HMMs for Complex Biologging Data

CANSSI Collaborative Research Team Project #22

Day 2

Arturo Esquivel Robert Zimmerman

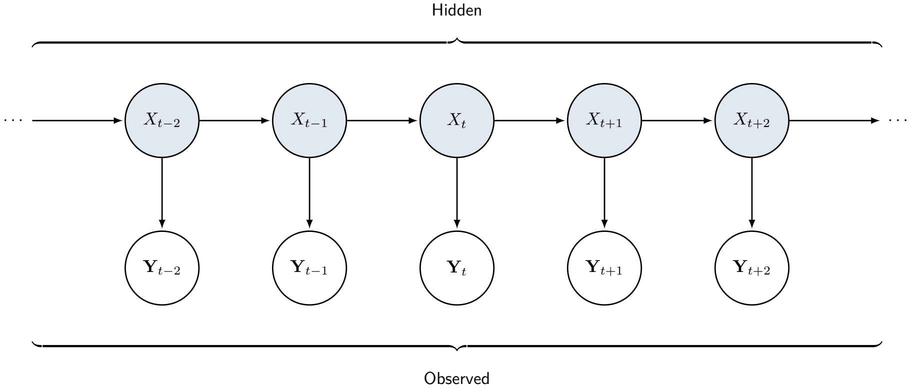
University of Toronto

November 9, 2022

Day 1 Recap

- Finite mixture models
- Markov chains
- Hidden Markov models
- Forward and backward variables
- Likelihood computations

Day 1 Recap


Hidden
$\cdots$
X_{t-2}
X_{t-1}
X_t
X_{t+1}
X_{t+2}
$\cdots$
$\mathbf{Y}_{t-2}$
$\mathbf{Y}_{t-1}$
$\mathbf{Y}_t$
$\mathbf{Y}_{t+1}$
$\mathbf{Y}_{t+2}$
Observed

Decoding

- Once the parameters $\boldsymbol{\eta}$ of the model have been estimated, we can do several things with the estimates

- While the estimates provide information about the data-generating process, they can also help us determine the states $x_{1:T}$ which were most likely to give rise to the observed data $Y_{1:T}$

- Classifying (or *decoding*) the unknown state sequence $X_{1:T}$ can be accomplished using several algorithms

Local and Global Decoding

- *Local decoding* refers to the classification of each X_t individually by setting

$$\hat{X}_t = \underset{x \in \mathcal{X}}{\text{argmax}} \, \mathbb{P}_{\hat{\boldsymbol{\eta}}} \left(X_t = x \mid Y_{1:T} = y_{1:T} \right), \quad t = 1, \ldots, T$$

- In words, we choose the state that maximizes the *a posteriori* state membership probability

- In contrast, *global decoding* classifies the entire state sequence at once:

$$\widehat{X_{1:T}} = \underset{x_{1:T} \in \mathcal{X}^T}{\text{argmax}} \, \mathbb{P}_{\hat{\boldsymbol{\eta}}} \left(X_{1:T} = x_{1:T} \mid Y_{1:T} = y_{1:T} \right)$$

- Here, we focus on local decoding (in practice, both methods tend to produce similar classifications)

Interlude: Backward Variables

- Yesterday, we introduced the *forward variables* $\alpha_{x,t} = f_{\hat{\boldsymbol{\eta}}}\left(Y_{1:t} = y_{1:t}, X_t = x\right)$

- Now, we will also need the *backward variables* $\beta_{x,t} = f_{\hat{\boldsymbol{\eta}}}\left(Y_{(t+1):T} = y_{(t+1):T} \mid X_t = x\right)$

- As with the forward variables, the vectors of backward variables $\boldsymbol{\beta}_t = (\beta_{1,t}, \ldots, \beta_{K,t})^\top$ can be computed recursively in polynomial time via dynamic programming

- In particular, if

$$\mathbf{P}(y) = \begin{pmatrix} f_1(y) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_K(y) \end{pmatrix},$$

  then it is easily shown that $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\mathbf{P}(y_{t+1})\boldsymbol{\beta}_{t+1}$ for $t = 1, 2, \ldots, T-1$

Local Decoding

- Local decoding is accomplished using forward variables

$$\alpha_{x,t} = f_{\hat{\boldsymbol{\eta}}}\left(Y_{1:t} = y_{1:t},\, X_t = x\right)$$

and backward variables

$$\beta_{x,t} = f_{\hat{\boldsymbol{\eta}}}\left(Y_{(t+1):T} = y_{(t+1):T} \mid X_t = x\right)$$

- Since

$$\alpha_{x,t} \cdot \beta_{x,t} = f_{\hat{\boldsymbol{\eta}}}\left(Y_{1:T} = y_{1:T},\, X_t = x\right),$$

Bayes' rule yields

$$\mathbb{P}_{\hat{\boldsymbol{\eta}}}\left(X_t = x \mid Y_{1:T} = y_{1:T}\right) = \frac{f_{\hat{\boldsymbol{\eta}}}\left(Y_{1:T} = y_{1:T},\, X_t = x\right)}{f_{\hat{\boldsymbol{\eta}}}\left(Y_{1:T} = y_{1:T}\right)} = \frac{\alpha_{x,t} \cdot \beta_{x,t}}{L_T}$$

Local Decoding

- Local decoding thus sets

$$\hat{X}_t = \underset{x \in \mathcal{X}}{\text{argmax}} \left(\frac{\alpha_{x,t} \cdot \beta_{x,t}}{L_T} \right) = \underset{x \in \mathcal{X}}{\text{argmax}} \left(\alpha_{x,t} \cdot \beta_{x,t} \right), \quad t = 1, \ldots, T$$

- The computation of these conditional probabilities via forward and backward variables is known as the *forward-backward algorithm*

Covariates in State Dependent Distributions

- The basic HMM may be too simplistic a model for certain applications
- Occasionally, we might want certain parameters in the model to depend on covariates (for example, an animal's sex, weight, age, etc.)
- For example, the state-dependent mean θ_x might depend linearly on some fixed vector $\mathbf{z} \in \mathbb{R}^p$, perhaps through some link function g :

  $$g(\theta_x) = g\left(\mathbb{E}\left[Y_t \mid X_t = x\right]\right) = \boldsymbol{\beta}_x^\top \mathbf{z},$$

  where $\boldsymbol{\beta}_x^\top = (\beta_{x,1}, \ldots, \beta_{x,p})$ is a vector of regression coefficients
- In other words, each state-dependent distribution carries its own generalized linear model

Covariates in Transition Probabilities

- Alternatively, we may incorporate covariates into each of the $K \cdot (K-1)$ transition probabilities

- This is typically accomplished by applying a multinomial logistic regression model to each row of the transition matrix:

$$\gamma_{j,x} = \mathbb{P}_{\boldsymbol{\eta}}\left(X_t = x \mid X_{t-1} = j\right) = \frac{e^{\boldsymbol{\beta}_{x|j}^{\top}\mathbf{z}}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_{k|j}^{\top}\mathbf{z}}}, \quad x, j \in \mathcal{X}$$

with $\boldsymbol{\beta}_{K|j} = \mathbf{0}$ for all $j \in \mathcal{X}$

More on Covariates

- In either case, the $\boldsymbol{\beta}_x$'s and/or $\boldsymbol{\beta}_{x|j}$'s are incorporated into the likelihood function and inference proceeds as usual
- We might also want to include covariates $\mathbf{z}_t$ that depend on time (for example, $\mathbf{z}_t$ could include the number of hours an animal has been awake at time t)
- In this case, inference proceeds in a similar fashion; however...
- Including time-varying covariates in the transition probabilities $\gamma_{j,x}$ destroys the assumption of time homogeneity, so each of the initial probabilities $\pi_x = \mathbb{P}_{\boldsymbol{\eta}}\left(X_1 = x\right)$ must also be estimated

Mixed HMMs

- We may have *multiple* time series — say S of them — available for inference
- When the time series are believed to be iid, they can be pooled together in a straightforward manner
- More realistically, the S time series are not iid, but still arise from HMMs with common features (such as the same underlying set of states $\mathcal{X}$)
- When the time series arise from the same parametric model (but with series-specific parameters), there can be up to $S \cdot \text{length}(\boldsymbol{\eta})$ parameters to estimate, which is cumbersome
- For example, there would be S state-dependent parameters for state j: $\theta_{j,1}, \ldots, \theta_{j,S}$

Random Effects

- Instead, one could regard the $\theta_{j,s}$'s as continuous random variables: $\theta_{j,1}, \ldots, \theta_{j,S} \overset{iid}{\sim} g_{\boldsymbol{\sigma}_j}$
- That is, each $\theta_{j,s}$ is a *random effect* with distribution $g_{\boldsymbol{\sigma}_j}$
- Each inclusion of such a random effect in the model reduces the number of parameters to estimate by $S - \text{length}(\boldsymbol{\sigma}_j)$
- The drawback, however, is that the $\theta_{j,s}$ must be integrated out of the likelihood:

$$\mathcal{L}(\ldots, \boldsymbol{\sigma}_j) = \int \cdots \int \mathcal{L}(\ldots, \theta_{j,1}, \ldots, \theta_{j,S}) \prod_{s=1}^{S} \left(g_{\boldsymbol{\sigma}_j}(\theta_{j,s}) \, \mathrm{d}\theta_{j,s} \right)$$

Discrete Random Effects

- Even for the simplest distributions $g_{\boldsymbol{\sigma}_j}$, such integrals are never available in closed form and must be computed numerically (which is difficult in high dimensions)

- Alternatively, one can assume the $\theta_{j,s}$'s to be *discrete* random variables on a finite sample space $\mathcal{M}$

- This makes for a much simpler likelihood computation:

$$\mathcal{L}(\ldots,\boldsymbol{\sigma}_j) = \sum_{s=1}^{S} \sum_{m\in\mathcal{M}} \mathcal{L}(\boldsymbol{\eta},\theta_{j,1},\ldots,\theta_{j,S}) \cdot \mathbb{P}_{\boldsymbol{\sigma}_j}\left(\theta_{j,s}=m\right)$$

- However, the applicability of such models may be limited

- The same ideas can be extended to dependent random effects, in which two or more parameters in the model follow a joint distribution

Multivariate Observations

- Until now we have assumed that each Y_t is a random variable
- However, everything discussed so far applies verbatim if the observations are d-dimensional *random vectors* $\mathbf{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})$
- An often-used simplifying assumption is that of *contemporaneous conditional independence*:

$$\mathbf{Y}_t \mid (X_t = x) \sim f_x(\mathbf{y}) = \prod_{h=1}^{d} f_{x,h}(y_h)$$

- In other words, the components of $\mathbf{Y}_t$ are assumed to be independent

Multivariate Observations

- The assumption of contemporaneous conditional independence makes inference almost as easy as that for univariate HMMs
- However, it is sometimes too strong (for example, occupancy)
- In such cases, one can choose a non-factorial multivariate distribution to better model the dependence between the components of $\mathbf{Y}_t$
- The drawback is that for most such distributions, inference can be challenging
- Parameter estimates are available in closed form for the multivariate normal distribution, but little else
- One can use copulas to model arbitrarily complex dependence structures, although parameter estimation then requires new techniques [Zimmerman et al., 2022]

Bayesian Inference

- One can also perform Bayesian inference on HMMs
- To do so, one must choose an appropriate prior distribution $\pi(\boldsymbol{\eta})$ for the unknown parameters of the model
- The rows of the transition matrix $\boldsymbol{\Gamma}_k$ and the initial distribution $\boldsymbol{\pi}$ are traditionally assigned Dirichlet priors (which are conjugate to the multinomial distribution)
- Priors for the parameters θ_x of the state-dependent distributions are chosen on a case-by-case basis

Bayesian Inference

- The posterior distribution

$$\pi(\boldsymbol{\eta} \mid y_{1:T}) \propto \pi(\boldsymbol{\eta}) \cdot \mathcal{L}(\boldsymbol{\eta})$$

  is never available in closed form and is impossible to sample from directly

- Thus, Markov chain Monte Carlo (MCMC) methods are typically required to sample from it

- A popular choice of MCMC method for HMMs is Hamiltonian Monte Carlo (or variants thereof), as implemented in the Stan programming language

- Although written in C++, Stan has an R interface which is accessed through the `rstan` library

Quantifying Uncertainty

- As in all statistical inference, it is always of interest to quantify uncertainty in estimates of unknown parameters
- For frequentist inference, asymptotic normality of the MLE has been proven under mild regularity conditions [Bickel et al., 1998]
- The observed information matrix — which itself is a consistent estimator of the Fisher information — can be approximated numerically, and this yields standard errors and confidence intervals for parameter estimates
- In the Bayesian setup, credible intervals can be obtained from posterior distributions using standard techniques

Thank you!

References

Bickel, P. J., Ritov, Y., and Ryden, T. (1998).
Asymptotic normality of the maximum-likelihood estimator for general hidden markov models.
The Annals of Statistics, 26(4):1614–1635.

Zimmerman, R., Craiu, R. V., and Leos-Barajas, V. (2022).
Copula modelling of serially correlated multivariate data with hidden structures.
arXiv preprint arXiv:2207.04127.