Notes

November 17, 2023

1 Methodology

1.1 Discrete Nonhomogeneous Semi-Markov Process

Let $\mathcal{S}$ denote a discrete state set. Map its element to integers in $[\mathcal{S}] = [m] := \{1, 2, \cdots, m\}$. Denote a sequence of r.v. take value in $[m]$ as $\{X_t, t \in \mathbb{N}\}$ representing the state of system at time point t. Denote $T_0 = 0$, $T_n = min\{t, X_t \neq X_{T_{n-1}}\}, n \geq 1$. and $J_n = X_{T_n}, \tau_n = T_n - T_{n-1}$. We call $\{X_t, t \in \mathbb{N}\}$ a discrete-time multi-state process. And J_n and T_n are nth state and corresponding transition time respectively. In the context of hypnogram modeling, we always assume there is a deterministic initial state $J_0 = s_0$. We also assume there is an absorbing state denoted as s_a.

The semi-markov property holds if and only if,

$$Pr\{J_n = j, \tau_n = t | J_{n-1} = i, T_{n-1}, ..., T_1, J_0\} = Pr\{J_n = j, \tau_n = t | J_{n-1} = i, T_{n-1}\}, \forall n. \tag{1.1}$$

Further, if the joint conditional probability of J_n, τ_n given J_{n-1}, T_{n-1} is the same for any n, we call $\{X_t, t \in \mathbb{N}\}$ a **discrete nonhomogeneous semi-markov process(dnsmp)**.

Due to the semi-markov property, it's straightforward to verify that $\{X_t, t \in \mathbb{N}\}$ can to specified by the conditional joint distribution of J_n, τ_n given J_{n-1}, T_{n-1} which is denoted as,

$$p(j, t | i, T_{n-1}) := Pr(J_n = j, \tau_n = t | J_{n-1} = i, T_{n-1}). \tag{1.2}$$

An alternative to specify the process is by transition intensity function defined as,

$$\lambda_{ij}(t | T_{n-1}) := Pr(\tau_n = t, J_n = j | J_{n-1} = i, \tau_n \geq t, T_{n-1}).$$

It's straightforward to verify that the transition intensity function can derive $p(j, t|i, T_{h-1})$. Dfine the conditional survival function of τ_n as

$$S_i(t | T_{n-1}) = Pr(\tau_n \geq t | j_{n-1} = i, T_{n-1}).$$

Then we have $p(j, t|i, T_{n-1}) = \lambda_{ij}(t|T_{n-1}) S_i(t|T_{n-1})$.

1.2 Multinomial Representation

1.2.1 Multinomial Reponse as Member of Multivariate Exponential Family and VGM

Suppose $\widetilde{\mathbf{y}}^T = (y_1, y_2, ..., y_m) \sim M(1, \lambda_1, \lambda_2, \cdots, 1 - \sum_{k=1}^{m-1} \lambda_k,)$. Denote $\lambda_m = 1 - \sum_{k=1}^{m-1} \lambda_k$. The distribution function of $\widetilde{\mathbf{y}}$ is,

$$\begin{aligned} f(\widetilde{\mathbf{y}}) &= \prod_{k=1}^{m} (\lambda_k)^{y_k} \\ &= exp\big[\sum_{k=1}^{m-1} y_k log(\lambda_k) + (1 - \sum_{k=1}^{m-1} y_k) log(\lambda_m) \big] \\ &= exp\big[\sum_{k=1}^{m-1} y_k log(\frac{\lambda_k}{\lambda_m}) + log(\lambda_m) \big] \end{aligned}$$

Define $\mathbf{y}^T = (y_1, \cdots, y_{m-1})$, it belongs to the exponential family since,

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\theta}) &= exp[\mathbf{y}^T \boldsymbol{\theta} + c(\boldsymbol{\theta})], \\ \text{where } \boldsymbol{\theta}^T &= (log(\frac{\lambda_1}{\lambda_m}), \cdots, log(\frac{\lambda_{m-1}}{\lambda_m})), \end{aligned}$$

here $\boldsymbol{\theta}$ is called natural parameter, and c is a function of natural parameter in the context of exponential family. The conditionial mean of $\mathbf{y}$ is $\boldsymbol{\mu} = (\lambda_1, \cdots, \lambda_{m-1})^T$. As long as we define the predictor part $\boldsymbol{\eta} = (\eta_1(\mathbf{x}), \cdots, \eta_{m-1}(\mathbf{x}))$, and find a link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$, we complete the parametrization of the model.

We can see that it differs from the GLM and GAM in that $\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\mu}$ are all multivariate. We generally call it **Vector Generalized Model(VGM)** for we cannot decide yet whether the predictor part η is linear w.r.t $\mathbf{x}$.

1.2.2 Multinomial Representation of Discrete SMP

Consider **one day's** sleep trajectory $\{J_n, T_n\}_{n=0}^N$. Define $\tau_n = T_n - T_{n-1}$. The likelihood is,

$$\begin{aligned} \mathcal{L} = \prod_{n=1}^{N} p(J_n, \tau_n | T_{n-1}, J_{n-1}) &= \prod_{n=1}^{N} \lambda_{J_{n-1}J_n}(\tau_n | T_{n-1}) S_{J_{n-1}}(\tau_n | T_{n-1}) \\ &= \prod_{n=1}^{N} \lambda_{J_{n-1}J_n}(\tau_n | T_{n-1}) \prod_{t \le \tau_n - 1} [1 - \sum_{k \neq J_{n-1}} \lambda_{J_{n-1}k}(t | T_{n-1})]. \end{aligned}$$

The second equation holds due to the semi-markov property. And the third equation holds due to the relationship between survival and hazard function.

Now define Y_{nt} as,

$$Y_{nt} = \begin{cases} J_n & \text{if } \tau_{n+1} > t, \\ k & \text{if } \tau_{n+1} = t \text{ and } J_{n+1} = k. \end{cases}$$

Define $\mathbf{y}_{nt}$ to be the one-hot encoder of Y_{nt} where,

$$\mathbf{y}_{nt}^T = (y_{nt0}, y_{nt1}, \ldots, y_{ntk}, \ldots, y_{ntm}) = (0, \ldots, 1, \ldots, 0),$$

where $y_{ntk} = 1$ for $Y_{nt} = k$.

Using $\mathbf{y}_{nt}$, we can rewrite the likelihood as

$$\mathcal{L} = \prod_{n=1}^{N}\prod_{t=1}^{\tau_n}\Big\{[\prod_{\substack{k=1\\k\neq J_{n-1}}}^{m}\lambda_{J_{n-1}k}(t|T_{n-1})^{y_{n-1tk}}][1-\sum_{\substack{k=1\\k\neq J_{n-1}}}^{m}\lambda_{J_{n-1}k}(t|T_n)]^{y_{n-1t0}}\Big\}$$

The final step is to swap the order of multiplication. We introduce indicator $\mathbf{1}_{J_n=i}$ and I_i represents the index set of all $n \in \{0,1,\cdots,N-1\}$ which satisfies that $J_n = i$. Then write likelihood as $\mathcal{L} = \prod_{i\in[m]}\mathcal{L}_i$,where

$$\begin{aligned}
\mathcal{L}_i &= \prod_{n=1}^{N}\Big\{\prod_{t=1}^{\tau_n}[\prod_{\substack{k=1\\k\neq J_{n-1}}}^{m}\lambda_{ik}(t|T_{n-1})^{y_{n-1tk}}][1-\sum_{\substack{k=1\\k\neq J_{n-1}}}^{m}\lambda_{J_{n-1}k}(t|T_{n-1})]^{y_{n-1t0}}\Big\}^{\mathbf{1}_{J_{n-1}=i}}\\
&= \prod_{n\in I_i}\mathcal{L}_n^i\\
\mathcal{L}_n^i &= \prod_{t=1}^{\tau_{n+1}}\Big\{[\prod_{\substack{k=1\\k\neq i}}^{m}\lambda_{ik}(t|T_n)^{y_{ntk}}][1-\sum_{\substack{k=1\\k\neq i}}^{m}\lambda_{ik}(t|T_n)]^{y_{nti}}\Big\}. \qquad (1.3)
\end{aligned}$$

Firstly, since the likelihood can be decomposed into different $\mathbf{L}_i$, and each $\mathbf{L}_i$ will not contain joint parameters as we illustrate later. We can deal with them seperately.

Furthermore, We can see that for each $n \in I$, equation 1.3 is the same as the likelihood for the τ_{n+1} observations $\mathbf{y}_{n1},\ldots\mathbf{y}_{n\tau_{n+1}}$ of a multinomial response model. The indicator y_{ntk} variables actually represent the distributions given that a specific epoch(t) is reached. Given that after system reaches $J_n = i$ and exactly t epoch has passed, then the response is multinomially distributed with $\mathbf{y}_{nt}^T = (y_{nt1},\ldots,y_{nti},\ldots,y_{ntm}) \sim M\big(1,\lambda_{i1}(t|T_n),\ldots,1-\sum_{\substack{k=1\\k\neq i}}\lambda_{ik}(t|T_n),\ldots,\lambda_{im}(t|T_n)\big)$.

In conclusion, now for each i we have independent response variable $\{Y_{nt}\}_{n\in I,t\in[\tau_{n+1}]}$ or equally $\{\mathbf{y}_{nt}\}_{n\in I,t\in[\tau_{n+1}]}$. And its corresponding covariate(or predictor) is t and T_n as illustrated in equation 1.3. We use Vector Generalized Model in section 1.2.1 to model the regression relationship between Y_{nt} and (t,T_n).

1.2.3 Vector Generalized Semi-parametric Parametrization

To simplify notation, we only show how we parameterize samples from I_m. Now we parameterize the model by designing the predictor part $\boldsymbol{\eta}_{nt}^m = (\eta_1^m(t,T_n),\ldots,\eta_{m-1}^m(t,T_n))^T$. The most generalized way is to assume η_k^m is an unspecified function of t and T_n. However, it's redundant, especially for t to perceive it as a continuous variable thus incorporating nolinear component w.r.t. t. Notice in reality, t represents the number of episodes system stays after it jumps to a certain state, therefore t only takes a finite number of values. We regulate the model and assume support of τ_n is finite $\{1,2,\ldots,r\}$. Therefore we can perceive t as a categorical variable and further one-hot-encode it as $\mathbf{t} \in \{0,1\}^{r-1}$.

For T_n, there's no need to assume linearity for η_k^m w.r.t. T_n, and we incorporate nonlinear part into it. Finally, we parameterize the predictor part as,

$$\eta_k^m(t,T_n) := \mathbf{t}^T\boldsymbol{\beta}_k^m + f_k^m(T_n), k = 1,\ldots,m-1$$

The predictor is partially linear w.r.t. the covariate $\mathbf{x}_{nt} = (\mathbf{t}^T,T_n)^T$, and is semi-parametric. So we classify our model in the literature of **Vector Generalized Semi-parametric(or partially linear) model**.

Finally, corresponding the standard VGM modeling introduced in section 1.2.1, we only need to specify the link function which is the canonical link,

$$
\begin{aligned}
g &= h^{-1} \\
h(\boldsymbol{\eta}) &= (h_1(\boldsymbol{\eta}), \cdots, h_{m-1}(\boldsymbol{\eta}))^T \\
h_k(\boldsymbol{\eta}) &= \frac{exp(\eta_k)}{1 + \sum_{j=1}^{m-1} exp(\eta_j)}
\end{aligned}
$$

We summarise our model as follows, for all samples $\{Y_{nt}, \mathbf{x}_{nt}\}_{n \in I_i, t \in [\tau_{n+1}]}$, we have,

$$
\begin{aligned}
&Pr(Y_{nt} = k) = \begin{cases} \lambda_{ik}(t|T_n), & \textbf{if } k \neq i \\ 1 - \sum_{j=1}^{m} \lambda_{ij}(t|T_n), k = i \end{cases}, \ \textbf{if } t = 1, 2, \cdots, \tau_n \\
&\lambda_{ik}(t|T_n) = h_k(\boldsymbol{\eta}_{nt}^i), \\
&\boldsymbol{\eta}_{nt}^i = \left(\eta_1^i(\mathbf{t}, T_n), \ldots, \eta_{i-1}^i(\mathbf{t}, T_n), \eta_{i+1}^i(\mathbf{t}, T_n), \ldots, \eta_{m-1}^i(\mathbf{t}, T_n)\right)^T \\
&\eta_k^i(\mathbf{t}, T_n) = \mathbf{t}^T \boldsymbol{\beta}_k^i + f_k^i(T_n), k \neq i \\
&h_k(\boldsymbol{\eta}_{nt}^i) = \frac{exp[\eta_k^i(\mathbf{t}, T_n)]}{1 + \sum_{j=1, j \neq i}^{m} exp[(\eta_j^i(\mathbf{t}, T_n)]}.
\end{aligned} \tag{1.4}
$$

The log-likelihood of our model for each starting state i is,

$$
l_i = \sum_{n \in I_i} \sum_{t=1}^{\tau_{n+1}} \Big\{ \sum_{\substack{k=1 \\ k \neq i}}^{m} y_{ntk} [log[h_k(\boldsymbol{\eta}_{nt}^i)] + y_{nti} log(1 - \sum_{\substack{j=1 \\ j \neq i}}^{m} h_j(\boldsymbol{\eta}_{nt}^i)] \Big\}.
$$

1.2.4 Discussion

We categorize our model into **Vector Generalized Semi-parametric Model**. In statistical literature, generalized linear model is a widely used model for regression of non-gaussian response variables. Soon the model is extended to the setting where covariates have nonlinear effect and generalized additive model(GAM) is brought up as a simple way to deal with smoothing multiple predictors. However, **our model is between them for we both have linear and nonlinear component**. In statistical literature, it is mostly refered to generalized partially nonlinear model or generalized semi-parametric model.

Another thing to notice is that most of the GLM or GAM or GPLM literature considers the case when the response variable is univariate, however, a multinomial distribution is a special case in exponential family because its response is multivariate. Therefore, our model **also differs in that the reponse is a vector**.

1.3 Estimation

Since we introduce nonlinear(or nonparametric) part in equation 1.4, we have to consider how to estimate $\mathbf{f}_k^i$(here i is the subscript for different starting state and k is the state system jumps to). When estimating functions, it involves literature of **smoothing**. From my knowledge, there are

Here, we talk about two broad categories of smoothers:

- Regression or series smoothers(regression spline, polynomial series).
- Smoothing spline.

1.3.1 Regression Spline

In the first category, we show how to estimate the model through regression spline, which is also the simplest way in three categories.

To the best of our knowledge, regression spline is simply approximating f_k^i with a linear combination of spline bases $\sum \gamma_s B_s()$. We simply evaluate $B_s()$ in given points and incorporate them into design matrix. In conclusion, estimation procedure using regression is the same as Vector GLM.

For vector GLM, the estimation of parameters is through Iteratively Reweighted Least Square(IRLS)

1.3.2 Smoothing Spline

Different from regression spline, smoothing spline add smoothing penalty to the loss function and estimate the coefficient.

1.3.3 IRLS computation

1.4 Change-point Estimation

References