The categorical distribution in hmmTMB

Théo Michelot

2023-02-08

```
# Load package
library(hmmTMB)
# Set random seed for reproducibility
set.seed(2901)
```

This vignette focuses on using the categorical distribution for the observation process of a hidden Markov model (HMM) in the package hmmTMB. Use of this distribution is a little different from others. If you are new to hmmTMB, a better starting point would be the general vignette 'Analysing time series data with hidden Markov models in hmmTMB'.

Statistical background

We consider an observation process (Z_t) where, for any time step t, Z_t can take one of K discrete values ($Z_t \in \{1, 2, \dots, K\}$). We refer to these values as "categories", and note that they are not quantitative.

We model Z_t using a state-dependent categorical distribution (where the state is determined by a Markov chain, following the usual HMM formulation). In state $j \in \{1, 2, \dots, J\}$, the observation process is

$$Z_t \mid \{S_t = j\} \sim \text{Cat}(p_{1j}, p_{2j}, \dots, p_{Kj}),$$

where the parameter $p_{kj} \in [0, 1]$ is the probability of category k in state j. That is, we have $\Pr(Z_t = k \mid S_t = j) = p_{kj}$. By definition, the probabilities must sum to one (because Z_t has to be in one of the categories), i.e., $\sum_{k=1}^{K} p_{kj} = 1$ in each state.

This formulation is useful in cases where the proportion of each category is expected to change through time, which can be captured by the HMM.

Analysis in hmmTMB

The categorical distribution is called `cat` in hmmTMB, and here we demonstrate the simple case of a model with two states ($K = 2$) and three categories ($J = 3$). For this example, we use simulated data, i.e., the true parameters are known.

Simulating data

We simulate data from a 2-state categorical HMM with transition probability matrix

$$\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.03 & 0.97 \end{pmatrix}$$

and with observation parameters $(p_{11}, p_{21}, p_{31}) = (0.1, 0.2, 0.7)$ in state 1, and $(p_{12}, p_{22}, p_{32}) = (0.4, 0.5, 0.1)$ in state 2. We label the three categories as `group1`, `group2`, and `group3`.

```r
# True transition probability matrix (for simulation)
tpm <- matrix(c(0.9, 0.1,
                0.03, 0.97),
              nrow = 2, byrow = TRUE)
# True observation parameters (for simulation)
p <- matrix(c(0.1, 0.2, 0.7,
              0.4, 0.5, 0.1),
            nrow = 2, byrow = TRUE)
cats <- c("group1", "group2", "group3")

# Setup simulation
n <- 1e4
s <- rep(1, n)
x <- rep(NA, n)
# Initial observation
x[1] <- sample(cats, size = 1, prob = p[s[1],])
# Loop over time steps
for(i in 2:n) {
  # Sample new state
  s[i] <- sample(1:2, size = 1, prob = tpm[s[i-1],])
  # Sample new observation
  x[i] <- sample(cats, size = 1, prob = p[s[i],])
}
# Put observations into data frame for hmmTMB
```

```r
data <- data.frame(x = x)

head(data)
```

```
       x
1 group3
2 group2
3 group3
4 group3
5 group3
6 group3
```

In this example, the observations are character strings (`"group1"`, `"group2"`, `"group3"`), but they could also be factors, or integers. The package automatically transforms them to integers between 1 and the number of categories.

Model fitting

Model specification in hmmTMB requires creating two objects, from the `MarkovChain` and `Observation` classes, respectively, for the hidden state and the observation model. The observation model takes a few arguments, including `dists` (list of observation distributions) and `par` (initial observation parameters). The categorical distribution is called `"cat"`, so this is what we pass for `dists`. The initial parameters are starting values for the state-dependent category probabilities. Here, there are three categories in the data, and this is automatically detected by the package. We only need to pass two probabilities for each state, because the third one can be deduced using $\sum_{k=1}^{K} p_{kj} = 1$. Specifically, hmmTMB uses the first category as the reference, i.e., the category for which the probability is not estimated (but instead derived from the others). So, we pass initial values for p_{21}, p_{31}, p_{22}, and p_{32}.

```r
# Initial parameters
par0 <- list(x = list(p2 = c(0.1, 0.4), p3 = c(0.5, 0.1)))

# Create observation model
obs <- Observation$new(data = data, dists = list(x = "cat"),
                       n_states = 2, par = par0)
```

We can now combine this `Observation` object with a `MarkovChain` object to create an HMM, and fit it to obtain parameter estimates.

```r
# Create hidden state model
hid <- MarkovChain$new(data = data, n_states = 2)

# Create HMM
hmm <- HMM$new(obs = obs, hid = hid)

# Fit HMM
hmm$fit(silent = TRUE)

# Look at parameter estimates
hmm$par()
```

```
$obspar
, , 1

       state 1   state 2
x.p2 0.1865736 0.5020485
x.p3 0.7061396 0.1064130


$tpm
, , 1

           state 1    state 2
state 1 0.90954956 0.09045044
state 2 0.02703292 0.97296708
```

We can compare the estimated parameters to the values used for simulation, to check that the estimation worked well.