Appendix to to "Multivariate discrete hidden Markov models for domain-based measurements and assessment of risk factors in child development" published in the Journal of Computation and Graphical Statistics

Qiang Zhang, Alison Snow Jones, Frank Rijmen, Edward H. Ip

January 20, 2010

1. Estimation of GLM

First, group all parameters into a single vector $\boldsymbol{\beta} = (\phi, \xi, \gamma, \sigma)$, where $\phi = (\phi_{jk})$, and $\xi = (\xi_{d(j)sk})$. Also, group each $\mathcal{Y}_{i\boldsymbol{I}}$ (defined in Lemma 1) accordingly to form a "new" $\mathcal{Y}_{i\boldsymbol{I}}$. Under the unified model, the conditional outcome parameter $\psi_{i\boldsymbol{I}} = \pi_{ijskl}$ in Lemma 1 is now a function of β. The transition component and the prior marginal component remain the same as in Lemma 1. The estimate for $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_i \sum_{\boldsymbol{I}} \mathcal{Y}_{i\boldsymbol{I}} \log(\psi_{i\boldsymbol{I}}(\beta)).$$

Denote the design matrix by $\boldsymbol{X}$. To illustrate how $\boldsymbol{X}$ can be constructed, we use the following example with the number of items being $J = 4$, the number of states being $S = 2$, the number of domains being $D = 2$, the number of predictor being $p = 2$, and the number of response categories being $K_j \equiv 2$ for all items. We further assume that the first two items are in the first domain and that the remaining two are in the second domain. The design matrix $\boldsymbol{X}$ is given by:

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N)', \\ \boldsymbol{X}_i &= \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & x_{i1} & x_{i2} & v_1 \\ 0 & 1 & 0 & 0 & 1 & 0 & x_{i1} & x_{i2} & v_1 \\ 0 & 0 & 1 & 0 & 0 & 1 & x_{i1} & x_{i2} & v_1 \\ 0 & 0 & 0 & 1 & 0 & 1 & x_{i1} & x_{i2} & v_1 \\ 1 & 0 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & v_1 \\ 0 & 1 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & v_1 \\ 0 & 0 & 1 & 0 & 0 & 0 & x_{i1} & x_{i2} & v_1 \\ 0 & 0 & 0 & 1 & 0 & 0 & x_{i1} & x_{i2} & v_1 \\ \ldots & & & & & & & & \\ 1 & 0 & 0 & 0 & 1 & 0 & x_{i1} & x_{i2} & v_q \\ 0 & 1 & 0 & 0 & 1 & 0 & x_{i1} & x_{i2} & v_q \\ 0 & 0 & 1 & 0 & 0 & 1 & x_{i1} & x_{i2} & v_q \\ 0 & 0 & 0 & 1 & 0 & 1 & x_{i1} & x_{i2} & v_q \\ 1 & 0 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & v_q \\ 0 & 1 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & v_q \\ 0 & 0 & 1 & 0 & 0 & 0 & x_{i1} & x_{i2} & v_q \\ 0 & 0 & 0 & 1 & 0 & 0 & x_{i1} & x_{i2} & v_q \end{pmatrix}; \end{aligned}$$

$\boldsymbol{\beta} = (\phi_{11}, \ldots, \phi_{41};\ \xi_{domain=1,s=1}, \xi_{domain=2,s=1}, \gamma_1, \gamma_2, \sigma)'$, x_{i1}, x_{i2} are the covariates of each subject; and $v_1, \ldots v_q$ are the quadrature points of $N(0,1)$. This technique for estimating random effects using Gauss-Hermite is discussed in Hinde (1982), and Fahrmeir and Tutz (1994, Section 7.4). Note that the s index in $\xi_{d(j)sk}$ stops at $S-1$ to ensure the full column rank of $\boldsymbol{X}$.

The Fisher scoring iteration uses the following general updating equation,

$$\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} + (\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{\beta}^{(n)})\boldsymbol{X})^{-1} s(\boldsymbol{\beta}^{(n)}),$$

where the weight matrix $\boldsymbol{W}(\beta)$ and the score function $s(\beta)$ are respectively,

$$W(\boldsymbol{\beta}) = \boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{V}^{-1}(\boldsymbol{\beta})D'(\boldsymbol{\beta}), \text{ and } s(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})).$$

Here, $\boldsymbol{D}$ is the derivative of the inverse link function – i.e. $\boldsymbol{D} = \partial h^{-1}(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$, where $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ is the linear predicator, $\boldsymbol{\mu}(\boldsymbol{\beta})$ is the mean structure $h^{-1}(\boldsymbol{X}\boldsymbol{\beta})$, $\boldsymbol{Y} = \text{vec}(\mathcal{Y}_{i\boldsymbol{I}})$, and $\boldsymbol{V}$ is the variance function. The specific forms of the derivative matrix $\boldsymbol{D}$ and the variance function $\boldsymbol{V}$ for the unified model will be provided below.

Matrices $\boldsymbol{W}$, $\boldsymbol{V}$, and $\boldsymbol{D}$, all take a block diagonal form – i.e.,

$$\boldsymbol{W} = diag(\boldsymbol{W}_m), \boldsymbol{V} = diag(\boldsymbol{V}_m), \boldsymbol{D} = diag(\boldsymbol{D}_m), m = 1, \ldots, M,$$

and $M = N \times S \times J \times q$. Each block of $\boldsymbol{W}_m$ is a $(K_j - 1) \times (K_j - 1)$ matrix, given by $\boldsymbol{D}_m \boldsymbol{V}_m^{-1} \boldsymbol{D}'_m$ (Fahrmeir and Tutz 1994, pp. 39-40).

Using the cumulative logit link function (14), the inverse link $h^{-1}(.)$ is:

$$\begin{aligned} \zeta_{ijs1l} = h_1^{-1}(\eta_{ijs1l}) &= \frac{e^{\eta_{ijs1l}}}{1 + e^{\eta_{ijs1l}}} \\ \zeta_{ijskl} = h_k^{-1}(\eta_{ijskl}) &= \frac{e^{\sum_{k'=1}^{k} e^{\eta_{ijsk'l}}}}{1 + e^{\sum_{k'=1}^{k} e^{\eta_{ijsk'l}}}}, k = 2, \ldots, K_j - 1, \end{aligned}$$

and thus $\boldsymbol{D}_m \in \mathbb{R}^{(K_j-1)\times(K_j-1)}$ is given by:

$$(\boldsymbol{D}_m)_{kk'} = \left(\frac{\partial h_k^{-1}}{\partial \eta_{ijsk'l}}\right) = \begin{cases} \zeta_{ijs1l}(1 - \zeta_{ijs1l}), & k = k' = 1, \\ e^{\eta_{ijskl}} \zeta_{ijskl}(1 - \zeta_{ijskl}), & k = k' > 1, \\ e^{\eta_{ijsk'l}} \zeta_{ijskl}(1 - \zeta_{ijskl}), & k > k', \\ 0, & k < k'. \end{cases}$$

The variance function $\boldsymbol{V}$ for the above mixed effect regression model on the conditional probabilties of the HMM (Section 3.4.1 of Fahrmeir and Tutz 1994) is given by:

$$\boldsymbol{V}_m = \frac{1}{\sum_t \sum_k \delta(y_{itj}, k) \tilde{\alpha}_{itsl}^{(n)}} [diag(\boldsymbol{\pi}_{ijsl}) - \boldsymbol{\pi}_{isjl} \boldsymbol{\pi}'_{ijsl}],$$

where $\boldsymbol{\pi}_{ijsl} = (\pi_{ijskl}) \in [0, 1]^{K_j - 1}$.

Because the EM algorithm does not directly operate on the marginal likelihood, it does not provide the observed information matrix necessary for computing standard errors. Several methods have been proposed to calculate standard errors for the HMM. Lystig and Hughes (2002) provide an overview and also propose an efficient method based upon a foward-backward algorithm. We have implemented two methods for computing standard errors. The first method was based upon the conditional expectation of the score function of the complete data, which can be shown to equal to the score function of the observed data (e.g., Meilijson 1989). Numerical differentiation was then used to compute the second derivative of the expected score function. The second method, described in Meilijson (1989), was based upon the sum of the outer product of the individual contributions to the score function. Both methods have led to similar standard error estimates for our application, and here we only report results from using the second method (see also Friedl and Kauermann 2000).

2. Proof of Lemma 2.

Start from the Laplace approximation of the Bayes factor in (26) in the original article. Take the logarithm of (26) and multiply by -2, the expression becomes a sum of four terms as follows:

$$-2\log p(\boldsymbol{Y}|H) = -d\log(2\pi) - 2\log|\Sigma|^{1/2} - 2\log p(\boldsymbol{Y}|\hat{\beta}, H) - 2\log(f(\hat{\beta}|H)), \tag{1}$$

where the estimate of the parameter vector $\hat{\beta} = (\hat{\alpha_1}, \hat{\tau}, \hat{\pi})$.

The third term, $-2\log p(\boldsymbol{Y}|\hat{\beta}, H)$, is the deviance. The second term involves the likelihood function, which can be used to directly derive $|\Sigma|$, the determinant of the covariance matrix $\Sigma = [-\boldsymbol{D}^2 l(\hat{\beta})]^{-1}$. By the definition of $l(\beta)$, we can separate it into two parts:

$$l(\hat{\beta}) = \log(p(\boldsymbol{Y}|\hat{\beta}, H) f(\hat{\beta}|H)) = \log(p(\boldsymbol{Y}|\hat{\beta}, H)) + \log(f(\hat{\beta}|H)),$$

where the first part is the log likelihood of observed data and the second part is the log likelihood of estimated parameters.

We specify independent prior distributions for each set of parameters, α_{1s}, τ_{rs}, and π_{sjk}, with the count parameter $\boldsymbol{\mu}$ respectively highlighted using the superscripts (α), (τ), and (π), respectively. Specifically, we have

$$\begin{aligned}
f_\alpha(\alpha_{11}, \ldots, \alpha_{1,S-1}; \mu_1^{(\alpha)}, \ldots, \mu_S^{(\alpha)}) &= \frac{1}{B(\boldsymbol{\mu}^{(\alpha)})} \prod_{s=1}^{S} \alpha_{1s}^{\mu_s^{(\alpha)}-1}, \\
f_\tau(\tau_{r1}, \ldots, \tau_{r,S-1}; \mu_1^{(\tau)}, \ldots, \mu_S^{(\tau)}) &= \frac{1}{B(\boldsymbol{\mu}^{(\tau)})} \prod_{s=1}^{S} \tau_{rs}^{\mu_s^{(\tau)}-1}, \\
f_\pi(\pi_{sj1}, \ldots, \pi_{sj,K-1}; \mu_1^{(\pi)}, \ldots, \mu_K^{(\pi)}) &= \frac{1}{B(\boldsymbol{\mu}^{(\pi)})} \prod_{k=1}^{K} \pi_{sjk}^{\mu_s^{(\pi)}-1},
\end{aligned} \tag{2}$$

where $B(.)$ is the beta function. The overall prior function is therefore

$$f(\beta; \boldsymbol{\mu}_\alpha, \boldsymbol{\mu}_\tau, \boldsymbol{\mu}_\pi | H) = f_\alpha \prod_{r=1}^{S} f_\tau \prod_{s=1}^{S} \prod_{j=1}^{J} f_\pi .$$

By choosing $\boldsymbol{\mu} = \mathbf{1}$ (Scott, James, and Sugar 2005) for each individual distribution function, the RHS of (2) simplifies to $1/B(\boldsymbol{\mu})$, and since the prior distribution is independent of β, the second derivative of $\log(f(\hat{\beta}|H))$ would be zero. Thus we can cancel out the second term in the RHS of (1) and equate Σ with $\tilde{\Sigma}$.

For the last term of (1), we use the definition of B:

$$B(\boldsymbol{\mu}^{(\alpha)}) = B(\boldsymbol{\mu}^{(\tau)}) = \frac{\prod_{i=1}^{S} \Gamma(\mu_i)}{\Gamma(\sum_{i=1}^{S} \mu_i)} = \frac{1}{\Gamma(S)} \text{ and } B(\boldsymbol{\mu}^{(\pi)}) = \frac{1}{\Gamma(K)}.$$

It follows that the explicit form for the prior is

$$-2 \log f(\beta|H) = 2(S+1)\log(\Gamma(S)) + 2SJ \log \Gamma(K). \tag{3}$$

Replace the last term on the RHS of (1) with the two terms on the RHS of (3), we complete the proof.

3. Instructions on Using the Computer Codes

We provide the MATLAB codes and the subset of NLSY data in a zip file, which is downloadable from the link: http://www.phs.wfubmc.edu/public/downloads/MHMM.zip. The purpose of the Supplementary Materials section is to encourage open scientific evaluation and discussion, and we have no intention to use it as a means to distribute the NLSY data set. Upon registration, the full NLSY data set is available at the website: http://www.bls.gov/nls/. Also, to run the codes, the Bayesian Network Toolbox (BNT) needs to be installed and can be downloaded at http://people.cs.ubc.ca/ murphyk/Software/BNT/bnt.html

Once the zip file has been downloaded, follow the steps:

1. Unzip the zip file.

2. In MATLAB, set the path to the installed BNT directory.

3. Open ”runIt.m” in MATLAB.

4. Change the two boolean variables, “doFixedEffect” and “doRandomEffect”, to the specified model. Model I corresponds to both variables specified as 0, while the user should specify “doFixedEffect” as 0 and “doRandomEffect” as 1 in Model II, “doFixedEffect” as 1 and “doRandomEffect” as 0 in Model III and both variables as 1 in Model IV.

5. Run the matlab program and wait until the log likelihood converges. The computation time might be in the order of hours for the mixed-effects model and the results might slightly vary from the one published, given the different initial values. Multiple runs are suggested by changing the value of variable ”R” and also experimenting with different initial values of β is strongly encouraged.