = 0.5$.

Figure 5.16 displays the results for $n = 10$ and $T = 75, 100, 200, 300$.

Here, the posterior mode in predictive step equal 0 due to the fact that we supposed the function F is equal 0 in the state equation. Hence the posterior mode in WEKF is $a_{t|t-1} = F(a_{t|t}, \mathbf{u}_i(t), \gamma)$, and it follows that $a_{t|t-1} = 0$.

As seen in Figure 5.16 the results are good, this proves the efficiency of our approach.

Figure 5.15: The graphs for $n = 10$ individuals



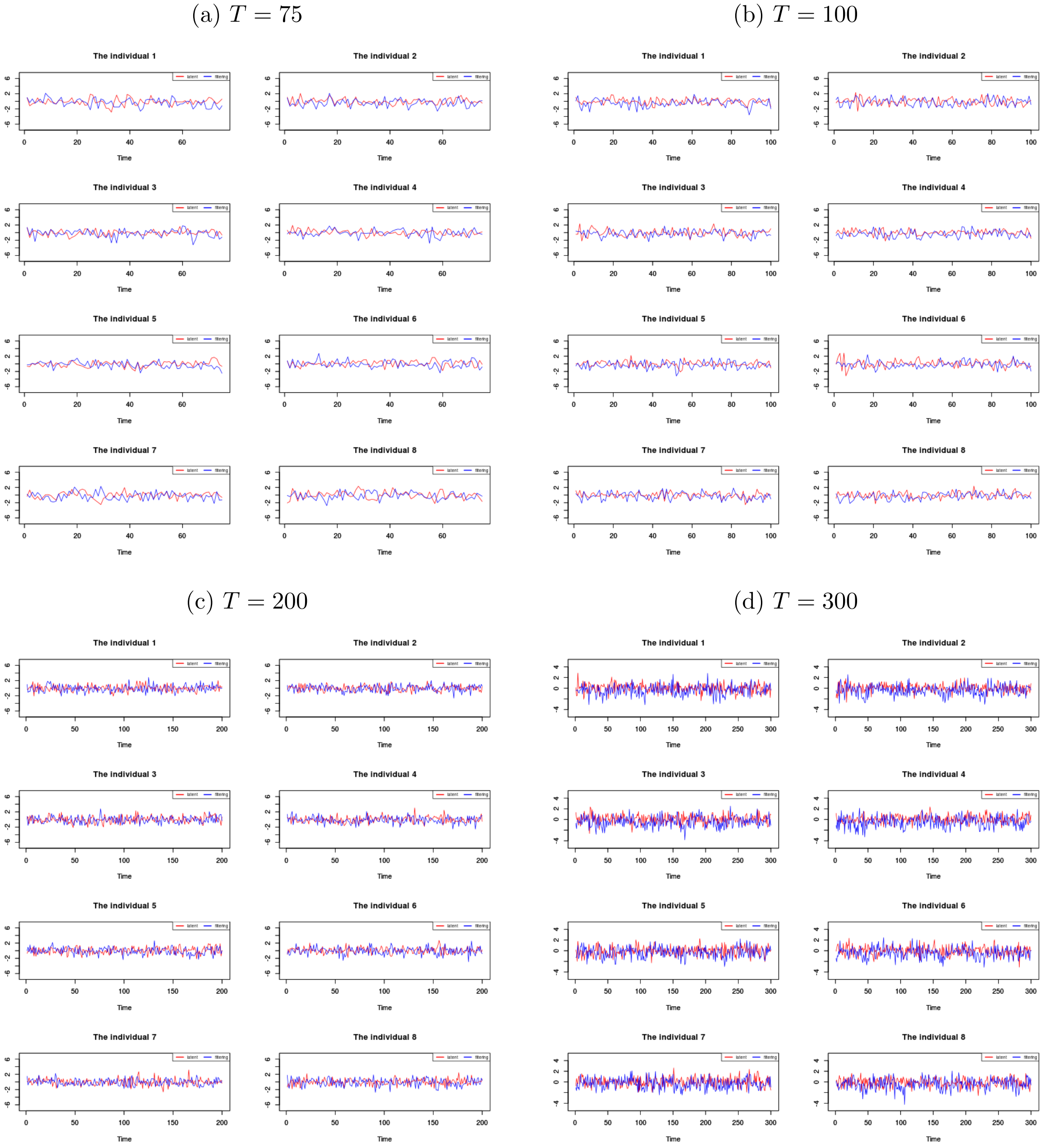
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.16: The graphs for $n = 10$ individuals



(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

5.2.2.6 The exponential state noise with AR(1) model

In this subsection, the simulation II outlined precedently is also Performed. We considered an AR(1) model for state equation, but with state noise from the exponential distribution $\varepsilon_i(t) \sim exp(\lambda)$. In step (1) of the simulation II outline, set $\lambda = 1$, to generate the state noise. The state noise $\varepsilon_i(t)$ is centralized by the following

$$\frac{\varepsilon_i(t) - \mathrm{E}(\varepsilon_i(t))}{\sqrt{\mathrm{Var}(\varepsilon_i(t))}}$$

In steps (2-5), the initial values are supposed $\lambda^{(0)} = 0.6$, and $\beta^{(0)}$ as the Gaussian distribution.
To apply the working Kalman filtering recursions, we supposed $\rho^{(p)} = 1/\lambda^{(p)}$ and $R_t^{(p)} = 1/\lambda^{(p)2}$.
As the previous section, there are two parts. First, $(n \leq 15)$ with occasions $(T > 75)$. Second $(n > 75)$ with occasions $(T \leq 10)$.

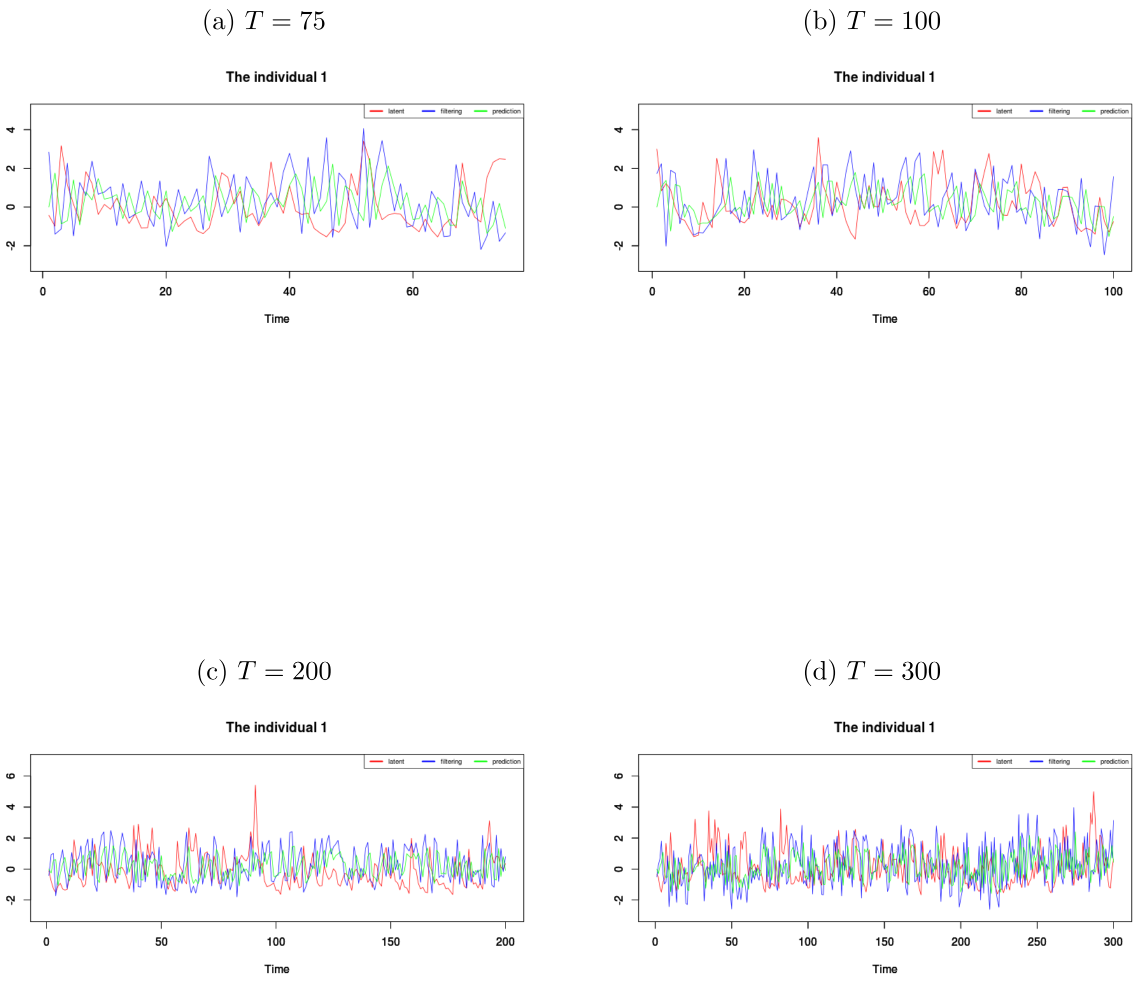
- **n small with T large**

As the gaussian state noise we consider $n = 1, 2, 3, 4, 10$, and $T = 75, 100, 200, 300$.
As seen in Figures 5.17-5.18 some sharp jumps are appearing in the curve of latent variables, this leads to distancing values of posterior mode from the latent variables. Figures 5.19-5.21 show the results are good.

- **n large with T small**

In this part,we consider $n = 75, 100, 200, 300$, and $T = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. As the gaussian state noise at instant $T = 1, 2, 3, 4$, the red circle refers to the latent variables, the blue circle refers to the posterior mode via the filtering step is , and the posterior mode via the prediction step is the green circle.
Figures 5.22-5.29 show at instant $T = 1, 2, 3, 4$, the results are excellent, but at $T = 5, 10$ there are some of the spaces in the curves, but the results are acceptable.

Figure 5.17: The graphs for $n = 1$ individual



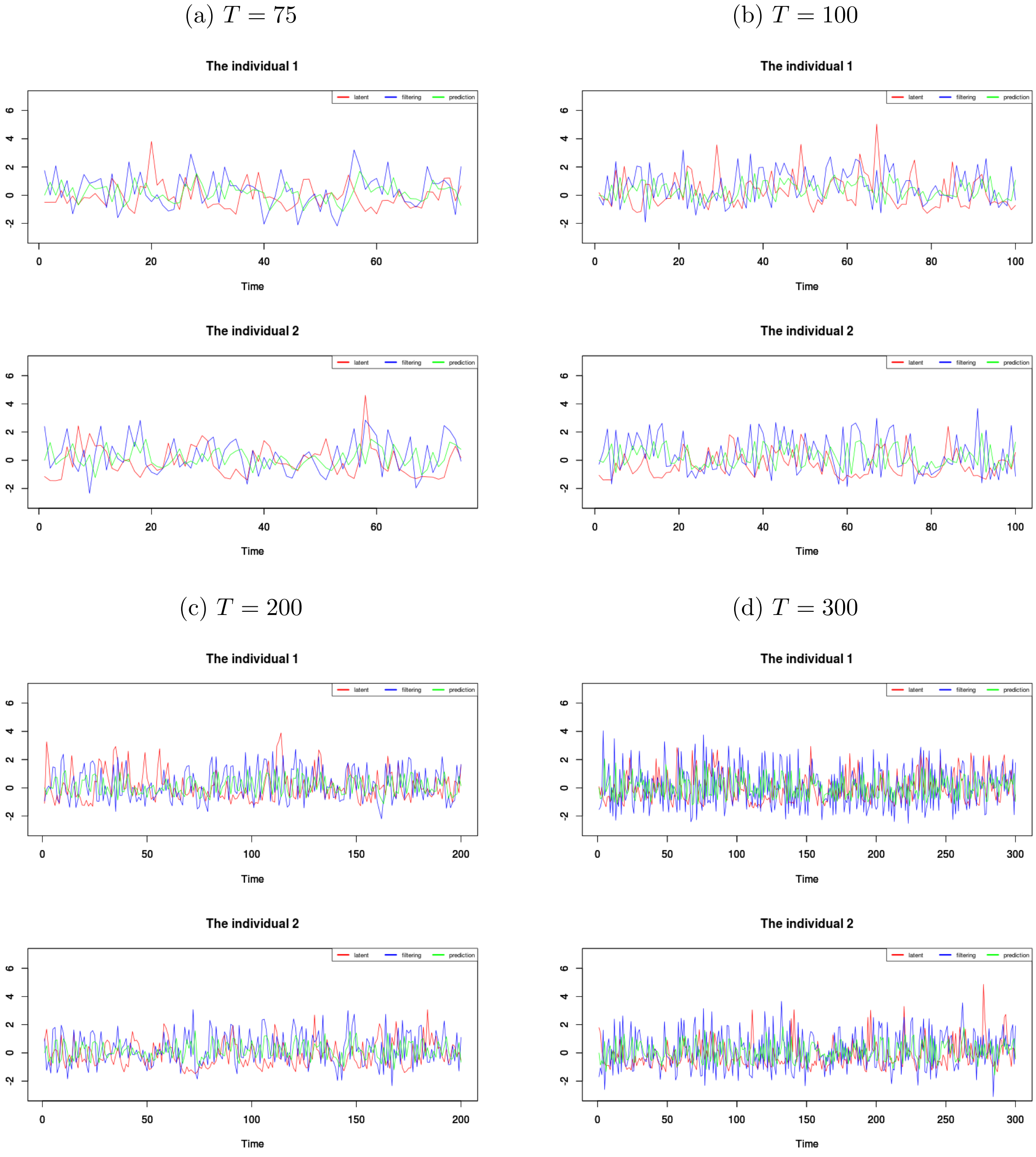
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.18: The graphs for $n = 2$ individuals



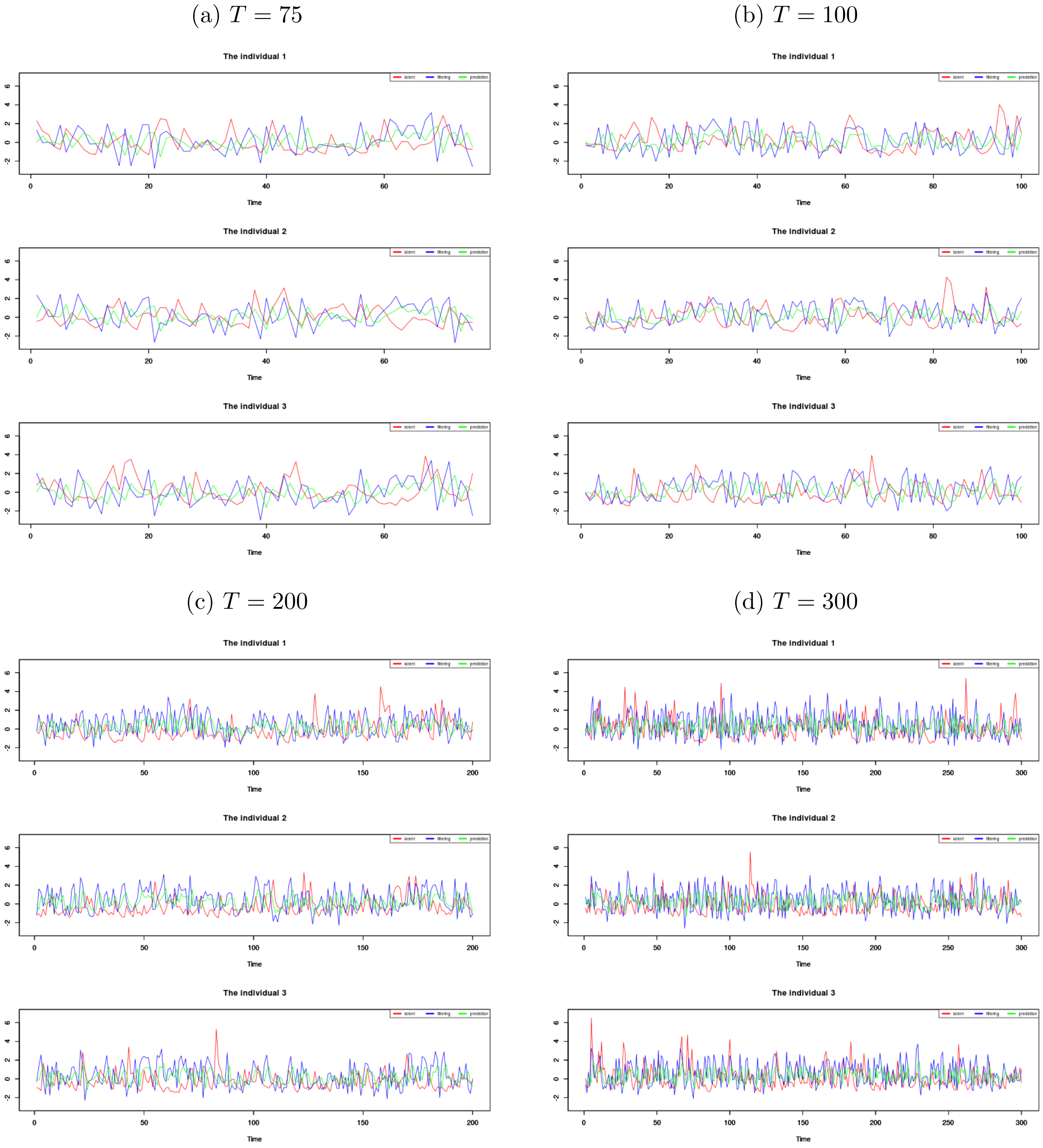
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.19: The graphs for $n = 3$ individuals

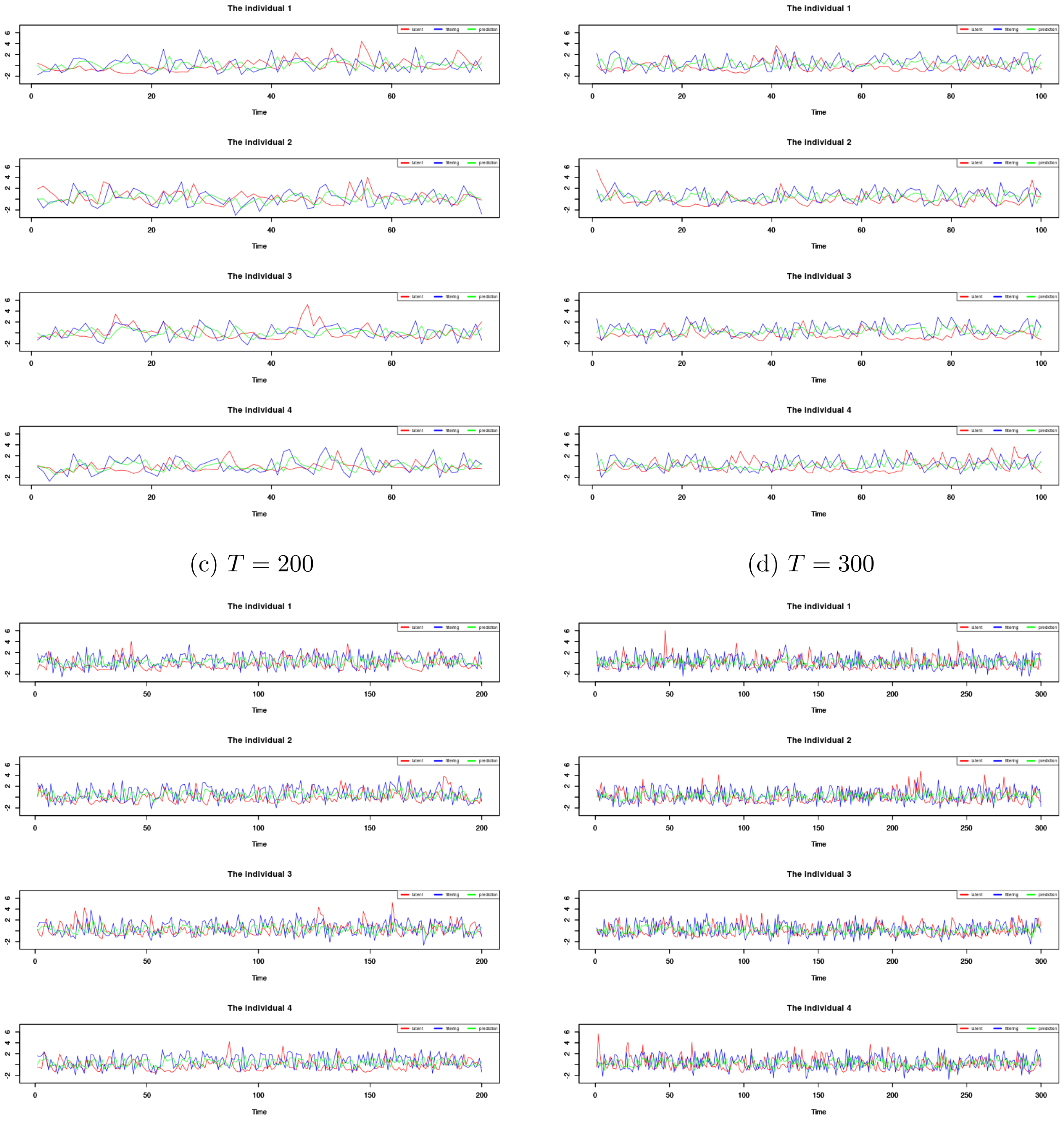

(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.20: The graphs for $n = 4$ individuals

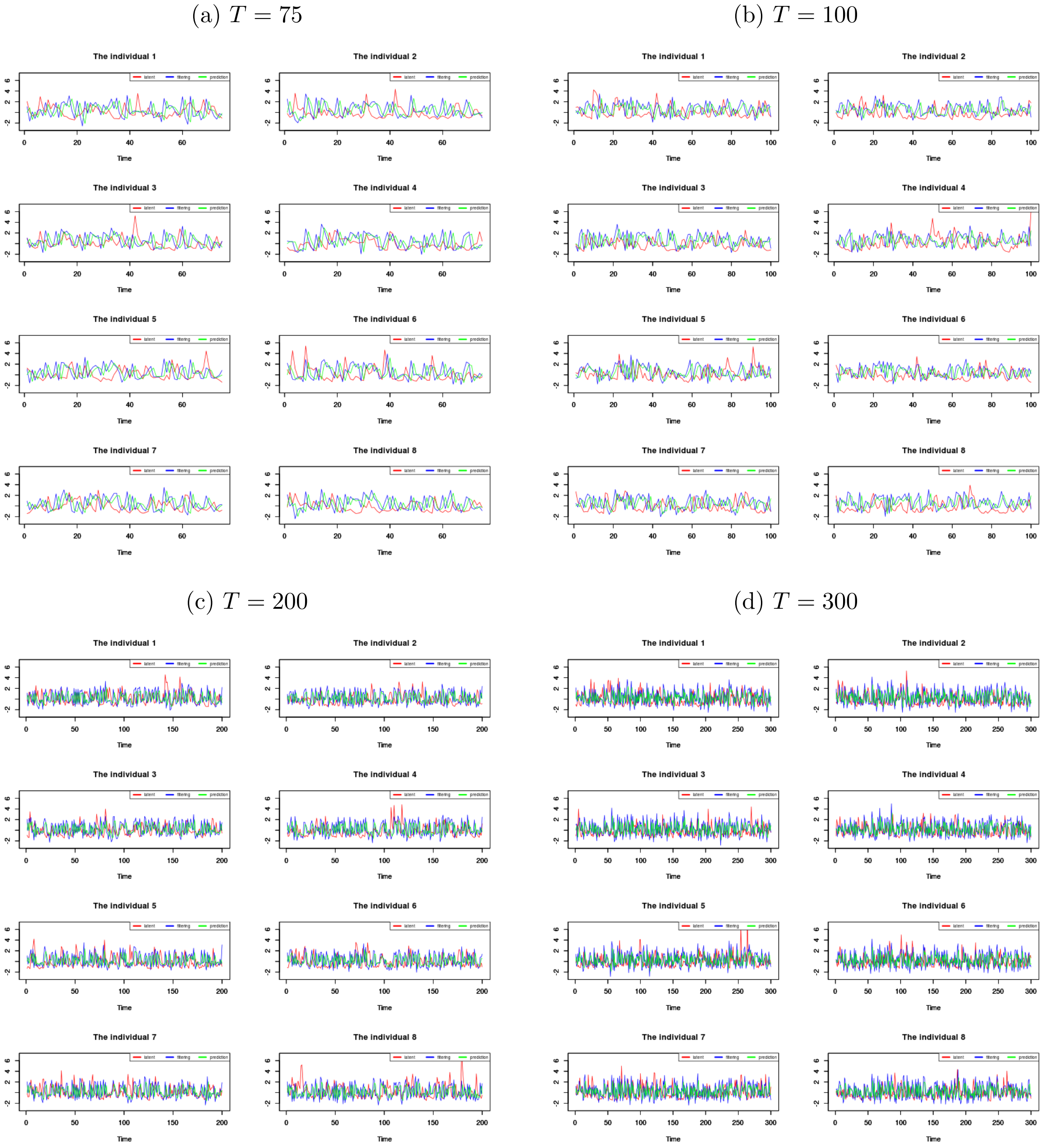

(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.21: The graphs for $n = 10$ individuals



(a) $T = 75$

(b) $T = 100$

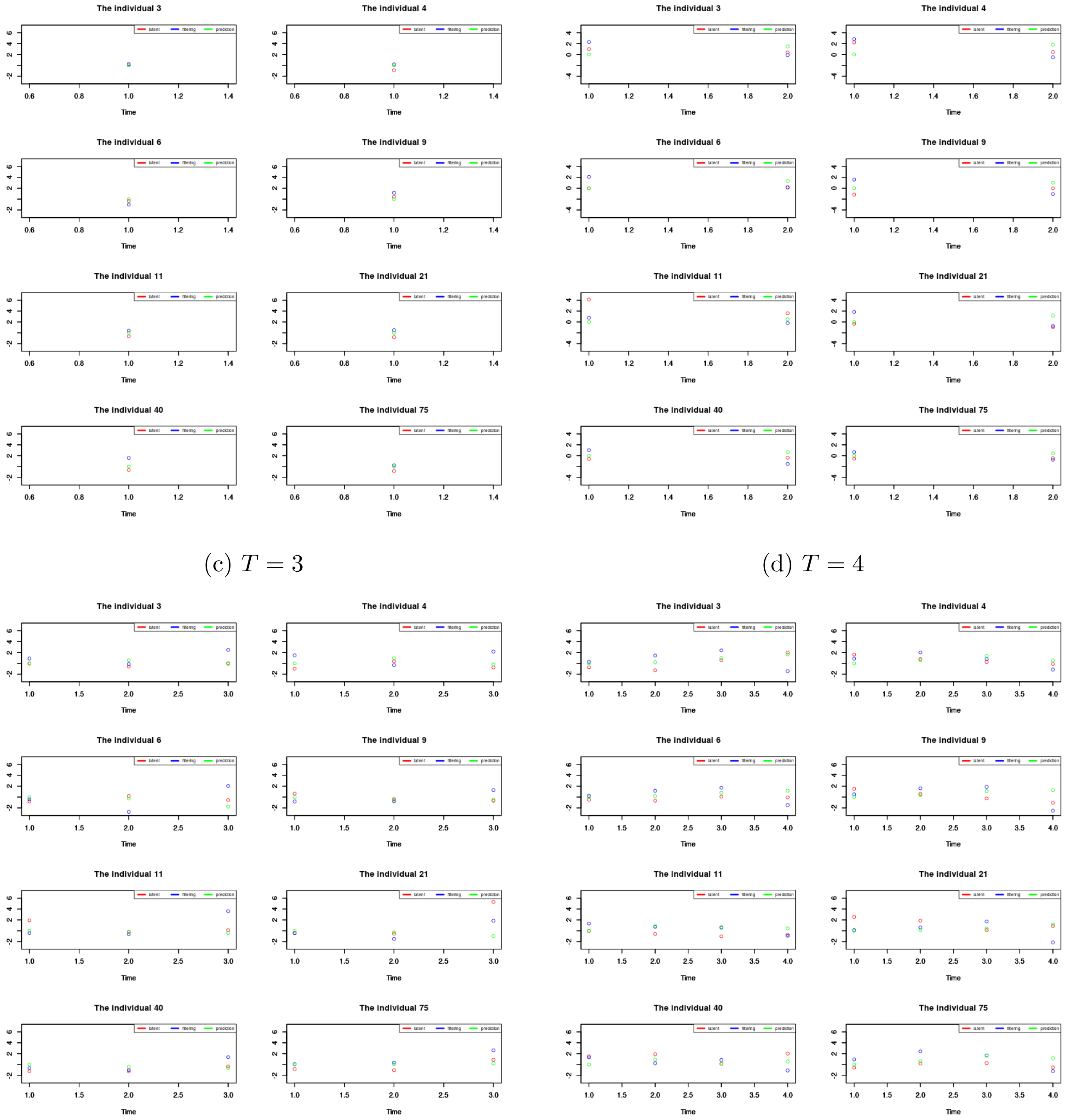
(c) $T = 200$

(d) $T = 300$

Figure 5.22: The graphs for $n = 75$ individuals with $T = 1, 2, 3$ and 4

(a) $T = 1$

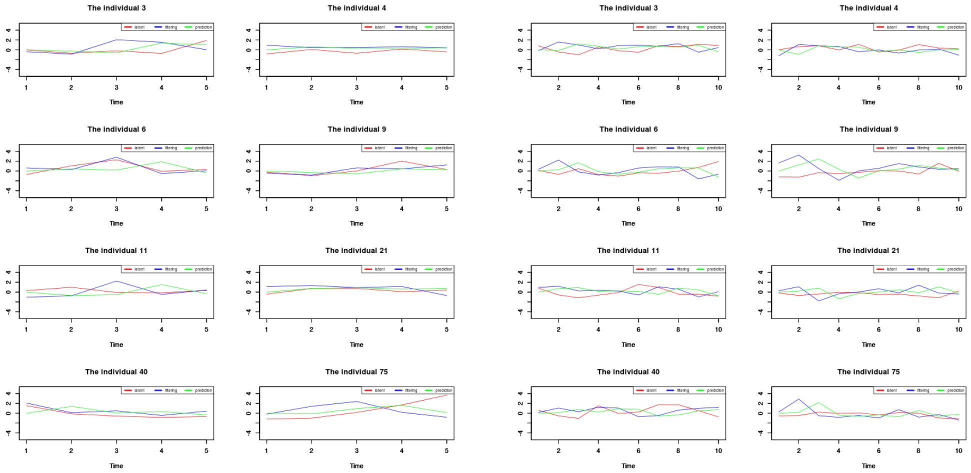
(b) $T = 2$



(c) $T = 3$

(d) $T = 4$

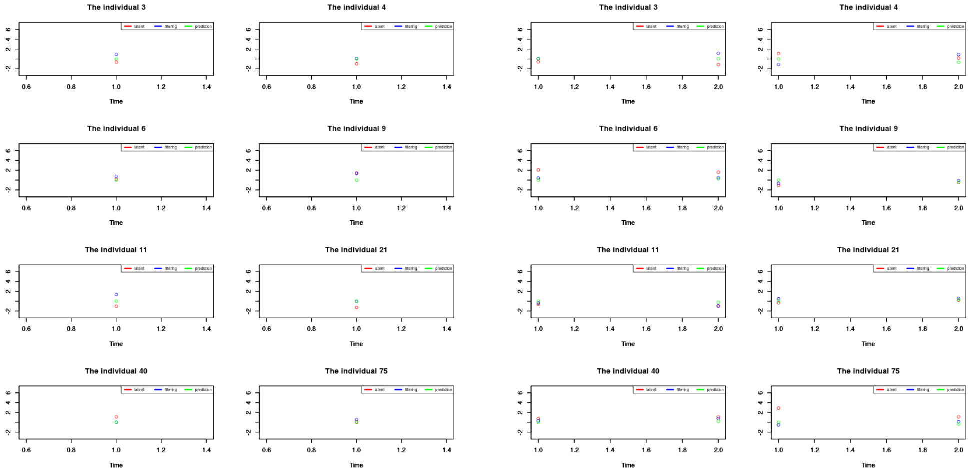
Figure 5.23: The graphs for $n = 75$ individuals with $T = 5$ and 10



(a) $T = 5$

(b) $T = 10$

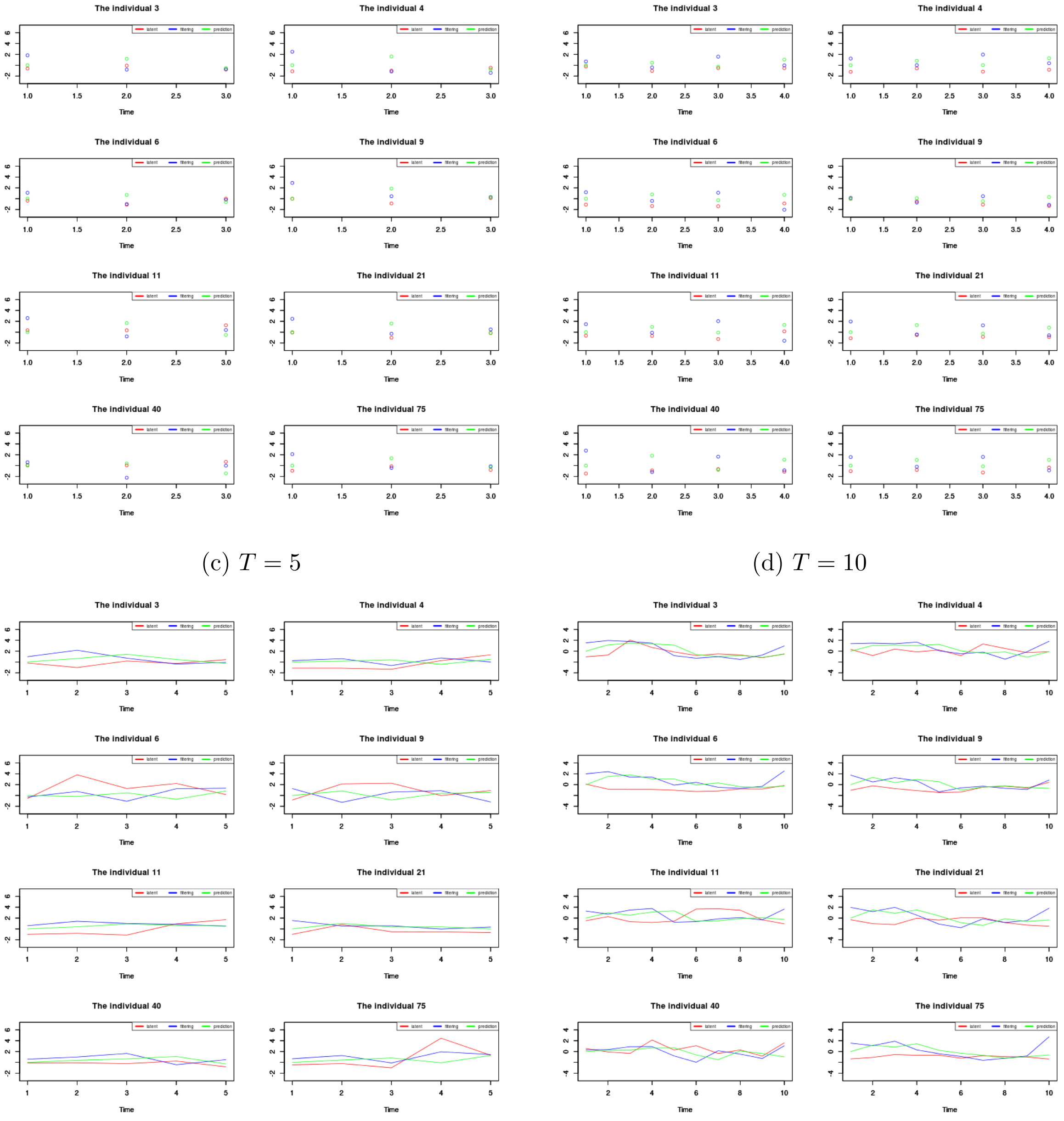
Figure 5.24: The graphs for $n = 100$ individuals with $T = 1$ and 2



(a) $T = 1$

(b) $T = 2$

Figure 5.25: The graphs for $n = 100$ individuals with $T = 3, 4, 5$ and 10



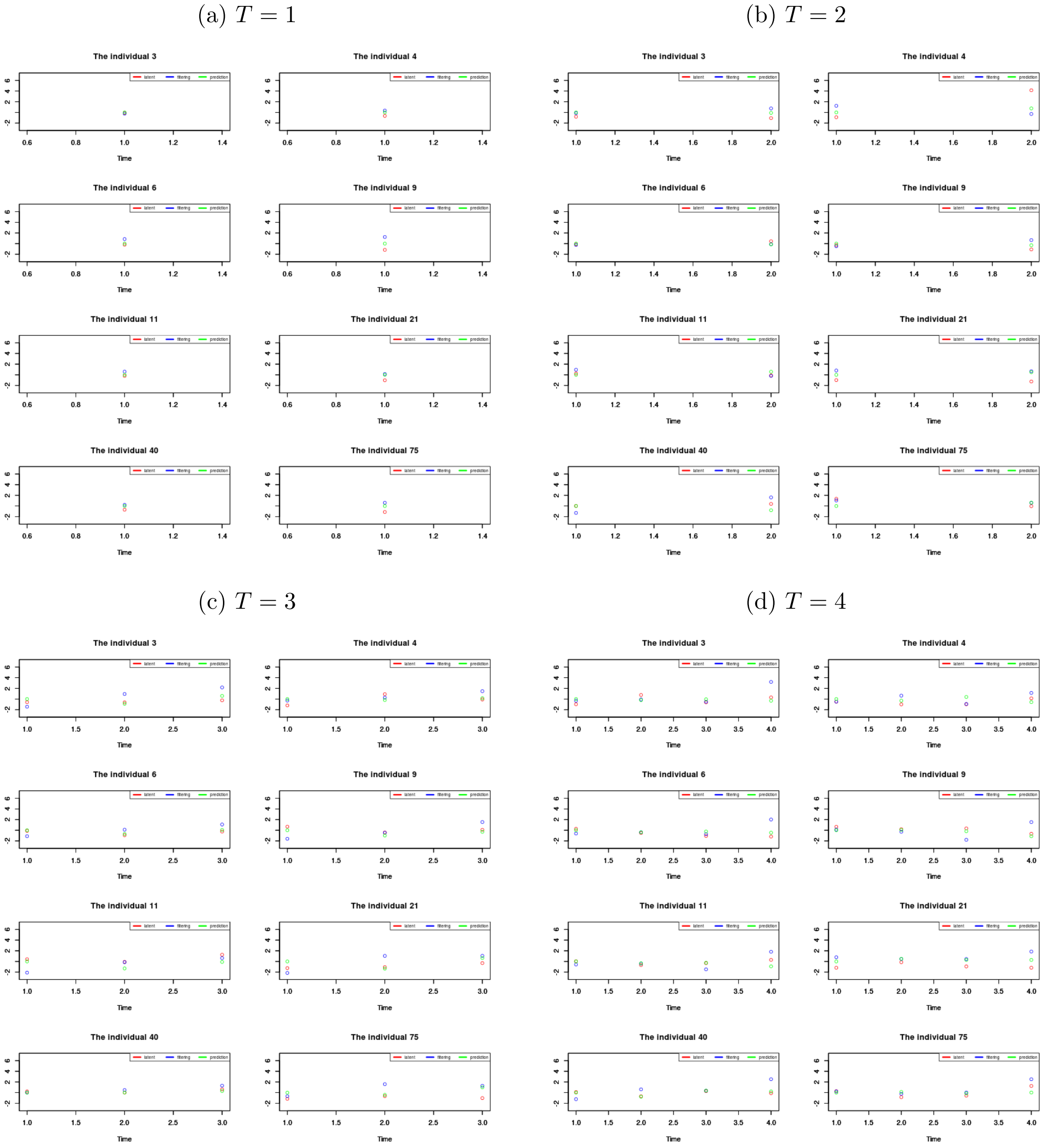
(a) $T = 3$

(b) $T = 4$

(c) $T = 5$

(d) $T = 10$

Figure 5.26: The graphs for $n = 200$ individuals with $T = 1, 2, 3$ and 4



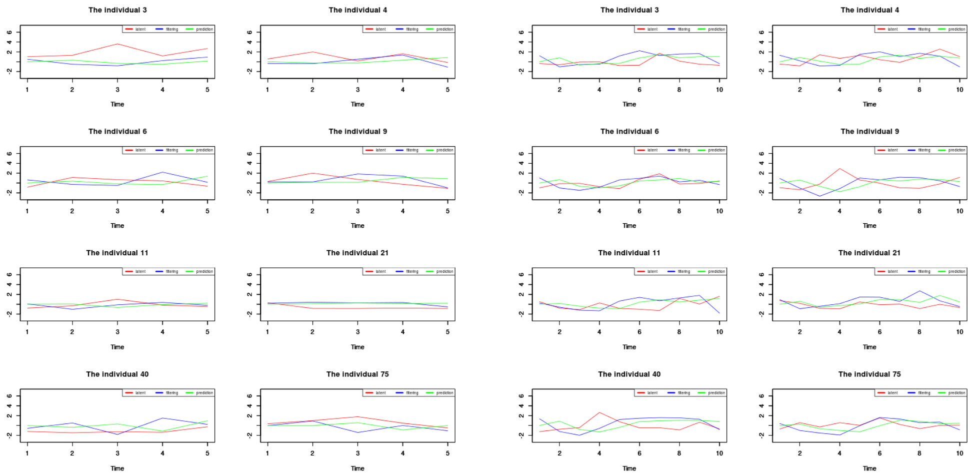
(a) $T = 1$

(b) $T = 2$

(c) $T = 3$

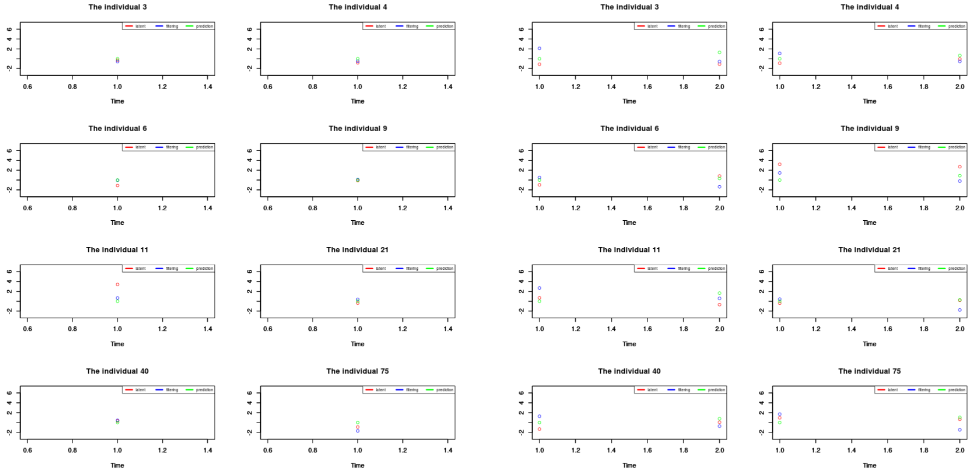
(d) $T = 4$

Figure 5.27: The graphs for $n = 200$ individuals with $T = 5$ and 10


(a) $T = 5$

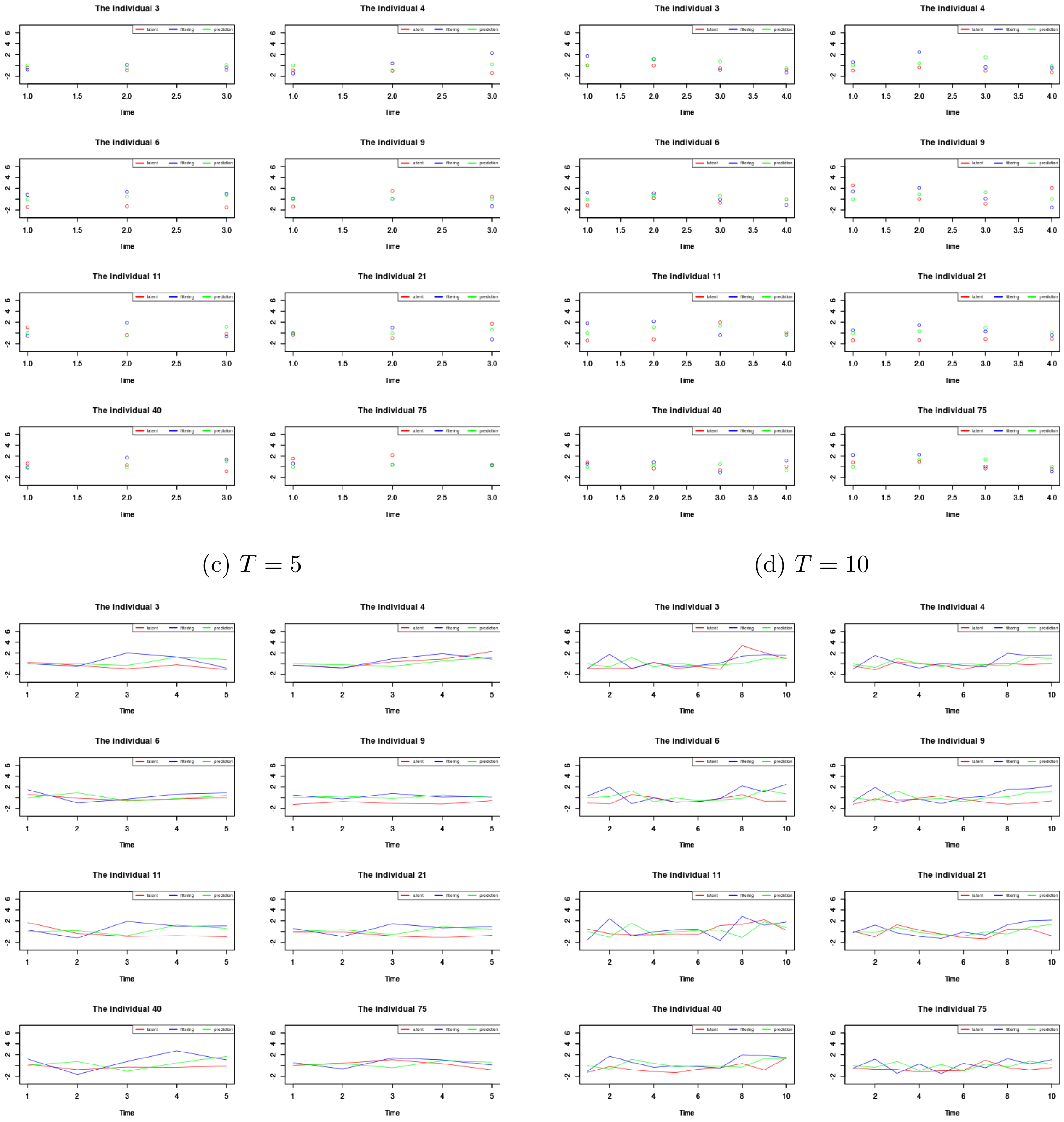
(b) $T = 10$

Figure 5.28: The graphs for $n = 300$ individuals with $T = 1$ and 2


(a) $T = 1$

(b) $T = 2$

Figure 5.29: The graphs for $n = 300$ individuals with $T = 3, 4, 5$ and $T = 10$



(a) $T = 3$

(b) $T = 4$

(c) $T = 5$

(d) $T = 10$

5.2.2.7 The exponential state noise with CHARN(1,1)

The state space equation is described as follows

$$X_i(t) = \rho_1 X_i(t-1) + \sqrt{\rho_1 + \rho_2 X_i^2(t-1)}\varepsilon_i(t), \varepsilon_i(t) \sim exp(\lambda). \tag{5.2.7}$$

Here, also in step(1) of the simulation II outline suppose $\rho_1 = 0.4, \rho_2 = 0.1$ and $\lambda = 1$. Implement the steps (2-5) with $\rho_1^{(0)} = 0.5, \rho_2^{(0)} = 0.3$, and $\lambda^{(0)} = 0.6$.
Figure 5.30 show the results are good as the case of a CHARN(1,1) model with gaussian state noise.

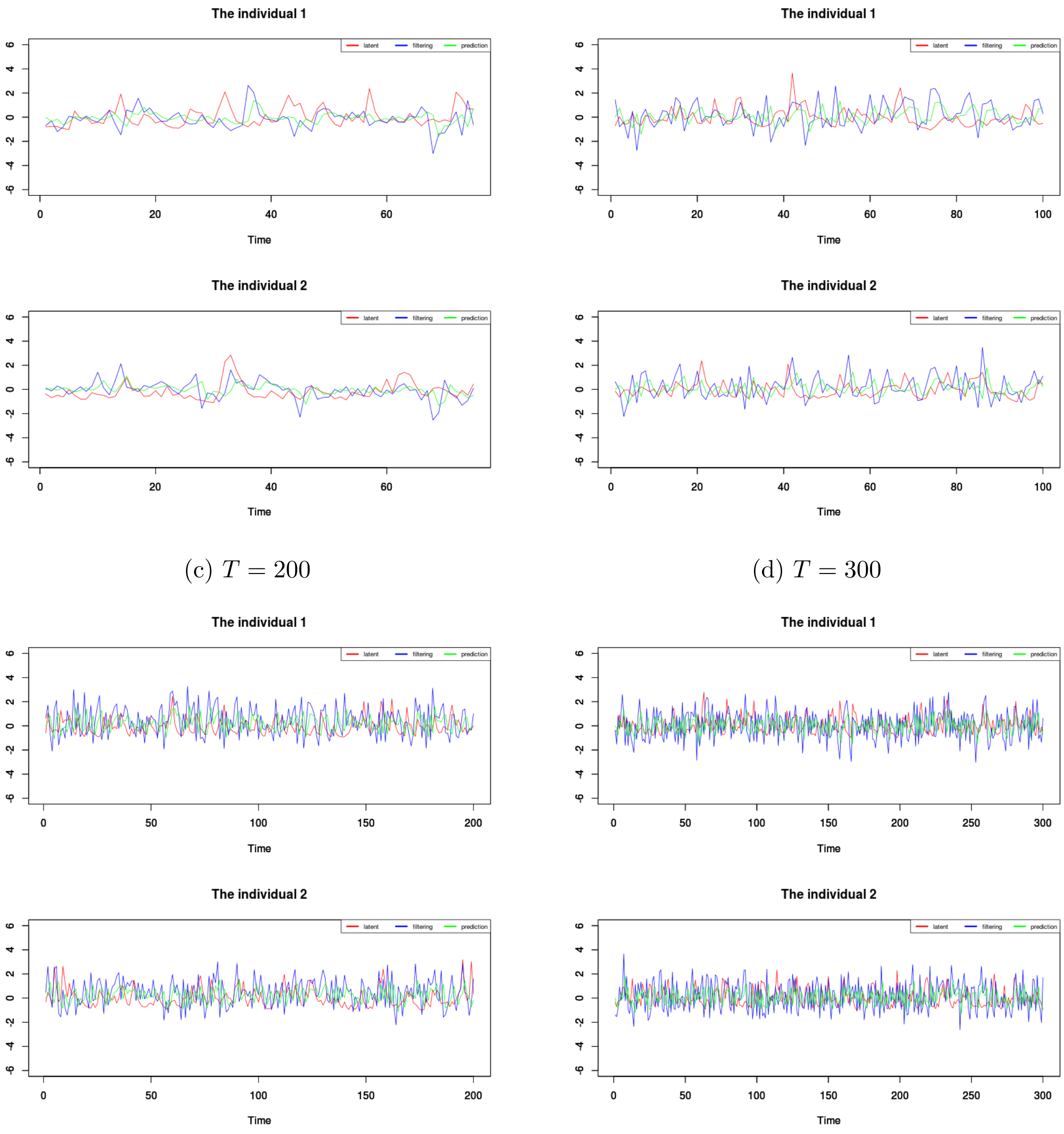
5.2.2.8 The exponential state noise with CHARN(0,1)

The state space equation is considered as

$$X_i(t) = \sqrt{\rho_1 + \rho_2 X_i^2(t-1)}\varepsilon_i(t), \varepsilon_i(t) \sim exp(\lambda). \tag{5.2.8}$$

In step(1) of the simulation II outline suppose $\rho_1 = 0.4, \rho_2 = 0.1$ and $\lambda = 1$, the steps (2-5) are performed with $\rho_1^{(0)} = 0.5, \rho_2^{(0)} = 0.3$, and $\lambda^{(0)} = 0.6$.
Figure 5.31 display the results are good as the previous cases.
As seen in Figure 5.31 the prediction equal 0 as we explained in gaussian state space.

Figure 5.30: The graphs for $n = 2$ individuals



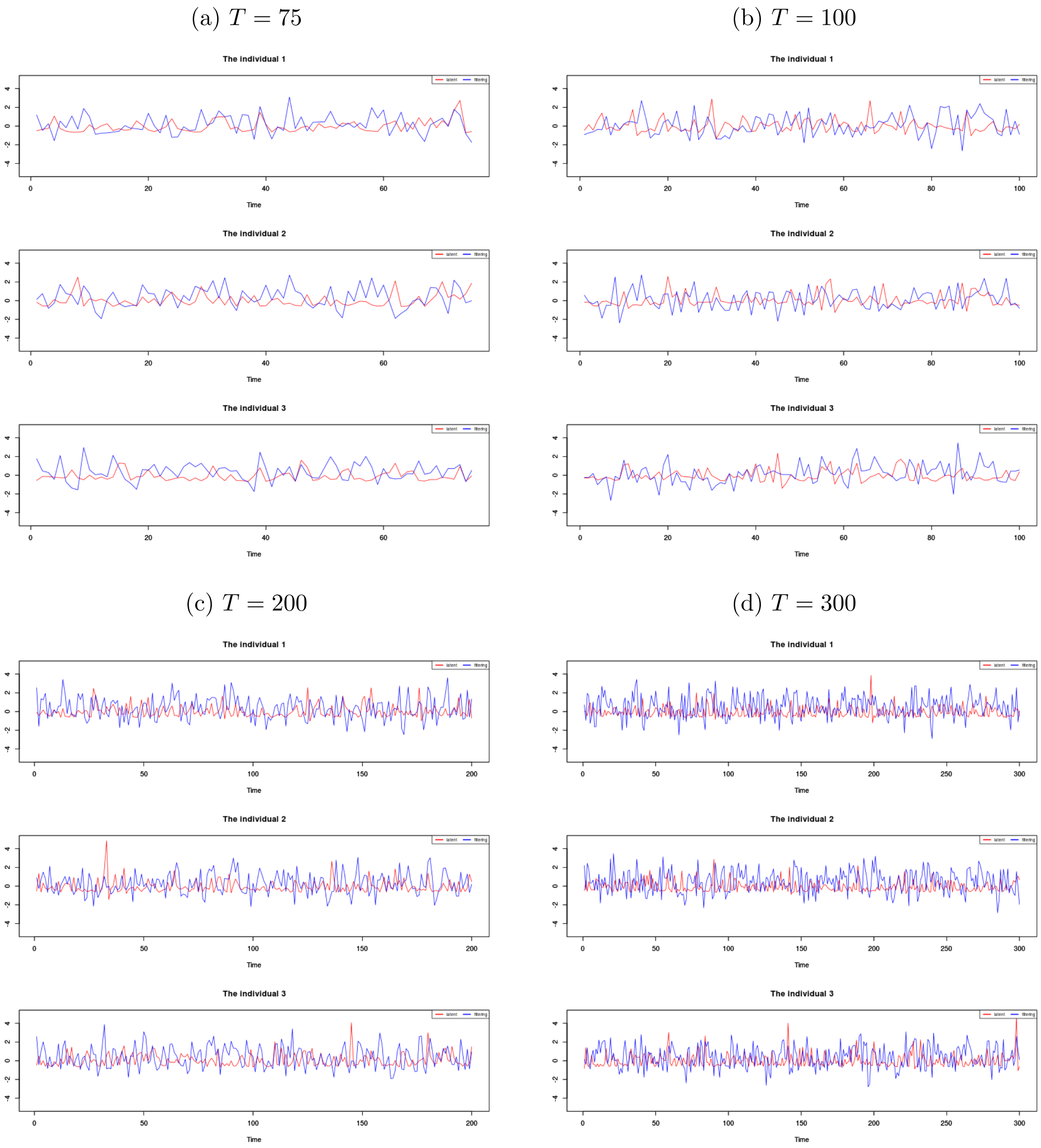
(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

Figure 5.31: The graphs for $n = 3$ individuals



(a) $T = 75$

(b) $T = 100$

(c) $T = 200$

(d) $T = 300$

5.2.3 Discussion of the results

A disadvantage of EM algorithm is its slow convergence as we have seen when applying the simulation II. Dempster et al.(1977) proved that converges only linearly to the maximum likelihood estimate. Then we used the numerical analysis through the Fisher scoring method. The Fisher scoring method is an iterative algorithm for obtaining the maximum of likelihood. It started with initial values for the parameters. The iterations continue until the maximum likelihood estimates are achieved.
The key for performing the working Kalman filter recursions is using the parameters which are calculated in the EM algorithm as initial values to estimate of the posterior mode. In some cases, we observed that the calculated values of the parameters achieved the condition of convergence for likelihood in equation (5.2.2). But it is not optimal initial values for the working extended Kalman filtering. So this explains the existence of some unsatisfactory results.
We conclude that the optimal choice for the initial values of the parameters leads to the best estimators for the state variables.

5.3 Application to real data

Rotonda *et al.* (2011) presented a study on quality of life. Their aim investigate factors correlated with cancer-related fatigue for women surgered for cancer. In this study, 502 patients were recruited from September 2008 until September 2010. Three French cancer centres 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, France. The patients filled the questionnaire several times, the patients' questionnaires were completed at their clinic visits, or they were given a postage-paid envelope to return the questionnaires. The questionnaires considered personality traits which was completed before the surgery. The LOT "life Orientation Test" questionnaire and the trait section of the STAI-B " State-Trait Anxiety Inventory" instrument were used, the Table 5.1 shows that the French version (STAI-B) consists of twenty items. 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 children's age, education, employment status, the group chemotherapy, the step of treatment, and the distance between patient's home and hospital are collected at the baseline assessment.

5.3.1 Data analysis

We obtained data from the above mentioned centres. Some of the patients' informations are missed (Not registered). So these patients' informations are removed from the data, and it remains 435 patients with complete information. Each patient is asked to fill the questionnaire at 10 instants. The questionnaire has 20 items. We selected 3 covariates: the marital status, the step of treatment, and the group of chemotherapy. These are said by the expert to be related to the latent variable studied that is the fatigue of the patient. They are:
marital status are scored : 1= "single", 2= "cohabitation", 3= "bride", 4="widow", 5="divorced", and 6= "bride/cohabitation". The scores of the step of treatment are : 1= "step I", 2= "step II", and 3= "step III". The scores of the group chemotherapy are : 1= " the group without chemotherapy ", and 2= "the group with chemotherapy ".

Table 5.1: The STAI (State-Trait Anxiety Questionnaire)

	Non	Plutôt non	Plutôt oui	Oui
1- je me sens calme				
2-je me sens en sécurité				
3- je me sens tendue				
4- je me sens surmenée				
5- je me sens tranquille,bien dans ma peau				
6- je me sens bouleversée				
7- je me sens préoccupée par tout les malheurs possibles				
8- je me sens comblée				
9- je me sens effrayée				
10- je me sens à l'aise				
11- je me sens que j'ai confiance en moi				
12- je me sens nerveuse				
13- je suis affolée				
14- je me sens indécise				
15- je suis détendue				
16- je me sens satisfaite				
17- je suis préoccupée				
18- je ne sais plus où j'en suis, je me sens perturbée				
19- je me sens solide, posée				
20- je me sens de bonne humeur, aimable				

5.3.1.1 The modeling

The observation equations is as follows

$$\Pr[Y_{ik}(t) = y^1_{ik}(t), \cdots y^{c_k}_{ik}(t) \mid X_i(t) = x_i(t)] = \prod_{s=1}^{c_k} [\pi^s_{ik}(t)]^{y^s_{ik}(t)}, \qquad (5.3.1)$$

where

$$\pi^s_{ik}(t) = \frac{\exp[\eta^s_{ik}(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta^j_{ik}(t)]} \quad . \qquad (5.3.2)$$

The link function $\eta^s_{ik}(t)$ is defined to be the logit, namely

$$\eta^s_{ik}(t) = logit(\pi^s_{ik}(t)) = \log\left[\frac{\pi^s_{ik}(t)}{\pi^{c_k}_{ik}(t)}\right] = \mathbf{u}_i^\top(t)\beta^s_k + X_i(t), s = 1, \cdots, 4.$$

The marital status is classified into two classes. First, $u_{m1_i}(t)$ has the values of " bride/cohabitation" = (2,3,6). Second, $u_{m2_i}(t)$ has the other values. The step of treatment is classified into three classes. $u_{t1_i}(t)$ has the value "step I "=1, $u_{t2_i}(t)$ has the value "step II "=2, and $u_{t3_i}(t)$ has the value "step III "=3. The group of chemotherapy for two classes, $u_{g1_i}(t)$ has value "the group without chemotherapy"=1, and $u_{g2_i}(t)$ has value "the group with chemotherapy"=2.
The link function can be rewritten with the selected covariates as follows

$$\begin{aligned}\eta^s_{ik}(t) = \beta^s_{0k} &+ \beta^s_{m1k}u_{m1_i}(t) + \beta^s_{m2k}u_{m2_i}(t) + \beta^s_{t1k}u_{t1_i}(t) + \beta^s_{t2k}u_{t2_i}(t)\\ &+\beta^s_{t3k}u_{t3_i}(t) + \beta^k_{g1k}u_{g1_i}(t) + \beta^s_{g2k}u_{g2_i}(t) + X_i(t) \qquad (5.3.3)\end{aligned}$$

where

- β^s_{0k} is the intercept parameter.
- $\beta^s_{m1k}, \beta^s_{m2k}$ are the parameters of the marital status classes.
- $\beta^s_{t1k}, \beta^s_{t2k}, \beta^s_{t3k}$ are the parameters of the step of treatment classes.
- $\beta^s_{g1k}, \beta^s_{g2k}$ are the parameters of the group of chemotherapy classes.
- $X_i(t)$ the fatigue of the individual i at instant t, is not observed (latent variable) and is assumed to have numerical values, and to a follow a CHARN model described by the equation:

$$X_i(t) = F(X_i(t-1), \mathbf{u}_i(t), \gamma) + H(X_i(t-1), \mathbf{u}_i(t), \delta)\varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0, R_t), \quad (5.3.4)$$

5.3.2 The outline of data analysis

The following steps are performed to find the fatigue of the individuals

1. Read the data from excel.
2. Compute the posterior distribution $p(\mathbf{X} \mid \mathbf{Y})$ via the Auxiliary Iterated Extended Kalman Particle Filter (AIEKPF) algorithm.
3. Set iteration $p = 0$, 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 $\beta^{p+1}, \gamma^{p+1}, \rho^{p+1}$ via EM algorithm. Set the initial values of parameters R_t^0, ρ^0, and $\beta_k^{s(0)}$
5. Implement the Working Extended Kalman Filtering Recursions (WEKF) to compute the posterior mode $a_i^{p+1}(t)$. If $\mid a_i^{p+1}(t) - a_i^p(t) \mid < 0.001$, STOP, else set $p = p + 1$ and go to step 3.

   As we mentioned in preceding Chapters , the working extended Kalman filter recursions depends on $F(.,.,.)$ and $H(.,.,.)$ functions of the state equation. Then, two cases to the CHARN model are performed. First, the AR(1) model. Second, the CHARN(1,1) model.

5.3.3 AR(1) model

The steps of outline are implemented with AR(1) as in equation (5.3.4), set the initial values of parameters as:

$R_t^0 = 0.5, \rho^0 = 0.6$, and $\beta_{rk}^{s(0)}$ is as follows

$$\begin{aligned}
\beta_{01}^{s(0)} = \beta_{02}^{s(0)} = \cdots = \beta_{0q}^{s(0)} &= \begin{pmatrix} 3 & 3 & 0.1 & -2 \end{pmatrix} \\
\beta_{m11}^{s(0)} = \beta_{m12}^{s(0)} = \cdots = \beta_{m1q}^{s(0)} &= \begin{pmatrix} 1.5 & 0.1 & 0.1 & 1.7 \end{pmatrix} \\
\beta_{m21}^{s(0)} = \beta_{m22}^{s(0)} = \cdots = \beta_{m2q}^{s(0)} &= \begin{pmatrix} -0.5 & 0.9 & 1.5 & 0.1 \end{pmatrix} \\
\beta_{t11}^{s(0)} = \beta_{t12}^{s(0)} = \cdots = \beta_{t1q}^{s(0)} &= \begin{pmatrix} 0.2 & 0.7 & 0.8 & 0.6 \end{pmatrix} \\
\beta_{t21}^{s(0)} = \beta_{t22}^{s(0)} = \cdots = \beta_{t2q}^{s(0)} &= \begin{pmatrix} 2 & 2 & 0.1 & -0.6 \end{pmatrix} \\
\beta_{t31}^{s(0)} = \beta_{t32}^{s(0)} = \cdots = \beta_{t3q}^{s(0)} &= \begin{pmatrix} -0.9 & -0.7 & 0.2 & 0.2 \end{pmatrix} \\
\beta_{g11}^{s(0)} = \beta_{g12}^{s(0)} = \cdots = \beta_{g1q}^{s(0)} &= \begin{pmatrix} 2 & 2 & 0.4 & 0.1 \end{pmatrix} \\
\beta_{g21}^{s(0)} = \beta_{g22}^{s(0)} = \cdots = \beta_{g2q}^{s(0)} &= \begin{pmatrix} 0.1 & 0.1 & 2 & 2 \end{pmatrix}
\end{aligned}$$

The Working Kalman filter recursions is performed. The fatigue of individuals is estimated by posterior mode, Figures 5.32 - 5.36 display the filter and predictive steps of the first ten individuals in the ranking for every hundred individuals. We recall that the patient fatigue is varying over time around a mean value assumed to be nil in our work.

Figure 5.32 shows the curves of prediction and filter steps for the 1st individual to the fifth individual are identical , and the sixth individual to the tenth individual are similar. One observes the same thing with Figures 5.33- 5.36. The reason is that the effects of the covariates are identical which make the values of link functions are also similar. Consequently, the predictive and filter values are almost equal for each group of individuals. The variance-covariance matrix for predictive, filter $P_{t|t-1}, P_{t|t}$ respectively have values less than 0.001.

In general, on Figures 5.32- 5.36, a positive value indicates that the patient is rested and a negative value indicates that he is tired.

From Figure 5.32 we can see that the individual 1 is tired at $T = 1$, and he feels rested at $T = 2, 3, 4$. Then, he returns tired at $T = 5$, while he feels rested at $T = 6, 7, 8, 9, 10$.

The individual 6 is tired at $T = 1$, and he feels rested at $T = 2, 3, 4$. Then, he is tired at $T = 5$, but he feels rested at $T = 6, 7$, whereas at $T = 8$ he is tired. Finally, at $T = 9, 10$ he feels rested.

Figure 5.32: The graphs for the 1st individual to 10th individual with AR(1) model

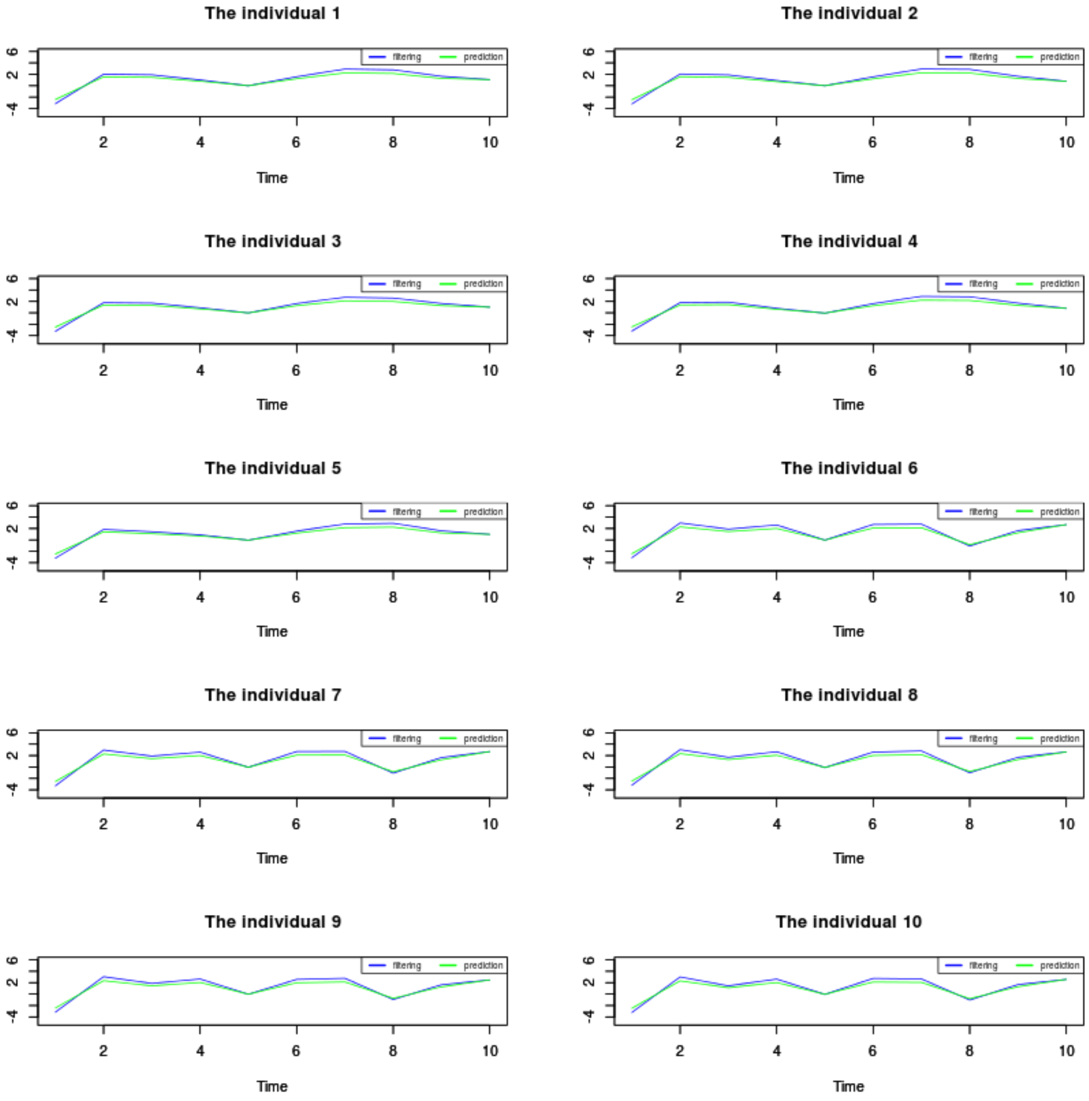


Figure 5.33: The graphs for the 100th individual to 109th individual with AR(1) model

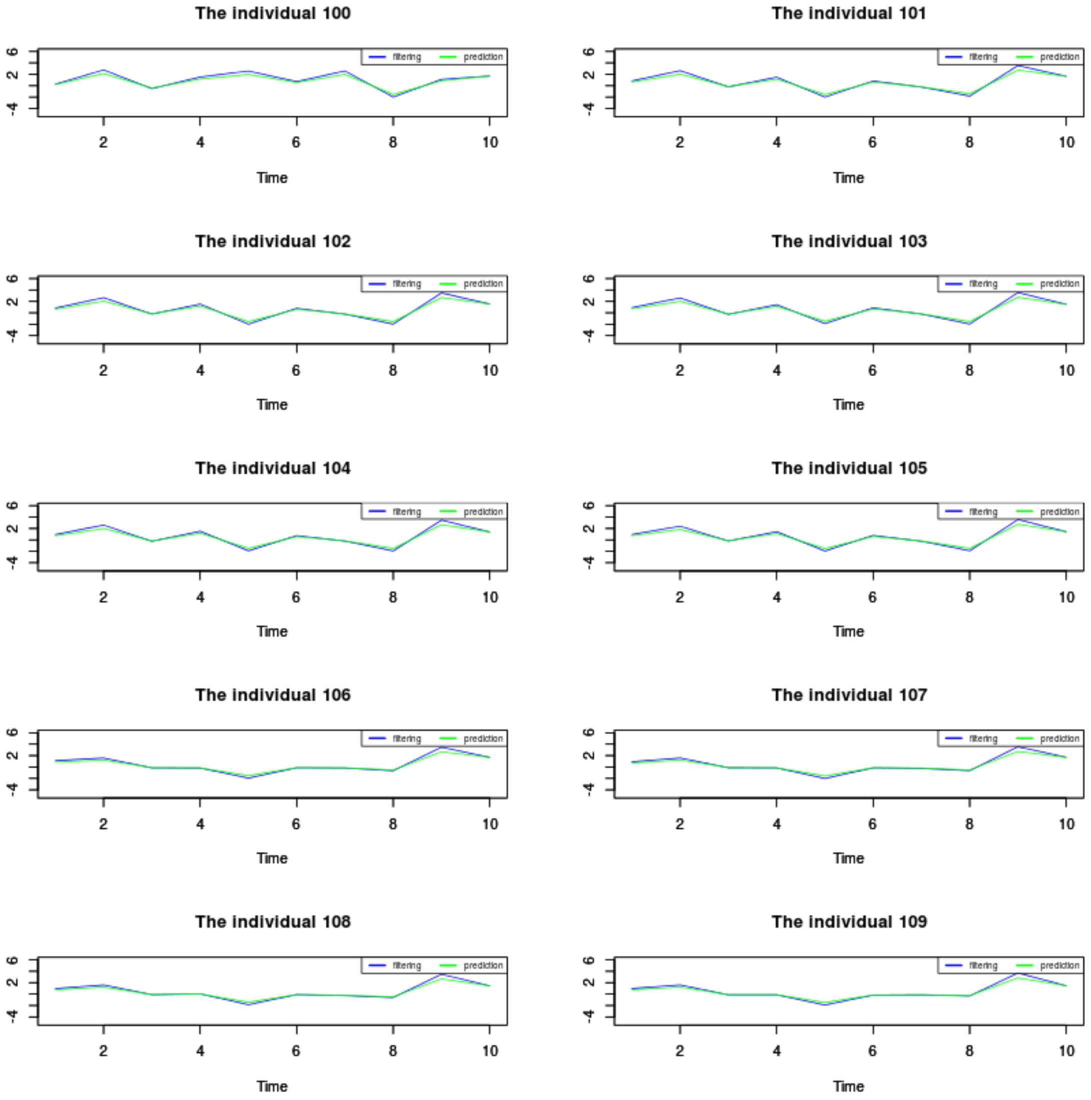


Figure 5.34: The graphs for the 200th individual to 209th individual with AR(1) model

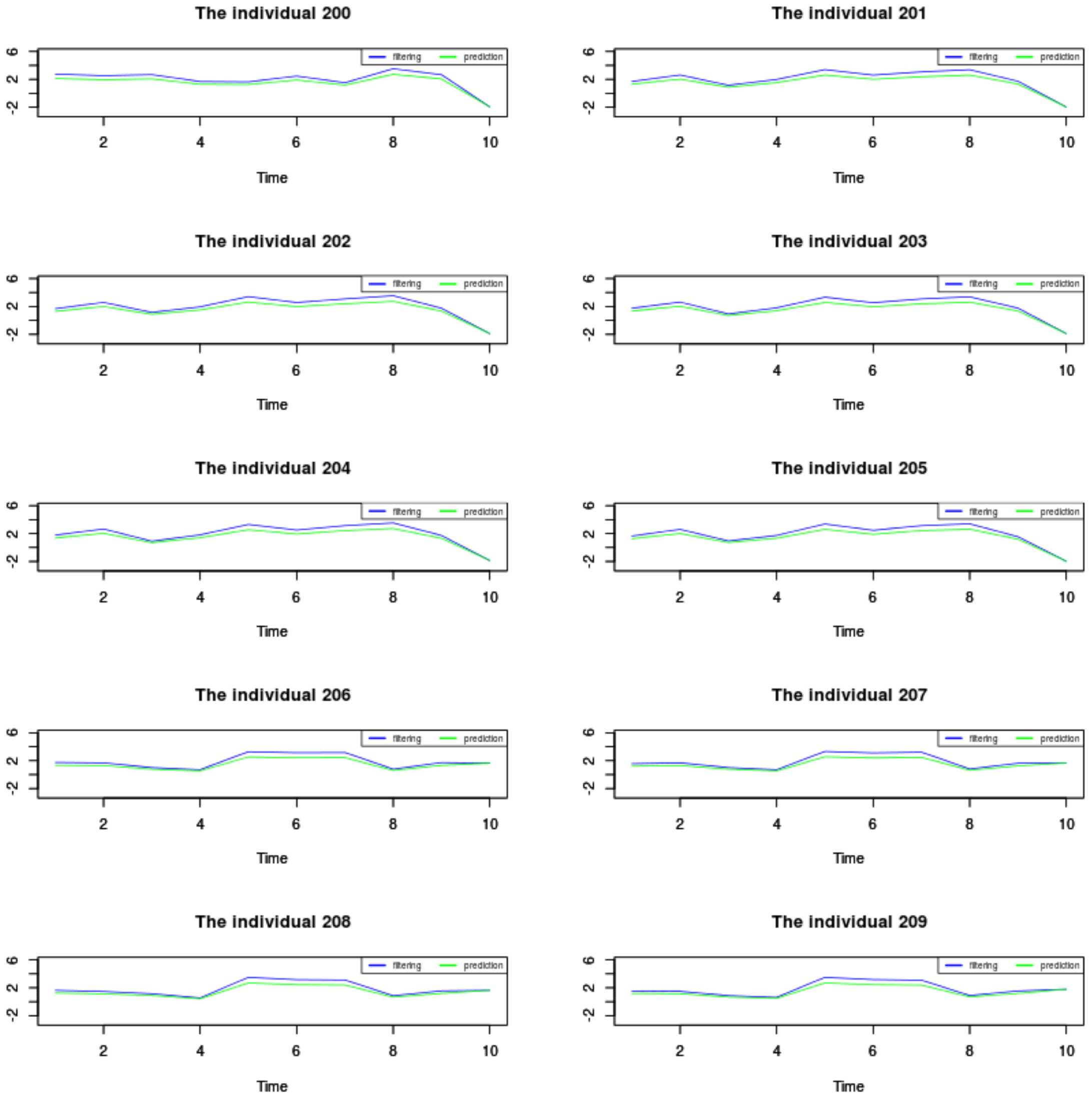


Figure 5.35: The graphs for the 300th individual to 309th individual with AR(1) model

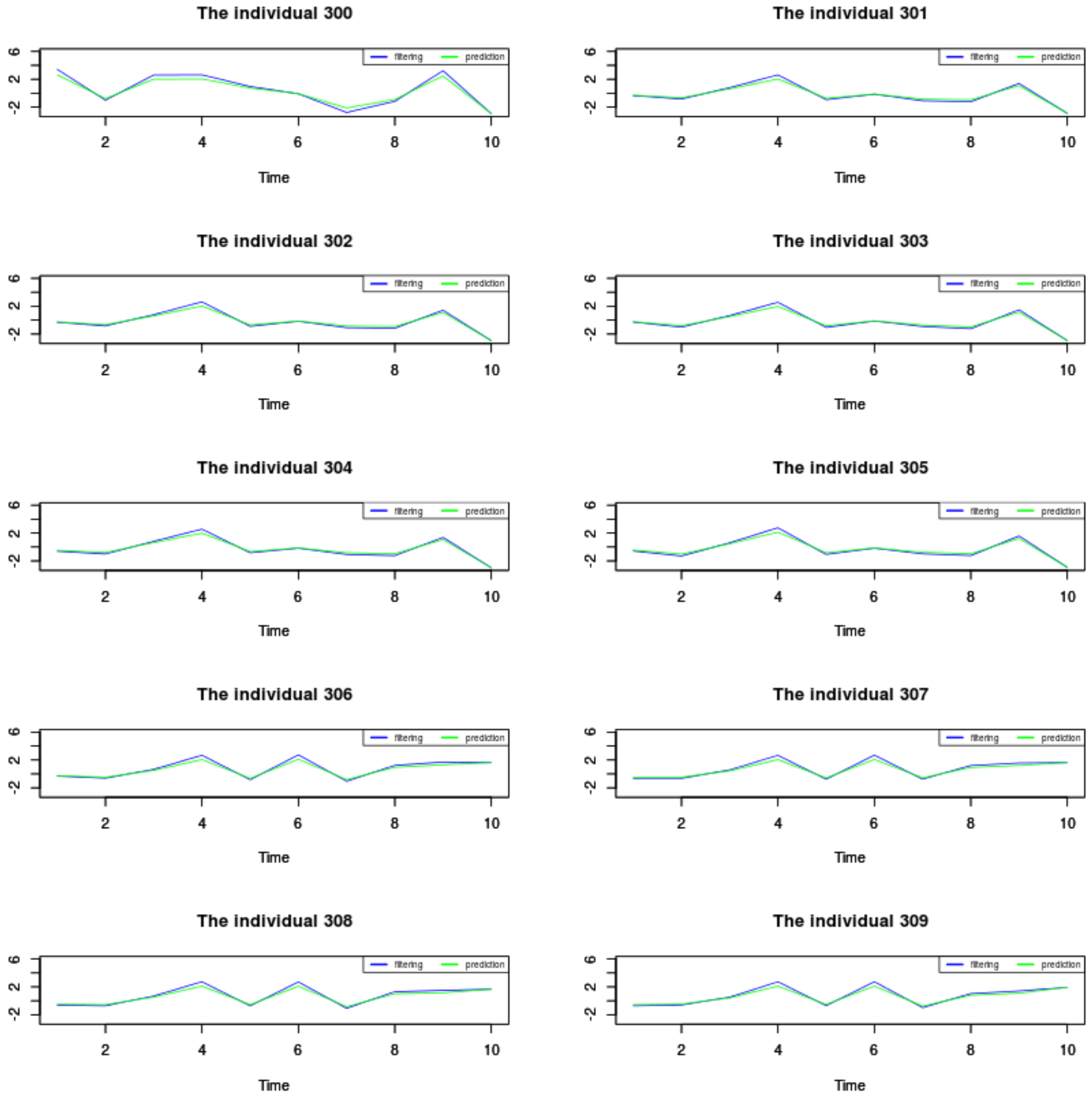
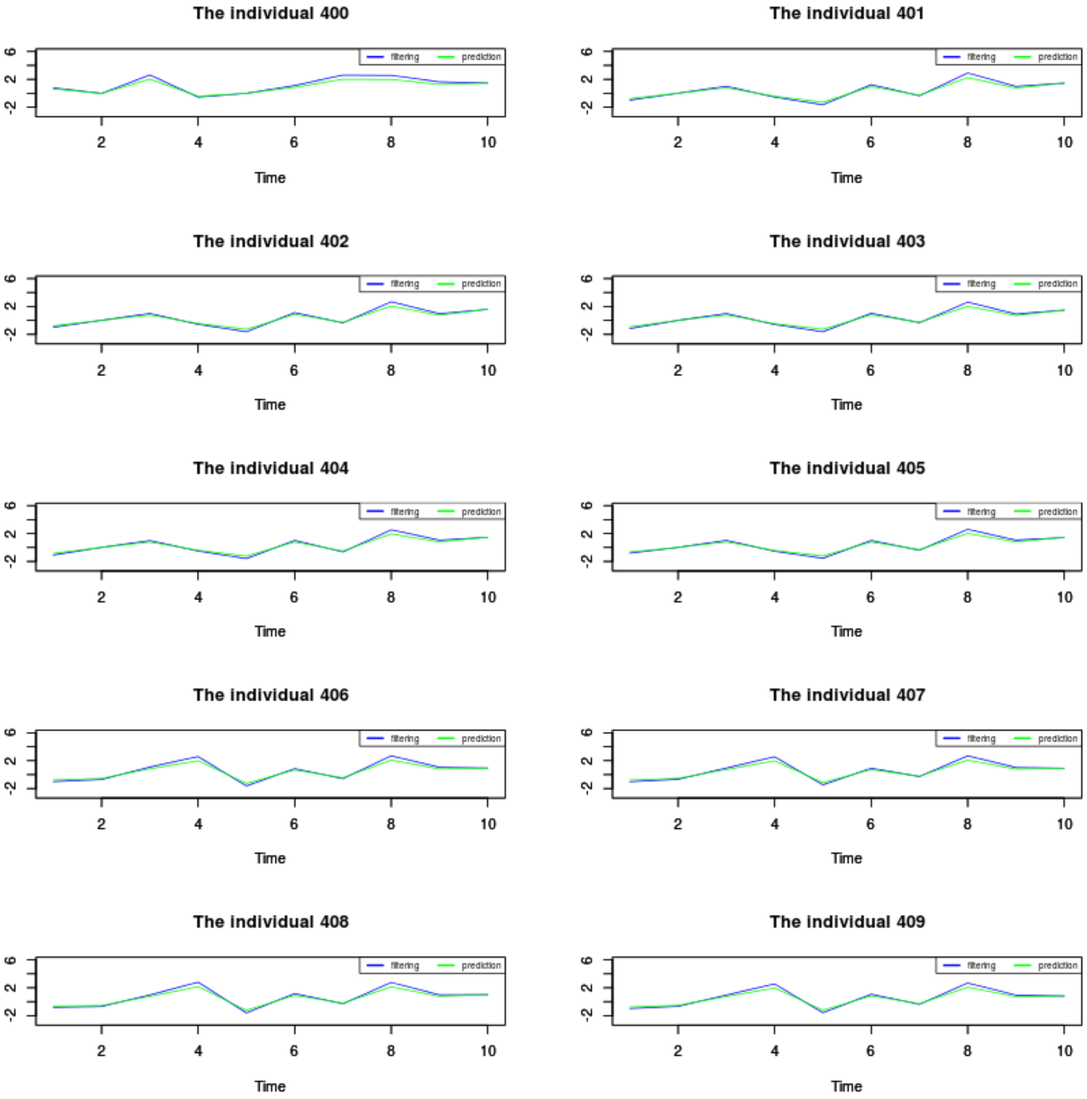


Figure 5.36: The graphs for the 400th individual to 409th individual with AR(1) model


The individual 400

The individual 401

The individual 402

The individual 403

The individual 404

The individual 405

The individual 406

The individual 407

The individual 408

The individual 409

5.3.4 CHARN(1,1) model

As in the preceding subsection, the outline of the analysis of data is performed with CHARN(1,1) as in equation (5.2.5). We set the initial values for β^0 and R_t^0 as in the AR(1) model, and set $\rho_1^0 = 0.6, \rho_2^0 = 0.4$.

Figures 5.37 - 5.41 display the filter and predictive steps for the first ten individuals of every hundred individuals.
Figures 5.37- 5.41 show that the curves of predictive and filter steps for each individual with CHARN(1,1) model are the same as the curves for AR(1) model. In other words, the effect of CHARN(1,1) is similar to that the AR(1). We can explain that the predictive and the filter values depend on the function $F = \rho X_i(t-1)$, which is supposed that it is the same in two model. The variance-covariance matrix for predictive, filter $P_{t|t-1}, P_{t|t}$ respectively also have values less than 0.001.
On Figures 5.37 - 5.41 too, positive values the rested patient and negative values the tired patient. Figure 5.37 shows the curves of prediction and filter steps for the 1st individual to the fifth individual are identical. We can also see that the individual 1 is tired at $T = 1$, and he feels rested at $T = 2, 3, 4$, and he returns tired at $T = 5$. Finally, he feels rested at $T = 6, 7, 8, 9, 10$.
The curves of prediction and filter steps for the sixth individual to the tenth individual are similar. We also see that the individual 6 is tired at $T = 1$, and he feels rested at $T = 2, 3, 4$. Then, he is tired at $T = 5$, but he feels rested at $T = 6, 7$, while at $T = 8$ he is tired. Eventually, at $T = 9, 10$ he feels rested.

Figure 5.37: The graphs for the 1st individual to 10th individual with CHARN(1,1) model

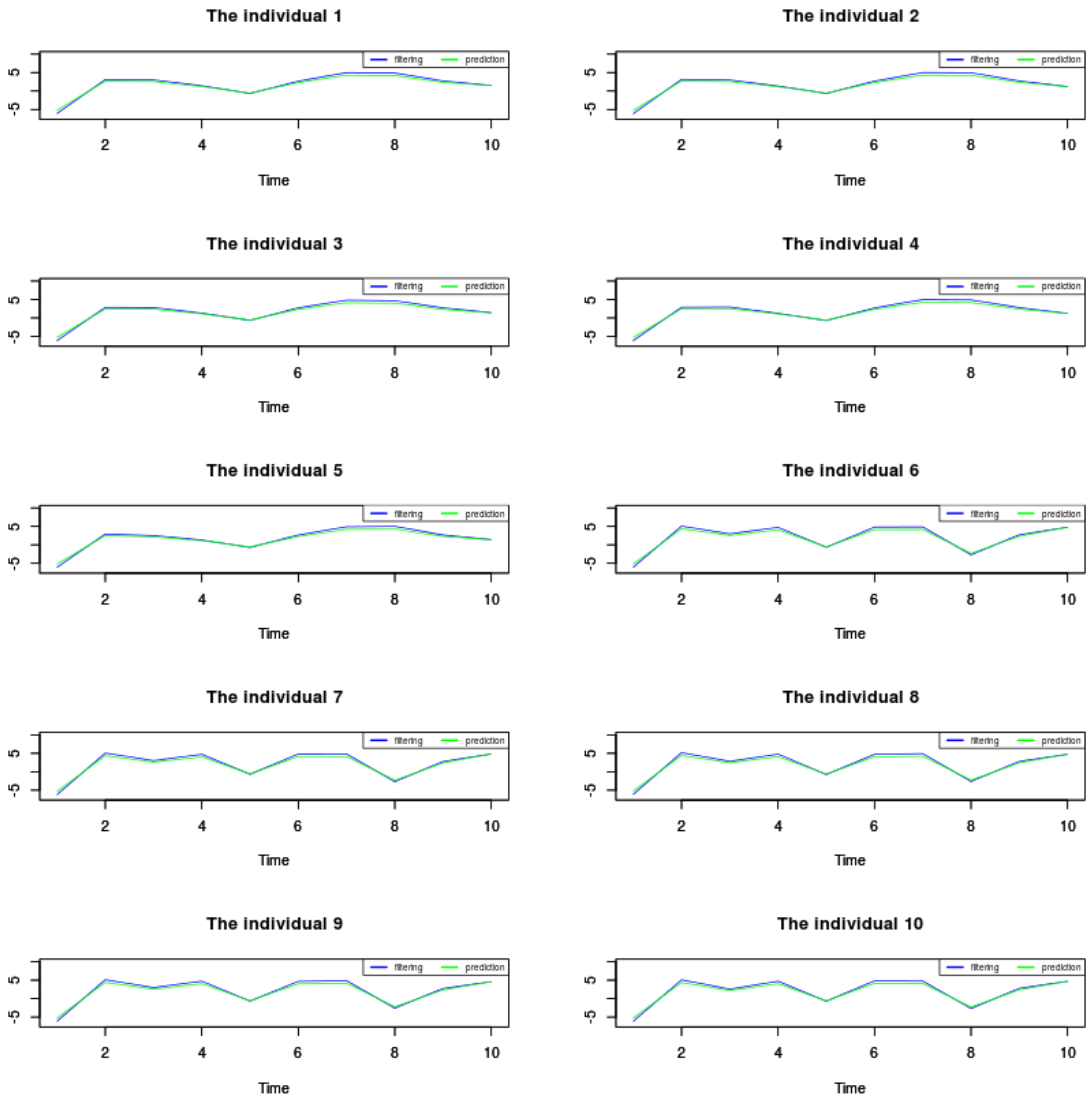
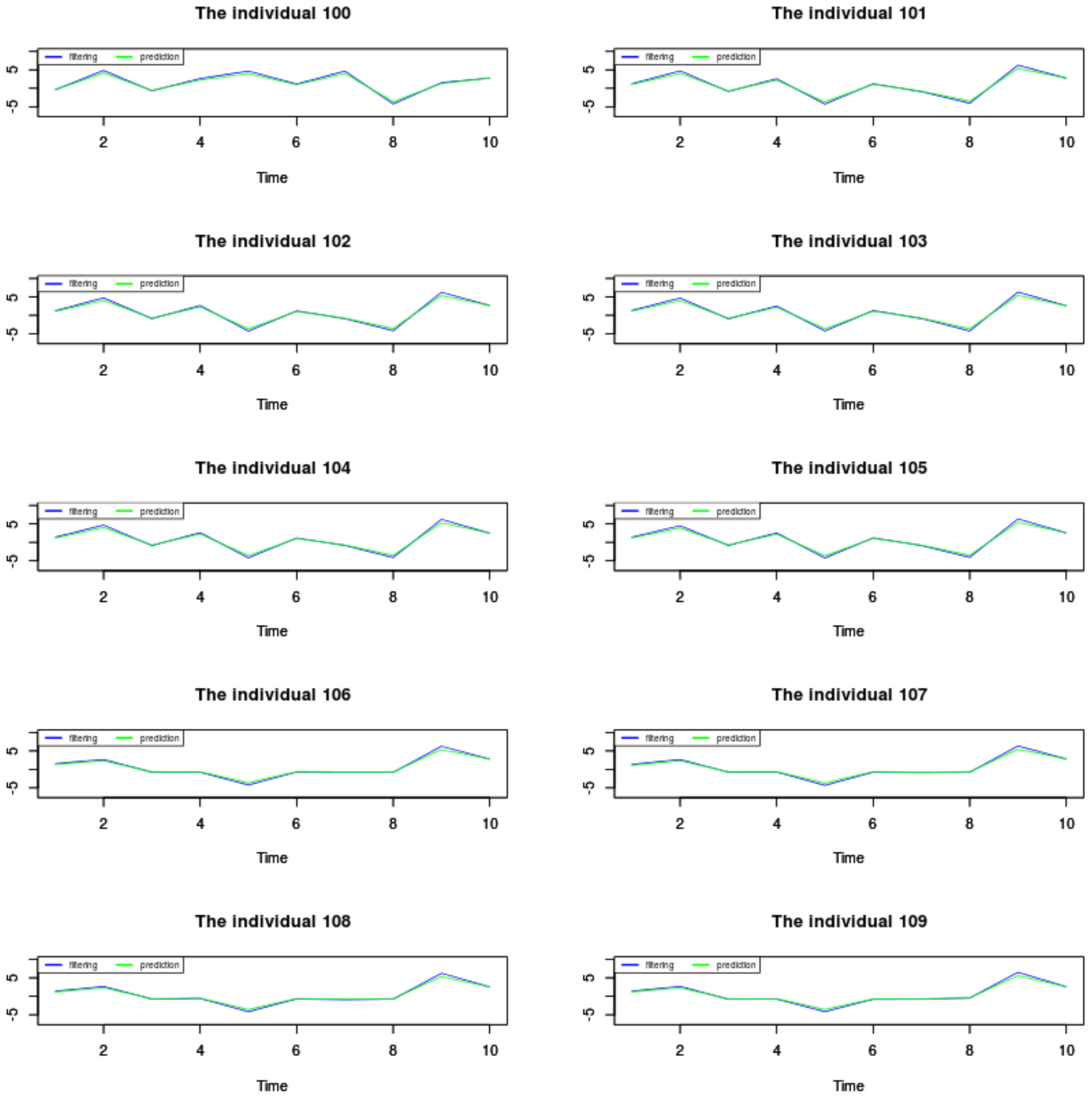


Figure 5.38: The graphs for the 100th individual to 109th individual with CHARN(1,1) model


The individual 100

The individual 101

The individual 102

The individual 103

The individual 104

The individual 105

The individual 106

The individual 107

The individual 108

The individual 109

Figure 5.39: The graphs for the 200th individual to 209th individual with CHARN(1,1) model

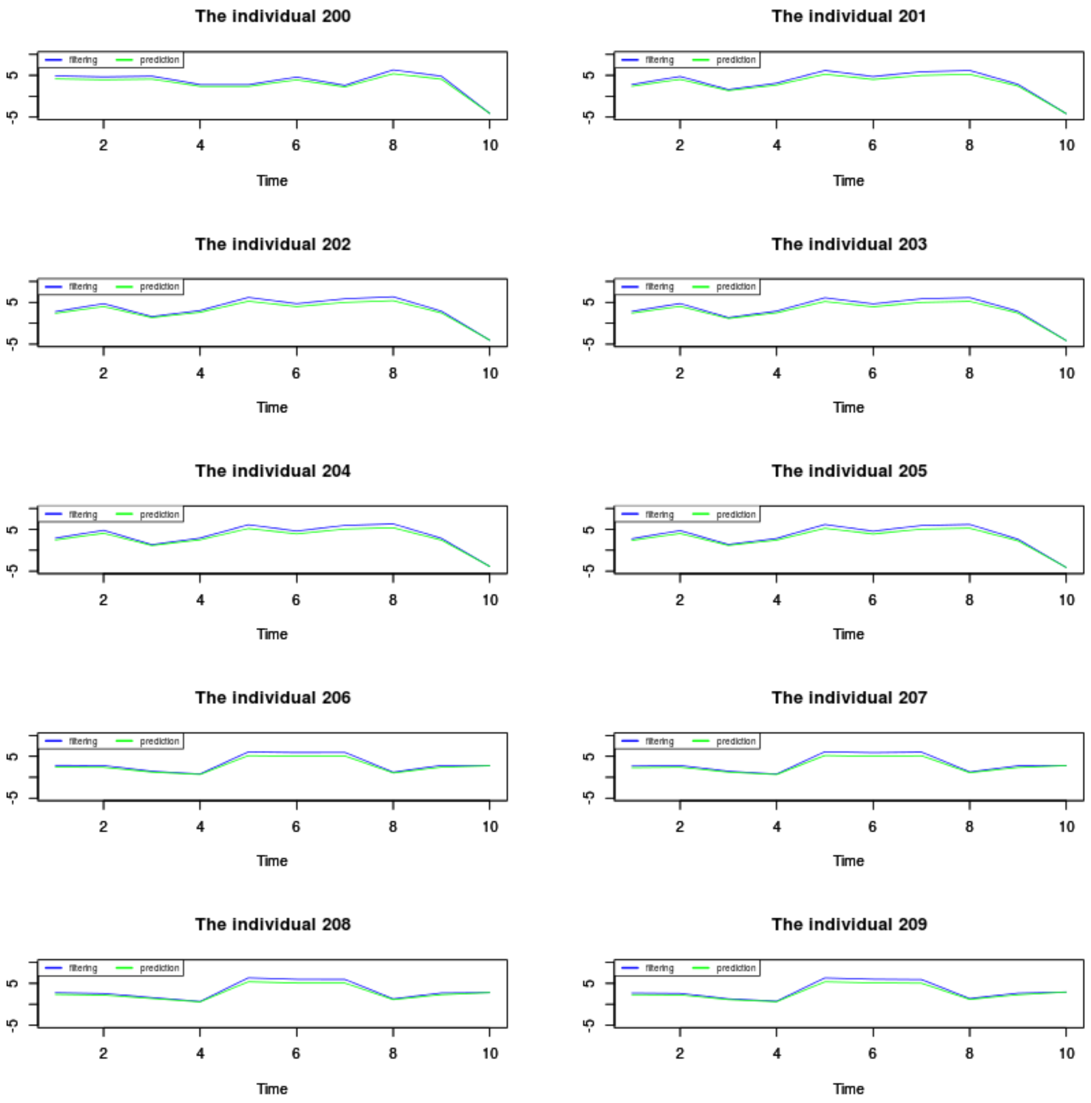
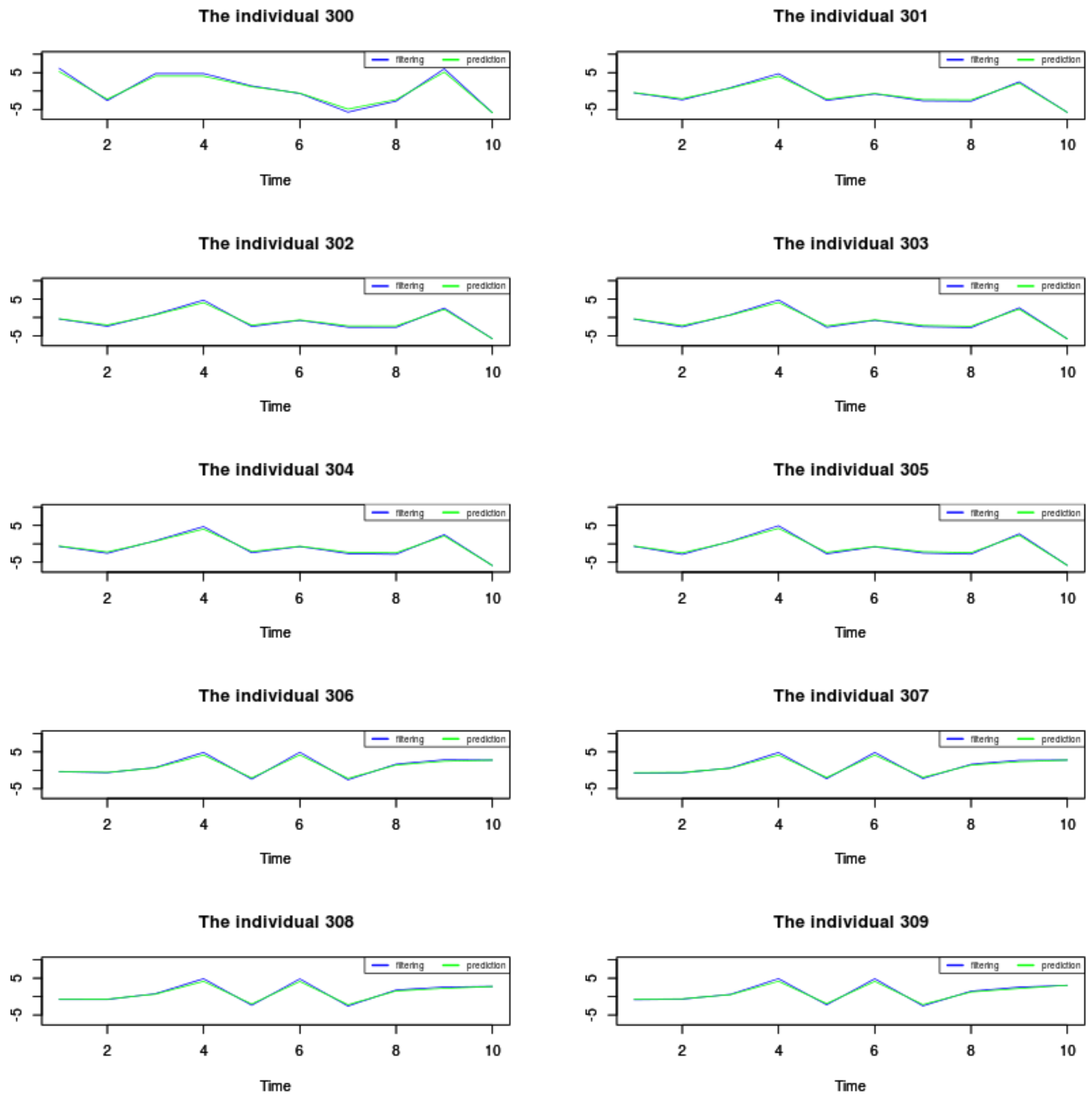


Figure 5.40: The graphs for the 300th individual to 309th individual with CHARN(1,1) model



The individual 300

The individual 301

The individual 302

The individual 303

The individual 304

The individual 305

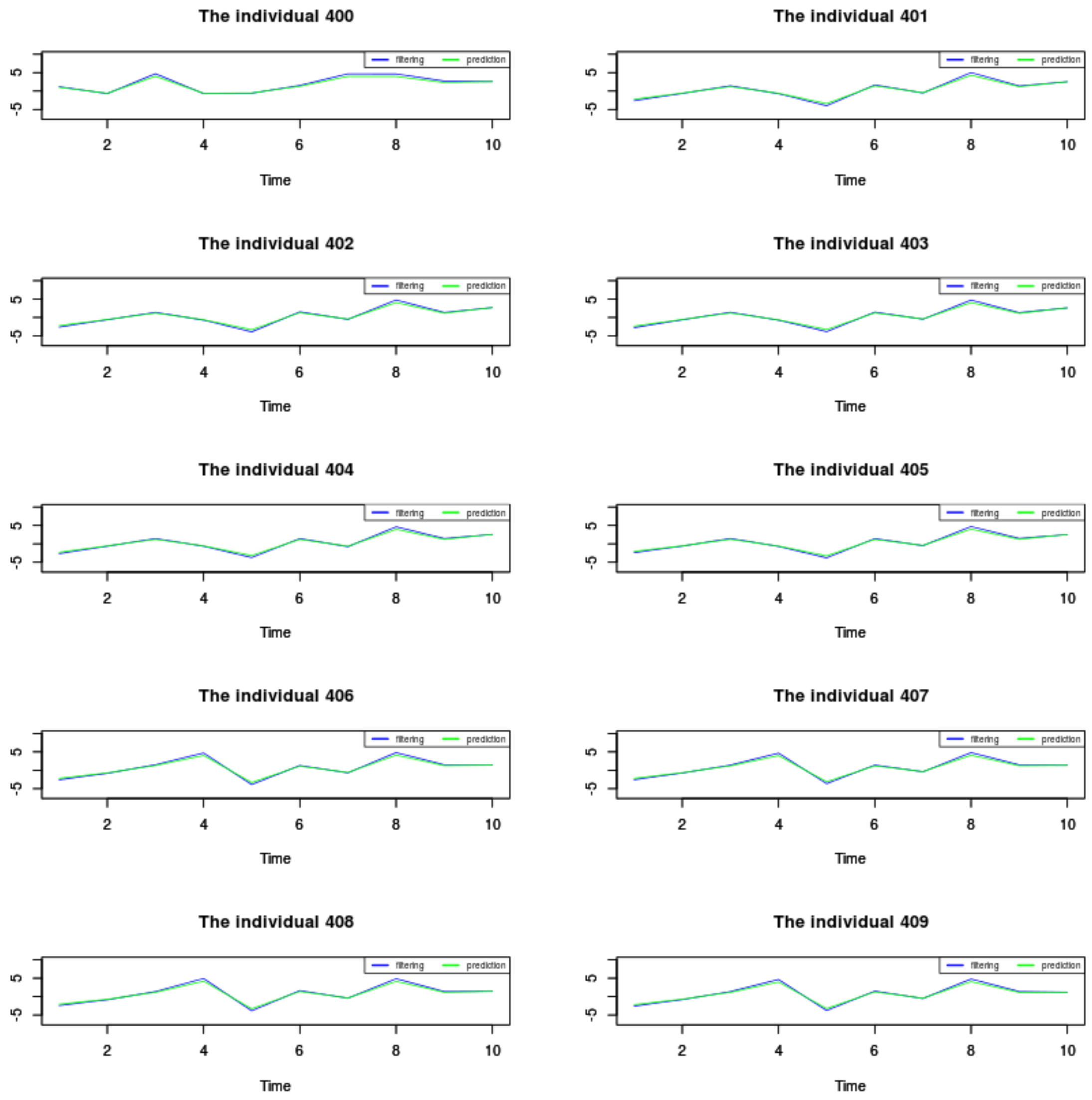
The individual 306

The individual 307

The individual 308

The individual 309

Figure 5.41: The graphs for the 400th individual to 409th individual with CHARN(1,1) model



The individual 400

The individual 401

The individual 402

The individual 403

The individual 404

The individual 405

The individual 406

The individual 407

The individual 408

The individual 409

Chapter 6

Conclusions and further studies

6.1 Introduction

This thesis focuses on the generalization of Kalman filter and smoother recursions with generalized state space models. We took a practical model, the observations are multi-categorical and longitudinal, and the state equation is described by a CHARN models with noise from exponential family distributions. This chapter presents the contributions of the study, conclusions and some recommendation for future studies.

6.2 Conclusion

In this thesis, we have investigated and discussed the generalization of Kalman filter and smoother recursions with generalized state space models. The main purpose of this study is to estimate the state by generalized Kalman filter recursions. This thesis also discusses the Kalman filter recursion with linear and nonlinear state space model, EM algorithm for estimating the model parameters. We propose two approaches in this study to estimating the state variables. The first, Working Extended Kalman Filtering Recursion (WEKFS). The second, the maximum a posteriori (MAP) via auxiliary Iterated Kalman particle filter.
To achieve this objective, we apply the working extended Kalman filtering recursions with various simulation studies have been designed. The simulation studies are designed with different characteristics such as size of samples (number of individuals and length of occasions), the form of the distribution of state noise (gaussian or exponential), and the type of state equation (AR(1) model, CHARN(1,1), and CHARN(0,1)). In addition to the simulation, an application to real data in the quality of life of patients surged for breast cancer. In order to evaluate our approach in this study, we present the figures that appear the state

variables which are simulated with their predictive and filter values. Another criteria we use the $P_{t|t-1}, P_{t|t}$ are the variance-covariance matrix to predictive and filter respectively, where each matrix have values less than 0.001 refer to the best estimator.

The first objective is to propose a new multicategorical longitudinal in quality of life, where we use the state space model approach. The observation equation is a conditional probability of $Y_{ik}(t)$ given the state variables $X_i(t)$ described by a multinomial distribution. The state equation is described by conditional heteroskedastic autoregressive nonlinear (CHARN) model.

The second objective of this study is to propose an alternative and generalisation of existing methods for estimating the latent trait in quality of life.

6.3 Areas of Future Studies

In this section, some future recommendations would be mentioned based on our current study, which can play an important role in the future research. We can summarize the suggested topic for future studies as follows.

- Our approach succeeds for estimating a latent variable in the health field. Consequently, our approach can be applied in the following fields: the economic field for estimating the business confidence or morale of customers, in the industrial field for estimating the level of anxiety due to the machines or robots on workers in factories.

- In chapter 5, we apply the first approach the working Kalman filter recursion through created the R-packages, we can create the R-package for the maximum a posteriori (MAP) via the Auxiliary Iterated Kalman Particle Filter (AIEKPF) method. One can compare between two approaches which are proposed.

- In real data analysis, we found the data Incomplete, where missed the covariates or the responses for the individuals. We found challenging to analyse this data. Therefore we ignored the individuals have missing data. One can develop our approach to estimating the latent trait for the Incomplete data.

- In our study, we proposed the link function is linear predictor $\eta^s_{ik}(t) = \mathbf{u}_i^\top(t)\beta^s_k + X_i(t)$. Hastie and Tibshirani (1990) introduced the class of generalized additive models which replaces the linear form by a sum of smooth function $\sum_j s_j(u_j)$, where $s_j(.)^\top$s are unspecified functions that are estimated using a scatterplot smoother. By the similar way one can be developed our model with $\eta^s_{ik}(t) = s_i(\mathbf{u}_i^\top(t)) + X_i(t)$.

Appendix A

Kalman Filter and Smother Recursions

A.1 Derivation of the Kalman filter and smoother recursions

In this appendix, a brief derivation of the Kalman filter (Kalman (1960)) and the fixed interval smoothing algorithm (Bierman (1977)) are presented.

A.1.1 One-step-Ahead prediction

At first we define $\mathbf{Y}_t = (Y_1, Y_2, \cdots, Y_t)^\top$. From the state equation $X_t = F_t X_{t-1} + H_t \varepsilon_t$, we obtain

$$\begin{aligned} X_{t|t-1} &= \mathrm{E}(X_t \mid \mathbf{Y}_{t-1}) \\ &= \mathrm{E}(F_t X_{t-1} + H_t \varepsilon_t \mid \mathbf{Y}_{t-1}) \\ &= F_t \mathrm{E}(X_{t-1} \mid \mathbf{Y}_{t-1}) \\ &= F_t X_{t-1|t-1}, \end{aligned} \tag{A.1.1}$$

$$\begin{aligned} V_{t|t-1} &= \mathrm{E}(X_t - X_{t|t-1})^2 \\ &= \mathrm{E}(F_t(X_{t-1} - X_{t-1|t-1} + H_t \varepsilon_t)^2 \\ &= F_t \mathrm{E}(X_{t-1} - X_{t-1|t-1})^2 F_t^\top + H_t \mathrm{E}(\varepsilon_t)^2 H_t^\top \\ &= F_t V_{t-1|t-1} F_t^\top + H_t R_t^2 H_t^\top, \end{aligned} \tag{A.1.2}$$

A.1.1.1 Filter

The prediction error of Y_t is denoted by η_t; then from $Y_t = G_t X_t + \xi_t$, can be written as

$$\begin{aligned}\eta_t &\equiv Y_t - \mathrm{E}(Y_t \mid \mathbf{Y}_{t-1}) \\ &= G_t X_t + \xi_t - \mathrm{E}(G_t X_t + \xi_t \mid \mathbf{Y}_{t-1}) \\ &= G_t X_t + \xi_t - G_t \mathrm{E}(X_t \mid \mathbf{Y}_{t-1}) \\ &= G_t(X_t - X_{t|t-1}) + \xi_t \quad . \qquad \text{(A.1.3)}\end{aligned}$$

Therefore, obtain

$$\begin{aligned}\mathrm{Var}(\eta_t) &= G_t V_{t|t-1} G_t^\top + \Sigma_t \qquad \text{(A.1.4)} \\ \mathrm{Cov}(X_t, \eta_t) &= \mathrm{Cov}(X_t, G_t(X_t - X_{t|t-1}) + \xi_t) \\ &= \mathrm{Var}(X_t - X_{t|t-1}) G_t^\top \\ &= V_{t|t-1} G_t^\top . \qquad \text{(A.1.5)}\end{aligned}$$

By using the facts that $\mathbf{Y}_t = \{\mathbf{Y}_{t-1}, Y_t\} = \mathbf{Y}_{t-1} \oplus \eta_t \equiv \mathbf{Y}_{t-1} + \eta_t$, where $\oplus$ refer to a direct sum, in the case of the normal distribution, the conditional expectation of X_t given Y_t is expressible by orthogonal projection as follows

$$\begin{aligned}X_{t|t} = \mathrm{E}(X_t \mid \mathbf{Y}_t) &= \mathrm{Proj}(X_t \mid \mathbf{Y}_t) \\ &= \mathrm{Proj}(X_t \mid \mathbf{Y}_{t-1}, \eta_t) \\ &= \mathrm{Proj}(X_t \mid \mathbf{Y}_{t-1}) + \mathrm{Proj}(X_t \mid \eta_t). \qquad \text{(A.1.6)}\end{aligned}$$

Due to $\mathrm{Proj}(X_t \mid \eta_t)$ is obtained by fall back X_t on η_t, from (A.1.4) and (A.1.5),

$$\begin{aligned}\mathrm{Proj}(X_t \mid \eta_t) &= \mathrm{Cov}(X_t, \eta_t)\mathrm{Var}(\eta_t)^{-1}\eta_t \\ &= V_{t|t-1} G_t^\top (G_t V_{t|t-1} G_t^\top + \Sigma_t)^{-1} \eta_t \\ &= K_t \eta_t . \qquad \text{(A.1.7)}\end{aligned}$$

Therefore, we express

$$X_{t|t} = X_{t|t-1} + K_t \eta_t . \qquad \text{(A.1.8)}$$

In contrast, from

$$\begin{aligned}V_{t|t-1} &= \mathrm{E}(X_t - X_{t|t-1})^2 \\ &= \mathrm{E}(X_t - X_{t|t} + K_t \eta_t)^2 \\ &= V_{t|t} + K_t \mathrm{Var}(\eta_t) K_t^\top , \qquad \text{(A.1.9)}\end{aligned}$$

we obtain

$$\begin{aligned}V_{t|t} &= V_{t|t-1} - K_t G_t V_{t|t-1} \\ &= (I - K_t G_t) V_{t|t-1} . \qquad \text{(A.1.10)}\end{aligned}$$

A.1.2 Smoothing

$\psi_{t+1} \equiv X_{t+1} - X_{t+1|t}$ is supposed to be the prediction error of X_{t+1}. Define $\mathbf{Z}_t$ by

$$\mathbf{Z}_t \equiv \mathbf{Y}_t \oplus \psi_{t+1} \oplus \{\xi_{t+1}, \cdots, \xi_T, \varepsilon_{t+1}, \varepsilon_{t+1}, \cdots, \varepsilon_T\}.$$

Then, the decomposition as follows

$$\begin{aligned} Z_t &\equiv \mathrm{Proj}(X_t \mid \mathbf{Z}_t) \\ &= \mathrm{Proj}(X_t \mid \mathbf{Y}_t) + \mathrm{Proj}(X_t \mid \psi_{t+1}) + \mathrm{Proj}(X_t \mid \xi_{t+1}, \cdots, \xi_T, \varepsilon_{t+1}, \cdots, \varepsilon_T), \end{aligned} \tag{A.1.12}$$

Consequently, we find that

$$\begin{aligned} \mathrm{Proj}(X_t \mid \mathbf{Y}_t) &= X_{t|t} \\ \mathrm{Proj}(X_t \mid \psi_{t+1}) &= \mathrm{Cov}(X_t, \psi_{t+1})\mathrm{Var}(\psi_{t+1})^{-1}\psi_{t+1} \\ \mathrm{Proj}(X_t &\mid \xi_{t+1},, \xi_T, \varepsilon_{t+1}, \cdots, \varepsilon_T) = 0. \end{aligned} \tag{A.1.13}$$

Furthermore, we have

$$\begin{aligned} \mathrm{Var}(\psi_{t+1}) &= V_{t+1|t}, \\ \mathrm{Cov}(X_t, \psi_{t+1}) &= \mathrm{Cov}(X_t, F_{t+1}(X_t - X_{t|t}) + H_{t+1}\varepsilon_{t+1}) \\ &= \mathrm{E}(X_t - X_{t|t})^2 F_{t+1}^\top \\ &= V_{t|t}F_{t+1}^\top. \end{aligned} \tag{A.1.14}$$

Therefore, by putting

$$A_t = V_{t|t}F_{t+1}^\top V_{t+1|t}^{-1},$$

can be obtained

$$Z_t = X_{t|t} + A_t(X_{t+1} - X_{t+1|t}). \tag{A.1.15}$$

Here, considering that $\mathbf{Z}_T$ produces $\mathbf{Y}_T$, , Obtains

$$\begin{aligned} X_{t|T} &= \mathrm{Proj}(X_t \mid \mathbf{Y}_t) \\ &= \mathrm{Proj}(\mathrm{Proj}(X_t \mid \mathbf{Z}_T) \mid \mathbf{Y}_T) \\ &= \mathrm{Proj}(X_t \mid \mathbf{Y}_t) \\ &= X_{t|t} + A_t(X_{t+1|T} - X_{t+1|t}). \end{aligned} \tag{A.1.16}$$

Furthermore, using

$$X_t - X_{t|T} + A_tX_{t+1|T} = X_t - X_{t|t} + A_tX_{t+1|t}, \tag{A.1.17}$$

and

$$\mathrm{E}\left\{(X_t - X_{t|T})X_{t+1|T}^\top\right\} = \mathrm{E}\left\{(X_t - X_{t|t})X_{t+1|t}^\top\right\} = 0,$$

obtains

$$V_{t|T} + A_t \mathrm{E}\left\{X_{t+1|T}X_{t+1|T}^\top\right\} A_t^\top = V_{t|t} + A_t \mathrm{E}\left\{X_{t+1|t}X_{t+1|t}^\top\right\} A_t^\top. \qquad \text{(A.1.18)}$$

Here, using

$$\begin{aligned} \mathrm{E}\left\{(X_t - X_{t|T})X_{t+1|T}^\top\right\} &= 0 \\ \mathrm{E}\left\{(X_t - X_{t|t})X_{t+1|t}^\top\right\} &= 0, \end{aligned}$$

yields

$$\begin{aligned} &\mathrm{E}\left\{X_{t+1|T}X_{t+1|T}^\top\right\} \\ &\quad = \mathrm{E}\left\{(X_{t+1|T} - X_{t+1} + X_{t+1})(X_{t+1|T} - X_{t+1} + X_{t+1})^\top\right\} \\ &\quad = V_{t+1|T} + \mathrm{E}\{X_{t+1}X_{t+1}^\top\} + 2\mathrm{E}\{(X_{t+1|T} - X_{t+1})X_{t+1}^\top\} \\ &\quad = V_{t+1|T} + \mathrm{E}\{X_{t+1}X_{t+1}^\top\} - 2\mathrm{E}\{(X_{t+1|T} - X_{t+1})(X_{t+1|T} - X_{t+1})^\top\} \\ &\quad = \mathrm{E}\{X_{t+1}X_{t+1}^\top\} - V_{t+1|T}, \qquad \text{(A.1.19)} \end{aligned}$$

and

$$\mathrm{E}\{X_{t+1|t}X_{t+1|t}^\top\} = \mathrm{E}\{X_{t+1}X_{t+1}^\top\} - V_{t+1|t}. \qquad \text{(A.1.20)}$$

Substituting this into (A.1.18) yields

$$V_{t|T} = V_{t|t} + A_t(V_{t+1|T} - V_{t+1|t})A_t^\top. \qquad \text{(A.1.21)}$$

A.2 The derivation of the Extended Kalman filter Recursions

Let the observation process $Y_t : Y_t \in R^\ell$. Then the observation equation as follows:

$$Y_t = G(\mathbf{u}_t, X_t, \lambda) + \xi_t. \qquad \text{(A.2.1)}$$

The state process $X_t : X_t \in R^k$, the state equation as follows:

$$X_t = F(\mathbf{u}_t, X_{t-1}, \gamma) + H(\mathbf{u}_t, X_{t-1}, \delta) \cdot \varepsilon_t. \qquad \text{(A.2.2)}$$

Initially, the information certainly available is the mean, X_0^a, and the covariance, V_0 of the initial state, then the initial optimal estimate X_0^a and the covariance of error V_0^a as follows

$$\begin{aligned} x_0^a &= \mathrm{E}[X_0] \qquad &\text{(A.2.3)} \\ V_0^a &= \mathrm{E}[(X_0 - X_0^a)(X_0 - X_0^a)^\top] \qquad &\text{(A.2.4)} \end{aligned}$$

Prediction suppose an optimal estimate is

$$x_{t-1}^a \equiv \mathrm{E}[X_{t-1} \mid \mathbf{Y}_{t-1}]$$

with V_{t-1} covariance at time $t-1$. The predictable part of X_t is written as:

$$\begin{aligned} x_t^f &\equiv \mathrm{E}[X_t \mid \mathbf{Y}_{t-1}] \qquad (A.2.5) \\ &= \mathrm{E}[F(X_{t-1}, \mathbf{u}_t, \gamma) + H(X_{t-1}, \mathbf{u}_t; \delta).\varepsilon_t \mid \mathbf{Y}_{t-1}] \\ &= E[F(X_{t-1}, \mathbf{u}_t, \gamma) \mid \mathbf{Y}_{t-1}] \end{aligned}$$

By using first-order Taylor expansion can be linearized the functions in the equation (A.2.2)

(a) Expanding $F(X_{t-1}, \mathbf{u}_t, \gamma)$ in Taylor series about X_{t-1}^a

$$F(X_{t-1}, \mathbf{u}_t, \gamma) = F(x_{t-1}^a, \mathbf{u}_t, \gamma) + A_t(X_{t-1} - x_{t-1}^a) + H.O.T \qquad (A.2.6)$$

where

$$A_t = \frac{\partial F(x, \mathbf{u}_t, \gamma)}{\partial x} \mid_{x=X_{t-1}^a}$$

and (H.O.T) the higher order expressions are considered negligible.

(b) For a function that depends on two variables, x and y, first -order Taylor series about the point (a, b) as follows

$$f(x,y) = f(a,b) + \frac{\partial f(x,y)}{\partial x} \mid_{x=a} (x-a) + \frac{\partial f(x,y)}{\partial y} \mid_{y=b} (y-b) + H.O.T \qquad (A.2.7)$$

Now we Expanding $H(X_{t-1}, \mathbf{u}_t, \delta)\varepsilon_t$, about two points $(x_{t-1}^a, \hat{\varepsilon}_t)$ by using (A.2.7), where

$$\varepsilon_t \sim \mathcal{N}(0, R_t).$$

and

$$\hat{\varepsilon}_t = \mathrm{E}[\varepsilon_t] = 0$$

Then, obtains

$$\begin{aligned} H(X_{t-1}, \mathbf{u}_t, \delta)\varepsilon_t &= H(X_t^a, \mathbf{u}_t, \delta) \underbrace{\hat{\varepsilon}_t}_{=0} \\ &+ \frac{\partial H(x, \mathbf{u}_t, \delta)}{\partial x} \mid_{x=x_{t-1}^a} \underbrace{\hat{\varepsilon}_t}_{=0} (X_{t-1} - X_{t-1}^a)) \\ &+ H(X_{t-1}^a, \mathbf{u}_t, \delta)(\varepsilon_t - 0) + \underbrace{H.O.T}_{=0} \\ &= C_t \varepsilon_t \qquad (A.2.8) \end{aligned}$$

where

$$C_t = H(X_{t-1}^a, \mathbf{u}_t, \delta)$$

since $e_{t-1}^a \equiv X_{t-1} - X_{t-1}^a$, then

$$\mathrm{E}[F(X_{t-1}, \mathbf{u}_t, \gamma) \mid \mathbf{Y}_{t-1}] = F(X_{t-1}^a, \gamma) + A_t E[e_{t-1} \mid \mathbf{Y}_{t-1}] \tag{A.2.9}$$

and $\mathrm{E}[e_{t-1} \mid \mathbf{Y}_{t-1}] = 0$. Consequently, the forecast value of X_t yields

$$x_t^f \approx F(X_{t-1}^a, \mathbf{u}_t, \gamma) \tag{A.2.10}$$

Substituting (A.2.6) and (A.2.8) in the forecast error equation

$$\begin{aligned} e_t^f &\equiv X_t - X_t^f \\ &= F(X_{t-1}, \mathbf{u}_t, \gamma) + H(X_{t-1}, \mathbf{u}_t, \delta)\varepsilon_t \\ &\quad -F(X_{t-1}^a, \mathbf{u}_t, \gamma) \\ &\approx F(X_{t-1}^a, \mathbf{u}_t, \gamma) + A_t e_{t-1} + C_t \varepsilon_t \\ &\quad -F(X_{t-1}^a t, \mathbf{u}_t, \gamma) \\ &\approx A_t e_{t-1}^a + C_t \varepsilon_t. \end{aligned} \tag{A.2.11}$$

The forecast error covariance is specified by

$$\begin{aligned} V_t^f &= \mathrm{E}[e_t^f (e_t^f)^\top] \\ &= A_t \mathrm{E}[e_{t-1}^a (e_{t-1}^a)^\top] A_t^\top + C_t \mathrm{E}[\varepsilon_t \varepsilon_t^\top] C_t^\top \\ &= A_t^\top V_{t-1} A_t^\top + C_t R C_t^\top. \end{aligned} \tag{A.2.12}$$

Filtering At time t have two value of input: the forecast value x_t^f with the covariance V_t^f and the observation Y_t with conditional covariance Σ_t. Our objective is to compute the best unbiased estimate. In the least squares concept, x_t^a of X_t, one way is to suppose the estimate is a linear combination of both x_t^f and Y_t as follows

$$x_t^a = a + K_t Y_t. \tag{A.2.13}$$

From the unbiasedness definition

$$\begin{aligned} 0 &= \mathrm{E}[X_t - x_t^a \mid \mathbf{Y}_t] \\ &= \mathrm{E}[(x_t^f + e_t^f) - (a + K_t Y_t) \mid \mathbf{Y}_t] \\ &= \mathrm{E}[(x_t^f + e_t^f) - (a + K_t G(X_t, \mathbf{u}_t, \lambda) + K_t \xi_t) \mid \mathbf{Y}_t] \\ &= x_t^f - a - K_t E[G(X_t, \mathbf{u}_t, \lambda)] + K_t \underbrace{E(\upsilon_t)}_{=0} \mid \mathbf{Y}_t] \\ a &= x_t^f - K_t E[G(X_t, \mathbf{u}_t, \lambda)] \mid \mathbf{Y}_t] \end{aligned} \tag{A.2.14}$$

Substituting (A.2.14) in (A.2.13):

$$x_t^a = x_t^f + K_t(Y_t - E[G(X_t, \mathbf{u}_t, \lambda)]) \tag{A.2.15}$$

As in model forecast step, expanding $G(X_t, \mathbf{u}_t, \lambda)$ in Taylor series about x_t^f obtains

$$G(X_t, \mathbf{u}_t, \lambda) = G(x_t^f, \mathbf{u}_t, \lambda) + B_t(X_t - x_t^f) + H.O.T \tag{A.2.16}$$

where

$$B_t = \frac{\partial G(x, \mathbf{u}_t, \lambda)}{\partial x} \mid_{x=x_t^f}$$

Then the expectation on both sides of (A.2.16) given $\mathbf{Y}_t$

$$\mathrm{E}[G(X_t, \mathbf{u}_t, \lambda) \mid \mathbf{Y}_t] \approx G(x_t^f, \mathbf{u}_t, \lambda) + B_t \mathrm{E}[e_t^f \mid \mathbf{Y}_t] \tag{A.2.17}$$

where $\mathrm{E}[e_t^f \mid \mathbf{Y}_t] = 0$. Substituting (A.2.17) in (A.2.15), the state estimate (posterior mode) yields:

$$x_t^a \approx x_t^f + K_t[Y_t - G(x_t^f, \mathbf{u}_t, \lambda)] \tag{A.2.18}$$

The error in the estimate x_t^a yields:

$$\begin{aligned} e_t^a &\equiv X_t - x_t^a \\ &= [F(X_{t-1}, \mathbf{u}_t, \gamma) + H(X_{t-1}, \mathbf{u}_t, \delta).\varepsilon_t] \\ &\quad -x_t^f - K_t[Y_t - G(x_t^f, \mathbf{u}_t, \lambda)] \\ &= [F(X_{t-1}, \mathbf{u}_t, \gamma) + A_t e_{t-1}^a + C_t \varepsilon_t] \\ &\quad -F(x_{t-1}^a, \mathbf{u}_t, \gamma) - K_t[Y_t - G(x_t^f, \mathbf{u}_t, \lambda) + \xi_t] \\ &\approx A_t e_{t-1}^a + C_t \varepsilon_t - K_t(B_t e_t^f + \xi_t) \\ &\approx A_t e_{t-1}^a + C_t \varepsilon_t - K_t B_t (A_t e_{t-1}^a (t-1) \\ &\quad +C_t \varepsilon_t) - K_t \xi_t \\ &\approx (I - K_t B_t) A_t e_{t-1}^a \\ &\quad +(I - K_t B_t) C_t \varepsilon_t) - K_t \xi_t \end{aligned} \tag{A.2.19}$$

Then, the posterior covariance as follows

$$\begin{aligned} V_t^a &\equiv \mathrm{E}[e_t^a e_t^{a\top}] \\ &= (I - K_t B_t) A_t V_{t-1} A_t^\top (I - K_t B_t)^\top \\ &\quad +(I - K_t B_t) C_t R C_t^\top (I - K_t B_t)^\top \\ &\quad +K_t \Sigma_t K_t^\top \end{aligned} \tag{A.2.20}$$

By using equation (A.2.12) obtains

$$\begin{aligned} V_t^a &= (I - K_t B_t) V_t^f (I - K_t B_t)^\top + K_t \Sigma_t K_t^\top \\ &= V_t^f - K_t B_t V_t^f - V_t^f B_t^\top K_t^\top \\ &\quad + K_t B_t V_t^f B_t^\top K_t^\top + K_t \Sigma_t K_t^\top \end{aligned} \tag{A.2.21}$$

Such as the standard Kalman filter, the posterior covariance formula for any K_t can be computed by minimizing $tr(V_t^a)$ w.r.t K_t [1].

$$\begin{aligned} 0 &= \frac{\partial tr(V_t^a)}{\partial K_t} \\ &= -(B_t V_t^f)^\top - V_t^f B_t^\top \\ &\quad + 2 K_t B_t V_t^f B_t^\top + 2 K_t \Sigma_t \end{aligned} \tag{A.2.22}$$

where the Kalman gain is:

$$K_t = V_t^f B_t^\top [B_t V_t^f B_t + \Sigma_t]^{-1} \tag{A.2.23}$$

Substituting this back in (A.2.20), obtains

$$\begin{aligned} V_t^a &= (I - K_t B_t) V_t^f - (I - K_t B_t) V_t^f B_t^\top K_t^\top \\ &\quad + K_t \Sigma_t K_t^\top \\ &= (I - K_t B_t) V_t^f \\ &\quad - \left[V_t^f B_t^\top - K_t B_t) V_t^f B_t^\top - K_t \Sigma_t \right] K_t^\top \\ &= (I - K_t B_t) V_t^f \\ &\quad - [V_t^f B_t^\top - K_t \{ B_t V_t^f B_t^\top + \Sigma_t \}] K_t^\top \\ &= (I - K_t B_t) V_t^f - \underbrace{[V_t^f B_t - V_t^f B_t] K_t}_{=0} \\ &= (I - K_t B_t) V_t^f \end{aligned} \tag{A.2.24}$$

Summary

Initialization x_0^a with error covariance V_0^a

Predictor

$$X_t^f \approx F(x_{t-1}^a, \mathbf{u}_t, \gamma) \tag{A.2.25}$$

$$V_t^f = A_t V_{t-1} A_t^\top + C_t R C_t^\top \tag{A.2.26}$$

[1] The aim to minimizing the expected value of the square of this vector $E \parallel X_t - \hat{x}_{t|t} \parallel^2$, this equivalent to minimizing the trace of the posterior covariance matrix V_t^a, which is minimized when its matrix derivative with respect to the gain matrix is zero.

Filtering

$$\begin{aligned} x_t^a &\approx x_t^f + K_t[Y_t - G(x_t^f, \mathbf{u}_t, \lambda)] & \text{(A.2.27)} \\ K_t &= V_t^f B_t^\top [B_t V_t^f B_t + \Sigma_t]^{-1} & \text{(A.2.28)} \\ V_t^a &= (I - K_{Bt}) V_t^f & \text{(A.2.29)} \end{aligned}$$

A.3 The Monte Carlo Filter

This Appendix presented a derivation of the Monte Carlo filter algorithm . The readers are indicated to Kitagawa (1996) and Doucet *et al.* (2001) for details.

A.3.1 One-step-Ahead Prediction

Suppose that m particles $\{f_{t-1}^{(1)}, \cdots, f_{t-1}^{(m)}\}$ that can be considered as m independent realizations from the conditional density $p(X_{t-1} \mid \mathbf{Y}_{t-1})$ of the state X_{t-1}, and m particles $\{\varepsilon_t^{(1)}, \cdots, \varepsilon_t^{(m)}\}$ from the system noise distribution $p(\varepsilon)$ are known. Namely, assume the following

$$f_{t-1}^{(i)} \sim p(X_{t-1} \mid \mathbf{Y}_{t-1}), \quad \varepsilon_t^{(i)} \sim q(\varepsilon). \tag{A.3.1}$$

Then, the one-step-ahead predictive distribution of X_t is written as

$$\begin{aligned} p(X_{t-1} \mid \mathbf{Y}_{t-1}) &= \int\int p(X_t, X_{t-1}, \varepsilon_t \mid \mathbf{Y}_{t-1}) dX_{t-1} d\varepsilon_t & \text{(A.3.2)} \\ &= \int\int p(X_t \mid X_{t-1}, \varepsilon_t, \mathbf{Y}_{t-1}) p(\varepsilon_t \mid X_{t-1}, \mathbf{Y}_{t-1}) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} d\varepsilon_t. \end{aligned}$$

where the system noise ε_t is independent of previous states and observations, the conditional distribution of the system noise yields

$$p(\varepsilon_t \mid X_{t-1}, \mathbf{Y}_{t-1}) = p(\varepsilon_t).$$

In contrast , since X_t depends only on X_{t-1} and ε_t,

$$p(X_t \mid X_{t-1}, \varepsilon_t, \mathbf{Y}_{t-1}) = p(X_t \mid X_{t-1}, \varepsilon_t) = \delta(X_t - F(X_{t-1}, \varepsilon_t)),$$

where $\delta(.)$ indicates the Dirac delta function, obtains

$$p(X_t \mid \mathbf{Y}_{t-1}) = \int\int \delta(X_t - F(X_{t-1}, \varepsilon_t)) p(\varepsilon_t) p(X_{t-1} \mid \mathbf{Y}_{t-1}) dX_{t-1} d\varepsilon_t. \tag{A.3.3}$$

Therefore, when realizations $\{\varepsilon_t^{(j)}\}$ of $p(\varepsilon_t)$ and $\{f_{t-1}^{(j)}\}$ of $p(X_t \mid \mathbf{Y}_{t-1})$ are obtained by

$$p_t^{(j)} = F(f_{t-1}^{(j)}, \mathbf{u}_t^{(j)}, \gamma) + H(f_{t-1}^{(j)}, \mathbf{u}_t^{(j)}, \delta) \varepsilon_t^{(j)}, \tag{A.3.4}$$

A.3.1.1 Filter

If $p_t^{(1)}, \cdots, p_t^{(m)}$ are given, which are m independent realizations of the distribution $p(X_t \mid \mathbf{Y}_{t-1})$, it is equivalent to approximation of the distribution $p(X_t \mid \mathbf{Y}_{t-1})$ by the empirical distribution function

$$P_t(X) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{I}(X, p_t^{(i)}), \tag{A.3.5}$$

where $\mathbf{I}(X, a) = 0$ for $X < 0$ and $\mathbf{I}(X, a) = 1$ otherwise. In other words, that the predictive distribution $p(X_t \mid \mathbf{Y}_{t-1})$ is approximated by the probability function

$$\Pr(X_t = p_t^{(j)} \mid \mathbf{Y}_{t-1}) = \frac{1}{m}, \quad \text{for} \quad j = 1, \cdots, m. \tag{A.3.6}$$

Then, the observation y_t is given, the posterior distribution of X_t is computed by

$$\begin{aligned} \Pr(X_t = p_t^{(j)} \mid \mathbf{Y}_{t-1}) &= \Pr(X_t = p_t^{(j)} \mid \mathbf{Y}_{t-1}, y_t) \\ &= \lim_{\triangle y \to 0} \frac{\Pr(X_t = p_t^{(j)}, y_t \le y \le y_t + \triangle\, y \mid \mathbf{Y}_{t-1})}{\Pr(y_t \le y \le y_t + \triangle\, y \mid \mathbf{Y}_{t-1})} \\ &= \frac{p(y_t \mid p_t^{(j)}) \Pr(X_t = p_t^{(j)} \mid \mathbf{Y}_{t-1})}{\sum_{i=1}^{m} p(y_t \mid p_t^{(i)}) \Pr(X_t = p_t^{(i)} \mid \mathbf{Y}_{t-1})} \\ &= \frac{\alpha_t^{(j)} \cdot \frac{1}{m}}{\sum_{i=1}^{m} \alpha_t^{(i)} \cdot \frac{1}{m}} = \frac{\alpha_t^{(j)}}{\sum_{i=1}^{m} \alpha_t^{(i)}}. \end{aligned} \tag{A.3.7}$$

The cumulative distribution function yields

$$\frac{1}{\sum_{i=1}^{m} \alpha_t^{(i)}} \sum_{i=1}^{m} \alpha_t^{(i)} \mathbf{I}(X, p_t^{(i)}), \tag{A.3.8}$$

By using the formula (A.3.8), the approximation of the distribution of the filter can be obtained . We can facilely to re-approximate it by m particles $f_t^{(1)}, \cdots, f_t^{(m)}$ with equal weights to perform the calculation of the prediction (A.3.1) in the next time step. This corresponds to appearing the distribution of (A.3.8) by the empirical distribution function

$$\frac{1}{m} \sum_{i=1}^{m} \mathbf{I}(X, f_t^{(i)}). \tag{A.3.9}$$

The m realizations $\{f_t^{(1)}, \cdots, f_t^{(m)}\}$ can be obtained by re-sampling $\{p_t^{(1)}, \cdots, p_t^{(m)}\}$ with probabilities

$$\Pr(f_t^{(j)} = p_t^{(i)} \mid \mathbf{Y}_t) = \frac{\alpha_t^{(i)}}{\alpha_t^{(1)} + \cdots + \alpha_t^{(m)}}, \quad j = 1, \cdots, m. \tag{A.3.10}$$

A.3.1.2 Smoothing

For smoothing, assume that

$$\Pr(X_1 = s^{(j)}_{1|t-1}, \cdots, X_{t-1} = s^{(j)}_{t-1|t-1} \mid \mathbf{Y}_{t-1}) = \frac{1}{m},$$

and $\varepsilon^{(j)}_t \sim q(\varepsilon)$ and define $(p^{(j)}_{1|t-1}, \cdots, p^{(j)}_{t|t-1})$ as follows:

$$p^{(j)}_{i|t-1} = \begin{cases} s^{(j)}_{i|t-1}, & \text{for } i = 1, \cdots, t-1 \\ F(s^{(j)}_{t-1|t-1}, \varepsilon^{(j)}_t), & \text{for } i = t. \end{cases}$$

Then, $(p^{(j)}_{1|t-1}, \cdots, p^{(j)}_{t|t-1})$ can be considered as a realization from the joint distribution of $(X_1, \cdots, X_t)$ when the observation $\mathbf{Y}_{t-1}$ is given. Next, given the observation y_t, the distribution $\Pr(X_1 \leq p^{(j)}_{1|t-1}, \cdots, X_t \leq p^{(j)}_{t|t-1} \mid \mathbf{Y}_{t-1})$ can be updated as follows:

$$\begin{aligned} \Pr(X_1 &= p^{(j)}_{1|t-1}, \cdots, X_t = p^{(j)}_{t|t-1} \mid \mathbf{Y}_t) \\ &= \Pr(X_1 = p^{(j)}_{1|t-1}, \cdots, X_t = p^{(j)}_{t|t-1} \mid \mathbf{Y}_{t-1}, y_t) \\ &= \frac{p(y_t \mid X_1 = p_{1|t-1}, \cdots, X_t = p^{(j)}_{t|t-1})}{p(y_t \mid \mathbf{Y}_{t-1})} \\ &\quad \times \Pr(X_1 = p^{(j)}_{1|t-1}, \cdots, X_t = p^{(j)}_{t|t-1} \mid \mathbf{Y}_{t-1}) \\ &= \frac{p(y_t \mid p^{(j)}_{t|t-1}) \Pr(X_1 = p^{(j)}_{1|t-1}, \cdots, X_t = p^{(j)}_{t|t-1} \mid \mathbf{Y}_{t-1})}{p(y_t \mid \mathbf{Y}_{t-1})}. \quad (A.3.11) \end{aligned}$$

Since $p^{(j)}_{t|t-1}$ is the same as the particle $p^{(j)}_t$ of the filter algorithm (A.3.5), the smoothing distribution $p(X_1, \cdots, X_t \mid \mathbf{Y}_t)$ can be obtained by re-sampling m $n-$ dimensional vectors $p^{(j)}_{1|t-1}, \cdots, p^{(j)}_{t|t-1})', j = 1, \cdots, m$ with the same weights as the filter.

A.4 The derivative of the Forward-Backward smoothing

The marginal smoother can be obtained using forward-backward smoother as

$$p(X_t \mid \mathbf{Y}_T) = p(X_t \mid \mathbf{Y}_t) \int \frac{p(X_{t+1} \mid \mathbf{Y}_t) p(X_{t+1} \mid X_t)}{p(X_{t+1} \mid \mathbf{Y}_t)} dX_{t+1},$$

where, $p(X_t \mid \mathbf{Y}_t)$ and $p(X_{t+1} \mid \mathbf{Y}_t)$ are the filtering density and one step ahead predictive density respectively, at time t. Thus, starting with $p(X_T \mid \mathbf{Y}_T), p(X_t \mid \mathbf{Y}_T)$ can be recursively obtained from $p(X_{t+1} \mid \mathbf{Y}_T)$. Using the above recursion, the marginal smoothing distribution can now be approximated by the weighted particle cloud. Here, one starts with the forward filtering pass for computing the filtered distribution at each time step using particle filter as eq (3.4.1), then one performs the backward smoothing pass as given by (A.4.1) to approximate the smoothing distribution

$$\hat{p}(X_t \mid \mathbf{Y}_T) = \sum_{m=1}^{N} w_{t|T}^{(m)} \delta(\mathbf{X}_t - \mathbf{x}_t^m) \tag{A.4.1}$$

Where δ denoted the Dirac delta function. The smoothing weights are obtained through the following backward recursion:

$$w_{t|T}^{(m)} \quad = \quad w_t^{(m)} \sum_{j=1}^{N} \left[w_{t+1|T}^{(j)} \frac{p(X_{t+1}^{(j)} \mid X_{t+1}^{(m)})}{\sum_{r=1}^{N} p(X_{t+1}^{(j)} \mid X_t^{(r)}) w_t^{(r)}} \right]$$

with $w_{T|T}^{(m)} = w_T^{(m)}$. It is important to note that the forward-backward smoother keeps the same particle support as used in filtering step and re-weights the particles to obtain the approximated particle based smoothed distribution. To obtain the marginal MAP smoother, one needs the posterior density $p(X_t \mid \mathbf{Y}_T)$ from the above cloud representation. Here, we proceed as follows. Using the Bayes' rule, one can write the one step ahead predictive density in equation (A.4.1) as

$$p(X_{t+1} \mid \mathbf{Y}_t) \quad = \quad \frac{p(X_{t+1} \mid \mathbf{Y}_{t+1}) p(Y_{t+1} \mid \mathbf{Y}_t)}{p(Y_{t+1} \mid X_{t+1})}$$

(A.4.2)

Then equation (A.4.1) becomes

$$\begin{aligned} p(X_t \mid \mathbf{Y}_T) &= p(X_t \mid \mathbf{Y}_t) \int \frac{p(X_{t+1} \mid \mathbf{Y}_T) p(X_{t+1} \mid X_t) p(Y_{t+1} \mid X_{t+1})}{p(X_{t+1} \mid \mathbf{Y}_{t+1}) p(Y_{t+1} \mid \mathbf{Y}_t)} dX_{t+1} \\ &= \frac{p(X_t \mid \mathbf{Y}_t)}{p(Y_{t+1} \mid \mathbf{Y}_t)} \int \left[\frac{p(X_{t+1} \mid X_t) p(Y_{t+1} \mid X_{t+1})}{p(X_{t+1} \mid \mathbf{Y}_{t+1})} \right] p(X_{t+1} \mid \mathbf{Y}_T) dX_{t+1} \\ &= \frac{p(X_t \mid \mathbf{Y}_t)}{p(Y_{t+1} \mid \mathbf{Y}_t)} \int \left[\frac{p(X_{t+1} \mid X_t) p(Y_{t+1} \mid X_{t+1})}{p(X_{t+1} \mid \mathbf{Y}_{t+1})} \right] \\ &\quad \times \hat{P}(X_{t+1} \mid \mathbf{Y}_T). \end{aligned} \tag{A.4.3}$$

Approximating the above integration by Monte Carlo integration method, one obtains

$$p(X_t \mid \mathbf{Y}_T) \;\approx\; \frac{p(X_t \mid \mathbf{Y}_t)}{p(Y_{t+1} \mid \mathbf{Y}_t)} \sum_{j=1}^{N} \left[\frac{p(X_{t+1}^{(j)} \mid X_t) p(Y_{t+1} \mid X_{t+1}^{(j)})}{p(X_{t+1}^{(j)} \mid \mathbf{Y}_{t+1})} \right] w_{t+1|T}^{(j)}.$$

Further approximating the filtered density $p(X_{t+1} \mid \mathbf{Y}_{t+1})$ from the running particle filter as

$$p(X_{t+1} \mid \mathbf{Y}_{t+1}) \approx \frac{p(Y_{t+1} \mid X_{t+1}) \sum_r p(X_{t+1} \mid X_t^{(r)}) w_t^{(r)})}{p(Y_{t+1} \mid \mathbf{Y}_{t+1})} \tag{A.4.4}$$

we can rewrite equation (A.4.4) as

$$p(X_t \mid \mathbf{Y}_T) \;\approx\; p(X_t \mid \mathbf{Y}_t) \sum_{j=1}^{N} \left[\frac{p(X_{t+1}^{(j)} \mid X_t)}{\sum_{r=1}^{N} p(X_{t+1}^{(j)} \mid X_t^{(r)}) w_t^{(r)}} \right] w_{t+1|T}^{(j)}.$$

The MAP estimate of the marginal smoothing density , $p(X_t \mid \mathbf{Y}_T)$ can then be obtained by finding the location of its global maxima. This maximization can be performed using different optimization method. We maximize along the particles by using the equation (A.4.5), this leads to the approximate particle based MAP estimate as

$$X_{t|T}^{MAP} \;\approx\; \arg\max_{X_t^{(m)}} p(X_t^{(m)} \mid \mathbf{Y}_t) \sum_{j=1}^{N} \left[\frac{p(X_{t+1}^{(j)} \mid X_t)}{\sum_{r=1}^{N} p(X_{t+1}^{(j)} \mid X_t^{(r)}) w_t^{(r)}} \right] w_{t+1|T}^{(j)}. \tag{A.4.5}$$

Where $m = 1, \cdots, N$ and N is the number of particles in the cloud at each time step. By using equation (A.4.2), the estimator can be further simplified to

$$x_{t|T}^{MAP} = \arg\max_{x_t^{(m)}} p(X_t^{(m)} \mid \mathbf{Y}_t) \frac{w_{t|T}^{(m)}}{w_t^{(m)}}, \tag{A.4.6}$$

where the filtered density $p(X_t \mid \mathbf{Y}_t)$ at the particle cloud $\{X_t^{(m)}\}_{m=1}^{N}$ can be evaluated during the forward filtering step as

$$p(X_t \mid \mathbf{Y}_t) \approx \frac{p(Y_t \mid X_t^{(m)}) \sum_j p(X_t^{(m)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}}{p(Y_t \mid \mathbf{Y}_{t-1})} \tag{A.4.7}$$

Since $p(Y_t \mid \mathbf{Y}_{t-1})$ in equation (A.4.7) is independent of $X_t^{(m)}$, to obtain $X_{t|T}^{MAP}$, one can replace $p(X_t^{(m)} \mid \mathbf{Y}_t)$ in equation(A.4.6) by the un-normalized filtered density

$$q(X_t^{(m)} \mid \mathbf{Y}_t) \;=\; p(Y_t \mid X_t^{(m)}) \sum_j p(X_t^{(m)} \mid X_{t-1}^{(j)}) w_{t-1}^{(j)}. \tag{A.4.8}$$

Appendix B

B.1 The Multinomial distribution

B.1.1 The first- order derivative of parameter β for the multinomial logistic regression model

Hence $Y_{ik}(t) = (y_{ik}^1(t), \cdots, y_{ik}^{c_k}(t))^\top$, has a vector of conditional probabilities $\pi_{ik}(t) = (\pi_{ik}^1(t), \cdots, \pi_{ik}^{c_k}(t))^\top$, where $\sum_{s=1}^{c_k} y_{ik}^s(t) = 1$ and $\sum_{s=1}^{c_k} \pi_{ik}^s(t) = 1$. The likelihood is a product of the multinomial probabilities

$$l(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{q}\prod_{t=1}^{T}\prod_{s=1}^{c_k}[\pi_{ik}^s(t)]^{y_{ik}^s(t)}$$

the log-likelihood is given by

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\sum_{s=1}^{c_k} y_{ik}^s(t)\log[\pi_{ik}^s(t)] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\left[y_{ik}^1(t)\log\pi_{ik}^1(t) + \cdots + y_{ik}^{c_k}(t)\log\pi_{ik}^{c_k}(t)\right] \qquad \text{(B.1.1)} \end{aligned}$$

By substituting

$$y_{ik}^{c_k}(t) = 1 - \sum_{s=1}^{c_k-1} y_{ik}^s(t) = 1 - y_{ik}^1(t) - y_{ik}^2(t) - \cdots - y_{ik}^{c_k-1}(t)$$

Then

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\{y_{ik}^1(t)\log\pi_{ik}^1(t) + \cdots \\ &\quad + [1 - y_{ik}^1(t) - y_{ik}^2(t) - \cdots - y_{ik}^{c_k-1}(t)]\log\pi_{ik}^{c_k}(t)\} \end{aligned}$$

we putting the same coefficients together

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\{y_{ik}^{1}(t)\log\left[\frac{\pi_{ik}^{1}(t)}{\pi_{ik}^{c_k}(t)}\right]+y_{ik}^{2}(t)\log\left[\frac{\pi_{ik}^{2}(t)}{\pi_{ik}^{c_k}(t)}\right]+\cdots \\ &\quad +y_{ik}^{c_k-1}(t)\log\left[\frac{\pi_{ik}^{c_k-1}(t)}{\pi_{ik}^{c_k}(t)}\right]+\log\pi_{ik}^{c_k}(t)\} \end{aligned}$$

Where $\eta_{ik}^{s}(t)=\log\left[\frac{\pi_{ik}^{s}(t)}{\pi_{ik}^{c_k}(t)}\right]$ and $\pi_{ik}^{c_k}(t)=\frac{1}{1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^{j}(t)]}$
The likelihood is rewritten by

$$l(\beta)=\sum_{k=1}^{q}\sum_{t=1}^{T}\sum_{s=1}^{c_k-1}y_{ik}^{s}(t)\eta_{ik}^{s}(t)-\log\left(1+\sum_{j=1}^{c_k}\exp\eta_{ik}^{j}(t)\right)$$

To calculate the partial score, recall the chain rule for multivariate functions to obtain

$$\frac{\partial l(\beta)}{\partial\beta_k^{s\top}}=\frac{\partial l(\beta)}{\partial\eta_{ik}^{s\top}(t)}\frac{\partial\eta_{ik}^{s}(t)}{\partial\pi_{ik}^{s\top}(t)}\frac{\partial\pi_{ik}^{s}(t)}{\partial\eta_{ik}^{s\top}(t)}\frac{\partial\eta_{ik}^{s}(t)}{\partial\beta_k^{s\top}}, \tag{B.1.2}$$

We calculate the partial score, uses it only to about the double sum i and t:

$$\begin{aligned} \frac{\partial l(\beta)}{\partial\eta_{ik}^{s\top}(t)} &= \frac{\partial}{\partial\eta_{ik}^{s}(t)}\left[\sum_{i=1}^{n}\sum_{k=1}^{q}\sum_{t=1}^{T}\sum_{s=1}^{c_k-1}y_{ik}^{s}(t)\eta_{ik}^{s}(t)-\log\left(1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^{j}(t)\right)\right] \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T}y_{ik}^{s}(t)-\frac{1}{\left(1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^{j}(t)]\right)}\times\exp(\eta_{ik}^{s}(t)) \\ &= \sum_{i=1}^{n}\ \sum_{t=1}^{T}y_{ik}^{s}(t)-\frac{\exp(\eta_{ik}^{s}(t))}{\left(1+\sum_{j=1}^{c_k}\exp[\eta_{ik}^{j}(t)]\right)} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T}[y_{ik}^{s}(t)-\pi_{ik}^{s}(t)]^{\top} \end{aligned} \tag{B.1.3}$$

and

$$\frac{\partial\eta_{ik}^{s}(t)}{\partial\pi_{ik}^{s\top}(t)}=\Sigma_{ik}^{-1}(X_t) \tag{B.1.4}$$

and $\Sigma_{ik}(t)=cov(y_{ik}^{s}(t)$ with generic elements

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

To solve $\frac{\partial\pi_{ik}^{s}(t)}{\partial\eta_{ik}^{s\top}(t)}$ we will using the quotient rule for differentiating of division two equations:

$$\left(\frac{f}{g}\right)'(a)=\frac{g(a).f'(a)-f(a).g'(a)}{[g(a)]^2}$$

where

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

$$\begin{aligned}
\frac{\partial \pi^s_{ik}(t)}{\partial \eta^s_{ik}(t)} &= \frac{(1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]).(\exp[\eta^s_{ik}(t)]) - (\exp[\eta^s_{ik}(t)])^2}{(1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)])^2} \\
&= \frac{\exp[\eta^s_{ik}(t)]\{(1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]) - (\exp[\eta^s_{ik}(t)])\}}{(1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)])^2} \\
&= \frac{\exp[\eta^s_{ik}(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]} \times \frac{(1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]) - \exp[\eta^s_{ik}(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]} \\
&= \pi^s_{ik}(t) \times \left\{\frac{1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]} - \frac{\exp[\eta^s_{ik}(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]}\right\} \\
&= \pi^s_{ik}(t) \times \left\{1 - \frac{\exp[\eta^s_{ik}(t)]}{1+\sum_{j=1}^{c_k}\exp[\eta^j_{ik}(t)]}\right\} \\
&= \pi^s_{ik}(t)(1-\pi^s_{ik}(t)) \\
&= D_{ik}(X_t) \qquad \text{(B.1.5)}
\end{aligned}$$

Since $\eta^s_{ik}(t) = \mathbf{u}_i^\top(t)\beta^s_k + X_i(t)$, we obtain

$$\frac{\partial \eta^s_{ik}(t)}{\partial \beta^{s\top}_k} = \mathbf{u}_i^\top(t) \qquad \text{(B.1.6)}$$

Substitution of equations (B.1.3) ,(B.1.4), (B.1.5), (B.1.6) into (B.1.2) shows that

$$\frac{\partial l(\beta)}{\partial \beta^{s\top}_k} = \sum_{i=1}^{n}\sum_{t=1}^{T}[y^s_{ik}(t) - \pi^s_{ik}(t)]^\top \mathbf{u}_i^\top(t) D^\top_{ik}(X_t)\Sigma^{-1}_{ik}(X_t) \qquad \text{(B.1.7)}$$

Then we using this result to find the score function of β^s_k with EM algorithm:

$$\begin{aligned}
\beta^{s(p)}_k &= \sum_{i=1}^{n}\sum_{t=1}^{T}\int\cdots\int\left\{\mathbf{u}_i^\top(t)D^\top_{ik}(X_t)\Sigma^{-1}_{ik}(t)[y^s_{ik}(t)-\pi^s_{ik}(t)]^\top\right\} \\
&\quad \times p(X_i(t) \mid Y_i(t))d(X_i). \qquad \text{(B.1.8)}
\end{aligned}$$

B.1.2 Second-order derivative of parameter β for the multinomial logistic regression model

$$\frac{\partial^2 l(\beta)}{\partial \beta^s_k \partial \beta^{s\top}_k} = \frac{\partial}{\partial \beta^s_k}\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{u}_i^\top(t)D_{ik}(X_t)\Sigma^{-1}_{ik}(X_t)[y^s_{ik}(t)-\pi^s_{ik}(t)]^\top\right]$$

To calculate the second partial score, recall the chain rule for multivariate functions to obtain

$$\frac{\partial^2 l(\beta)}{\partial \beta_k^{s\top} \partial \beta_k^s} = \frac{\partial^2 l(\beta_k^s)}{\partial \pi_{ik}^s(t)} \frac{\partial \pi_{ik}^s(t)}{\partial \beta_k^{s\top}}, \tag{B.1.9}$$

Where

$$\frac{\partial^2 l(\beta_k^s)}{\partial \pi_{ik}^s(t)} = -\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{u}_i^\top(t) D_{ik}(X_t) \Sigma_{ik}^{-1}(X_t) \tag{B.1.10}$$

and we can find the derivation for the probability to β_k^s, where

$$\frac{d}{d\beta} e^{\eta_{ik}^s(t)} = e^{\eta_{ik}^s(t)} \mathbf{u}_i^\top(t)$$

$$\begin{aligned}
\frac{\partial \pi_{ik}^s(t)}{\partial \beta_k^s} &= \left[\frac{(1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]).(\exp[\eta_{ik}^s(t)]) \times \mathbf{u}_i^\top(t) - (\exp[\eta_{ik}^s(t)])^2 \times \mathbf{u}_i^\top(t)}{(1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)])^2} \right] \\
&= \mathbf{u}_i(t) \left[\frac{\exp[\eta_{ik}^s(t)] \{ (1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]) - (\exp[\eta_{ik}^s(t)]) \}}{(1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)])^2} \right] \\
&= \mathbf{u}_i(t) \left[\frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \times \frac{(1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]) - \exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \right] \\
&= \mathbf{u}_i(t) \left[\pi_{ik}^s(t) \times \left\{ \frac{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} - \frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \right\} \right] \\
&= \mathbf{u}_i(t) \left[\pi_{ik}^s(t) \times \left\{ 1 - \frac{\exp[\eta_{ik}^s(t)]}{1 + \sum_{j=1}^{c_k} \exp[\eta_{ik}^j(t)]} \right\} \right] \\
&= \mathbf{u}_i(t) \left[\pi_{ik}^s(t)(1 - \pi_{ik}^s(t)) \right] \\
&= \mathbf{u}_i(t) D_{ik}(X_t) \qquad \text{(B.1.11)}
\end{aligned}$$

Substitution of equations (B.1.10) ,(B.1.11) into (B.1.9) shows that

$$\frac{\partial^2 l(\beta)}{\partial \beta_k^s \partial \beta_k^{s\top}} = -\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{u}_i^\top(t) D_{ik}^\top(X_t) \Sigma_{ik}^{-1}(X_t) D_{ik}(X_t) \mathbf{u}_i(t) \tag{B.1.12}$$

Appendix C

The exponential family distributions

C.1 Members of the exponential family distributions

C.1.1 Binomial distribution

$$\begin{aligned} p(x) &= \tbinom{n}{x}\pi^x(1-\pi)^{n-x}, \qquad\qquad x \in \{0,1,2,\cdots,n\}, \\ &= \exp\left\{x\log(\frac{\pi}{1-\pi}) + n\log(1-\pi) + \log\tbinom{n}{x}\right\}. \end{aligned}$$

$$\nu = \log\left(\frac{\pi}{1-\pi}\right), \quad \phi = 1, \quad b(\nu) = -n\log(1-\pi), \quad c(x,\phi) = \log\tbinom{n}{x}$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n}\{[x_i\log\frac{x_i}{1-x_i} - x_i\log\frac{\hat{\pi}_i}{1-\hat{\pi}_i}] - [n_i\log\frac{1}{1-x_i} - n_i\log\frac{1}{1-\hat{\pi}_i}]\} \\ &= 2\sum_{i=1}^{n}\{x_i\log\frac{x_i}{\hat{\pi}_i} - x_i\log\frac{1-x_i}{1-\hat{\pi}_i} + n_i\log\frac{1-x_i}{1-\hat{\pi}_i}\} \\ &= 2\sum_{i=1}^{n}\{x_i\log\frac{x_i}{\hat{\pi}_i} + (n_i - x_i)\log\frac{1-x_i}{1-\hat{\pi}_i}\} \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi} = \frac{Dev}{1} = Dev$$

C.1.2 Negative Binomial distribution

$$\begin{aligned} p(x) &= \tbinom{x-1}{r-1}\,\pi^r(1-\pi)^{x-r}, \quad x=\{r,r+1,\cdots\}, \\ &= \exp\left\{x\log(1-\pi)+r\log\left(\frac{\pi}{1-\pi}\right)+\log\tbinom{x-1}{r-1}\right\}. \end{aligned}$$

$$\nu=\log(1-\pi), \quad \phi=1, \quad b(\nu)=-r\log\left(\frac{\pi}{1-\pi}\right), \quad c(x,\phi)=\log\tbinom{x-1}{r-1}$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i-x_i\hat{\nu}_i]-[b(\tilde{\nu}_i)-b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n}\{[x_i\log(1-x_i)-x_i\log(1-\hat{\pi}_i)]-[-r_i\log\frac{x_i}{1-x_i}+r_i\log\frac{\hat{\pi}_i}{1-\hat{\pi}_i}]\} \\ &= 2\sum_{i=1}^{n}\{x_i\log\frac{(1-x_i)}{(1-\hat{\pi}_i)}+r_i\log\frac{x_i}{\hat{\pi}_i}+r_i\log\frac{(1-x_i)}{(1-\hat{\pi}_i)}\} \end{aligned}$$

$$Dev^*=\frac{Dev}{\phi}=\frac{Dev}{1}=Dev$$

C.1.3 Bernoulli distribution

$$\begin{aligned} p(x) &= \pi^x(1-\pi)^{1-x}, \qquad x\in\{0,1\}, \\ &= \exp\left\{x\log(\frac{\pi}{1-\pi})+\log(1-\pi)\right\}. \end{aligned}$$

$$\nu=\log\left(\frac{\pi}{1-\pi}\right), \quad \phi=1, \quad b(\nu)=-\log(1-\pi), \quad c(x,\phi)=0$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i-x_i\hat{\nu}_i]-[b(\tilde{\nu}_i)-b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n}\{[x_i\log\frac{x_i}{1-x_i}-x_i\log\frac{\hat{\pi}_i}{1-\hat{\pi}_i}]-[\log\frac{1}{1-x_i}-\log\frac{1}{1-\hat{\pi}_i}]\} \\ &= 2\sum_{i=1}^{n}\{x_i\log\frac{x_i}{\hat{\pi}_i}-x_i\log\frac{1-x_i}{1-\hat{\pi}_i}+\log\frac{1-x_i}{1-\hat{\pi}_i}\} \\ &= 2\sum_{i=1}^{n}\{x_i\log\frac{x_i}{\hat{\pi}_i}+(1-x_i)\log\frac{1-x_i}{1-\hat{\pi}_i}\} \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi} = \frac{Dev}{1} = Dev$$

C.1.4 Multinomial distribution

$$\begin{aligned} p(x) &= \prod_{k=1}^{q} \pi_k^{x_k}, \qquad x \in \{0,1\}, \quad \sum_{k=1}^{q} x_k = 1, \\ &= \exp\left\{ \sum_{k=1}^{q-1} x_k \log(\frac{\pi_k}{\pi_q}) + \log(\pi_q) \right\}. \end{aligned}$$

where $\sum_{k=1}^{q} \pi_k = 1$,

$$\nu = \left(\log \frac{\pi_1}{\pi_q}, \log \frac{\pi_2}{\pi_q}, \cdots, \log \frac{\pi_q}{\pi_q} \right)^\top, \quad \phi = 1, \quad b(\nu) = -\log\left(\pi_q\right), \quad c(x,\phi) = 0$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n} \{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n} \{\sum_{k=1}^{q} x_{ik} \log x_{ik} - \sum_{k=1}^{q} x_{ik} \log \hat{\pi}_{ik}\} \\ &= 2\sum_{i=1}^{n} \sum_{k=1}^{q} x_{ik} \log \frac{x_{ik}}{\hat{\pi}_{ik}} \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi} = \frac{Dev}{1} = Dev$$

C.1.5 Poisson distribution

$$\begin{aligned} p(x) &= \frac{e^{-\lambda}}{x!}\lambda^x, \qquad x \in \{0,1,2,\cdots,n\}, \\ &= \exp\left\{x\log\lambda - \lambda - \log x!\right\}. \end{aligned}$$

$$\nu = \log\lambda, \quad \phi = 1, \quad b(\nu) = \lambda, \quad c(x,\phi) = -\log(x!)$$

$$
\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\{[x_i \log x_i - x_i \log \hat{\lambda}_i] - [x_i - \hat{\lambda}_i]\} \\
&= 2\sum_{i=1}^{n}\{x_i \log \frac{x_i}{\hat{\lambda}_i} - [x_i - \hat{\lambda}_i]\}
\end{aligned}
$$

$$
Dev^* = \frac{Dev}{\phi} = \frac{Dev}{1} = Dev
$$

C.1.6 Dirichlet distribution

$$
\begin{aligned}
p(x) &= \frac{\Gamma(\sum_{k=1}^{q}\pi_k)}{\prod_{k=1}^{q}(\Gamma\pi_k)}\prod_{k=1}^{q} x_k^{\pi_k - 1}, \qquad x \in \{0,1\}, \quad \sum_{k=1}^{q} x_k = 1, \\
&= \exp\left\{(1-\pi_k)\log x_k + \log[\Gamma(\sum_{k=1}^{q}\pi_k)] - \log[\sum_{k=1}^{q}(\Gamma\pi_k)]\right\}.
\end{aligned}
$$

$$
\nu = (1-\pi_k), \quad \phi = 1, \quad b(\nu) = \log[\Gamma(\sum_{k=1}^{q}\pi_k)] - \log[\sum_{k=1}^{q}(\Gamma\pi_k)], \quad c(x,\phi) = 0
$$

$$
\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\{[(1-x_{ik})\log x_{ik} - (1-\hat{\pi}_{ik})\log x_{ik}] \\
&\quad -[\log[\Gamma(\sum_{k=1}^{q} x_{ik})] - \log[\sum_{k=1}^{q}(\Gamma x_{ik})] - \log[\Gamma(\sum_{k=1}^{q}\hat{\pi}_{ik})] - \log[\sum_{k=1}^{q}(\Gamma\hat{\pi}_{ik})]]\} \\
&= 2\sum_{i=1}^{n}\left\{(x_{ik} - \hat{\pi}_{ik})\log x_{ik} - \log\frac{\Gamma(\sum_{k=1}^{q} x_{ik})}{\Gamma(\sum_{k=1}^{q}\hat{\pi}_{ik})} + \log\frac{\sum_{k=1}^{q}(\Gamma x_{ik})}{\sum_{k=1}^{q}(\Gamma\hat{\pi}_{ik})}\right\}
\end{aligned}
$$

$$
Dev^* = \frac{Dev}{\phi} = \frac{Dev}{1} = Dev
$$

C.1.7 Normal distribution

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \qquad\qquad x \in \mathbb{R}, \\ &= \exp\left\{\frac{\mu x - \mu^2/2}{\sigma^2} + [-\frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}]\right\}. \end{aligned}$$

$$\nu = \mu, \quad \phi = \sigma^2, \quad b(\nu) = \frac{\mu^2}{2}, \quad c(x,\phi) = \frac{-1}{2}\log(2\pi\sigma^2) - x^2/(2\sigma^2)$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n}\{x_i^2 - x_i\hat{\mu}_i - \frac{x_i^2}{2} + \frac{\hat{\mu}_i^2}{2}\} \\ &= \sum_{i=1}^{n}\{x_i^2 - 2x_i\hat{\mu}_i + \hat{\mu}_i^2\} = \sum_{i=1}^{n}(x_i - \hat{\mu}_i)^2 \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi_i} = \sum_{i=1}^{n}\frac{(x_i - \hat{\mu}_i)^2}{\sigma_i^2}$$

C.1.8 Log-Normal distribution

$$\begin{aligned} p(x) &= \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\log x-\mu)^2}{2\sigma^2}\right], \qquad\qquad x \in \mathbb{R}^+, \\ &= \exp\left\{\frac{\mu\log x - \mu^2/2}{\sigma^2} + [-\frac{1}{2}\log(2\pi\sigma^2) - \log x - \frac{\log x^2}{2\sigma^2}]\right\}. \end{aligned}$$

$$\nu = \mu, \quad \phi = \sigma^2, \quad b(\nu) = \frac{\mu^2}{2}, \quad c(x,\phi) = \frac{-1}{2}\log(2\pi\sigma^2) - \log x - \log x^2/(2\sigma^2)$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n}\{\log x_i^2 - \log x_i\hat{\mu}_i - \frac{\log x_i^2}{2} + \frac{\hat{\mu}_i^2}{2}\} \\ &= \sum_{i=1}^{n}\{\log x_i^2 - 2\log x_i\hat{\mu}_i + \hat{\mu}_i^2\} = \sum_{i=1}^{n}(\log x_i - \hat{\mu}_i)^2 \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi_i} = \sum_{i=1}^{n} \frac{(\log x_i - \hat{\mu}_i)^2}{\sigma_i^2}$$

C.1.9 Inverse-Gaussian distribution

$$\begin{aligned} p(x) &= \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x \in (0, +\infty) \\ &= \exp\left\{\frac{-\lambda x}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2x} + \frac{\log\lambda}{2} - \left[\frac{1}{2}\log(2\pi) - 3\log x\right]\right\} \\ &= \exp\left\{\frac{x(\frac{-1}{\mu^2}) + 2/\mu}{2/\lambda} - \frac{\lambda}{2x} + \frac{\log\lambda - \log(2\pi) - 3\log x}{2}\right\}. \end{aligned}$$

$$\nu = -1/\mu^2, \quad \phi = 2/\lambda, \quad b(\nu) = \frac{-2}{\mu}, \quad c(x,\phi) = -\frac{\lambda}{2x} + \frac{\log\lambda - \log(2\pi) - 3\log x}{2}$$

$$\begin{aligned} Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\ &= 2\sum_{i=1}^{n}\{[x_i(\frac{-1}{x_i^2}) - x_i(\frac{-1}{\hat{\mu}_i^2})] - [-2\frac{1}{x_i} + 2\frac{1}{\hat{\mu}_i}] \\ &= 2\sum_{i=1}^{n}\frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 x_i} \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi_i} = \lambda\sum_{i=1}^{n}\frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 x_i}$$

C.1.10 Gamma distribution

$$\begin{aligned} p(x) &= \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\lambda}\right)^{\alpha} x^{\alpha-1} e^{-\frac{\alpha x}{\lambda}}, \qquad x \in (0,\infty), \\ &= \exp\left\{(-x\frac{\alpha}{\lambda}) + (\alpha-1)\log x + \alpha\log(\alpha) - \alpha\log(\lambda) - \log(\Gamma(\alpha))\right\} \\ &= \exp\left\{\frac{-x(\frac{1}{\lambda}) - \log(\lambda)}{\frac{1}{\alpha}} + (\alpha-1)\log x + \alpha\log\alpha - \log(\Gamma(\alpha))\right\}. \end{aligned}$$

$$\nu = -\frac{1}{\lambda}, \quad \phi = \frac{1}{\alpha}, \quad b(\nu) = \log(\lambda), \quad c(x,\phi) = (\alpha-1)\log(x) + \alpha\log\alpha - \log(\Gamma(\alpha))$$

$$
\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\left\{[x_i(\frac{-1}{x_i}) - x_i(\frac{-1}{\hat{\lambda}_i})] - [\log(x_i) - \log\hat{\lambda}_i]\right\} \\
&= 2\sum_{i=1}^{n}\left\{\frac{(x_i - \hat{\lambda}_i)}{\hat{\lambda}_i} - \log\frac{x_i}{\hat{\lambda}_i}\right\}
\end{aligned}
$$

$$
Dev^* = \frac{Dev}{\phi_i} = 2\hat{\alpha}_i\sum_{i=1}^{n}\left\{\frac{(x_i - \hat{\lambda}_i)}{\hat{\lambda}_i} - \log\frac{x_i}{\hat{\lambda}_i}\right\}
$$

C.1.11 Exponential distribution

A gamma distribution with shape parameter $\alpha = 1$ and scale parameter λ is an exponential (λ) distribution.

$$
\begin{aligned}
p(x) &= \lambda\exp\left[-\lambda x\right], \quad x \in (0, \infty), \\
&= \exp\left[-\lambda x + \log\lambda\right],
\end{aligned}
$$

$$
\nu = -\lambda, \quad \phi = 1, \quad b(\nu) = -\log(\lambda), \quad c(x, \phi) = 0
$$

$$
\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\left\{\frac{(x_i - \hat{\lambda}_i)}{\hat{\lambda}_i} - \log\frac{x_i}{\hat{\lambda}_i}\right\}
\end{aligned}
$$

C.1.12 Beta distribution

$$
\begin{aligned}
p(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma\alpha\Gamma\beta}x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0, 1), \\
&= \exp\left\{(\alpha - 1)\log x - (\beta - 1)\log(1 - x) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma\alpha\Gamma\beta}\right\}.
\end{aligned}
$$

$$
\nu = \binom{(\alpha-1)}{(\beta-1)}, \quad \phi = 1, \quad b(\nu) = -\log\frac{\Gamma(\alpha+\beta)}{\Gamma\alpha\Gamma\beta}, \quad c(x, \phi) = 0
$$

$$
\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\{[\begin{bmatrix}\log x_i & \log(1-x_i)\end{bmatrix}\begin{bmatrix}\log x_i \\ \log(1-x_i)\end{bmatrix} - \begin{bmatrix}\log x_i & \log(1-x_i)\end{bmatrix}\begin{bmatrix}(\hat{\alpha}_i - 1) \\ (\hat{\beta} - 1)\end{bmatrix}] \\
&\quad -[-\log\frac{\Gamma(\log x_i) + (\log(1-x_i))}{\Gamma(\log x_i)\Gamma(\log(1-x_i))} + \log\frac{\Gamma\hat{\alpha} + \hat{\beta}}{\Gamma\hat{\alpha}\Gamma\hat{\beta}}]
\end{aligned}
$$

$$
Dev^* = \frac{Dev}{\phi_i} = \frac{Dev}{1} = Dev
$$

C.1.13 Pareto distribution

$$
\begin{aligned}
f(x) &= \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, \quad x > x_m, \quad x \in (x_m, +\infty), \\
&= \exp\{\log\alpha + \alpha\log x_m - (\alpha+1)\log x\}.
\end{aligned}
$$

$$
\nu = (\alpha+1), \quad \phi = 1, \quad b(\nu) = \log\alpha + \alpha\log x_m, \quad c(x,\phi) = 0
$$

$$
\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\{[x_i(x_i+1) - x_i(\hat{\alpha}_i+1)] - [(\log x_i + x_i\log x_m) - (\log\hat{\alpha} + \hat{\alpha}_i\log x_m)]\} \\
&= 2\sum_{i=1}^{n}\left\{(x_i^2 - x_i\hat{\alpha}_i) + \log\frac{\hat{\alpha}_i}{x_i} + (x_i + \hat{\alpha}_i)\log x_m\right\}
\end{aligned}
$$

$$
Dev^* = \frac{Dev}{\phi_i} = \frac{Dev}{1} = Dev
$$

C.1.14 Weibull distribution

with known shape α

$$
\begin{aligned}
f(x) &= \frac{\alpha x^{\alpha-1}}{\lambda^{\alpha}}\exp\left[-(\frac{x}{\lambda})^{\alpha}\right], \quad x \in [0, +\infty), \\
&= \exp\left\{-x^{\alpha}\lambda^{-\alpha} - \alpha\log\lambda + \log\alpha + (\alpha-1)\log x\right\}.
\end{aligned}
$$

$$
\nu = \lambda^{-\alpha}, \quad \phi = 1, \quad b(\nu) = \alpha\log\lambda, \quad c(x,\phi) = (\alpha-1)\log x - \log\alpha
$$

$$\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\left\{[x_i^{\alpha}x_i^{-\alpha} - x_i^{\alpha}(\hat{\lambda}_i^{-\alpha})] - [\alpha\log x_i - \alpha\log\hat{\lambda}_i]\right\} \\
&= 2\sum_{i=1}^{n}\left\{1 + (\frac{x_i}{\hat{\lambda}_i})^{\alpha} + \alpha\log\frac{\hat{\lambda}_i}{x_i}\right\}
\end{aligned}$$

$$Dev^* = \frac{Dev}{\phi_i} = \frac{Dev}{1} = Dev$$

C.1.15 Laplace distribution

The mean μ is known

$$\begin{aligned}
f(x) &= \frac{1}{2\sigma}\exp\left\{\frac{\mid x-\mu \mid}{\sigma}\right\}, \quad x \in (-\infty;+\infty), \\
&= \exp\left\{\frac{\mid x-\mu \mid}{\sigma} - 2\log\sigma\right\}.
\end{aligned}$$

$$\nu = \frac{1}{\sigma}, \quad \phi = 1, \quad b(\nu) = -2\log\sigma, \quad c(x,\phi) = 0$$

let $z_i =\mid x_i - \mu_i \mid$, then

$$\begin{aligned}
Dev &= 2\sum_{i=1}^{n}\{[x_i\tilde{\nu}_i - x_i\hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)]\} \\
&= 2\sum_{i=1}^{n}\{[z_iz_i - z_i\hat{\sigma}_i] - [-2(\log z_i) + 2(\log\hat{\sigma}_i)]\} \\
&= 2\sum_{i=1}^{n}\left\{(z_i^2 - z_i\hat{\sigma}_i) + 2\log\frac{z_i}{\hat{\sigma}_i}\right\}
\end{aligned}$$

$$Dev^* = \frac{Dev}{\phi_i} = \frac{Dev}{1} = Dev$$

C.1.16 Chi-Squared distribution

$$\begin{aligned} f(x) &= \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} x^{k/2-1} e^{-x/2}, \quad x \in (0, +\infty), \\ &= \exp\left\{ (\frac{k}{2} - 1) \log x - \frac{x}{2} - (\frac{k}{2}) \log 2 - \log \Gamma(\frac{k}{2}) \right\}. \\ &= \exp\left\{ \frac{(k-2) \log x - k \log 2 - 2 \log(\frac{k}{2})}{2} - \frac{x}{2} \right\}. \end{aligned}$$

$$\nu = (k-2), \quad \phi = 2, \quad b(\nu) = k \log 2 - 2 \log \Gamma(\frac{k}{2}), \quad c(x, \phi) = -\frac{x}{2}$$

$$\begin{aligned} Dev &= 2 \sum_{i=1}^{n} \{ [x_i \tilde{\nu}_i - x_i \hat{\nu}_i] - [b(\tilde{\nu}_i) - b(\hat{\nu}_i)] \} \\ &= 2 \sum_{i=1}^{n} \{ \Big[\log x_i \log x_i - \log x_i (\hat{k}_i - 2) \Big] \\ &\quad - \left[\left(\log x_i \log 2 - 2 \log \Gamma(\frac{\log x_i}{2}) \right) - \left(\hat{k} \log 2 - 2 \log \Gamma(\frac{\hat{k}}{2}) \right) \right] \} \\ &= 2 \sum_{i=1}^{n} \left\{ (\log x_i)^2 - (\hat{k} - 2) \log x_i] + [(x_i + \hat{k}_i) \log 2 + 2 \log \frac{\Gamma x_i / 2}{\Gamma \hat{k} / 2}] \right\} \end{aligned}$$

$$Dev^* = \frac{Dev}{\phi_i} = \frac{Dev}{2}$$

C.2 Estimating the dispersion parameter via the Deviance and Pearson Estimation

C.2.1 The Deviance estimation

McCullagh and *Nelder* (1989) presented a method of estimating the dispersion parameter by the goodness-of-fit criterion. $g(X_i(t), \nu_i(t), \phi_i(t))$ is the density function of an individual $X_i(t)$, where $\nu_i(t)$ is function of $\mu_i(t)$ then the log-likelihood, expressed as a function of the mean-value parameter $\mu_i(t)$ and the dispersion parameter is just

$$l(X_i(t), \mu_i(t), \phi_i(t)) = \log g[X_i(t); \nu[\mu_i(t)], \phi_i(t)].$$

The log likelihood for an individual based on a set of independent variables $X_i = (X_i(1), \cdots, X_i(T))'$ is just the sum of the variables, so that

$$l(X_i, \mu_i, \phi_i) = \sum_{t=1}^{T} \log g_i[X_i(t), \nu[\mu_i(t)], \phi_i(t)].$$

where $\mu_i = (\mu_1(1), \cdots, \mu_1(T))'$,and $\phi_i = (\phi_1(1), \cdots, \phi_1(T))'$. There are advantages in using as the goodness-of-fit criterion, not the log likelihood $l(X_i, \mu_i, \phi_i)$ but a particular linear function, namely

$$\begin{aligned} Dev^*(X_i, \mu_i) &= 2l(X_i, \phi_i, X_i) - 2l(X_i, \mu_i, \phi_i) \\ &= \sum_{t=1}^{T} \left\{ \frac{X_i(t)[\tilde{\nu}_i(t) - \hat{\nu}_i(t)] - b[\tilde{\nu}_i(t)] + b[\hat{\nu}_i(t)]}{\phi_i(t)} \right\} \end{aligned}$$

where $\tilde{\nu}_i(t) = \nu_i[X_i(t)]$ and $\hat{\nu}_i(t) = \nu[\hat{\mu}_i(t)]$, which they called the *scaled deviance*. we present $Dev^*(X_i, \mu_i)$ in **Appendix (F)** for the exponential family distributions. Note that, for the exponential family distributions considered here, $l(X_i, \phi_i, X_i)$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to latent variables, they consider maximizing $l(X_i, \mu_i, \phi_i)$ is equivalent to minimizing $Dev^*(X_i, \mu_i)$ with respect to μ_i. Deviance, as a measure of goodness of fit, is defined as following

$$Dev(X_i, \mu_i) = \phi_i . Dev^*(X_i, \mu_i) \tag{C.2.1}$$

where $Dev(X_i, \mu_i)$ is known as the deviance for the current model and is a function of the data only. We can estimate ϕ_i by the mean scaled Deviance

$$\hat{\phi}_i^d = \frac{Div(X_i, \mu_i)}{N - r} \tag{C.2.2}$$

where $N = (n \times T)$ is the sample size and r is the total number of unknown parameters. $Dev_i^* = Dev_i/\phi_i$ is approximately $\chi^2(N-r)$ distributed with expectation $(N - r)$. Thus $\hat{\phi}_d$ is approximately unbiased.

Example 1: we have for an individual

$$X_i(t) \sim \mathcal{N}(\mu_i(t), V_i(t))$$

$$f(X_i(t), \mu_i(t), V_i(t)) = \frac{1}{(\sqrt{2\pi\sigma^2})^T} \exp\left\{ -\sum_{t=1}^{T} \frac{(X_i(t) - \mu_i(t))^2}{2V_i(t)} \right\},$$

so that the log-likelihood is

$$l(X_i, \mu_i(t), V_i(t)) = -\frac{T}{2} \log(2\pi V_i(t)) - \sum_{i=1}^{T} \frac{(X_i(t) - \mu_i(t))^2}{2V_i(t)},$$

setting $\mu_i(t) = X_i(t)$ gives the maximum likelihood achievable log likelihood, namely

$$l(X_i(t), V_i(t), X_i(t)) = -\frac{T}{2}\log(2\pi V_i(t)),$$

so that the scaled deviance function is

$$Dev^*(X_i;\mu_i) = 2\{l(X_i(t), V_i(t), X_i(t)) - l(X_i(t), \mu_i(t), V_i(t)\} = \sum_{i=1}^{T}\frac{(X_i(t)-\mu_i(t))^2}{V_i(t)},$$

or

$$\begin{aligned} Dev^*(X_i;\mu_i) &= 2\sum_{t=1}^{T}\left\{\frac{X_i(t)[\tilde{\nu}_i(t)-\hat{\nu}_i(t)] - b[\tilde{\nu}_i(t)] + b[\hat{\nu}_i(t)]}{\phi_i(t)}\right\} \\ &= 2\sum_{t=1}^{T}\left\{\frac{X_i(t)[X_i(t)-\hat{\mu}_i(t)] - \frac{1}{2}X_i^2(t) + \frac{1}{2}\hat{\mu}_i^2(t)}{V_i(t)}\right\} \\ &= 2\sum_{t=1}^{T}\left\{\frac{\frac{1}{2}X_i^2(t) - X_i(t)\hat{\mu}_i(t) + \frac{1}{2}\hat{\mu}_i^2(t)}{V_i(t)}\right\} \\ &= \sum_{t=1}^{T}\left\{\frac{X_i^2(t) - 2X_i(t)\hat{\mu}_i(t) + \hat{\mu}_i^2(t)}{V_i(t)}\right\} \\ &= \sum_{t=1}^{T}\left\{\frac{[X_i(t)-\hat{\mu}_i(t)]^2}{V_i(t)}\right\} \end{aligned}$$

and

$$\begin{aligned} Dev(X_i,\mu_i) &= V_i(t)\times\left(\sum_{t=1}^{T}\left\{\frac{[X_i(t)-\hat{\mu}_i(t)]^2}{V_i(t)}\right)\right. \\ &= \sum_{t=1}^{T}\left\{[X_i(t)-\hat{\mu}_i(t)]^2\right. \end{aligned} \tag{C.2.3}$$

Then

$$\hat{\phi}_d = \hat{V}_i(t) = \frac{\sum_{t=1}^{T}[X_i(t)-\hat{\mu}_i(t)]^2}{T-2}$$

Example 2: we have for an individual

$$X_i(t) \sim Gamma(\alpha, \frac{\lambda}{\alpha}),$$

where $\nu_i(t) = \frac{-1}{\lambda_i(t)}$, $\phi_i(t) = \frac{1}{\alpha_i(t)}$, $b[\nu_i(t)] = -\log[\nu_i(t)]$, $\mu_i(t) = \frac{\alpha_i(t)}{\lambda_i(t)}$, $V_i(t) = \frac{\alpha_i(t)}{\lambda_i^2(t)}$, then

$$\begin{aligned}
Dev^*(X_i;\mu_i) &= 2\sum_{t=1}^{T}\left\{\frac{X_i(t)[\tilde{\nu}_i(t)-\hat{\nu}_i(t)]-b[\tilde{\nu}_i(t)]+b[\hat{\nu}_i(t)]}{\phi_i(t)}\right\} \\
&= 2\sum_{t=1}^{T}\left\{\frac{[X_i(t)\tilde{\nu}_i(t)-X_i(t)\hat{\nu}_i(t)]-[\log(-\tilde{\nu}_i(t))+\log(-\hat{\nu}_i(t))}{1/\alpha_i(t)}\right\} \\
&= 2\sum_{t=1}^{T}\alpha_i(t)\left\{[X_i(t)(-\frac{1}{X_i(t)})-X_i(t)(-\frac{1}{\hat{\lambda}_i(t)})]-[\log(X_i(t))+\log[\hat{\lambda}_i(t)]\right\} \\
&= 2\sum_{t=1}^{T}\alpha_i(t)\left\{\frac{X_i(t)-\hat{\lambda}_i(t)}{\hat{\lambda}_i(t)}-\log\left[\frac{X_i(t)}{\hat{\lambda}_i(t)}\right]\right\}
\end{aligned}$$

and

$$\begin{aligned}
Dev(X_i,\mu_i) &= \frac{1}{\alpha_i(t)}\times 2\sum_{t=1}^{T}\alpha_i(t)\left\{\frac{X_i(t)-\hat{\lambda}_i(t)}{\hat{\lambda}_i(t)}-\log\left[\frac{X_i(t)}{\hat{\lambda}_i(t)}\right]\right\} \\
&= \sum_{t=1}^{T}\left\{\frac{X_i(t)-\hat{\lambda}_i(t)}{\hat{\lambda}_i(t)}-\log\left[\frac{X_i(t)}{\hat{\lambda}_i(t)}\right]\right\}
\end{aligned}$$

Then

$$\hat{\phi}_d=\frac{1}{\hat{\alpha}_i(t)}=\frac{1}{T-2}\sum_{t=1}^{T}\left\{\frac{X_i(t)-\hat{\lambda}_i(t)}{\hat{\lambda}_i(t)}-\log\left[\frac{X_i(t)}{\hat{\lambda}_i(t)}\right]\right\}$$

C.2.2 Scaled Pearson estimation

The Pearson chi-squared statistics which take the form where for an individual $Z_i(t)$[1] has an exponential family distribution

$$X^2=\sum_{t=1}^{T}\frac{[Z_i(t)-\hat{\mu}_i(t)]^2}{V[\hat{\mu}_i(t)]}, \qquad \text{(C.2.4)}$$

where $V[\hat{\mu}_i(t)]$ is the estimated variance function for the distribution of $Z_i(t)$, where

$$Var[Z_i(t)]=\phi_i(t).V[\mu_i(t)].$$

The scaled Pearson chi-squared statistic is defined as

$$X_s^2=\frac{X^2}{\phi_i(t)}. \qquad \text{(C.2.5)}$$

[1] we write the latent variable $Z_i(t)$ to not mixed with the Pearson chi-squared statistics.

It turns out that, if the model is specified correctly,

$$X_s^2 \sim \chi^2_{T-r},$$

asymptotically, where T is the sample size and r is the number of unknown parameters in the model. Since the mean of a χ^2_{T-r} random variable is $T-r$, we can use the approximation

$$X_s^2 \approx T-r,$$

we can estimate $\phi_i(t)$

$$\hat{\phi}_i(t) = \frac{X^2}{T-r}. \tag{C.2.6}$$

Examples

(a) Normal distribution

$$Var[Z_i(t)] = V[\mu_i(t)]\phi_i(t) = 1.V_i(t),$$

so $V[\mu_i(t)] = 1$

$$X^2 = \sum_{i=1}^{T}[Z_i(t) - \hat{\mu}_i(t)]^2.$$

Therefore,

$$\hat{\phi}_i(t) = \hat{V}_i(t) = \frac{\sum_{i=1}^{T}[Z_i(t) - \hat{\mu}_i(t)]^2}{T-2}$$

This is the usual unbiased estimator of $V_i(t)$.(The MLE which has T rather than $T-2$ in the denominator, is biased.)

(b) Gamma distribution where

$$\mu_i(t) = \frac{\alpha_i(t)}{\lambda_i(t)}$$

$$\begin{aligned} Var[Z_i(t)] &= V[\mu_i(t)]\phi_i(t) = \frac{\alpha_i(t)}{\lambda_i^2(t)}, \\ &\equiv \left[\frac{\alpha_i^2(t)}{\lambda_i^2(t)}\right] \times \frac{1}{\alpha_i(t)} = [\mu_i(t)]^2 \times \phi_i(t) \end{aligned} \tag{C.2.7}$$

so $V[\mu_i(t)] = \mu_i^2(t)$

$$X^2 = \sum_{i=1}^{T}\frac{[Z_i(t) - \mu_i(t)]^2}{\mu_i^2(t)}.$$

Therefore,

$$\hat{\phi}_i(t) = \frac{1}{\hat{\alpha}_i(t)} = \sum_{i=1}^{T}\frac{[Z_i(t) - \mu_i(t)]^2}{\mu_i^2(t)(T-2)}$$

Bibliography

[1] Agresti, A. (2002). Categorical Data Analysis. Second Edition, John Wiley & Sons, Inc.

[2] Arulampalam, M. S., Maskell, S., Gordon, N., Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on signal processing, 50(2), 174-188.

[3] Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach. (Vol. 904). John Wiley Sons.

[4] 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.

[5] Bartolucci, F., Bacci, S., & Pennoni, F. (2014). 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.

[6] Bartolucci, F., Farcomeni, A., & Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. arXiv preprint arXiv:1003.2804.

[7] Bartolucci, F., Farcomeni, A., & Pennoni, F. (2012). A note on the application of the Oakes' identity to obtain the observed information matrix of hidden Markov models. arXiv preprint arXiv:1201.5990.

[8] Bartolucci, F., & Farcomeni, A. (2015). Information matrix for hidden Markov models with covariates. Statistics and Computing, 25(3), pp.515-526.

[9] Bierman, G. J., Thornton, C. L. (1977). Numerical comparison of Kalman filter algorithms: Orbit determination case study. Automatica, 13(1), pp.23-35.

[10] Biller, C. (1997). Posterior mode estimation in dynamic generalized linear mixed models. AStA Advances in Statistical Analysis, 1(85), pp.27-43.

[11] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), pp.29-51.

[12] Bousseboua, M. & Mesbah, M. (2013). Longitudinal Rasch Process with Memory Dependence. Pub.Inst. Stat.Univ.Paris, 57, fasc. 1-2, 2013, pp.45-58.

[13] Bousseboua, M., & Mesbah, M. (2010). Processus de markov longitudinal latent rasch observable. Publ Inst Stat Univ Paris, fasc, 1-2.

[14] Briegel, T., & Tresp, V. (1999). Fisher scoring and a mixture of modes approach for approximate inference and learning in nonlinear state space models. In Advances in Neural Information Processing Systems, pp. 403-409.

[15] Brockwell, P. J., Davis, R. A. (1991). Time series: data analysis and theory. Springer, New York.

[16] Brockwell, P. J., & Davis, R. A. (2013). Time series: theory and methods. Springer Science Business Media.

[17] Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting. springer.

[18] Casella, G., & Berger, R. L. (2002). Statistical inference. (Vol. 2). Pacific Grove, CA: Duxbury.

[19] Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. The Annals of Statistics, 32(6), pp.2385-2411.

[20] Creal, D. D. (2017). A Class of Non-Gaussian State Space Models with Exact Likelihood Inference. Journal of Business Economic Statistics, 1-13.

[21] Cox, D. R., & Hinkley, D. V. (1979). Theoretical statistics. CRC Press.

[22] Czado, C., Song, P. X. K. (2008). State space mixed models for longitudinal observations with binary and binomial responses. Statistical Papers, 49(4),pp. 691-714.

[23] Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep. net/stat/mlelr. pdf.

[24] Davis, R. A., Dunsmuir, W. T., Wang, Y. (1999). Modeling time series of count data. Statistics Textbooks and Monographs, 158,pp. 63-114.

[25] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal statistical Society, 39(1), pp.1-38.

[26] Douc, R., & Moulines, E. (2007). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. In ESAIM: Proceedings (Vol. 19, pp. 101-107). EDP Sciences.

[27] Douc, R., Moulines, E., & Olsson, J. (2009). Optimality of the auxiliary particle filter. Probability and Mathematical Statistics, 29(1), pp.1-28.

[28] Doucet, A., Godsill, S., Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and computing, 10(3), pp. 197-208.

[29] Doucet, A., De Freitas, N., Murphy, K., Russell, S. (2000, June). Rao-Blackwellised particle filtering for dynamic Bayesian networks. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence (pp. 176-183). Morgan Kaufmann Publishers Inc..

[30] Durbin, J., & Koopman, S. J. (2000). 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.

[31] Dunsmuir, W. T., Scott, D. J. (2015). The glarma package for observation driven time series regression of counts. Journal of Statistical Software, 67(7), pp.1-36.

[32] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society, pp.987-1007.

[33] Fahrmeir, L., & Kaufmann, H. (1991). On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. Metrika, 38(1), pp.37-60.

[34] Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. Journal of the American Statistical Association, 87(418), pp. 501-509.

[35] Fahrmeir, L., & Tutz, G. (2013). Multivariate statistical modelling based on generalized linear models. Springer Science & Business Media.

[36] Fahrmeir, L., & Wagenpfeil, S. (1997). Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. Computational Statistics Data Analysis, 24(3), pp.295-320.

[37] Feddag, M. L., & Mesbah, M. (2005). Generalized estimating equations for longitudinal mixed Rasch model. Journal of statistical planning and inference, 129(1), pp.159-179.

[38] Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta psychologica, 37(6), pp.359-374.

[39] Forthofer, R. (2012). Public program analysis: a new categorical data approach. Springer Science & Business Media.

[40] Godsill, S., Doucet, A., & West, M. (2001). Maximum a posteriori sequence estimation using Monte Carlo particle filters. Annals of the Institute of Statistical Mathematics, 53(1), pp.82-96.

[41] Gordon, N. J., Salmond, D. J., Smith, A. F. (1993, April). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In IEE Proceedings F (Radar and Signal Processing) (Vol. 140, No. 2, pp. 107-113). IET Digital Library.

[42] Hartigan, J. A. (1969). Linear bayesian methods. Journal of the Royal Statistical Society. Series B (Methodological), pp.446-454.

[43] astie, T., & Tibshirani, R. (1990). Generalized additive models. John Wiley & Sons, Inc..

[44] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97-109.

[45] Henson, M. A., & Seborg, D. E. (1997). Feedback linearizing control. Nonlinear process control, pp.149-231.

[46] Johansen, A. M., & Doucet, A. (2008). A note on auxiliary particle filters. Statistics & Probability Letters, 78(12), pp.1498-1504.

[47] Jones, R. H. (1993). Longitudinal data with serial correlation: a state-space approach. CRC Press.

[48] Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. Psychometrika, 43, pp.443-477.

[49] Kedem, B., & Fokianos, K. (2005). Regression models for time series analysis (Vol. 488). John Wiley Sons.

[50] Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. The Annals of Statistics, pp.79-98.

[51] Klein, B. M. (2003). State space models for exponential family data (Doctoral dissertation, Syddansk Universitet).

[52] Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian non-linear state space models. Journal of computational and graphical statistics, 5(1), pp. 1-25.

[53] Kitagawa, G. (2010). Introduction to time series modeling. CRC press.

[54] Kroese, D. P., Rubinstein, R. Y. (2012). Monte carlo methods. Wiley Interdisciplinary Reviews: Computational Statistics, 4(1), 48-58.

[55] Lehmann, E. L., Casella, G. (1998). Theory of Point Estimation, Springer-Verlag. New York.

[56] Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, pp.963-974.

[57] Laird, N. M., Beck, G. J., & Ware, J. H. (1991). Mixed models for serial categorical response. Quted in: Ekholm, A, 7.

[58] Liang-Qun, L., Hong-Bing, J., & Jun-Hui, L. (2005, October). The iterated extended Kalman particle filter. In Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on (Vol. 2, pp. 1213-1216). IEEE.

[59] Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. Journal of the Royal Statistical Society. Series B (Methodological), pp.3-40.

[60] Lindstrom, M. J., & Bates, D. M. (1988). Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association, 83(404), pp.1014-1022.

[61] Louis, T. A. (1988). General methods for analysing repeated measures. Statistics in Medicine, 7(12), pp.29-45.

[62] Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), pp.149-174.

[63] McCullagh, P., Nelder, J. A. (1989). Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability.

[64] Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. Psychometrika, 61(4), pp.629-645.

[65] Mesbah, M. (2010). Statistical quality of life. Encyclopedia of Statistical Sciences.

[66] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6), 1087-1092.

[67] Shumway, R. H., Stoffer, D. S. (1991). Dynamic linear models with switching. Journal of the American Statistical Association, 86(415), pp.763-769.

[68] So, M. K. (2003). Posterior mode estimation for nonlinear and non-Gaussian state space models. Statistica Sinica, pp.255-274.

[69] Morrell, D. (1997). Extended Kalman Filter Lecture Notes. EEE 581-Spring.

[70] Moustaki, I., & Knott, M. (2000). Generalized latent trait models. Psychometrika, 65(3), pp.391-411.

[71] Ngatchou-Wandji, J. (2008). Estimation in a class of nonlinear heteroscedastic time series models. Electronic Journal of Statistics, 2, pp.40-62.

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

[73] Poyiadjis, G., Doucet, A., & Singh, S. S. (2005). Maximum likelihood parameter estimation in general state-space models using particle methods. In Proc of the American Stat. Assoc.

[74] Poyiadjis, G., Doucet, A., & Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. Biometrika, 98(1), pp.65-80.

[75] Rotonda, C., Guillemin, F., Bonnetain, F., & Conroy, T. (2011). 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.

[76] Saha, S., Mandal, P. K., Bagchi, A., Boers, Y., & Driessen, H. (2008). On the Monte Carlo marginal MAP estimator for general state space models.

[77] Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. ETS Research Report Series.

[78] Shumway, R. H., & Stoffer, D. S. (2000). Time series analysis and its applications. Studies In Informatics And Control, 9(4), pp.375-376.

[79] Terejanu, G. A. (2008). Extended kalman filter tutorial. Online.

[80] Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49(4), pp.501-519.

[81] Tiao, G. C., & Tsay, R. S. (1989). Model specification in multivariate time series. Journal of the Royal Statistical Society. Series B (Methodological), pp.157-213.

[82] Tierney, L. (1994). Markov chains for exploring posterior distributions. the Annals of Statistics, 1701-1728.

[83] Tjϕstheim, D. (1986). Estimation in nonlinear time series models. Stochastic Processes and their Applications, 21(2), pp.251-273.

[84] Tjϕstheim, D. (1994). Non-linear time series: a selective review. Scandinavian Journal of Statistics, pp.97-130.

[85] Tong, H.(1990). Non-linear Time series. A Dynamic System Approach. Oxford: Oxford University Press.

[86] Tsay, R. S. (1984). Regression models with time series errors. Journal of the American Statistical Association, 79(385), pp.118-124.

[87] Turkman, K., Scotto, M. G., & de Zea Bermudez, P. (2014). Non-linear time series: extreme events and integer value problems. Springer.

[88] West, M., Harrison, P. J., & Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. Journal of the American Statistical Association, 80(389), pp.73-83.

[89] Whiteley, N., & Johansen, A. M. (2010). Recent developments in auxiliary particle filtering. Barber, Cemgil, and Chiappa, editors, Inference and Learning in Dynamic Models. Cambridge University Press, 38, pp.39-47.

[90] Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, pp.95-103.

[91] Xi, Y., Peng, H., Kitagawa, G., & Chen, X. (2015). The auxiliary iterated extended Kalman particle filter. Optimization and Engineering, 16(2), pp.387-407.

[92] Zeger, SL. (1987). The analysis of discrete longitudinal data: Commentary. Statistics in Medicine. 7,pp.161-168.

[93] Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, pp.121-130.

[94] Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, pp.1049-1060.

[95] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American statistical Association, 57(298), pp.348-368.