Latent State Modeling in Mobile Health and Diagnostic Classification: Recent Advances in the MCMC Approach

Tze Leung Lai, *Stanford University*

February 28, 2019

Outline

- Hidden Markov models (HMMs) and MCMC methods for
  - Latent state modeling in mHealth
  - Diagnostic classification models in psychometric testing
- A new MCMC approach: MCMC with state substitutions
- Applications to psychometric diagnostic classification
- Applications to adaptive filtering and control in mHealth

Latent State Modeling in mHealth and Diagnostic Classification

- Dempsey et al. (2017) use continuous-time HMMs to model time-to-event outcomes (e.g., lapse to alcohol or illicit drug use) from multimodal data (e.g., sensor streams along with self-report), with interpretable states representing behavioral constructs such as stress and craving, noisy measurements relating the observations to the states, and parameters capturing domain knowledge.
- DiBello, Henson and Stout (2015) introduce generalized diagnostic classification models for psychometric multiple-choice tests that replace 0/1 dichotomous scoring (wrong/right) by polytomous scoring to classify examinee's response strategies into G (guess) and C (cognitive response), including the option of incorrect/incomplete cognitive thinking or lack of certain skills from wrong answers using C. They use a Bayesian approach to model the latent states and parameters.

Particle Filters and Sequential Monte Carlo

- Let $\{X_t, t \geq 1\}$ be a Markov chain with latent/hidden states and let $Y_1, Y_2, \ldots$ be conditionally independent given $\{X_t, t \geq 1\}$, such that $X_t \sim p_t(\cdot|X_{t-1})$, $Y_t \sim g_t(\cdot|X_t)$ in which p_t and g_t are density functions with respect to measures ν_X and ν_Y. Gordon, Salmond and Smith (1993) introduced particle filters for estimating the hidden states in HMM, and the monograph of Jun Liu (2001) provides a collection of techniques that have been developed up to that time. Chan and Lai (2013) have developed a comprehensive methodology to compute $E[\psi(X_T)|Y_1, \ldots, Y_T]$, summarized in Algorithm 1 below.
- The density function $\widetilde{p}_T$ of $X_{1:T} = (X_1, \ldots, X_T)$ conditional on $Y_{1:T} = (Y_1, \ldots, Y_T)$ is $\widetilde{p}_T(x_{1:T}|Y_{1:T}) \propto \prod_{t=1}^{T}[p_t(x_t|x_{t-1})g_t(Y_t|x_t)]$.This conditional distribution is often difficult to sample from and the normalizing constant is also difficult to compute for high-dimensional or complicated state spaces, and particle filters use sequential Monte Carlo (SMC) that involves importance sampling and resampling to circumvent this difficulty. Let $X_{1:t-1}^m$ denote the sample path of the mth particle (trajectory), $1 \leq m \leq M$. The algorithm uses importance sampling from a proposal density q_t to circumvent this difficulty and updates not only the particles $X_{1:t-1}^m$ but also the associated weights w_{t-1}^m and ancestor A_{t-1}^m of X_t^m. It is initialized with $A_0^m = m$ and $w_0^m = 1/m$.

Algorithm 1: SMC updating procedure for M particles

1. Sample $(\widetilde{X}^m_{1:t-1}, A^m_t)$, $1 \leq m \leq M$, with replacement from $\{(X^m_{1:t-1}, A^m_{t-1}) : 1 \leq m \leq M\}$, using probability $w^m_{t-1} / \sum_{i=1}^M w^i_{t-1}$ for $(X^m_{1:t-1}, A^m_{t-1})$
2. Sample X^m_t from $q_t(\cdot|\widetilde{X}^m_{1:t-1})$, $1 \leq m \leq M$
3. Update $w^m_t = p_t(X^m_t|\widetilde{X}^m_{t-1}) g_t(Y_t|X^m_t) / q_t(X^m_t|\widetilde{X}^m_{1:t-1}), X^m_{1:t} = (X^m_{1:t-1}, X^m_t)$, $1 \leq m \leq M$

- Step 1 of the algorithm is often referred to as "bootstrap sampling", while step 2 as corresponds to importance sampling, yield the importance weight w^m_t.
- The SMC estimate of $\psi = E[\psi(X_T)|Y_1, \ldots, Y_T]$ is

$$\hat{\psi}_T = (M\bar{w}_T)^{-1} \sum_{m=1}^M w^m_T \psi(X^m_T), \text{ where } \bar{w}_T = m^{-1} \sum_{m=1}^M w^m_T.$$

- By using martingale theory, Chan and Lai (2013) have also shown under weak regularity conditions that $\sqrt{M}(\widehat{\psi}_T - \psi) \Rightarrow N(0, \sigma^2)$, with a consistent estimate of σ^2 that involves the ancestral origin A^m_{t-1}.

Adaptive Particle Filters

- In practice, the HMM has unknown parameters: $X_t \sim p_{t,\theta}(\cdot|X_{t-1})$, $Y_t \sim g_{t,\theta}(\cdot|X_t)$.
- Naive approach: Treat θ as a hidden state and proceed with usual particle filters, with multiple θ's generated within the same particle filter. However, repeated resampling of the static parameter results in a few distinct values of θ, i.e., degeneracy of particles.
- A standard Bayesian approach is to estimate θ by MCMC and then apply the estimate $\hat{\theta}$ to the particle filter, but this is computationally prohibitive.

Particle MCMC

The basic idea behind PMCMC (Andrieu, Doucet & Holenstein (2010)) is to use MCMC iterations to approximate the distribution of $(\theta, \mathcal{X}_T)$ given $\mathcal{Y}_T$.
At the k^{th} iteration

- Sample θ^* from proposal density $q(\theta^*|\theta^{(k-1)})$.
- Use SMC involving M particles to approximate $p_{\theta^*}(\mathcal{X}_T|\mathcal{Y}_T)$
- The approximating particle filter is used as proposal distribution for a Metropolis-Hastings update for $(\theta, \mathcal{X}_T)$ given $\mathcal{Y}_T$.

General PMCMC and Convergence Theory

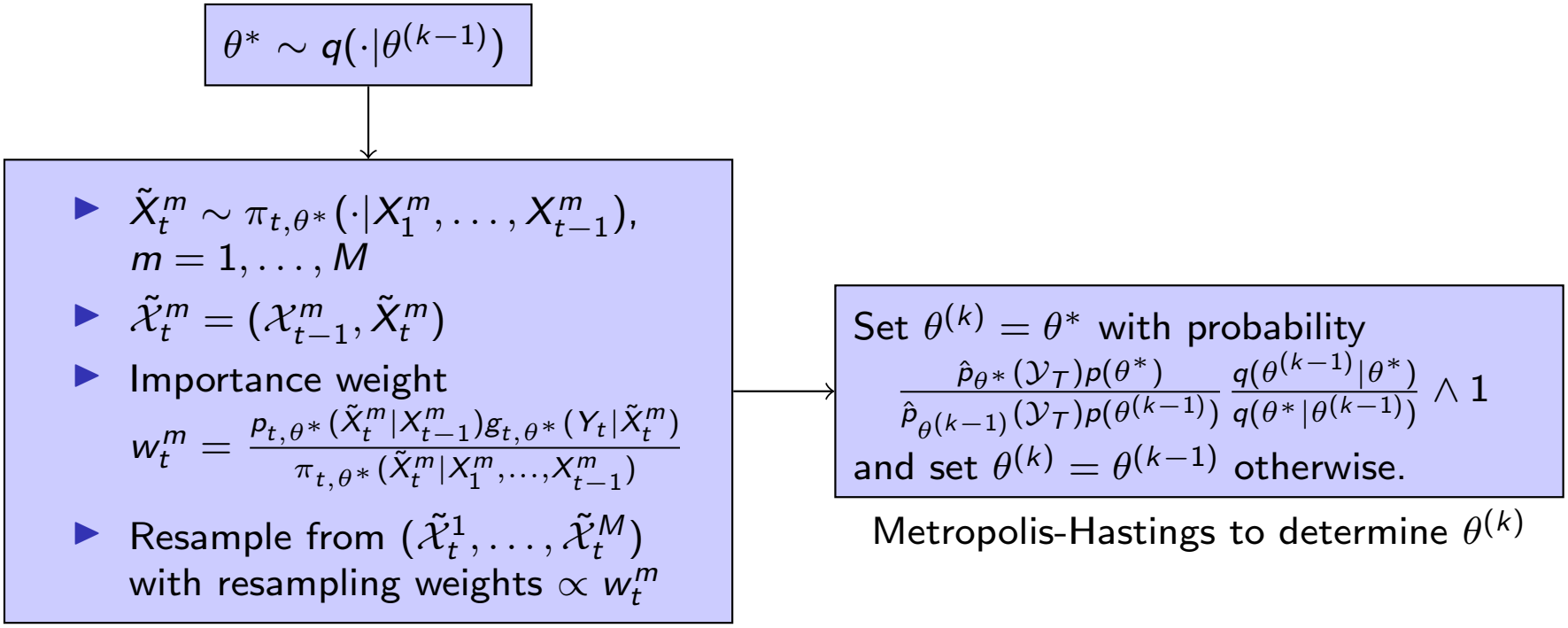
- Andrieu, Doucet & Holenstein also propose PMCMC to approximate a distribution with density $\pi(\theta, \mathcal{X}_T) = p(\theta)p_\theta(\mathcal{X}_T)$, for which SMC is used to approximate $p_{\theta^*}(\mathcal{X}_T)$ at the k^{th} iteration. Assuming that the Metropolis-Hastings sampler with proposal density $q(\theta^*|\theta)$ is irreducible and aperiodic and satisfies other regularity conditions, they show that the PMCMC sampler converges to π in total variation norm as $K \to \infty$, where K is the number of iterations.
- Besides Metropolis-Hastings, they also consider using the Gibbs sampler that leads to the particle Gibbs method, for which they prove similar results.

SMC^2 (Chopin, Jacob & Papaspiliopoulis, 2013)

This is a method for joint state and parametric estimation in HMM, which is a variant of PMCMC.

- To each of N "θ-particles" attach a particle filter that propagates N_x "X-particles". The θ-particles are generated by Chopin's (2002) iterated batch importance sampling (IBIS) algorithm, which is an SMC procedure to approximate the posterior distribution of θ given $\mathcal{Y}_T$, with MCMC (Metropolis-Hastings) steps to rejuvenate the current set of θ-particles when a degeneracy criterion is met.
- It is argued that "SMC^2 may offer several advantages over PMCMC" since it "calibrates automatically its tuning parameters", allowing N_x to change with θ. However, other than a couple of examples, there is no theory nor practical guideline on the choice of N_x.

Flow Chart of PMCMC



- Iterate till convergence in distribution, but don't know when and don't have *se* estimate of $\theta^{(k)}$ when termination occurs at k^{th} iteration.
- Even when the actual value of θ is known, variance of the state estimate is of order constant times M^{-1} (Chan & Lai, 2013). In ignorance of θ, PMCMC rejects many choices of θ^* in the MH iterations.
- Moreover, the difference between θ^* and $\theta^{(k)}$ introduces bias in state estimate.

A New MCMC Scheme

- PMCMC (or SMC^2) basically uses the likelihood of the parameter vector θ, computed by SMC, to update the posterior. This computation is expensive and therefore wasteful if a proposed move is not used in the MH step.
- We propose to approximate a target distribution of θ by a set of N "representative" atoms, chosen sequentially by an MCMC scheme so that the empirical distribution of the atoms converges weakly to the target distribution as $K(= \#\text{ of iterations}) \to \infty$.
- $S_k = \{\theta_k^1, \ldots, \theta_k^N\} \subset \Theta$ set of atoms chosen at the k^{th} iteration, $\Theta = \{\theta : p(\theta) > 0\}$, p: target density.
- Each MCMC iteration "substitutes" the rejected atoms by a new one according to an MH-type procedure.

Algorithm 2: MCMC-SS updating procedure for $\Theta_{b,k}$ from $\Theta_{b,k-1}$

---

1. Sample $\widetilde{\theta}$ from $q(\cdot|\gamma_{b,k-1})$ as candidate atom.
2. Let $\theta^b_{\nu+1,k-1} = \widetilde{\theta}$ and compute

$$\lambda^b_{i,k} = q(\theta^b_{i,k-1}|\gamma_{b,k-1})/f(\theta^b_{i,k-1}),\ i = 1,\ldots,\nu+1.$$

3. Sample $\widetilde{J}$ from $\{1,\ldots,\nu+1\}$ with probability $\pi^b_{i,k} = \lambda^b_{i,k} \Big/ \left(\sum_{j=1}^{\nu+1} \lambda^b_{j,k} \right)$ for i.
4. If $\widetilde{J} = \nu+1$, let $\Theta_{b,k} = \Theta_{b,k-1}$. Otherwise let $\Theta_{b,k} = (\Theta_{b,k-1} \cup \{\widetilde{\theta}\}) \setminus \{\theta^b_{\widetilde{J},k-1}\}$.

- Initialize by choosing $\gamma_0 \in \Gamma^o$ and generating νB i.i.d. $\theta^1_{1,0},\ldots,\theta^\nu_{1,0};\ldots;\theta^1_{B,0},\ldots,\theta^\nu_{B,0}$ from the proposal distribution $q(\cdot|\gamma)\,\mathrm{d}m$, thereby forming the B disjoint sets $\Theta_{b,0} = \{\theta^1_{b,0},\ldots,\theta^\nu_{b,0}\}$.
- At the kth iteration (stage), use the procedure to update atom set in bth block, $b = 1,\ldots,B$.

Central Limit Theorem and Consistent Estimator of σ^2

Define

$$\widehat{\psi}_{b,k} = \frac{1}{\nu} \sum_{\theta \in \Theta_{b,k}} \psi(\theta), \; \widehat{\psi} = \frac{1}{B(K-\kappa)} \sum_{b=1}^{B} \sum_{k=\kappa+1}^{K} \widehat{\psi}_{b,k},$$

$$\widehat{\sigma}^2 = \frac{1}{B(K-\kappa)} \sum_{b=1}^{B} \sum_{k=\kappa+1}^{K} \frac{1}{\nu - 1} \sum_{\theta \in \Theta_{b,k}} (\psi(\theta) - \widehat{\psi}_{b,k})^2,$$

where κ represents an initial burn-in period that is asymptotically negligible as $\kappa = o(K)$.

Main Theorem: Let $N = B(K-\kappa)$. Letting $\lambda(\theta|\gamma) = q(\theta|\gamma)/p(\theta)$, assume

$$\text{(C) } \alpha \leq \lambda(\theta|\gamma) \leq \beta \text{ for all } \theta \in \Theta, \gamma \in \Gamma, \text{ and some } \beta > \alpha > 0.$$

Then as $K \to \infty$ and $B \to \infty$ such that $B = \mathcal{O}(K)$,

$$\sqrt{N}(\widehat{\psi} - \mu) \Rightarrow N(0, \sigma^2),$$

where $\sigma^2 = \nu^{-1}\mathrm{Var}_p(\psi(\theta))$, in which Var_p denotes the variance when θ has density function p with respect to m. Moreover, σ^2 can be consistently estimated by $\widehat{\sigma}^2$.

Choice of $q(\cdot|\gamma)$ to satisfy (C) and $\gamma_{b,k-1}$

- Suppose the target density is a member of a parametric family $q(\cdot|\gamma)$ with $p(\cdot) = q(\cdot|\gamma(p))$, where $\gamma : \mathcal{P} \to \Gamma$, $\mathcal{P}$ is the space of probability measures on Θ, and Γ is a compact, convex measurable subset of $\mathbb{R}^d$, with $q(\cdot|\gamma)$ satisfying

  (C1) $\sup_{\theta\in\Theta,\gamma\in\Gamma^o} \|\nabla_\gamma q(\cdot|\gamma)\|_1 / V(\theta) < \infty$,
  (C2) $q(\theta|\gamma+\delta) - q(\theta|\gamma) - \delta^\top \nabla_\gamma q(\theta|\gamma) = O(\|\delta\|_1^2 V(\theta))$

  uniformly over $\theta \in \Theta$, $\gamma \in \Gamma^o$ and $\gamma + \delta \in \Gamma$, where $\|\mathbf{x}\|_1 = \sum |x_i|$ denotes the ℓ_1-norm of $\mathbf{x} \in \mathbb{R}^d$. Then (C) is satisfied.
- Define $\gamma_{b,k-1}$ to be the average of $\{\gamma(\theta) : \theta \in \Theta_{b,k-1}\}$ that involves the transitions in the bth block up to stage $k-1$, for $k \le \kappa$ (before burn-in period). On the other hand, for $k > \kappa$, pool across blocks to estimate $\gamma(p)$ consistently, i.e., let $\tilde{\gamma}_{k-1} = B^{-1} \sum_{b=1}^{B} \gamma_{b,k-1}$ and use it as the modified $\gamma_{b,k-1}$ for all blocks.

Oracle Property of MCMC-SS

- Since $\tilde{\gamma}_{k-1} = \gamma(p) + o_p(1)$, with probability approaching 1 for large k, the candidate atom $\widetilde{\theta}$ substitutes some existing atom in $\Theta_{b,k-1}$.
- For large k, the conditional density function of $\Theta_{b,k}$ given $\Theta_{b,k-1}$ behaves like the ν-fold product density of p on Θ^{ν}, suggesting the asymptotic optimality of $\widehat{\psi}$.
- The limiting variance σ^2 has the weight $1/\nu$ for the contribution of each chosen atom for a block, hence each random variable generated in the MCMC-SS scheme asymptotically contributes weight $(N\nu)^{-1}$ to (a) the estimate $\widehat{\psi}$ of μ and (b) the asymptotic variance of $\widehat{\psi}$.
- There is considerable flexibility in the choice of K and B in $N = B(K - \kappa)$ that determines the scaling factor in the CLT. The theorem highlights the case $B = \mathcal{O}(K)$ just to emphasize that K should not be chosen too small relative to B.

Case $p \notin \{q(\cdot|\gamma) : \gamma \in \Gamma\}$

- Let $I(q\|p) = \mathbb{E}_p\{\log(q(\theta)/p(\theta))\}$ be the Kullback-Leibler divergence and let $q_\gamma = q(\cdot|\gamma)$, $\widetilde{\gamma}(p) = \text{argmin}_{\gamma\in\Gamma} I(q_\gamma\|p)$. Let $L(\theta) = q_{\widetilde{\gamma}(p)}(\theta)/p(\theta)$. Let $w_{b,k}(\theta) = 1/\pi^b_{i,k}$ (defined in Step 3 of Updating Table) with $\theta = \theta^b_{i,k}$. Let

$$\widetilde{\psi}_{b,k} = \frac{1}{\nu}\left(\sum_{\theta\in\Theta_{b,k}} w_{b,k}(\theta)\psi(\theta)\right) \Big/ \left(\sum_{\theta\in\Theta_{b,k}} w_{b,k}(\theta)\right),$$

$$\widetilde{\psi} = \frac{1}{N}\sum_{b=1}^{B}\sum_{k=\kappa+1}^{K} \widetilde{\psi}_{b,k}.$$

- Since $q(\theta|\gamma)/p(\theta) = \left\{q_\gamma(\theta)/q_{\widetilde{\gamma}(p)}(\theta)\right\} L(\theta)$, application of main theorem shows that as $K \to \infty$ and $B \to \infty$ such that $B = \mathcal{O}(K)$, $\sqrt{N}(\widetilde{\psi} - \mu) \Rightarrow N(0, \sigma^2)$, where $\mu = \mathbb{E}_p(\psi(\theta))$, and σ^2 can be consistently estimated by

$$\widetilde{\sigma}^2 = \frac{1}{N}\sum_{b=1}^{B}\sum_{k=\kappa+1}^{K} \frac{1}{\nu-1}\sum_{\theta\in\Theta_{b,k}} \left(\psi(\theta) - \widetilde{\psi}_{b,k}\right)^2.$$

Joint State and Parameter Estimation in Adaptive Particle Filters

- Fix a block, and denote for notational simplicity $\Theta_{b,k}$ (with b fixed) by Θ_k. Let $\theta \in \Theta_k$ be the vector (X_k, γ_k) consisting of the hidden state and sequential estimate of the parameter. MCMC-SS represents the posterior distribution of X_k (respectively, γ_k) through the empirical measure of M particles $X_{1:k}$ (respectively, N atoms) in Algorithm 2. The basic idea is to first approximate the posterior distribution of $\gamma(p)$ (in the case $p(\cdot) = q(\cdot|\gamma(p))$) or of $\widetilde{\gamma}(p)$ (if $p \notin \{q(\cdot|\gamma) : \gamma \in \Gamma\}$), and then use this posterior distribution to generate the particles by an obvious modification of Algorithm 1.
- The proof of the Main Theorem on Algorithm 2 uses exponential bounds on the total variation norm of the difference between the empirical measure and the target measure, under condition (C). We can relax (C) to accommodate the case where the exponential convergence rate depends on the starting value by using a weighted total variation norm, as in V-uniform ergodicity that generalizes uniform ergodicity in the theory of Markov chains.

Group Sequential Parameter Updates and Oracle Property

- The parameter updates should be carried out at stages $k_1 < \cdots < k_{\tau-1}$, with $k_\tau = K$, which is similar to group sequential (instead of fully sequential) methods in sequential analysis. Not only does it reduce the computation task, but it is also much more stable prior to the burn-in period. In fact, for particle filters in HMMs with known parameters, occasional resampling is preferable to resampling at every stage because it is computationally and also statistically (in the sense of variance reduction) more efficient; see Chan and Lai (2013).
- The adaptive particle filter has the same asymptotic property as the oracle particle filter for which the posterior distribution of γ is known at every stage. Moreover, unlike PMCMC or SMC^2, it is recursive, which is a major advantage of Kalman filters for linear state-space models.

Generalized Diagnostic Classification in Psychometric Testing

- For the generalized diagnostic classification models (GDCM) described in the third slide, there is a far-reaching generalization/integration of item response theory and restricted latent class analysis in psychometric testing. It involves a "Q matrix" that has a row per item for each response option, and uses a Bayesian approach to model the examinee's latent states, which it has implemented by a Metropolis-Hastings within Gibbs (MHG) sampling algorithm.
- Due to the large computational resources needed, Nvidia has agreed to provide UIUC's project with GPU hardware and also software assistance, hoping to use new computational power to solve these problems.
- The automatic MCMC-SS procedure opens up new possibilities to use these computational resources.

Continuous-time HMM in iSurvive

- The continuous-time HMM that Dempsey et al. (2017) use to model the time to event (e.g., survival time) via a Cox model with time-varying covariates X_t for the hazard function, as mentioned in the third slide, is called "iSurvive" and consists of:
  - Latent states $X_t = (X_t^1 \dots X_t^d)$ as the covariates
  - Multimodal observation process $O_j(t)$, $1 \leq j \leq J$ labeling the data stream, is related to X_t via the generalized linear model with mean $E(O_j(t)) = g^{-1}(\phi_0 + \phi_1^T X_t)$
  - Markov jump process dynamics for $X_t : X_t^k$ assumes a finite set (with cardinality n_k) of values. Letting $n = n_1 \dots n_d$, the jump process X_t has transition rate matrix $Q = (q_{ij})$ such that $q_i = -q_{ii}$ is the rate at which the process leaves state j after an exponentially distributed time with parameter q_i, and jumps to state j with probability q_{ij}/q_i.
- The model parameters are those for the Cox model and the generalized linear model.

Activity Recognition from Accelerometer Data

- Jiawei Bai's talk on Tuesday has given a review of accelerometer data and activity recognition (resting, walking, etc.) from these data in various projects at John Hopkins in which he is involved.
- Since the activity states are latent, HMMs have been used for dynamic diagnostic classification with uncertainty quantification, as an approach to activity quantification from accelerometer data (Ganti et al., 2006; Zappi et al., 2008; Lee and Cho, 2011).

Mobile Cloud Telemedicine

- "Mobile devices, due to their limited storage space and computational capability, might not be able to perform tasks which require intensive computing and extensive storage resources. On the other hand, cloud computing, because of its unique features like elasticity and scalability in both computation ability and storage space, can be combined with mobile computing to provide ubiquitous and personalized healthcare to patients." (Wang, Su and Jin, 2017; Wang et al., 2013; Jin and Chen, 2015).
- Wang, Xu and Jin (2017) develop a dynamic scheduling approach that uses HMM to optimize synergistically the mobile–cloud–based telemedicine applications toward multiple objectives: high accuracy, low latency, and long battery life.