The categorical distribution in hmmTMB

Théo Michelot

2024-04-17

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'.

1 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

$$Z_t \mid \{S_t = j\} \sim \text{Cat}(p_1^{(j)}, p_2^{(j)}, \dots, p_K^{(j)}),$$

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

2 Analysis in hmmTMB

The categorical distribution is called `cat` in hmmTMB, and here we demonstrate its use on a data set available in the mHMMbayes R package, on the non-verbal communication of patients and therapists (Aarts (2023)). This data set includes four categorical variables,

describing the verbalising and looking behaviour of both the patients and therapists at a 1-sec resolution. For more details about the variables, please consult the vignette "Multilevel HMM tutorial" for the package mHMMbayes, available at https://CRAN.R-project.org/package=mHMMbayes/vignettes/tutorial-mhmm.html, which presents an HMM analysis of this data set.

2.1 Data

We load the data set from mHMMbayes. It is already in the format required by hmmTMB, except the identifier column is called `id` (rather than `ID` as required by hmmTMB).

```
data("nonverbal", package = "mHMMbayes")
data <- as.data.frame(nonverbal)

# Rename ID column as required by hmmTMB
colnames(data)[1] <- "ID"
data$ID <- factor(data$ID)

head(data)
```

```
  ID p_vocalizing p_looking t_vocalizing t_looking
1  1            2         2            2         2
2  1            2         2            1         2
3  1            2         2            2         2
4  1            2         2            2         2
5  1            2         2            2         2
6  1            2         2            1         2
```

The data set includes four categorical variables, and the interpretation of each value is described in the mHMMbayes documentation (`?mHMMbayes::nonverbal`):

- `p_vocalizing`, $V_t^P \in \{1, 2, 3\}$
- `t_vocalizing`, $V_t^T \in \{1, 2, 3\}$
- `p_looking`, $L_t^P \in \{1, 2\}$
- `t_looking`, $L_t^T \in \{1, 2\}$

We want to fit the following model:

$$
\begin{aligned}
V_t^P \mid S_t = j &\sim \text{Cat}(p_{11}^{(j)}, p_{12}^{(j)}, p_{13}^{(j)}) \\
V_t^T \mid S_t = j &\sim \text{Cat}(p_{21}^{(j)}, p_{22}^{(j)}, p_{23}^{(j)}) \\
L_t^P \mid S_t = j &\sim \text{Cat}(p_{31}^{(j)}, p_{32}^{(j)}) \\
L_t^T \mid S_t = j &\sim \text{Cat}(p_{41}^{(j)}, p_{42}^{(j)})
\end{aligned}
$$

where $p_{ik}^{(j)}$ is the probability that variable i takes the value k in state j. For each i, j, we have the constraint that $\sum_k p_{ik}^{(j)} = 1$.

2.2 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. To create the hidden state model, we specify that we want to use two states, we include a random effect for `ID`, i.e., transition probabilities include a random intercept for each patient-therapist pair, and we indicate that the initial state distribution should be fixed to the stationary of the process rather than estimated (to reduce numerical instability).

```
# Load package
library(hmmTMB)

hid <- MarkovChain$new(data = data,
                       n_states = 2,
                       formula = ~ s(ID, bs = "re"),
                       initial_state = "stationary")
```

The observation model takes a few arguments, including `dists` (list of observation distributions) and `par` (starting values for 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.

The number of categories for each variable is automatically detected by the package. For `p_vocalizing` and `t_vocalizing` (which have three categories), we only need to pass two probabilities for each state, because the third one can be deduced using $\sum_k p_{ik}^{(j)} = 1$. Likewise, there is only one parameter for each state for `p_looking` and `t_looking`. 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). Expected parameter names for this distribution take the form `p2` (probability of category 2), `p3` (probability of category 3), etc.

```
# List of observation variables and distributions
dists <- list(p_vocalizing = "cat",
              t_vocalizing = "cat",
              p_looking = "cat",
              t_looking = "cat")

# List of initial parameter values for categorical probabilities
par0 <- list(p_vocalizing = list(p2 = c(0.3, 0.3),
                                 p3 = c(0.3, 0.3)),
             t_vocalizing = list(p2 = c(0.3, 0.3),
                                 p3 = c(0.3, 0.3)),
             p_looking = list(p2 = c(0.5, 0.5)),
             t_looking = list(p2 = c(0.5, 0.5)))

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

We can now combine the two model components to specify and fit the full HMM. Model fitting took around 30 seconds on a laptop.

```
hmm <- HMM$new(obs = obs, hid = hid)
hmm$fit(silent = TRUE)
```

2.3 Results

2.3.1 Observation parameters

We can extract all estimated observation parameters using `obs$par()` or `obs$par_alt()`. The latter will automatically compute the probabilities for the reference categories based on the estimated parameters.

```
# Observation parameters
round(as.data.frame(obs$par_alt("p_vocalizing")), 3)
```

```
      S1    S2
p1 0.865 0.008
p2 0.000 0.963
```

```
p3 0.135 0.028
```

```
round(as.data.frame(obs$par_alt("t_vocalizing")), 3)
```

```
      S1    S2
p1 0.062 0.711
p2 0.918 0.183
p3 0.020 0.106
```

```
round(as.data.frame(obs$par_alt("p_looking")), 3)
```

```
      S1    S2
p1 0.108 0.257
p2 0.892 0.743
```

```
round(as.data.frame(obs$par_alt("t_looking")), 3)
```

```
      S1    S2
p1 0.272 0.075
p2 0.728 0.925
```

We can read these results as follows:

- In State 1, `p_vocalizing` is most likely to take the value 1 (patient is not vocalizing), `t_vocalizing` is most likely to take the value 2 (therapist is vocalizing), `p_looking` is very likely to take the value 2 (patient is looking at therapist) and `t_looking` is somewhat likely to take the value 2 (therapist is looking at patient).
- In State 2, `p_vocalizing` is most likely to take the value 2 (patient is vocalizing), `t_vocalizing` is most likely to take the value 1 (therapist is not vocalizing), `p_looking` is somewhat likely to take the value 2 (patient is looking at therapist) and `t_looking` is very likely to take the value 2 (therapist is looking at patient).

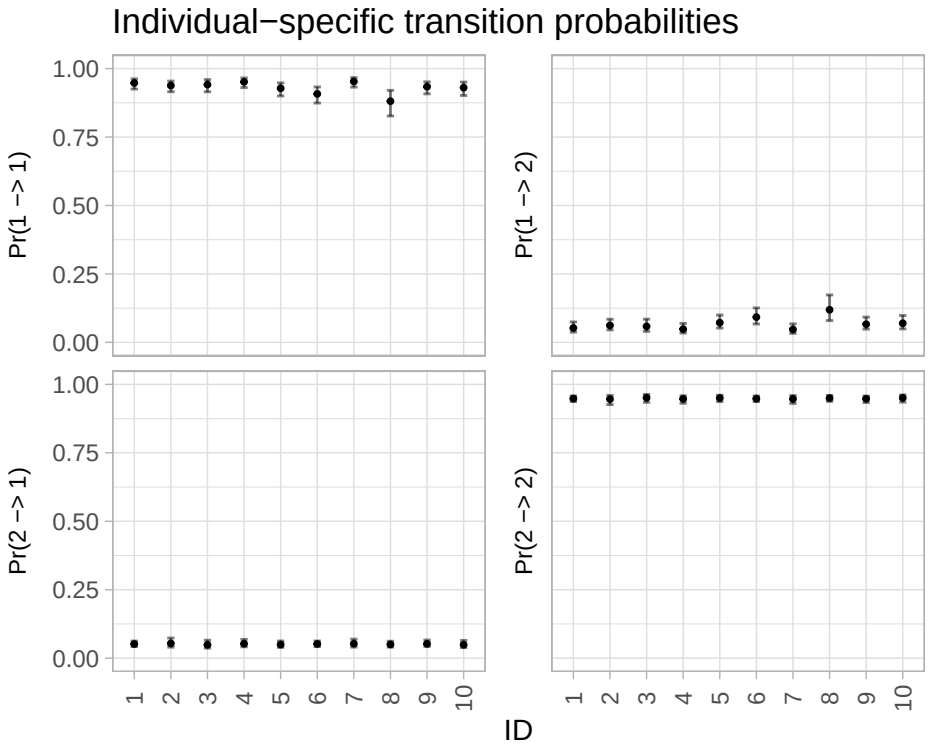
These are very similar to the two states identified in the mHMMbayes vignette "Multilevel HMM tutorial", and the authors conclude that "The resulting model indicates 2 well separated states: one in which the patient is speaking and one in which the therapist is speaking."

2.3.2 State process parameters and random effects

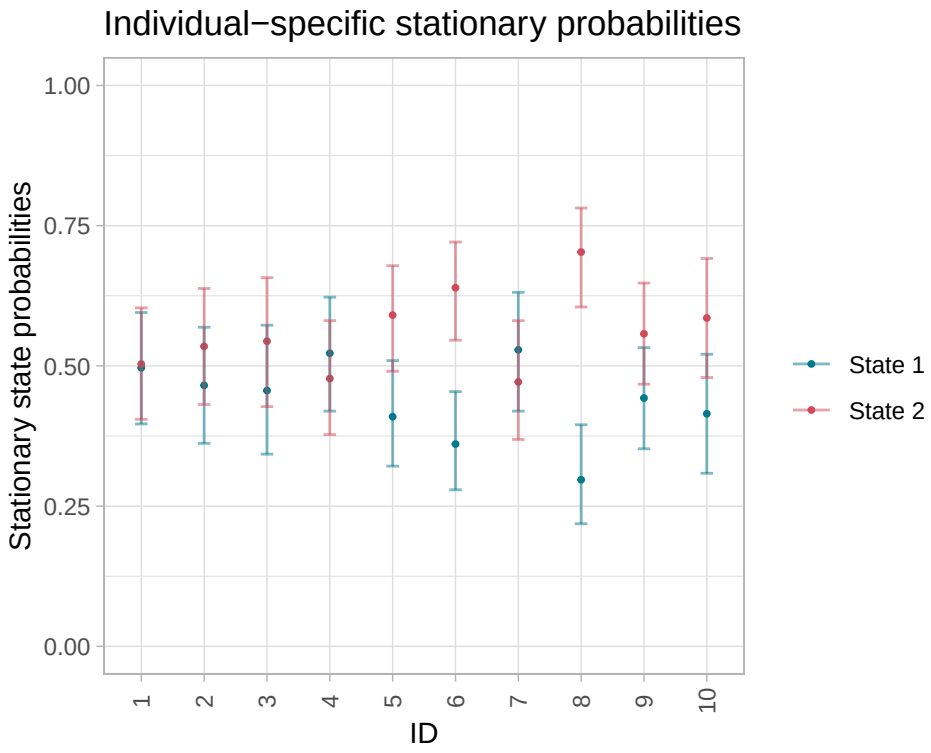
We can also visualise the estimated parameters of the state process, which included a random effect for the patient-therapist pair (10 levels). We can plot the transition probabilities directly, or the stationary state probabilities, which capture the long-term proportion of time spent in each state. These plots can be used to assess inter-group heterogeneity. For

example, here, it looks like patient-therapist pairs 6 and 8 tended to spend more time in state 2 ("patient is speaking") than others.

```
hmm$plot(what = "tpm", var = "ID") +
  labs(title = "Individual-specific transition probabilities")
```




```
hmm$plot(what = "delta", var = "ID") +
  labs(title = "Individual-specific stationary probabilities")
```



References

Aarts, Emmeke. 2023. *mHMMbayes: Multilevel Hidden Markov Models Using Bayesian Estimation*. https://CRAN.R-project.org/package=mHMMbayes.