

# A gentle tutorial on accelerated parameter and confidence interval estimation for hidden Markov models using Template Model Builder

Timothée Bacri<sup>1</sup> | Geir D. Berentsen<sup>2</sup> | Jan Bulla<sup>1,3</sup>  | Sondre Hølleland<sup>2,4</sup>

<sup>1</sup>Department of Mathematics, University of Bergen, Bergen, Norway

<sup>2</sup>Department of Business and Management Science, Norwegian School of Economics, Helleveien, Bergen, Norway

<sup>3</sup>Department of Psychiatry and Psychotherapy, University of Regensburg, Universitätsstraße, Regensburg, Germany

<sup>4</sup>Department of Pelagic Fish, Institute of Marine Research, Bergen, Norway

## Correspondence

Jan Bulla, Department of Mathematics, University of Bergen, P.O. Box 7803, 5020 Bergen, Norway.

Email: [jan.bulla@uib.no](mailto:jan.bulla@uib.no)

## Funding information

GENDER-Net Co-Plus Fund, Grant/Award Number: GNP-182; Financial Market Fund, Grant/Award Number: Norwegian Research Council project no. 309218



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## Abstract

A very common way to estimate the parameters of a hidden Markov model (HMM) is the relatively straightforward computation of maximum likelihood (ML) estimates. For this task, most users rely on user-friendly implementation of the estimation routines via an interpreted programming language such as the statistical software environment R. Such an approach can easily require time-consuming computations, in particular for longer sequences of observations. In addition, selecting a suitable approach for deriving confidence intervals for the estimated parameters is not entirely obvious, and often the computationally intensive bootstrap methods have to be applied. In this tutorial, we illustrate how to speed up the computation of ML estimates significantly via the R package TMB. Moreover, this approach permits simple retrieval of standard errors at the same time. We illustrate the performance of our routines using different data sets: first, two smaller samples from a mobile application for tinnitus patients and a well-known data set of fetal lamb movements with 87 and 240 data points, respectively. Second, we rely on larger data sets of simulated data of sizes 2000 and 5000 for further analysis. This tutorial is accompanied by a collection of scripts, which are all available in the Supporting Information. These scripts allow any user with moderate programming experience to benefit quickly from the computational advantages of TMB.

## KEYWORDS

confidence intervals, hidden Markov model, maximum likelihood estimation, TMB, tutorial

## 1 | INTRODUCTION

Hidden Markov models (HMMs) are a well-established, versatile type of model employed in many different applications. Since their first application in speech recognition (see, e.g., Baum & Petrie, 1966; Fredkin & Rice, 1992; Gales & Young,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH

2008), HMMs found wide usage in many applied sciences. To name only a few, biology and bioinformatics (Durbin, 1998; Eddy, 1998; Schadt et al., 1998), finance (Hamilton, 1989), ecology (McClintock et al., 2020), stochastic weather modeling (Ailliot et al., 2015; Lystig & Hughes, 2002), and engineering (Mor et al., 2021). Furthermore, the scientific literature on this topic in statistics is rich, as illustrated, for example, by the manuscripts of Bartolucci et al. (2012), Cappé et al. (2006), and Zucchini et al. (2016). The aforementioned sources contain, among many other aspects, detailed descriptions of parameter estimation for HMMs by maximization of the (log-)likelihood function. In short, on the one hand, maximum likelihood (ML) estimation is commonly achieved by a direct numerical maximization as introduced by Turner (2008) and later detailed by Zucchini et al. (2016), who also provided a collection of R (R Core Team, 2021) scripts that is widely used. Alternatively, Expectation Maximization (EM)-type algorithms as first described by Baum et al. (1970) or Dempster et al. (1977) serve for parameter estimation equally well. These algorithms possess several advantageous properties, for example, their robustness to poor initial values compared to direct numerical maximization via hill-climbing algorithms. For more details on the EM algorithm in the context of HMMs and a comparison of both approaches, see Bulla and Berzel (2008), who also describe a hybrid approach combining both algorithms.

Evaluating uncertainty and obtaining confidence intervals (CIs) constitute another essential aspect when working with HMMs—and it is less straightforward than parameter estimation. Although Cappé et al. (2006, ch. 12) showed that CIs could be obtained based on asymptotic normality of the ML estimates of the parameters under certain conditions, Frühwirth-Schnatter (2006, p. 53) points out that in independent mixture models, “the regularity conditions are often violated.” McLachlan and Peel (2004, p. 68) add that “In particular for mixture models, it is well known that the sample size  $n$  has to be very large before the asymptotic theory of maximum likelihood applies.” Lystig and Hughes (2002) show a way to compute the exact Hessian, and Zucchini et al. (2016) present an alternative way to compute the approximate Hessian and thus CIs, but admit that “the use of the Hessian to compute standard errors (and thence confidence intervals) is unreliable if some of the parameters are on or near the boundary of their parameter space.” In addition, Visser et al. (2000) report that computational problems arise when deriving CIs from the Hessian for sequences longer than 100 observations.

In this tutorial, we illustrate how to accelerate parameter estimation for HMMs with the help of Template Model Builder (TMB). As described by Kristensen et al. (2016), TMB is an R package developed for efficiently fitting complex statistical models to data. It provides exact calculations of first- and second-order derivatives of the (log-)likelihood of a model by automatic differentiation, which allows for efficient gradient- and/or Hessian-based optimization of the likelihood on the one hand. On the other hand, TMB permits to infer CIs for the estimated parameters by means of the Hessian. We show how to carry out this part for HMMs using a couple of simple examples. Then, we compare the Hessian-based CIs with CIs resulting from likelihood profiling and bootstrapping, which are both more computationally intensive.

The tutorial is accompanied by a collection of scripts (listed at [https://timothee-bacri.github.io/HMM\\_with\\_TMB/github.html#directory-structure](https://timothee-bacri.github.io/HMM_with_TMB/github.html#directory-structure)), which guide any user working with HMMs through the implementation of computationally efficient parameter estimation. The majority of scripts require only knowledge of R, just the computation of the (log-)likelihood function requires the involvement of C++. Moreover, we illustrate how TMB permits Hessian- or profile likelihood-based CIs for the estimated parameters at a very low computational cost. Naturally, the accelerated parameter estimation procedure may also serve for implementing computationally more efficient bootstrap CIs, an aspect we also make use of for our analyses.

## 2 | PRINCIPLES OF USING TMB FOR ML ESTIMATION

In order to keep this tutorial at acceptable length, all sections follow the same concept. That is, the reader is encouraged to consult the respective part of our [Supporting Information](https://timothee-bacri.github.io/HMM_with_TMB) in parallel to reading each section; it is available at [https://timothee-bacri.github.io/HMM\\_with\\_TMB](https://timothee-bacri.github.io/HMM_with_TMB). In particular, we recommend opening the file *Data supplements.Rproj* (available in the related repository at [https://github.com/timothee-bacri/HMM\\_with\\_TMB](https://github.com/timothee-bacri/HMM_with_TMB)) with R-Studio, which lets the reader have the correct working path set up automatically. This permits to copy-paste or download all the scripts presented in this tutorial for each section on the one hand. On the other hand, it allows for consistent maintenance of the code. Moreover, the repository also contains additional explanations, comments, and scripts.

### 2.1 | Setup

Execution of our routines requires the installation of the R-package TMB and the software `Rtools`, where the latter serves for compiling C++ code on Windows. In order to ensure reproducibility of all results involving the generation of random numbers, the `set.seed` function requires R version number 3.6.0 or greater. Our scripts were tested on a workstation

with 4 Intel(R) Xeon(R) Gold 6134 processors (3.7 GHz) running under the Linux distribution Ubuntu 18.04.6 LTS (Bionic Beaver) with 384 GB RAM and required about 1 week of computing time.

In particular for beginners, those parts of scripts involving C++ code can be difficult to debug because the code operates using a specific template. Therefore it is helpful to know that TMB provides a debugging feature, which can be useful to retrieve diagnostic error messages, in RStudio. Enabling this feature is optional and can be achieved by the command `TMB:::setupRStudio()` (requires manual confirmation and restarting RStudio).

## 2.2 | Linear regression example

We begin by demonstrating the principles of TMB, which we illustrate through the fitting procedure for a simple linear model. This permits, among other things, to show how to handle parameters subject to constraints, an aspect particularly relevant for HMMs. A more comprehensive tutorial on TMB presenting many technical details in more depth is available at [https://kaskr.github.io/adcomp/\\_book/Tutorial.html](https://kaskr.github.io/adcomp/_book/Tutorial.html).

Let  $\mathbf{x}$  and  $\mathbf{y}$  denote the predictor and response vector, respectively, both of length  $n$ . For a simple linear regression model with intercept  $a$  and slope  $b$ , the negative log-likelihood equals

$$-\log L(a, b, \sigma^2) = -\sum_{i=1}^n \log(\phi(y_i; a + bx_i, \sigma^2)),$$

where  $\phi(\cdot; \mu, \sigma^2)$  corresponds to the density function of the univariate normal distribution with mean  $\mu$  and variance  $\sigma^2$ .

The use of TMB requires the (negative) log-likelihood function to be coded in C++ under a specific template, which is then loaded into R. The minimization of this function and other postprocessing procedures are all carried out in R. Therefore, we require two files. The first file, named `linreg.cpp`, is written in C++ and defines the objective function, that is, the negative log-likelihood (nll) function of the linear model, as follows:

```
#include <TMB.hpp> //import the TMB template

template<class Type>
Type objective_function<Type>::operator() ()
{
  DATA_VECTOR(y); // Data vector y passed from R
  DATA_VECTOR(x); // Data vector x passed from R

  PARAMETER(a); // Parameter a passed from R
  PARAMETER(b); // Parameter b passed from R
  PARAMETER(tsigma); // Parameter sigma (transformed, on log-scale)
                      // passed from R

  // Transform tsigma back to natural scale
  Type sigma = exp(tsigma);

  // Declare negative log-likelihood
  Type nll = - sum(dnorm(y,
                        a + b * x,
                        sigma,
                        true));

  // Necessary for inference on sigma, not only tsigma
  ADREPORT(sigma);

  return nll;
}
```

Note that we define data inputs  $x$  and  $y$  using the `DATA_VECTOR()` declaration in the above code. Furthermore, we declare the `nll` as a function of the three parameters  $a$ ,  $b$ , and  $\log(\sigma)$  using the `PARAMETER()` declaration. In order to be able to carry out unconstrained optimization procedures in the following, the `nll` function is parameterized in terms of  $\log(\sigma)$ . While the parameter  $\sigma$  is constrained to be nonnegative,  $\log(\sigma)$  can be freely estimated. Alternatively, constraint optimization methods could be carried out, but we do not investigate such procedures. The `ADREPORT()` function is optional but useful for parameter inference at the postprocessing stage.

The second file needed is written in R and serves for compiling the `nll` function defined above and carrying out the estimation procedure by numerical optimization of the `nll` function. The `.R` file (shown below) carries out the compilation of the C++ file and minimization of the `nll` function:

```
# Loading TMB package
library(TMB)
# Compilation. The compiler returns 0 if the compilation of
# the cpp file was successful
TMB::compile("code/linreg.cpp")

## [1] 0

# Load the dynamic library
dyn.load(dynlib("code/linreg"))
# Generate the data for our test sample
set.seed(123)
data <- list(y = rnorm(20) + 1:20, x = 1:20)
parameters <- list(a = 0, b = 0, tsigma = 0)
# Instruct TMB to create the likelihood function
obj_linreg <- MakeADFun(data, parameters, DLL = "linreg",
                        silent = TRUE)
# Optimization of the objective function with nlminb
# Arguments are the initial parameters, the objective function to
# minimize, and the gradient and Hessian passed from TMB
mod_linreg <- nlminb(start = obj_linreg$par,
                    objective = obj_linreg$fn,
                    gradient = obj_linreg$gr,
                    hessian = obj_linreg$he)

mod_linreg$par

##           a           b           tsigma
## 0.31009251 0.98395536 -0.05814649
```

In addition to the core functionality presented above, different types of postprocessing of the results are possible as well. For example, the function `sdreport` returns the ML estimates and standard errors of the parameters in terms of which the `nll` is parameterized:

```
sdreport(obj_linreg, par.fixed = mod_linreg$par)

## sdreport(.) result
##           Estimate Std. Error
## a           0.31009251 0.43829087
## b           0.98395536 0.03658782
## tsigma     -0.05814649 0.15811383
## Maximum gradient component: 1.300261e-10
```

In principle, the argument `par.fixed = mod_linreg$par` is optional but recommended, because it ensures that the `sdreport` function is carried out at the minimum found by `nlminb` (Gay, 1990). Instead of `nlminb`, other optimization routines may be used as well. For example, Zucchini et al. (2016) rely on the R function `nlm`. We selected `nlminb` because

it offers a higher degree of flexibility, while having a syntax close to `nlm`. Note that the standard errors above are based on the Hessian matrix of the `nlm`.

From a practical perspective, it is usually desirable to obtain standard errors for the constrained variables, in this case  $\sigma$ . To achieve this, one can run the `summary` function with argument `select = "report"`:

```
summary(sdreport(obj_linreg, par.fixed = mod_linreg$par),
        select = "report")

##           Estimate Std. Error
## sigma 0.9435117  0.1491823
```

These standard errors result from the generalized delta method described by Kass and Steffey (1989), which is implemented within TMB. Note that full functionality of the `sdreport` function requires calling the function `ADREPORT` on the additional parameters of interest (i.e., those including transformed parameters, in our example  $\sigma$ ) in the C++ part. The `select` argument restricts the output to variables passed by `ADREPORT`. This feature is particularly useful when the likelihood has been reparameterized as above, and is especially relevant for HMMs. Following Zucchini et al. (2016), we refer to the original parameters as natural parameters, and to their transformed version as the working parameters.

Last, we display the estimation results from the `lm` function for comparison.

```
rbind(
  "lm" = lm(y ~ x, data = data)$coef, # linear regression using R
  "TMB" = mod_linreg$par[1:2] # intercept and slope from TMB fit
)

##      (Intercept)          x
## lm      0.3100925  0.9839554
## TMB      0.3100925  0.9839554
```

### 3 | PARAMETER ESTIMATION TECHNIQUES FOR HMMs

In this section we recall basic concepts underlying parameter estimation for HMMs via direct numerical optimization of the likelihood. In terms of notation, we stay as close as possible to Zucchini et al. (2016), where a more detailed presentation is available.

#### 3.1 | Basic notation and model setup

A large variety of modeling approaches is possible with HMMs, ranging from rather simple to highly complex setups. In a basic HMM, one assumes that the data-generating process corresponds to a time-dependent mixture of conditional distributions. More specifically, the mixing process is driven by an unobserved (hidden) homogeneous Markov chain. In this paper we focus on a Poisson HMM, but only small changes are necessary to adapt our scripts to models with other conditional distributions. For an example of the application of TMB in a more complex HMM setting, see Berentsen et al. (2022). Let  $\{X_t : t = 1, \dots, T\}$  and  $\{C_t : t = 1, \dots, T\}$  denote the observed and hidden process, respectively, where  $t$  denotes the (time) index ranging from one to  $T$ . For an  $m$ -state Poisson HMM, the conditional distributions with parameter  $\lambda_i$  are then specified through

$$p_i(x) = P(X_t = x | C_t = i) = \frac{e^{-\lambda_i} \lambda_i^x}{x!},$$

where  $i = 1, \dots, m$ . Furthermore, we let  $\Gamma = \{\gamma_{ij}\}$  and  $\delta$  denote the transition probability matrix (TPM) of the Markov chain and the corresponding stationary distribution, respectively. It is noteworthy that Markov chains in the context of HMMs are often assumed irreducible and aperiodic. For example, Grimmett et al. (2001, Lemma 6.3.5 on p. 225 and Theorem 6.4.3

on p. 227) show that irreducibility ensures the existence of the stationary distribution, and Feller (1968, p. 394) describes that aperiodicity implies that a unique limiting distribution exists and corresponds to the stationary distribution. These results are, however, of limited relevance for most estimation algorithms, because the elements of  $\mathbf{\Gamma}$  are in general strictly positive. Nevertheless, one should be careful when manually fixing selected elements of  $\mathbf{\Gamma}$  to zero.

### 3.2 | The likelihood function of an HMM

The likelihood function of an HMM requires, in principle, a summation over all possible state sequences. As shown, for example, by Zucchini et al. (2016, p. 37), a computationally convenient representation as a product of matrices is possible. Let  $X^{(t)} = \{X_1, \dots, X_t\}$  and  $x^{(t)} = \{x_1, \dots, x_t\}$  denote the “history” of the observed process  $X_t$  and the observations  $x_t$ , respectively, from time 1 up to time  $t$ . Moreover, let  $\theta$  denote the vector of model parameters, which consists of the parameters of the TPM and the parameters of the conditional probability density functions. Given these parameters, the likelihood of the observations  $\{x_1, \dots, x_T\}$  can then be expressed as

$$L(\theta) = \mathbf{P}(X^{(T)} = x^{(T)}) = \boldsymbol{\delta} \mathbf{P}(x_1) \mathbf{\Gamma} \mathbf{P}(x_2) \mathbf{\Gamma} \mathbf{P}(x_3) \dots \mathbf{\Gamma} \mathbf{P}(x_T) \mathbf{1}', \quad (1)$$

where

$$\mathbf{P}(x) = \begin{pmatrix} p_1(x) & & & 0 \\ & p_2(x) & & \\ & & \ddots & \\ 0 & & & p_m(x) \end{pmatrix}$$

corresponds to a diagonal matrix with the  $m$  conditional probability density functions evaluated at  $x$  (we will use the term density despite the discrete support), and  $\mathbf{1}$  denotes a vector of ones. The first element of the likelihood function, the so-called initial distribution, is given by the stationary distribution  $\boldsymbol{\delta}$  here. Alternatively, the initial distribution may be estimated freely, which requires minor changes to the likelihood function discussed in Section 4.1.

Note that the treatment of missing data is comparably straightforward in this setup. If  $x$  is a missing observation, one just has to set  $p_i(x) = 1$ , thus  $\mathbf{P}(x)$  reduces to the unity matrix as detailed in Zucchini et al. (2016, p. 40). Furthermore, this representation of the likelihood is quite natural from an intuitive point of view. From left to right, it can be interpreted as a pass through the observations: One starts with the initial distribution multiplied by the conditional density of  $x_1$  collected in  $\mathbf{P}(x_1)$ . This is followed by iterative multiplications with the TPM modeling the transition to the next observation, and yet another multiplication with contributions of the following conditional densities.

### 3.3 | Forward algorithm and backward algorithm

The pass through the observations described above actually forms the basis for an efficient evaluation of the likelihood function. More precisely, the so-called “forward algorithm” allows for a recursive computation of the likelihood. For setting up this algorithm, we need to define the vector  $\boldsymbol{\alpha}_t$  by

$$\begin{aligned} \boldsymbol{\alpha}_t &= \boldsymbol{\delta} \mathbf{P}(x_1) \mathbf{\Gamma} \mathbf{P}(x_2) \mathbf{\Gamma} \mathbf{P}(x_3) \dots \mathbf{\Gamma} \mathbf{P}(x_t) \\ &= \boldsymbol{\delta} \mathbf{P}(x_1) \prod_{s=2}^t \mathbf{\Gamma} \mathbf{P}(x_s) \\ &= (\alpha_t(1), \dots, \alpha_t(m)) \end{aligned}$$

for  $t = 1, 2, \dots, T$ . The name forward algorithm originates from the way of calculating  $\boldsymbol{\alpha}_t$ , that is,

$$\begin{aligned} \boldsymbol{\alpha}_0 &= \boldsymbol{\delta} \mathbf{P}(x_1) \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1} \mathbf{\Gamma} \mathbf{P}(x_t) \text{ for } t = 1, 2, \dots, T. \end{aligned}$$

After a pass through all observations, the likelihood results from

$$L(\theta) = \alpha_T \mathbf{1}'.$$

In a similar way, the “backward algorithm” also permits the recursive computation of the likelihood, but starting with the last observation. To formulate the backward algorithm, let us define the vector  $\beta_t$  for  $t = 1, 2, \dots, T$  so that

$$\begin{aligned} \beta_t' &= \Gamma \mathbf{P}(x_{t+1}) \Gamma \mathbf{P}(x_{t+2}) \dots \Gamma \mathbf{P}(x_T) \dots \mathbf{1}' \\ &= \left( \prod_{s=t+1}^T \Gamma \mathbf{P}(x_s) \right) \mathbf{1}' \\ &= (\beta_t(1), \dots, \beta_t(m)). \end{aligned}$$

The name backward algorithm results from the way of calculating  $\beta_t$ , that is,

$$\begin{aligned} \beta_T &= \mathbf{1}', \\ \beta_t &= \Gamma \mathbf{P}(x_{t+1}) \beta_{t+1} \text{ for } t = T - 1, T - 2, \dots, 1. \end{aligned}$$

Again, the likelihood can be calculated after a pass through all observations by

$$L(\theta) = \delta \beta_1.$$

In general, parameter estimation bases on the forward algorithm. The backward algorithm is, however, still useful because the quantities  $\alpha_t$  and  $\beta_t$  together serve for a couple of interesting tasks. For example, they are the basis for deriving a particular type of conditional distributions and for state inference by local decoding (Zucchini et al., 2016, ch. 5, pp. 81–93). We present details on local decoding in the [Supporting Information](#).

Last, it is well known that the execution of the forward (or backward) algorithm may quickly lead to underflow errors, because many elements of the vectors and matrices involved take values between zero and one. To avoid these difficulties, a scaling factor can be introduced. We follow the approach suggested by Zucchini et al. (2016, p. 48) and implement a scaled version of the forward algorithm, which directly provides the (negative) log-likelihood as result.

### 3.4 | Reparameterization of the likelihood function

The representation of the likelihood and the algorithms shown above rely on the data and the set of parameters  $\theta$  as input. The data are subject to several constraints:

- (i) Typically there are various constraints of the parameters in the conditional distribution. For the Poisson HMM, all elements of the parameter vector  $\lambda = (\lambda_1, \dots, \lambda_m)$  must be positive.
- (ii) In general, the parameters  $\gamma_{ij}$  of the TPM  $\Gamma$  have to be nonnegative, and the rows of  $\Gamma$  must sum up to one.

The constraints of the TPM can be difficult to deal with using constrained optimization of the likelihood. A common approach is to reparameterize the log-likelihood in terms of unconstrained “working” parameters  $\{\mathbf{T}, \eta\} = g^{-1}(\Gamma, \lambda)$ , as follows. A possible reparameterization of  $\Gamma$  is given by

$$\gamma_{ij} = \frac{\exp(\tau_{ij})}{1 + \sum_{k \neq i} \exp(\tau_{ik})}, \text{ for } i \neq j,$$

where  $\tau_{ij}$  are  $m(m - 1)$  real valued, thus unconstrained, elements of an  $m$  times  $m$  matrix  $\mathbf{T}$  with no diagonal elements. The diagonal elements of  $\Gamma$  follows implicitly from  $\sum_j \gamma_{ij} = 1 \forall i$  (Zucchini et al., 2016, p. 51). The corresponding reverse

transformation is given by

$$\tau_{ij} = \log\left(\frac{\gamma_{ij}}{1 - \sum_{k \neq i} \gamma_{ik}}\right) = \log(\gamma_{ij}/\gamma_{ii}), \text{ for } i \neq j.$$

For the Poisson HMM the intensities can be reparameterized in terms of  $\lambda_i = \exp(\eta_i)$ , and consequently the unconstrained working parameters are given by  $\eta_i = \log(\lambda_i)$ ,  $i = 1, \dots, m$ . Estimates of the “natural” parameters  $\{\Gamma, \lambda\}$  can then be obtained by maximizing the reparameterized likelihood with respect to  $\{\mathbf{T}, \boldsymbol{\eta}\}$  and then transforming the estimated working parameters back to natural parameters via the above transformations, that is,  $\{\hat{\Gamma}, \hat{\lambda}\} = g(\hat{\mathbf{T}}, \hat{\boldsymbol{\eta}})$ . Note that in general the function  $g$  needs to be one-to-one for the above procedure to work.

## 4 | USING TMB

In the following we show how ML estimation of the parameters of HMMs can be carried out efficiently via TMB.

### 4.1 | Likelihood function

Similar to the linear regression example presented in Section 2.2, the first and essential step is to define our nll function to be minimized later in a suitable C++ file. In our case, this function calculates the negative log-likelihood presented by Zucchini et al. (2016, p. 48), and our C++ code is analog to the R-code shown by Zucchini et al. (2016, pp. 331–333). This function, defined in the file named *poi\_hmm.cpp*, tackles our setting with conditional Poisson distributions only. An extension to, for example, Gaussian, binomial, and exponential conditional distributions is straightforward. It requires to modify the density function in the *poi\_hmm.cpp* file and the related functions for parameter transformation presented in Section 3.4. We illustrate the implementation of the Gaussian case in the [Supporting Information](#).

However, note that the number of possible modeling setups is very large: For example, the conditional distributions may vary from state to state, nested model specifications, the conditional mean may be linked to covariates, or the TPM could depend on covariates—to name only a few. Due to the very large number of possible extensions of the basic HMM, we refrain from implementing an R-package, but prefer to provide a proper guidance to the reader for building custom models suited to a particular application. As a small example, we illustrate how to implement a freely estimated initial distribution in the file *poi\_hmm.cpp*. This modification can be achieved by uncommenting a couple of lines only.

The file *poi\_hmm.cpp* is available at [https://github.com/timothee-bacri/HMM\\_with\\_TMB/blob/main/code/poi\\_hmm.cpp](https://github.com/timothee-bacri/HMM_with_TMB/blob/main/code/poi_hmm.cpp) and contains the following:

```
#include <TMB.hpp>
#include "../functions/utils.cpp"

// Likelihood for a poisson hidden markov model.
template<class Type>
Type objective_function<Type>::operator() ()
{
    // Data
    DATA_VECTOR(x); // timeseries vector
    DATA_INTEGER(m); // Number of states m

    // Parameters
    PARAMETER_VECTOR(tlambda); // conditional log_lambdas's
    PARAMETER_VECTOR(tgamma); // m(m-1) working parameters of TPM

    // Uncomment only when using a non-stationary distribution
    //PARAMETER_VECTOR(tdelta); // transformed stationary distribution,

    // Transform working parameters to natural parameters:
```



```

vector<Type> lambda = tlambda.exp();
matrix<Type> gamma = gamma_w2n(m, tgamma);

// Construct stationary distribution
vector<Type> delta = stat_dist(m, gamma);
// If using a non-stationary distribution, use this instead
// vector<Type> delta = delta_w2n(m, tdelta);

// Get number of timesteps (n)
int n = x.size();

// Evaluate conditional distribution: Put conditional
// probabilities of observed x in n times m matrix
// (one column for each state, one row for each datapoint):
matrix<Type> emission_probs(n, m);
matrix<Type> rowlvec(1, m);
rowlvec.setOnes();
for (int i = 0; i < n; i++) {
    if (x[i] != x[i]) { // f != f returns true if and only if f is NaN.
        // Replace missing values (NA in R, NaN in C++) with 1
        emission_probs.row(i) = rowlvec;
    }
    else {
        emission_probs.row(i) = dpois(x[i], lambda, false);
    }
}

// Corresponds to (Zucchini et al., 2016, p 333)
matrix<Type> foo, P;
Type mllk, sumfoo, lscale;

foo = (delta * vector<Type>(emission_probs.row(0)).matrix());
sumfoo = foo.sum();
lscale = log(sumfoo);
foo.transposeInPlace();
foo /= sumfoo;
for (int i = 2; i <= n; i++) {
    P = emission_probs.row(i - 1);
    foo = ((foo * gamma).array() * P.array()).matrix();
    sumfoo = foo.sum();
    lscale += log(sumfoo);
    foo /= sumfoo;
}
mllk = -lscale;

// Use adreport on variables for which we want standard errors
ADREPORT(lambda);
ADREPORT(gamma);
ADREPORT(delta);

// Variables we need for local decoding and in a convenient format
REPORT(lambda);
REPORT(gamma);
REPORT(delta);
REPORT(n);
REPORT(emission_probs);
REPORT(mllk);

return mllk;
}

```

TABLE 1 TYT data

```
653643566646646666646567655576565666567767666657616
021676665566250111231313011141412220
```

Note: Observations collected by the TYT app on 87 successive days (from left to right) for a single individual.

## 4.2 | Optimization

With the `nll` function available in C++, we can carry out the parameter estimation and all pre-/postprocessing in R. In the following we describe the steps to be carried out.

- (1) Loading of the necessary packages, compilation of the `nll` function with TMB and subsequent loading, and loading of the auxiliary functions for parameter transformation.

```
# Load TMB and optimization packages
library(TMB)
# Run the C++ file containing the TMB code
TMB::compile("code/poi_hmm.cpp")

## [1] 0

# Load it
dyn.load(dynlib("code/poi_hmm"))
# Load the parameter transformation function
source("functions/utils.R")
```

- (2) Loading of the observations. The data are part of a large data set collected with the “Track Your Tinnitus” (TYT) mobile application, a detailed description of which is presented in Pryss, Reichert, Langguth, et al. (2015) and Pryss, Reichert, Herrmann, et al. (2015). We analyze 87 successive days of the “arousal” variable recorded for a single individual. This variable is measured on a discrete scale, where higher values correspond to a higher degree of excitement and lower values to a more calm emotional state (for details, see Probst et al., 2016, 2017). Loading the “arousal” variable can be achieved simply with

```
load("data/tinnitus.RData")
```

Table 1 presents the raw data, which are also available for download in the [Supporting Information](#).

- (3) Initialization of the number of states and starting (or initial) values for the optimization. First, the number of states needs to be determined. As explained by Pohle et al. (2017b), Pohle et al. (2017a), and Zucchini et al. (2016, section 6) (to name only a few), usually one would first fit models with a different number of states. Then, these models are evaluated, for example, by means of model selection criteria (as carried out by Leroux & Puterman, 1992) or prediction performance (Celeux & Durand, 2008). The model selection procedure shows that both Akaike information criterion (AIC) and Bayesian information criterion (BIC) prefer a two-state model over a model with three or four states. Consequently, we focus on the two-state model in the following. The list object `TMB_data` contains the data and the number of states.

```
# Model with 2 states
m <- 2
TMB_data <- list(x = tinn_data, m = m)
```

Second, initial values for the optimization procedure need to be defined. Although we will apply unconstrained optimization, we initialize the natural parameters, because this is much more intuitive and practical than handling the

working parameters.

```
# Generate initial set of parameters for optimization
lambda <- c(1, 3)
gamma <- matrix(c(0.8, 0.2,
                  0.2, 0.8), byrow = TRUE, nrow = m)
```

- (4) Transformation from natural to working parameters. The previously created initial values are transformed and stored in the list `parameters` for the optimization procedure.

```
# Turn them into working parameters
parameters <- pois.HMM.pn2pw(m, lambda, gamma)
```

- (5) Creation of the TMB negative log-likelihood function with its derivatives. This object, stored as `obj_tmb` requires the data, the initial values, and the previously created DLL as input. Setting argument `silent = TRUE` disables tracing information and is only used here to avoid excessive output.

```
obj_tmb <- MakeADFun(TMB_data, parameters,
                    DLL = "poi_hmm", silent = TRUE)
```

This object also contains the previously defined initial values as a vector (`par`) rather than a list. The negative log-likelihood (`fn`), its gradient (`gr`), and Hessian (`he`) are functions of the parameters (in vector form) while the data are considered fixed:

```
obj_tmb$par

##      tlambda      tlambda      tgamma      tgamma
##  0.000000  1.098612 -1.386294 -1.386294

obj_tmb$fn(obj_tmb$par)

## [1] 228.3552

obj_tmb$gr(obj_tmb$par)

##           [,1]      [,2]      [,3]      [,4]
## [1,] -3.60306 -146.0336  10.52832 -1.031706

obj_tmb$he(obj_tmb$par)

##           [,1]      [,2]      [,3]      [,4]
## [1,]  1.902009 -5.877900 -1.3799682  2.4054017
## [2,] -5.877900 188.088247 -4.8501589  2.3434284
## [3,] -1.379968 -4.850159  9.6066700 -0.8410438
## [4,]  2.405402  2.343428 -0.8410438  0.7984216
```

- (6) Execution of the optimization. For this step we rely again on the optimizer implemented in the `nlminb` function. The arguments, that is, initial values for the parameters and the function to be optimized, are extracted from the

previously created TMB object.

```
mod_tmb <- nlmminb(start = obj_tmb$par, objective = obj_tmb$fn)
# Check that it converged successfully
mod_tmb$convergence == 0

## [1] TRUE
```

As mentioned previously, various alternatives to `nlminb` exist and could be used at this step instead.

- (7) Obtaining the ML estimates of the natural parameters together with their standard errors is possible by using the previously introduced command `sdreport`. Recall that this requires the parameters of interest to be treated by the `ADREPORT` statement in the C++ part. It should be noted that the presentation of the set of parameters gamma below results from a column-wise representation of the TPM.

```
summary(sdreport(obj_tmb, par.fixed = mod_tmb$par), "report")

##           Estimate Std. Error
## lambda 1.63641070 0.27758294
## lambda 5.53309626 0.31876141
## gamma  0.94980192 0.04374682
## gamma  0.02592209 0.02088689
## gamma  0.05019808 0.04374682
## gamma  0.97407791 0.02088689
## delta  0.34054163 0.23056401
## delta  0.65945837 0.23056401
```

Note that the table above also contains estimation results for  $\delta$  and accompanying standard errors, although  $\delta$  is not estimated, but derived from  $\Gamma$ . We provide further details on this aspect in Section 5.1.

The value of the nll function in the minimum found by the optimizer can also be extracted directly from the object `mod_tmb` by accessing the list element `objective`:

```
mod_tmb$objective

## [1] 168.5361
```

- (8) In the optimization above we already benefited from an increased speed due to the evaluation of the nll in C++ compared to the forward algorithm being executed entirely in R. However, the use of TMB also permits to introduce the gradient and/or the Hessian computed by TMB into the optimization procedure. This is in general advisable, because TMB provides an exact value of both gradient and Hessian up to machine precision, which is superior to approximations used by optimizing procedure. Similar to the nll, both quantities can be extracted directly from the TMB object `obj_tmb`:

```
# The negative log-likelihood is accessed by the objective
# attribute of the optimized object
mod_tmb <- nlmminb(start = obj_tmb$par, objective = obj_tmb$fn,
                  gradient = obj_tmb$gr, hessian = obj_tmb$he)
mod_tmb$objective

## [1] 168.5361
```

Note that passing the gradient and Hessian provided by TMB to `nlminb` leads to the same minimum, that is, value of the nll function, here.

### 4.3 | Basic nested model specification

In the context of HMMs (and other statistical models), nested models or models subject to certain parameter restrictions are commonly used. For example, it may be necessary to fix some parameters because of biological or physical constraints. TMB can be instructed to treat selected parameters as constants, or impose equality constraints on a set of parameters. For the practical implementation, it is noteworthy that such parameter restrictions should be imposed on the working parameters. However, it is also easily possible to impose restrictions on a natural parameter (e.g.,  $\lambda$ ), and then identify the corresponding restriction on the working parameter (i.e.,  $\log(\lambda)$ ). We illustrate a simple nested model specification by fixing  $\lambda_1$  to one in our two-state Poisson HMM, the other parameter components correspond to the previous initial values.

```
# Get the previous values, and fix some
fixed_par_lambda <- lambda
fixed_par_lambda[1] <- 1
```

We then transform these natural parameters into a set of working parameters.

```
# Transform them into working parameters
new_parameters <- pois.HMM.pn2pw(m = m,
                                lambda = fixed_par_lambda,
                                gamma = gamma)
```

For instructing TMB to treat selected parameters as constants, the `map` argument of the `MakeADFun` has to be specified in addition to the usual arguments. The `map` argument is a list consisting factor-valued vectors, which possess the same length as the working parameters and carry their names as well. The factor levels have to be unique for the regular parameters not subject to specific restrictions. If a parameter is fixed the corresponding entry of the `map` argument is filled with `NA`. In our example, this leads to:

```
map <- list(tlambda = as.factor(c(NA, 1)),
            tgamma = as.factor(c(2, 3)))
fixed_par_obj_tmb <- MakeADFun(TMB_data, new_parameters,
                              DLL = "poi_hmm",
                              silent = TRUE,
                              map = map)
```

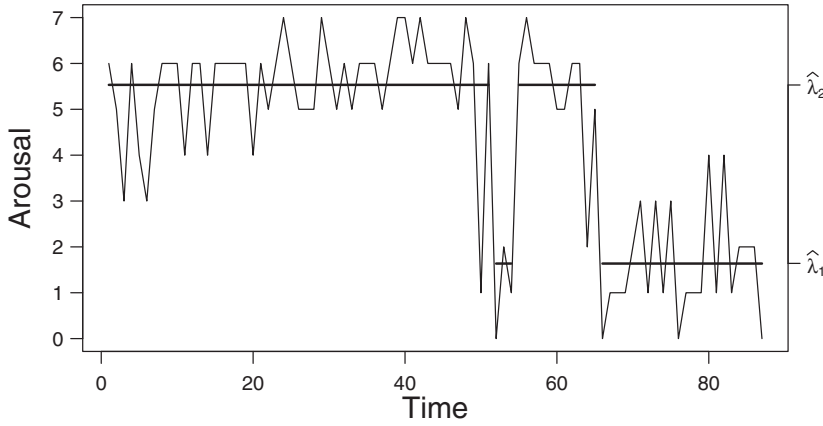
It is noteworthy that more complex constraints are possible as well. For example, to impose equality constraints (such as  $\gamma_{11} = \gamma_{22}$ ), the corresponding factor level has to be identical for the concerned entries. We refer to our [Supporting Information](#) for details.

Last, estimation of the remaining model parameters and extraction of the results is achieved as before.

```
fixed_par_mod_tmb <- nlminb(start = fixed_par_obj_tmb$par,
                           objective = fixed_par_obj_tmb$fn,
                           gradient = fixed_par_obj_tmb$gr,
                           hessian = fixed_par_obj_tmb$he)
summary(sdreport(fixed_par_obj_tmb), "report")

##           Estimate Std. Error
## lambda 1.000000000 0.000000000
## lambda 5.50164872 0.30963641
## gamma 0.94561055 0.04791050
## gamma 0.02655944 0.02133283
## gamma 0.05438945 0.04791050
## gamma 0.97344056 0.02133283
## delta 0.32810136 0.22314460
## delta 0.67189864 0.22314460
```

Note that the standard error of  $\lambda_1$  equals zero, because it is no longer considered a parameter and does not enter the optimization procedure.



**FIGURE 1** Plot of the TYT data  
*Note:* The solid horizontal lines correspond to the conditional mean of the inferred state at each time. See Table 5 for the values of  $\hat{\lambda}_i$

#### 4.4 | State inference and forecasting

After estimating an HMM by the procedures illustrated in Section 4.2, it is possible to carry out a couple analyses that provide insight into the interpretation of the estimated model. These include, for example, the so-called smoothing probabilities, which correspond to the probability of being in state  $i$  at time  $t$  for  $i = 1, \dots, m, t = 1, \dots, T$ , given all observations. These probabilities can be obtained by

$$P(C_t = i | X^{(T)} = x^{(T)}) = \frac{\alpha_t(i)\beta_t(i)}{L(\hat{\theta})},$$

where  $\hat{\theta}$  denotes the set of ML estimates. The derived smoothing probabilities then serve for determining the most probable state  $i_t^*$  at time  $t$  given the observations by

$$i_t^* = \arg \max_{i \in \{1, \dots, m\}} P(C_t = i | X^{(T)} = x^{(T)}).$$

Furthermore, the Viterbi algorithm determines the overall most probable sequence of states  $i_1^*, \dots, i_T^*$ , given the observations. This is achieved by evaluating

$$(i_1^*, \dots, i_T^*) = \arg \max_{i_1, \dots, i_T \in \{1, \dots, m\}} P(C_1 = i_1, \dots, C_T = i_T | X^{(T)} = x^{(T)}).$$

Other quantities of interest include the forecast distribution or  $h$ -step-ahead probabilities, which are obtained through

$$P(X_{T+h} = x | X^{(T)} = x^{(T)}) = \frac{\alpha_T \Gamma^h \mathbf{P}(x) \mathbf{1}'}{\alpha_T \mathbf{1}'} = \varphi_T \Gamma^h \mathbf{P}(x) \mathbf{1}',$$

where  $\varphi_T = \alpha_T / \alpha_T \mathbf{1}'$ .

All the quantities shown above and the related algorithms for deriving them are described in detail in Zucchini et al. (2016, Ch. 5). In order to apply these algorithms, it is only necessary to extract the quantities required as input from a suitable `MakeADFun` object. Note that most algorithms rely on scaled versions of the forward- and backward algorithm. This is illustrated in detail in the [Supporting Information](#). Figure 1 shows the TYT data together with the conditional mean values linked to the most probable state inferred by the smoothing probabilities.

## 5 | CIs

A common approach for deriving CIs for the estimated parameters of statistical models bases on finite-difference approximations of the Hessian. This technique is, however, not suited for most HMMs due to computational difficulties, as already pointed out by Visser et al. (2000). The same authors suggest likelihood profile CIs or bootstrap-based CIs as potentially better alternatives. Despite the potentially high computational load, bootstrap-based CIs have become

an established method in the context of HMMs (Bulla & Berzel, 2008; Zucchini et al., 2016) and found widespread application by practitioners.

In this section we illustrate how CIs based on the Hessian, likelihood profiling, and the bootstrap can be efficiently implemented by integrating TMB. This permits in particular to obtain Hessian-based and likelihood profile-based CIs at very low computational cost. For simplicity, we illustrate our procedures by means of the parameter  $\lambda_2$  of our two-state Poisson HMM. We will further address the resulting CIs for  $\Gamma$  and  $\lambda$  and performance-related aspects in Section 6.

## 5.1 | Wald-type CIs based on the Hessian

Since the negative log-likelihood function of HMMs typically depends on the working parameters, evaluation of the Hessian in the optimum found by numerical optimization only serves for inference about the working parameters. From a practical perspective, however, inference about the natural parameters usually is of interest. As the Hessian  $\nabla^2 \log L(\{\hat{\mathbf{T}}, \hat{\boldsymbol{\eta}}\})$  refers to the working parameters  $\{\mathbf{T}, \boldsymbol{\eta}\}$ , the delta method is suitable to obtain an estimate of the covariance matrix of  $\{\hat{\mathbf{T}}, \hat{\boldsymbol{\lambda}}\}$  by

$$\Sigma_{\hat{\mathbf{T}}, \hat{\boldsymbol{\lambda}}} = -\nabla g(\hat{\mathbf{T}}, \hat{\boldsymbol{\eta}}) (\nabla^2 \log L(\hat{\mathbf{T}}, \hat{\boldsymbol{\eta}}))^{-1} \nabla g(\hat{\mathbf{T}}, \hat{\boldsymbol{\eta}})', \quad (2)$$

with  $\{g(\hat{\mathbf{T}}, \hat{\boldsymbol{\lambda}}) = g(\hat{\mathbf{T}}, \hat{\boldsymbol{\eta}})$  as defined in Section 3.4. From a user's perspective, it is highly convenient that the entire right-hand side of Equation (2) can be directly computed via automatic differentiation in TMB. Moreover, it is particularly noteworthy that the standard errors of derived parameters can be calculated by the delta method similarly. For example, the stationary distribution  $\boldsymbol{\delta}$  is a function of  $\Gamma$  in our case, and TMB provides a straightforward way to obtain standard errors of  $\boldsymbol{\delta}$ . This is achieved by first defining  $\boldsymbol{\delta}$  inside the C++ file `poi_hmm.cpp` (or, in our implementation, the related `utils.cpp`, which gathers auxiliary functions). Second, it is necessary to call `ADREPORT` on  $\boldsymbol{\delta}$  within the `poi_hmm.cpp` file. To display the resulting estimates and corresponding standard errors in R, one can rely on the command shown previously in Section 4.2.

Subsequently, Wald-type CIs (Wald, 1943) follow in the usual manner. For example, the  $(1 - \alpha)\%$  CI for  $\lambda_1$  is given by  $\lambda_1 \pm z_{1-\alpha/2} * \sigma_{\lambda_1}$ , where  $z_x$  is the  $x$ -percentile of the standard normal distribution, and  $\sigma_{\lambda_1}$  is the standard error of  $\lambda_1$  obtained via the delta method. This part is easily implemented in R. We illustrate the calculation of these CIs for our two-state Poisson HMM in the [Supporting Information](#).

Finally, note that the reliability of Wald-type CIs may suffer from a singular Fisher information matrix, which can occur for many different types of statistical models, including HMMs. This also jeopardizes the validity of AIC and BIC criteria. For further details on this topic, see, for example, Drton and Plummer (2017).

## 5.2 | Likelihood profile-based CIs

The Hessian-based CIs presented above rely on asymptotic normality of the ML estimator. Properties of the ML estimator may, however, change in small samples. Moreover, symmetric CIs may not be suitable if the ML estimator lies close to a boundary of the parameter space. This occurs, for example, when states are highly persistent, which leads to entries close to one in the TPM. An alternative approach to construct CIs bases on the so-called profile likelihood (see, e.g., Meeker & Escobar, 1995; Venzon and Moolgavkar, 1988), which has also shown a satisfactory performance in the context of HMMs (Visser et al., 2000).

In the following, we illustrate the principle of likelihood profile based CIs by the example of the parameter  $\lambda_2$  in our two-state Poisson HMM. The underlying basic idea is to identify those values of our parameter of interest  $\lambda_2$  in the neighborhood of  $\hat{\lambda}_2$  that lead to a significant change in the log-likelihood, whereby the other parameters (i.e.,  $\Gamma, \lambda_1$ ) are considered nuisance parameters (Meeker & Escobar, 1995). The term “nuisance parameters” means that these parameters need to be re-estimated (by maximizing the likelihood) for any fixed value of  $\lambda_2$  different to  $\hat{\lambda}_2$ . That is, the profile likelihood of  $\lambda_2$  is defined as

$$L_p(\lambda_2) = \max_{\Gamma, \lambda_1} L(\Gamma, \lambda).$$

In order to construct profile likelihood-based CIs, let  $\{\hat{\mathbf{T}}, \hat{\boldsymbol{\lambda}}\}$  denote the ML estimate for our HMM computed as described in Section 4.2. Evaluation of the log-likelihood function in this point results in the value  $\log L(\{\hat{\mathbf{T}}, \hat{\boldsymbol{\delta}}\})$ . The deviation of the

likelihood of the ML estimate and the profile likelihood in the point  $\lambda_2^p$  is then captured by the following likelihood ratio:

$$R_p(\lambda_2) = -2[\log(L_p(\lambda_2)) - \log(L(\hat{\Gamma}, \hat{\lambda}))]. \quad (3)$$

As described above, the log-likelihood  $\log(L_p(\lambda_2))$  results from re-estimating the two-state HMM with fixed parameter  $\lambda_2$ . Therefore, this model effectively corresponds to a nested model of the full model with ML estimate  $\hat{\Gamma}, \hat{\lambda}$ . Consequently,  $R_p$  asymptotically follows a  $\chi^2$  distribution with one degree of freedom—the difference in degrees of freedom between the two models. Based on this, a CI for  $\lambda_2$  can be derived by evaluating  $R_p$  at many different values of  $\lambda_2^p$  and determining when the resulting value of  $R_p$  becomes “too extreme.” That is, for a given  $\alpha$ , one needs to calculate the  $1 - \alpha$  quantile of the  $\chi_1^2$  distribution (e.g., 3.841 for  $\alpha = 5\%$ ). The CI at level  $1 - \alpha$  for the parameter  $\lambda_2$  is then given by

$$\{\lambda_2 : R_p(\lambda_2) < \chi_{1,(1-\alpha)}^2\}. \quad (4)$$

For simplicity, the principles of likelihood profiling shown above rely on the natural parameters. Our `nll` function is, however, parameterized in terms of and optimized with respect to the working parameters. In practice, this aspect is easy to deal with. Once a profile CI for the working parameter (here  $\eta_2$ ) has been obtained following the procedure above, the corresponding CI for the natural parameter  $\lambda_2$  results directly from transforming the upper and lower boundary of the CI for  $\eta_2$  by the one-to-one transformation  $\lambda_2 = \exp(\eta_2)$ . For further details on the invariance of likelihood-based CIs to parameter transformations, see, for example, Meeker and Escobar (1995).

TMB provides an easy way to profiling through the function `tmbprofile`, which requires several inputs: first, the `MakeADFun` object called `obj_tmb` from our two-state Poisson HMM; second, the position of the (working) parameter to be profiled via the `name` argument. This position refers to the position in the parameter vector `obj_tmb$par`. Moreover, here the optional `trace` argument indicates how much information on the optimization is displayed. The following commands permit to profile the second working parameter  $\eta_2 = \log(\lambda_2)$ .

```
# Parameters and covariates
m <- 2
gamma <- matrix(0.2 / (m - 1), nrow = m, ncol = m)
diag(gamma) <- 0.8
lambda <- seq(from = quantile(tinn_data, 0.1),
              to = quantile(tinn_data, 0.9),
              length.out = m)

# Parameters & covariates for TMB
working_params <- pois.HMM.pn2pw(m, lambda, gamma)
TMB_data <- list(x = tinn_data, m = m)

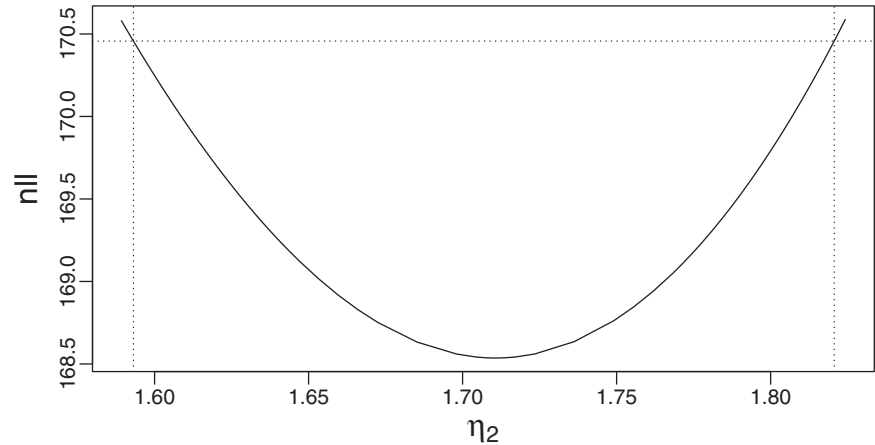
# Estimation
obj_tmb <- MakeADFun(TMB_data, parameters,
                    DLL = "poi_hmm", silent = TRUE)
mod_tmb <- nlminb(start = obj_tmb$par, objective = obj_tmb$fn,
                  gradient = obj_tmb$gr, hessian = obj_tmb$he)

# Profile
profile <- tmbprofile(obj = obj_tmb,
                     name = 2,
                     trace = FALSE)
par(mgp = c(2, 0.5, 0), mar = c(3, 3, 2.5, 1),
    cex.lab = 1.5)
plot(profile, level = 0.95,
     xlab = expression(eta[2]),
     ylab = "nll")
```



**FIGURE 2** Profile likelihood plot

*Note:* This figure shows the profile nll as function of the working parameter  $\eta_2$ . The vertical and horizontal lines correspond to the boundaries of the confidence interval and the critical value of the nll, respectively



Furthermore, Figure 2 obtained via the `plot` function shows the resulting profile nll as function of the working parameter  $\eta_2$ . The vertical and horizontal lines correspond to the boundaries of the 95% CI and the critical value of the nll derived from Equation (4), respectively. The CI for  $\eta_2$  can directly be extracted via the function `confint`:

```
# Confidence interval of tlambda
confint(profile, level = 0.95)

##           lower      upper
## tlambda 1.593141 1.820641
```

The corresponding CI for  $\lambda_2$  follows from:

```
# Confidence interval of lambda
exp(confint(profile, level = 0.95))

##           lower      upper
## tlambda 4.919178 6.175815
```

While simple linear combinations of variables can be profiled through the argument `lincomb` in the `tmbprofile` function, this is not possible for more complex functions of the parameters. This includes the stationary distribution  $\delta$ , for which CIs cannot be obtained by this method.

Note that the function `tmbprofile` carries out several optimization procedures internally for calculating profile CIs. If this approach fails, or one prefers a specific optimization routine, the necessary steps for profiling can also be implemented by the user. To do so, it would be—roughly speaking—necessary to compute  $R_p(\eta_2)$  for a sufficient number of  $\eta_2$  values to achieve the desired precision.

Last, it is noteworthy that calculating profile CIs for the elements of the TPM is not trivial. Since all elements in each row of the TPM depend on each other, the definition of the profile log-likelihood is already problematic. We apply a rather simplistic approach and compute CIs for each of the working parameters  $\tau_{i,j}$ ,  $i \neq j$ . The resulting upper and lower bounds of the profile CI are then transformed to their natural counterparts. However, other approaches may lead to improved results, see, for example, Fischer and Lewis (2021). Note that this approach seems to produce biased results for models with more than two states due to the dependence between row elements of the TPM. Hence, we advise to treat profile CIs for the TPM with care, in particular for HMMs with more than two states.

### 5.3 | Bootstrap-based CIs

The last approach for deriving CIs is the bootstrap, which is frequently applied by many practitioners. Efron and Tibshirani (1993) describe the underlying concepts of the bootstrap in their seminal manuscript. Many different bootstrap techniques have evolved since then, leading to an extensive treatment of this subject in the scientific literature.

**TABLE 2** Parameter estimates and corresponding nll of the two-state Poisson HMM with and without using TMB obtained for the TYT data

Par.	<i>DM</i>	<i>TMB<sub>0</sub></i>	<i>TMB<sub>G</sub></i>	<i>TMB<sub>H</sub></i>	<i>TMB<sub>GH</sub></i>
$\lambda_1$	1.636410931	1.636410932	1.636410933	1.636410932	1.636410997
$\lambda_2$	5.533095962	5.533095962	5.533095957	5.533095962	5.533095759
$\gamma_{11}$	0.949802041	0.949802041	0.949802042	0.949802041	0.949802094
$\gamma_{12}$	0.050197959	0.050197959	0.050197958	0.050197959	0.050197906
$\gamma_{21}$	0.025922044	0.025922044	0.025922044	0.025922044	0.025922038
$\gamma_{22}$	0.974077956	0.974077956	0.974077956	0.974077956	0.974077962
$\delta_1$	0.340541816	0.340541816	0.340541819	0.340541816	0.340541999
$\delta_2$	0.659458184	0.659458184	0.659458181	0.659458184	0.659458001
nll	168.536055869	168.536055869	168.536055869	168.536055869	168.536055869

A thorough overview of this subject would go beyond the scope of this paper. As pointed out by Härdle et al. (2003), the so-called parametric bootstrap is suitable in the context of time-series models. For further details on the bootstrap for HMMs including the implementation of a parametric bootstrap, we refer to Zucchini et al. (2016, ch. 3, pp. 56–60).

Basically all versions of the bootstrap have in common that some kind of resampling procedure needs to be carried out first. Second, the model of interest is reestimated for each of the resampled data sets. A natural way to accelerate the second part consists in the use of TMB for the model estimation by means of the procedures presented in Section 4.2. Our [Supporting Information](#) contains a detailed example illustrating the implementation of a parametric percentile bootstrap for our two-state Poisson HMM.

## 6 | APPLICATION TO DIFFERENT DATA SETS

This section aims to demonstrate the performance of TMB by means of a couple of practical examples that differ in terms of the number of observations and model complexity. These examples include the TYT data shown above, a data set of fetal lamb movements, and simulated data sets. For the performance comparisons, the focus lies on computational speed and the reliability of CIs. The R scripts necessary for this section may serve interested users for investigating their own HMM setting, and are all available in the [Supporting Information](#).

### 6.1 | TYT data

We begin by investigating the speed of five approaches for parameter estimation: one without the usage of TMB and four with TMB. In the following, *DM* denotes direct maximization of the log-likelihood through the optimization function `nllminb` without TMB. Furthermore, *TMB<sub>0</sub>*, *TMB<sub>H</sub>*, *TMB<sub>G</sub>*, and *TMB<sub>GH</sub>* denote direct maximization with TMB without using the gradient and Hessian provided by TMB, with the Hessian, with the gradient, and with both gradient and Hessian, respectively.

As a preliminary reliability check of our IT infrastructure and setup, we timed the fitting of our two-state HMM to the TYT data with the help of the `microbenchmark` package (Mersmann, 2021). For this data set, all five approaches converged to the same optimum and parameter estimates, apart from minor variations typical for numerical optimization (see Table 2).

Table 3 shows the resulting average time required for the parameter estimation and the number of iterations needed by each approach, measured over 200 replications. The results show that the use of TMB significantly accelerates parameter estimation in comparison with *DM*. The most substantial acceleration is achieved by *TMB<sub>G</sub>*, underlining the benefit of using the gradient provided by TMB. Moreover, *TMB<sub>GH</sub>* requires fewer iterations than the other approaches. However, the evaluation of the Hessian seems to increase the computational burden.

Next, we verified the reproducibility of the acceleration by TMB in a parametric bootstrap setting. More specifically, we simulated 200 bootstrap samples from the model estimated on the TYT data. Then, we reestimated the same model by our five approaches and derived acceleration ratios (with *DM* as reference approach) and their corresponding percentile CIs.

**TABLE 3** Average duration (in milliseconds) together with 95% CI and number of iterations required for fitting a two-state Poisson HMM to the TYT data

	<i>DM</i>	<i>TMB<sub>0</sub></i>	<i>TMB<sub>G</sub></i>	<i>TMB<sub>H</sub></i>	<i>TMB<sub>GH</sub></i>
Time (ms)	36.5	1.66	0.824	1.64	2.62
	(36.2, 36.9)	(1.6, 1.72)	(0.822, 0.826)	(1.6, 1.68)	(2.59, 2.65)
Iterations	13	13	13	13	7

Note: The CIs are of Wald-type and base on the standard error of the mean derived from 200 replications.

**TABLE 4** Acceleration and iterations for the TYT data

	<i>TMB<sub>0</sub></i>	<i>TMB<sub>G</sub></i>	<i>TMB<sub>H</sub></i>	<i>TMB<sub>GH</sub></i>
Acceleration ratio	21.8	41.8	21.8	16.7
	(20.7, 23)	(37.4, 46.8)	(20.8, 23.3)	(12.9, 22.1)
Iterations	12.6	12.6	12.6	5.07
	(9, 18)	(9, 19)	(9, 18)	(4, 8.03)

Note: The top lines show the acceleration ratios together with 95% percentile bootstrap CIs when using TMB in a bootstrap setting with 200 bootstrap samples. The bottom lines display the corresponding values for the number of iterations. The average parameter estimation duration without TMB is 37 milliseconds.

**TABLE 5** CIs for the TYT data set

Par.	Est.	Wald-type CI		Profile CI		Bootstrap CI		Coverage prob. (%)		
		L.	U.	L.	U.	L.	U.	Wald	Profile	Bootst.
$\lambda_1$	1.64	1.09	2.18	1.15	2.23	0.95	2.93	92.7	94.9	98.8
$\lambda_2$	5.53	4.91	6.16	4.92	6.18	4.88	6.31	93.5	93.5	98.4
$\gamma_{11}$	0.95	0.86	1.00	0.82	1.00	0.54	0.99	89.7	97.3	96.5
$\gamma_{12}$	0.05	0.00	0.14	0.00	0.18	0.01	0.46	89.7	97.3	96.5
$\gamma_{21}$	0.03	0.00	0.07	0.00	0.09	0.01	0.14	100.0	95.6	95.5
$\gamma_{22}$	0.97	0.93	1.00	0.91	1.00	0.86	0.99	100.0	95.6	95.5
$\delta_1$	0.34	0.00	0.79			0.07	0.79	86.2		95.8
$\delta_2$	0.66	0.21	1.00			0.21	0.93	86.2		95.8

Note: From left to right, the columns contain: the parameter name, parameter estimate, and lower (L.) and upper (U.) bound of the corresponding 95% CI derived via the Hessian provided by TMB, likelihood profiling, and percentile bootstrap. Then follow coverage probabilities derived for these three methods in a Monte Carlo study.

As shown in Table 4, all acceleration ratios take values significantly larger than one, whether the gradient and Hessian are passed from TMB or not. In addition, the findings from the single TYT data set are confirmed, with *TMB<sub>G</sub>* providing the most substantial acceleration and *TMB<sub>GH</sub>* reducing the number of iterations. This underlines that two factors are sources of the acceleration in Table 4: the use of C++ code on the one hand and computation of the gradient and/or Hessian by TMB on the other hand.

In order to obtain reliable results, we excluded all those bootstrap samples with a simulated state sequence not sojourning in each state of the underlying model at least once. This is necessary because such a constellation almost certainly leads to convergence problems on the one hand. On the other hand, even if the estimation algorithms converge, the estimated models are usually degenerate because of the lack of identifiability. Furthermore, for some very rare bootstrap samples, one or several of the estimation algorithms did not converge properly. In such cases, we discarded the results, generated an additional bootstrap sample, and reran the parameter estimation. Convergence problems mainly occurred due to *TMB<sub>0</sub>* and *TMB<sub>H</sub>* failing. Therefore, we recommend passing at least the gradient when optimizing with TMB for increased stability.

Last, we investigate CIs obtained for the TYT data by the three different methods described in Section 5, *TMB<sub>GH</sub>* served as the sole estimation approach. The columns to the left in Table 5 show the parameter estimates and the three types of 95% CIs obtained using the Hessian, likelihood profiling, and bootstrapping. For this data set, no major differences between the different CIs are visible. Furthermore, we assessed the accuracy of the different CIs by computing coverage probabilities, which are shown in the last three columns of Table 5. For calculating these coverage probabilities, we used

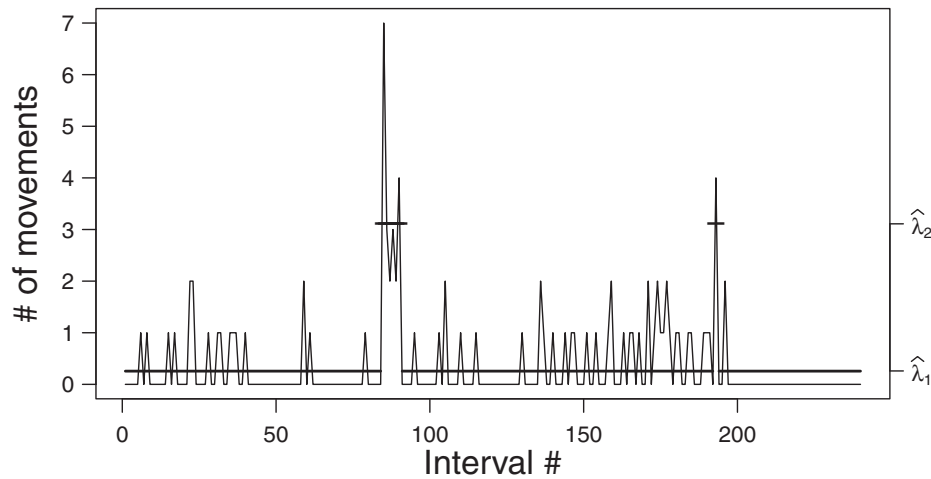


FIGURE 3 Plot of the lamb data.

Note: The solid horizontal lines correspond to the conditional mean of the inferred state at each time. See Table 7 for the values of  $\hat{\lambda}_i$

TABLE 6 Acceleration and iterations for the lamb data

	$TMB_0$	$TMB_G$	$TMB_H$	$TMB_{GH}$
Acceleration ratio	26.8	49.4	26.7	14.8
	(25, 29.1)	(44.2, 64.8)	(25.3, 29)	(9.17, 20.5)
Iterations	13.3	13.3	13.3	6.28
	(8.98, 33)	(8.98, 32)	(8.98, 33)	(4, 22)

Note: The top lines show the acceleration ratios together with 95% percentile bootstrap CIs when using TMB in a bootstrap setting with 200 bootstrap samples. The bottom lines display the corresponding values for the number of iterations. The average parameter estimation duration without TMB is 76 ms.

a Monte Carlo setting similar to the one described above. Samples that possessed state sequences not visiting all states or samples for which the estimation algorithm did not converge were replaced. Moreover, we also simulated a replacement sequence when the profile likelihood method failed to converge on any bound to ensure comparability of the results. The results, shown on the right of Table 5 indicate that all methods provide comparably reliable CI estimates, and neither outperforms the other for all parameters. For the Wald-type CIs, the coverage probabilities almost reach 100% for  $\gamma_{22}$ ,  $\gamma_{21}$ , but lie comparably low for both the other elements of the TPM and  $\delta_1$ ,  $\delta_2$ . However, profile likelihood-based CIs also take values above 95% for all elements of the TPM, and the coverage probabilities for bootstrap CIs are all above 95%. Altogether, it appears that the reliability of CIs should be investigated carefully when fitting HMMs to very short sequences of observations.

## 6.2 | Lamb data

The well-known data set presented in Leroux and Puterman (1992) serves as second example. This data set consists of the number of movements by a fetal lamb observed through ultrasound during 240 consecutive 5-s intervals, as shown in Figure 3. Since the results reported by Leroux and Puterman (1992) show that a two-state model is preferred by the BIC, we focus on this model only here—although other choices would be possible, for example, the AIC selects a three-state model.

We selected this data set for several reasons. First, the number of observations is larger than for the TYT data (but still comparably low). Second, according to the results of Leroux and Puterman (1992), the first state largely dominates the data-generating process, whereas the second state is not very persistent and linked to only a few observations. Third, the conditional means of the two states are not very different. The latter two aspects qualify these data as a “non-textbook example.”

Similar to the TYT data, all estimation algorithms converged to the same minimum of the nll, and provided almost identical parameter estimates on the original data set. A bootstrap experiment similar to the one described above for the TYT data led to comparable results, as shown in Table 6. The highest acceleration is achieved by  $TMB_G$ , whereas  $TMB_{GH}$

TABLE 7 CIs for the lamb data set

Par.	Est.	Wald-type CI		Profile CI		Bootstrap CI		Coverage prob. (%)		
		L.	U.	L.	U.	L.	U.	Wald	Profile	Bootst.
$\lambda_1$	0.26	0.18	0.34	0.15	0.33	0.08	0.33	95.0	96.2	95.7
$\lambda_2$	3.11	1.11	5.12	1.27	4.95	0.50	5.32	93.3	94.6	97.7
$\gamma_{11}$	0.99	0.97	1.00	0.93	1.00	0.71	1.00	99.9	96.1	97.9
$\gamma_{12}$	0.01	0.00	0.03	0.00	0.07	0.00	0.29	99.9	96.1	97.9
$\gamma_{21}$	0.31	0.00	0.67	0.04	0.68	0.10	1.00	93.3	97.6	98.8
$\gamma_{22}$	0.69	0.33	1.00	0.32	0.96	0.00	0.90	93.3	97.6	98.8
$\delta_1$	0.96	0.90	1.00			0.61	0.99	98.2		98.9
$\delta_2$	0.04	0.00	0.10			0.01	0.39	98.2		98.9

Note: From left to right, the columns contain: the parameter name, parameter estimate, and lower (L.) and upper (U.) bound of the corresponding 95% CI derived via the Hessian provided by TMB, likelihood profiling, and percentile bootstrap. Then follow coverage probabilities derived for these three methods in a Monte Carlo study.

achieves the lowest acceleration despite requiring a lower number of iterations than the other approaches. However, most ratios lie above those obtained for the TYT data, indicating an increased benefit of using TMB with an increasing number of observations.

Next, Table 7 shows parameter estimates and corresponding CIs in the columns to the left. Our estimates confirm the results of Leroux and Puterman (1992): The second state is not very persistent, and the conditional means  $\lambda_1$  and  $\lambda_2$  do not lie very far from each other. Concerning the CIs resulting from our three approaches, it is noticeable that the bootstrap CIs for the elements of the TPM are larger than those obtained by the other two approaches. The coverage probabilities presented in the columns to the right of Table 7 show similar patterns as observed for the TYT data. Wald-type CIs show certain deviations from 95% for the parameters related to the hidden state sequence, and the same is true for the profile CIs. Bootstrap CIs seem to be slightly too large for most parameters, leading to values greater than 95%. As for the TYT data, we recommend to be aware of the limited reliability of CIs for short sequences of observations.

On a minor note, some nonnegligible differences can be noted when comparing our estimation results to those reported by Leroux and Puterman (1992). The reasons for this are difficult to determine, but some likely explanations are given in the following. First, differences in the parameter estimates may result, for example, from the optimizing algorithms used and related setting (e.g., convergence criterion, number of steps, optimization routines used in 1992). Moreover, Leroux and Puterman (1992) seem to base their calculations on an altered likelihood, which is reduced by removing the constant term  $\sum_{i=1}^T \log(x_i!)$  from the log-likelihood. This modification may also possess an impact on the behavior of the optimization algorithm, as, for example, relative convergence criteria and step size could be affected.

### 6.3 | Simulation study

The two previously analyzed data sets are both of comparably small size. In order to systematically investigate the performance of TMB in the context of larger samples, we carried out a small simulation study. For this study, we simulated sequences of observations of length 2000 and 5000 from HMMs with two and three states, respectively. The parameters underlying the simulation are

$$\Gamma = \begin{pmatrix} 0.95 & 0.05 \\ 0.15 & 0.85 \end{pmatrix}, \quad \lambda = (1, 7),$$

for the two-state HMM and

$$\Gamma = \begin{pmatrix} 0.95 & 0.025 & 0.025 \\ 0.05 & 0.90 & 0.05 \\ 0.075 & 0.075 & 0.85 \end{pmatrix}, \quad \lambda = (1, 4, 7),$$

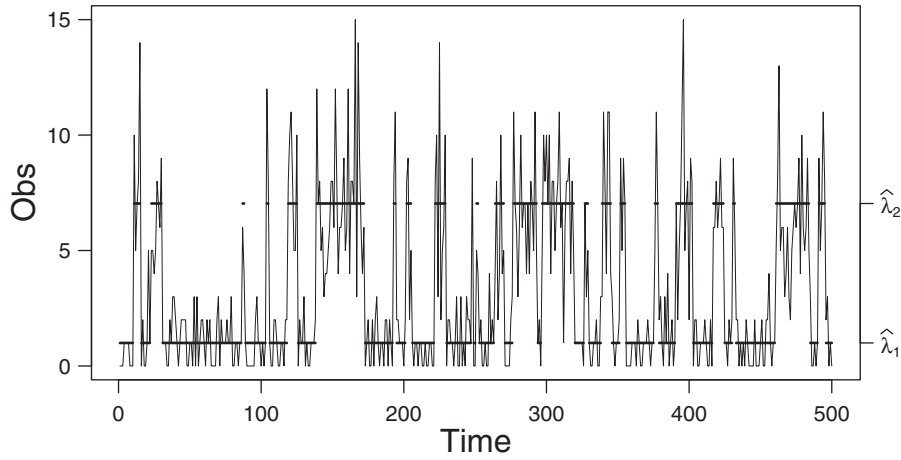


FIGURE 4 Plot of the first simulated data (of size 2000) generated by a two-state Poisson HMM

Note: The solid horizontal lines correspond to the conditional mean of the inferred state at each time. For readability, the graph is truncated to 500 data points. See Table 12 for the values of  $\hat{\lambda}_i$

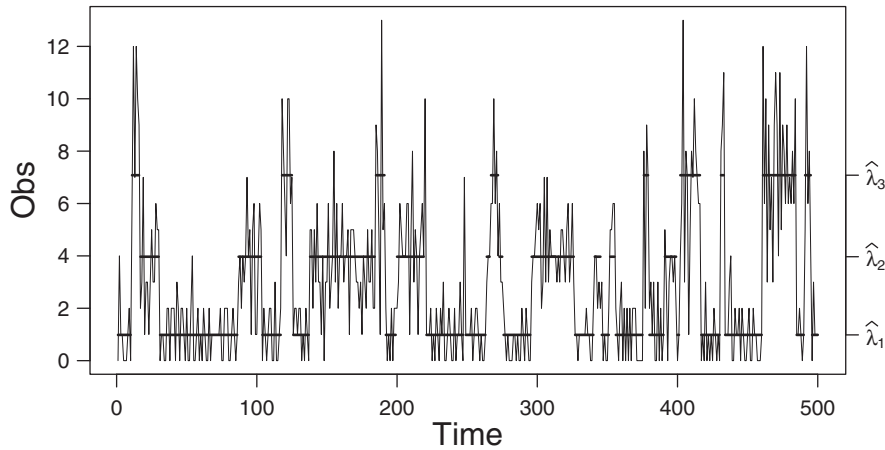


FIGURE 5 Plot of the second simulated data (of size 5000) generated by a three-state Poisson HMM

Note: The solid horizontal lines correspond to the conditional mean of the inferred state at each time. For readability, the graph is truncated to 500 data points. See Table 13 for the values of  $\hat{\lambda}_i$

for the HMM with three states. Figure 4 and Figure 5 display the first 500 observations of two exemplary sequences of observations generated for the two- and three-state model, respectively. We also simulated sequences of observations of length 2000 and 5000 from HMMs with four states. The parameters underlying both simulations are

$$\Gamma = \begin{pmatrix} 0.85 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.10 & 0.80 & 0.05 \\ 0.034 & 0.033 & 0.033 & 0.90 \end{pmatrix}, \quad \lambda = (1, 5, 9, 13).$$

The bootstrap setting described in the previous sections served for computing acceleration ratios for the simulated data sets. Table 8 displays the results obtained for the two-state model. Again,  $TMB_G$  provides the strongest acceleration compared to  $DM$ . Furthermore,  $TMB_{GH}$  achieves the second place, closely followed by the remaining two approaches. This confirms the good performance of  $TMB_G$  for large samples as well. Furthermore, it suggests a certain benefit from

TABLE 8 Acceleration and iterations for the first simulated data (of size 2000)

	$TMB_0$	$TMB_G$	$TMB_H$	$TMB_{GH}$
Acceleration ratio	22.5	41.3	22.4	23.7
	(20.6, 26.3)	(33, 50.6)	(20.6, 26.6)	(17.7, 31.2)
Iterations	22.3	22.6	22.3	4.08
	(17, 29)	(18, 28)	(17, 29)	(3, 6)

Note: The top lines show the acceleration ratios together with 95% percentile bootstrap CIs when using TMB in a bootstrap setting with 200 bootstrap samples. The bottom lines display the corresponding values for the number of iterations. The average parameter estimation duration without TMB is 832 ms.

TABLE 9 Acceleration and iterations for the second simulated data (of size 5000)

	$TMB_0$	$TMB_G$	$TMB_H$	$TMB_{GH}$
Acceleration ratio	14.2	54.1	14.2	31.4
	(12.7, 15.7)	(44.4, 65.4)	(12.6, 15.9)	(24.4, 39.5)
Iterations	42.4	42.6	42.4	4.28
	(34, 50)	(34, 51.1)	(34, 50)	(3, 6)

Note: The top lines show the acceleration ratios together with 95% percentile bootstrap CIs when using TMB in a bootstrap setting with 200 bootstrap samples. The bottom lines display the corresponding values for the number of iterations. The average parameter estimation duration without TMB is 8187 ms.

TABLE 10 Acceleration and iterations for the third simulated data (of size 2000)

	$TMB_0$	$TMB_G$	$TMB_H$	$TMB_{GH}$
Acceleration ratio	9.95	61.2	9.95	13.2
	(9.38, 10.8)	(52.8, 69.5)	(9.37, 10.7)	(8.12, 18.6)
Iterations	49.5	49.6	49.5	13
	(38, 66)	(39, 67)	(38, 66)	(6, 24)

Note: The top lines show the acceleration ratios together with 95% percentile bootstrap CIs when using TMB in a bootstrap setting with 200 bootstrap samples. The bottom lines display the corresponding values for the number of iterations. The average parameter estimation duration without TMB is 6777 ms.

employing the Hessian computed by TMB for longer observation sequences, since this method requires a much lower number of iterations than all other approaches.

The results for the three-state model, as shown in Table 9, basically underline all findings from the two-state model. Most importantly, again  $TMB_{GH}$  requires a much lower number of iterations than the other approaches (Tables 10 and 11).

Similar to the two previous sections, Table 12 to Table 15 show parameter estimates, CIs, and coverage probabilities. The four exemplary sequences of observations shown in Figure 4–7 served for deriving parameter estimates and CIs. The coverage probabilities result from a Monte Carlo study as the ones previously described. For the two-state model, the CIs

TABLE 11 Acceleration and iterations for the fourth simulated data (of size 5000)

	$TMB_0$	$TMB_G$	$TMB_H$	$TMB_{GH}$
Acceleration ratio	12.2	74.3	12.2	29.1
	(10.8, 13.7)	(62.9, 87.1)	(10.6, 13.6)	(21.5, 35.9)
Iterations	51.7	51.8	51.7	5.62
	(44, 58)	(44, 59)	(44, 58)	(4, 7.05)

Note: The top lines show the acceleration ratios together with 95% percentile bootstrap CIs when using TMB in a bootstrap setting with 200 bootstrap samples. The bottom lines display the corresponding values for the number of iterations. The average parameter estimation duration without TMB is 21358 ms.

TABLE 12 CIs for the first simulated data set (of size 2000)

Par.	Value	Est.	Wald-type CI		Profile CI		Bootstrap CI		Coverage prob. (%)		
			L.	U.	L.	U.	L.	U.	Wald	Profile	Bootst.
$\lambda_1$	1.00	0.99	0.93	1.04	0.93	1.04	0.93	1.04	95.1	95.4	94.9
$\lambda_2$	7.00	6.89	6.66	7.13	6.66	7.13	6.64	7.14	95.4	95.6	95.7
$\gamma_{11}$	0.95	0.94	0.93	0.95	0.92	0.95	0.92	0.95	95.1	95.1	95.1
$\gamma_{12}$	0.05	0.06	0.05	0.07	0.05	0.08	0.05	0.08	95.1	95.1	95.1
$\gamma_{21}$	0.15	0.16	0.13	0.20	0.13	0.20	0.13	0.20	94.1	95.1	94.5
$\gamma_{22}$	0.85	0.84	0.80	0.87	0.80	0.87	0.80	0.87	94.1	95.1	94.5
$\delta_1$	0.75	0.73	0.67	0.78			0.67	0.78	94.0		93.7
$\delta_2$	0.25	0.27	0.22	0.33			0.22	0.33	94.0		93.7

Note: From left to right, the columns contain: the parameter name, true parameter value, parameter estimate, and lower (L.) and upper (U.) bound of the corresponding 95% CI derived via the Hessian provided by TMB, likelihood profiling, and percentile bootstrap. Then follow coverage probabilities derived for these three methods in a Monte Carlo study.

TABLE 13 CIs for the second simulated data set (of size 5000)

Par.	Value	Est.	Wald-type CI		Profile CI		Bootstrap CI		Coverage prob. (%)		
			L.	U.	L.	U.	L.	U.	Wald	Profile	Bootst.
$\lambda_1$	1.000	1.016	0.97	1.06	0.97	1.06	0.98	1.06	95.4	95.3	95.6
$\lambda_2$	4.000	4.014	3.86	4.17	3.86	4.17	3.86	4.16	93.6	93.7	94.2
$\lambda_3$	7.000	7.329	7.06	7.60	7.07	7.60	7.07	7.59	94.3	94.6	94.8
$\gamma_{11}$	0.950	0.950	0.94	0.96	0.93	0.97	0.94	0.96	95.3	100.0	95.1
$\gamma_{12}$	0.025	0.026	0.02	0.03	0.02	0.04	0.02	0.03	93.8	93.4	93.7
$\gamma_{13}$	0.025	0.024	0.02	0.03	0.02	0.03	0.02	0.03	94.5	94.2	95.0
$\gamma_{21}$	0.050	0.050	0.03	0.06	0.04	0.06	0.04	0.07	95.2	93.1	95.2
$\gamma_{22}$	0.900	0.906	0.89	0.93	0.87	0.93	0.88	0.93	95.2	99.5	95.0
$\gamma_{23}$	0.050	0.045	0.03	0.06	0.03	0.06	0.03	0.06	94.1	93.3	94.8
$\gamma_{31}$	0.075	0.077	0.05	0.10	0.06	0.10	0.05	0.10	95.2	92.1	95.9
$\gamma_{32}$	0.075	0.089	0.06	0.12	0.06	0.12	0.06	0.13	94.9	93.5	95.2
$\gamma_{33}$	0.850	0.834	0.80	0.87	0.78	0.88	0.79	0.87	95.0	99.6	95.1
$\delta_1$	0.545	0.541	0.48	0.60			0.49	0.60	94.2		94.2
$\delta_2$	0.273	0.300	0.25	0.35			0.25	0.35	94.7		94.7
$\delta_3$	0.182	0.159	0.12	0.19			0.13	0.20	93.9		94.4

Note: From left to right, the columns contain: the parameter name, true parameter value, parameter estimate, and lower (L.) and upper (U.) bound of the corresponding 95% CI derived via the Hessian provided by TMB, likelihood profiling, and percentile bootstrap. Then follow coverage probabilities derived for these three methods in a Monte Carlo study.

obtained by our three methods are very similar for all parameters. The coverage probabilities lie comparably close to the theoretical level of 95% for all parameters of the two-state model. Moreover, no systematically too small or large CIs seem to result from any of the three methods. The same holds true for the three-state model, with minor exceptions for the profile CIs. For this method, the coverage probabilities of the CIs are close to 100% for diagonal elements of the TPM, while the corresponding probabilities for off-diagonal elements are less than 95%. This pattern amplifies for the four-state models, as shown in Table 15. Moreover, profile likelihood CIs of the elements of the TPM are not consistently provided by the `tmbprofile` function for the sequences of length 2000 from the four-state model. Therefore, the corresponding entries in Table 14 remain empty.

Last, we would like to point out that the runtime of our entire simulation study was approximately 1 week. Without the acceleration by TMB, several months would have been necessary. This illustrates not only the saving of time but also of resources (e.g., energy consumption, wear of IT infrastructure).



TABLE 14 CIs for the third simulated data set (of size 2000)

Par.	Value	Est.	Wald-type CI		Profile CI		Bootstrap CI		Coverage prob. (%)		
			L.	U.	L.	U.	L.	U.	Wald	Profile	Bootst.
$\lambda_1$	1.000	1.024	0.90	1.15			0.90	1.16	95.7	95.7	95.9
$\lambda_2$	5.000	5.298	5.01	5.59			4.99	5.61	93.7	93.9	94.4
$\lambda_3$	9.000	10.359	9.61	11.11			9.52	11.13	92.4	93.7	95.9
$\lambda_4$	13.000	13.451	12.98	13.92			13.01	14.00	95.1	95.3	96.1
$\gamma_{11}$	0.850	0.815	0.77	0.86			0.76	0.86	94.8		94.0
$\gamma_{12}$	0.050	0.061	0.03	0.10			0.03	0.10	93.8		94.7
$\gamma_{13}$	0.050	0.026	0.00	0.07			0.00	0.07	93.7		94.6
$\gamma_{14}$	0.050	0.097	0.05	0.14			0.05	0.15	93.2		94.9
$\gamma_{21}$	0.050	0.062	0.04	0.09			0.04	0.09	92.7		94.0
$\gamma_{22}$	0.850	0.847	0.81	0.89			0.80	0.88	95.4		95.4
$\gamma_{23}$	0.050	0.061	0.02	0.10			0.02	0.11	90.0		94.3
$\gamma_{24}$	0.050	0.031	0.00	0.06			0.00	0.07	92.7		94.1
$\gamma_{31}$	0.050	0.033	0.00	0.06			0.00	0.07	91.9		93.0
$\gamma_{32}$	0.100	0.107	0.05	0.16			0.06	0.17	93.9		94.7
$\gamma_{33}$	0.800	0.827	0.76	0.90			0.74	0.88	94.3		96.0
$\gamma_{34}$	0.050	0.033	0.00	0.07			0.00	0.09	89.5		95.4
$\gamma_{41}$	0.034	0.022	0.00	0.04			0.00	0.04	93.8		94.1
$\gamma_{42}$	0.033	0.023	0.00	0.05			0.00	0.05	91.7		93.4
$\gamma_{43}$	0.033	0.058	0.01	0.10			0.02	0.12	89.4		93.5
$\gamma_{44}$	0.900	0.896	0.86	0.93			0.85	0.93	95.1		94.4
$\delta_1$	0.223	0.173	0.13	0.22			0.13	0.22	93.1		93.2
$\delta_2$	0.266	0.278	0.22	0.34			0.21	0.34	94.1		94.8
$\delta_3$	0.177	0.231	0.15	0.31			0.15	0.32	93.8		96.9
$\delta_4$	0.333	0.318	0.22	0.42			0.21	0.42	93.6		93.8

Note: From left to right, the columns contain: the number of hidden states, parameter name, true parameter value, parameter estimate, and lower (L.) and upper (U.) bound of the corresponding 95% CI derived via the Hessian provided by TMB, likelihood profiling, and percentile bootstrap. Then follow coverage probabilities derived for these three methods in a Monte Carlo study.

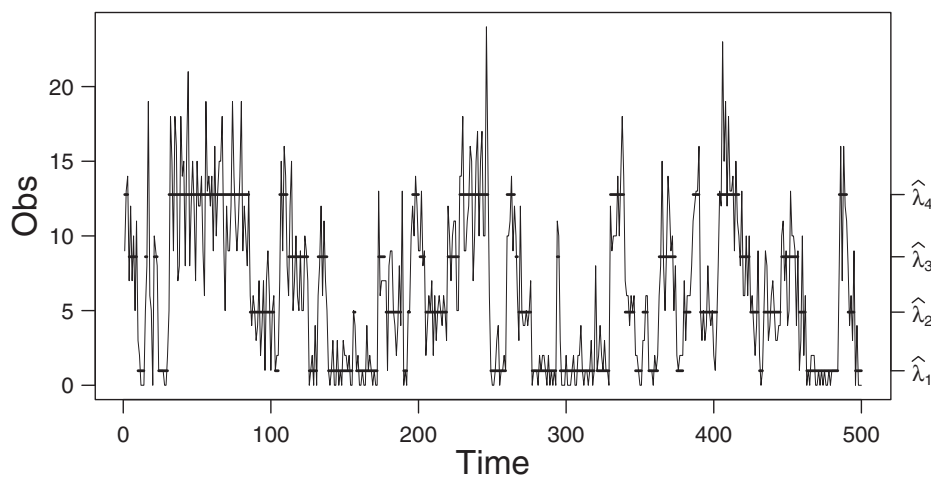


FIGURE 6 Plot of the third simulated data (of size 2000) generated by a four-state Poisson HMM

Note: The solid horizontal lines correspond to the conditional mean of the inferred state at each time. For readability, the graph is truncated to 500 data points. See Table 14 for the values of  $\hat{\lambda}_i$

TABLE 15 CIs for the fourth simulated data set (of size 5000)

Par.	Value	Est.	Wald-type CI		Profile CI		Bootstrap CI		Coverage prob. (%)		
			L.	U.	L.	U.	L.	U.	Wald	Profile	Bootst.
$\lambda_1$	1.000	0.955	0.89	1.02	0.89	1.02	0.89	1.03	96.0	95.9	95.7
$\lambda_2$	5.000	5.192	5.00	5.39	4.99	5.38	4.99	5.40	95.2	95.1	95.2
$\lambda_3$	9.000	9.253	8.65	9.86	8.66	9.86	8.63	9.85	93.4	93.6	94.6
$\lambda_4$	13.000	12.987	12.74	13.24	12.75	13.25	12.74	13.27	95.5	95.3	95.1
$\gamma_{11}$	0.850	0.840	0.82	0.86	0.78	0.89	0.82	0.86	95.1	99.9	95.0
$\gamma_{12}$	0.050	0.064	0.04	0.08	0.05	0.08	0.04	0.09	94.4	92.5	95.2
$\gamma_{13}$	0.050	0.044	0.02	0.07	0.02	0.06	0.02	0.07	93.4	90.4	93.8
$\gamma_{14}$	0.050	0.052	0.03	0.07	0.04	0.07	0.03	0.07	94.5	89.6	94.4
$\gamma_{21}$	0.050	0.058	0.04	0.08	0.04	0.07	0.04	0.08	95.0	91.2	95.3
$\gamma_{22}$	0.850	0.836	0.81	0.86	0.78	0.89	0.80	0.86	94.8	100.0	95.2
$\gamma_{23}$	0.050	0.049	0.02	0.08	0.03	0.08	0.02	0.08	94.7	92.2	95.1
$\gamma_{24}$	0.050	0.057	0.04	0.08	0.04	0.08	0.04	0.08	95.0	91.8	95.8
$\gamma_{31}$	0.050	0.057	0.02	0.09	0.03	0.08	0.03	0.09	96.1	90.0	95.9
$\gamma_{32}$	0.100	0.098	0.05	0.15	0.06	0.14	0.05	0.15	94.3	90.3	94.5
$\gamma_{33}$	0.800	0.786	0.73	0.84	0.68	0.89	0.71	0.84	94.4	100.0	95.2
$\gamma_{34}$	0.050	0.059	0.02	0.10	0.02	0.10	0.01	0.12	97.0	94.2	96.7
$\gamma_{41}$	0.034	0.029	0.02	0.04	0.02	0.04	0.02	0.04	94.5	91.2	94.3
$\gamma_{42}$	0.033	0.050	0.03	0.07	0.03	0.07	0.03	0.07	95.8	93.2	96.1
$\gamma_{43}$	0.033	0.028	0.01	0.05	0.01	0.05	0.01	0.06	95.0	94.7	96.2
$\gamma_{44}$	0.900	0.892	0.87	0.91	0.84	0.94	0.87	0.91	95.7	100.0	95.6
$\delta_1$	0.223	0.220	0.19	0.25			0.19	0.25	94.7		94.8
$\delta_2$	0.266	0.283	0.24	0.32			0.24	0.32	94.4		94.8
$\delta_3$	0.177	0.154	0.11	0.19			0.11	0.20	96.6		96.7
$\delta_4$	0.333	0.342	0.29	0.40			0.29	0.39	94.1		94.1

Note: From left to right, the columns contain: the parameter name, true parameter value, parameter estimate, and lower (L.) and upper (U.) bound of the corresponding 95% CI derived via the Hessian provided by TMB, likelihood profiling, and percentile bootstrap. Then follow coverage probabilities derived for these three methods in a Monte Carlo study.

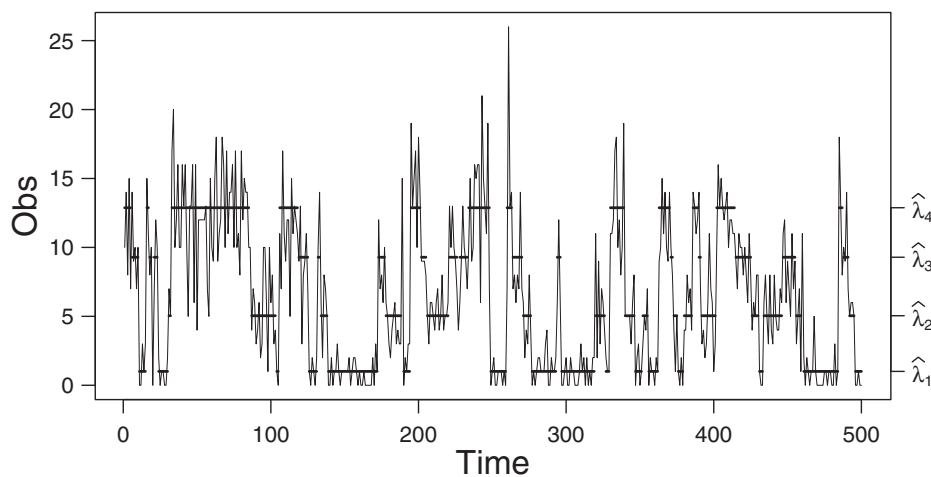


FIGURE 7 Plot of the fourth simulated data (of size 5000) generated by a four-state Poisson HMM

Note: The solid horizontal lines correspond to the conditional mean of the inferred state at each time. For readability, the graph is truncated to 500 data points. See Table 15 for the values of  $\hat{\lambda}_i$

## 7 | DISCUSSION

In this tutorial, we provide researchers from all applied fields with an introduction to parameter estimation for HMMs via TMB using R. Some procedures need to be coded in C++, which represents a certain requirement on the user and may not be beneficial if, for example, an HMM needs to be estimated only once. However, for users working more or less regularly with HMMs, the use of TMB accelerates existing parameter estimation procedures without having to carry out major changes to R code that is already in use. Moreover, after finishing the estimation procedure, TMB obtains standard errors for the estimated parameters at a very low computational cost.

We examined the performance of TMB in the context of Poisson HMMs through two small real data sets and in a simulation setting with longer sequences of observations. Overall, it is notable that the parameter estimation process is strongly accelerated on the one hand. This applies even to small data sets, and the highest acceleration is obtained when only the gradient is supplied by TMB (instead of both gradient and Hessian). On the other hand, the standard errors obtained through TMB are very similar to the standard errors obtained by profiling the likelihood and bootstrapping while being (much) less computationally intensive. This is novel since Hessian-based CIs did not seem to be reliable in the past, as illustrated, for example, by Visser et al. (2000).

Along with the tutorial character of this paper comes the shortcoming that we restricted ourselves to only one comparably simple HMM with Poisson conditional distributions. The extension to other distributions is, however, not overly complicated. We briefly illustrate the case of Gaussian conditional distributions in the [Supporting Information](#). Moreover, to keep the paper at acceptable length, we were not able to address a couple of potential extensions and research questions. For example, it would be interesting to investigate whether supplying derivatives provided by TMB to different optimizers has a positive impact on convergence properties. This may be of particular interest when considering the impact of different potentially poor initial values on the optimization results. Following Bulla and Berzel (2008), one could also consider a hybrid algorithm in this context to benefit from the advantageous properties of the EM algorithm, such as the resilience to poor initial values. When setting up this approach, the EM algorithm would most likely also benefit from acceleration by C++. The reliability of Wald-type CIs provided by TMB for other models and very long sequences with, for example, hundreds of thousands or millions of observations could also be of interest. In addition, improving the reliability of CIs for small samples (such as our TYT and lamb data) may be worth investigating. In a similar direction, it seems rather obvious to benefit from TMB for more complex settings such as panel data with random effects, where computationally efficient routines play an important role. Furthermore, in the bootstrap analysis of the TYT and in particular of the lamb data set, we noted that the TMB-internal function `tmbprofile` sometimes fails to provide CIs for parameters very close to a boundary. For these two data sets, this was the case for the elements of the TPM which take values close to one: Approximately 7% (TYT) and 29% (lamb), respectively, of the generated bootstrap samples were affected. The same problem showed for the four-state model with 2000 observations (here more than 50% of the generated data sets). It remains to clarify whether this problem can be solved by a suitable modification of `tmbprofile`, or if the underlying difficulties require an entirely different approach. Last, profile likelihood CIs for the elements of the TPM seems to be subject to bias for models with more than two states. This may be due to the strong dependence of all elements of the TPM in the same row, which is problematic for a proper definition of the profile likelihood function  $L_p(\gamma_{ij})$  difficult (see, e.g., Fischer & Lewis, 2021). Therefore, we recommend treating profile CIs for the elements of the TPM with care, in particular for models with more than two states, and further research in this direction is needed.

From an application perspective, the use of TMB allows executing estimation procedures at a significantly reduced cost compared to the execution of plain R. Such a performance gain could be of interest when repeatedly executing statistical procedures on mobile devices. It seems plausible to enrich the TYT app (or similar apps collecting a sufficient amount of data) by integrating a warning system that detects when the user enters a new state, inferred through a suitable HMM or another procedure accelerated via TMB in real time. This new state could, for example, represent an improvement or worsening of a predefined medical condition and recommend the user to contact the consulting physician if the change persists for a certain period. Provided the agreement of the user, this collected information could also be preprocessed and transferred automatically to the treating physician and allow to identify personalized treatment options.

## ACKNOWLEDGMENTS

The work of J. Bulla was supported by the GENDER-Net Co-Plus Fund (GNP-182). Furthermore, this work was supported by the Financial Market Fund (Norwegian Research Council project no. 309218). We thank the University of Regensburg and European School for Interdisciplinary Tinnitus Research (ESIT) for providing access to the TYT data. Special thanks go to J. Simões for data preparation, and W. Schlee and B. Langguth for helpful comments and suggestions. J. Bulla's

profound thanks for moral support go to the highly skilled, very helpful, and always friendly employees of Vestbrygg AS (Org. ID 912105954).


### CONFLICT OF INTEREST

The authors have declared no conflict of interest.

### DATA AVAILABILITY STATEMENT

The real data sets considered in this tutorial are available on the WWW under [https://urldefense.com/v3/\\_\\_https://timothee-bacri.github.io/HMM\\_with\\_TMB\\_!!N11eV2iwtfsls9aLuIPwskav3-zb0zKlTwa0ZrsmBrbQsrZv0GoophnwVWHp\\_SZ\\_TdgGdWB9NeOAMRUqEU2i-VEV\\_Gw\\$.](https://urldefense.com/v3/__https://timothee-bacri.github.io/HMM_with_TMB_!!N11eV2iwtfsls9aLuIPwskav3-zb0zKlTwa0ZrsmBrbQsrZv0GoophnwVWHp_SZ_TdgGdWB9NeOAMRUqEU2i-VEV_Gw$.)

### OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge **“Reproducible Research”** for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

### ORCID

Jan Bulla  <https://orcid.org/0000-0003-3308-0652>

### REFERENCES

- Ailliot, P., Allard, D., Monbet, V., & Naveau, P. (2015). Stochastic weather generators: An overview of weather type models. *Journal de la société française de statistique*, *156*, 101–113.
- Bartolucci, F., Farcomeni, A., & Pennoni, F. (2012). *Latent Markov models for longitudinal data*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Taylor & Francis.
- Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*, 1554–1563.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*, 164–171.
- Berentsen, G. D., Bulla, J., Maruotti, A., & Støve, B. (2022). Modeling clusters of corporate defaults: Regime-switching models significantly reduce the contagion source. *Journal of the Royal Statistical Society: Series C*. <https://rss.onlinelibrary.wiley.com/doi/10.1111/rssc.12551>
- Bulla, J., & Berzel, A. (2008). Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, *23*, 1–18.
- Cappé, O., Moulines, E., & Ryden, T. (2006). *Inference in hidden Markov models*. Springer Science & Business Media.
- Celex, G., & Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, *23*, 541–564.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*, 1–22.
- Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 323–380.
- Durbin, R. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*, 755–763.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- Feller, W. (1968). *An introduction to probability theory and its applications*. Wiley.
- Fischer, S. M., & Lewis, M. A. (2021). A robust and efficient algorithm to find profile likelihood confidence intervals. *Statistics and Computing*, *31*, 38.
- Fredkin, D. R., & Rice, J. A. (1992). Bayesian restoration of single-channel patch clamp recordings. *Biometrics*, *48*, 427–448.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer-Verlag.
- Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing*, *1*, 195–304.
- Gay, D. M. (1990). Usage summary for selected optimization routines. *Computing Science Technical Report*, *153*, 1–21.
- Grimmett, G., Grimmett, G. R., Grimmett, P. o. M. S. G., Stirzaker, D., & Stirzaker, M. I. D. R. (2001). *Probability and random processes*. OUP Oxford.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*, 357–384.
- Härdle, W., Horowitz, J., & Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review*, *71*, 435–459.

- Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, *84*, 717–726.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21.
- Leroux, B. G., & Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, *48*, 545–558.
- Lystig, T. C., & Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, *11*, 678–689.
- McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R., & Patterson, T. A. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, *23*, 1878–1903.
- McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
- Meeker, W. Q., & Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, *49*, 48–53.
- Mersmann, O. (2021). *microbenchmark: Accurate timing functions*. R package version 1.4.9.
- Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. *Archives of Computational Methods in Engineering*, *28*, 1429–1448.
- Pohle, J., Langrock, R., van Beest, F., & Schmidt, N. M. (2017a). Selecting the number of states in hidden Markov models—Pitfalls, practical challenges and pragmatic solutions. <https://arxiv.org/abs/1701.08673>
- Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017b). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, *22*, 270–293.
- Probst, T., Pryss, R., Langguth, B., & Schlee, W. (2016). Emotion dynamics and tinnitus: Daily life data from the “TrackYourTinnitus” application. *Scientific Reports*, *6*, 31166.
- Probst, T., Pryss, R. C., Langguth, B., Rauschecker, J. P., Schobel, J., Reichert, M., Spiliopoulou, M., Schlee, W., & Zimmermann, J. (2017). Does tinnitus depend on time-of-day? An ecological momentary assessment study with the “TrackYourTinnitus” application. *Frontiers in Aging Neuroscience*, *9*, 253.
- Pryss, R., Reichert, M., Herrmann, J., Langguth, B., & Schlee, W. (2015). Mobile crowd sensing in clinical and psychological trials—A case study. In *2015 IEEE 28th international symposium on computer-based medical systems*, IEEE, pp. 23–24.
- Pryss, R., Reichert, M., Langguth, B., & Schlee, W. (2015). Mobile crowd sensing services for tinnitus assessment, therapy, and research. In *2015 IEEE international conference on mobile services*, IEEE, pp. 352–359.
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Schadt, E. E., Sinsheimer, J. S., & Lange, K. (1998). Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Research*, *8*, 222–233.
- Turner, R. (2008). Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics & Data Analysis*, *52*, 4147–4160.
- Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *37*, 87–94.
- Visser, I., Raijmakers, M. E. J., & Molenaar, P. C. M. (2000). Confidence intervals for hidden Markov model parameters. *British Journal of Mathematical and Statistical Psychology*, *53*, 317–327.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.
- Zucchini, W., MacDonald, I., & Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R* (2nd ed.). Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Bacri, T., Berentsen, G. D., Bulla, J., & Hølleland, S. (2022). A gentle tutorial of accelerated parameter and confidence interval estimation for hidden Markov models using Template Model Builder. *Biometrical Journal*, 1–29. <https://doi.org/10.1002/bimj.202100256>