Fast and scalable inference in hidden Markov models with Gaussian fields

Jan-Ole Fischer*
Department of Business Administration and Economics, Bielefeld University

November 3, 2026

Abstract

Hidden Markov models (HMMs) are powerful tools for analyzing time series data that depend on discrete underlying but unobserved states. As such, they have gained prominence across numerous empirical disciplines, in particular ecology, medicine, and economics. However, the increasing complexity of empirical data is often accompanied by additional latent structure such as spatial effects, temporal trends, or measurement perturbations. Gaussian fields provide an attractive building block for incorporating such structured latent variation into HMMs.

Fast inference methods for Gaussian fields have emerged through the stochastic partial differential equation (SPDE) approach. Due to their sparse representation, these integrate well with novel frequentist estimation methods for random-effects models via the use of automatic differentiation and the Laplace approximation. Scaling to high dimensions requires tools such as `(R)TMB` to exploit sparsity in the Hessian w.r.t. the latent variables — a property satisfied by SPDE fields but violated by the HMM likelihood. We present a modified forward algorithm to compute the HMM likelihood, constructing sparsity in the Hessian and consequently enabling fast and scalable inference.

We demonstrate the practical feasibility and the usefulness through simulations and two case studies exploring the detection of stellar flares as well as modeling the movement of lions.

Keywords: Markov-switching, state-space models, Gaussian process, Laplace approximation, automatic differentiation

*Email: jan-ole.fischer@mailbox.org. ORCID: 0009-0004-1556-9053

1 Introduction

Over the past decade, hidden Markov models (HMMs) have become a powerful tool for modeling noisy time series driven by unobserved, serially correlated states. Their versatility has led to widespread adoption across disciplines such as ecology (McClintock, Langrock, et al., 2020; Mews, Langrock, et al., 2022), finance (Liu et al., 2012; Zhang et al., 2019), medicine (Amoros et al., 2019; Soper et al., 2020), and sports (Ötting et al., 2023; Michels, Bajons, et al., 2026). Recent advances have focused on replacing restrictive and potentially unrealistic parametric assumptions with more flexible model components. A notable development is the integration of penalized splines into HMMs, first proposed by Langrock, Kneib, Sohn, et al. (2015). Subsequent work (Langrock, Kneib, Glennie, et al., 2017; Chen and Finkenstädt, 2023; Feldmann et al., 2023; Koslik, Dupont, et al., 2025; Ammann et al., 2026) has expanded this approach and focussed on selecting an adequate degree of smoothness through a mixed model representation (Koslik, 2024; Michelot, 2025) as well as multivariate extensions (Koslik, 2025; Michels and Langrock, 2025).

While penalized splines have significantly advanced the flexibility of HMMs beyond traditional parametric formulations, their low-rank function approximations involve inherent trade-offs. Current implementations often rely on calculations with dense matrices, which can become computationally demanding as the dimensionality of the problem increases. Additionally, the penalties in penalized splines are primarily designed to enforce global function smoothness, which may not always capture more complex or domain-specific patterns in the data. On the other hand, Gaussian processes (GPs; O'Hagan, 1978) — also referred to as Gaussian *fields* (GFs) in spatial applications (Cressie, 2015) — offer a different, rich framework for function approximation. Through the choice of covariance kernels, GPs can encode prior knowledge about the underlying structure of the function — such as periodicity (e.g. Foreman-Mackey et al., 2017), spatial correlations, or non-stationarity —

making them particularly well-suited for applications where the data exhibits such intricate patterns. This adaptability opens up new possibilities for modeling complex phenomena that may not be fully captured by penalized splines alone. While historically GPs have been limited by their poor scalability, Lindgren, Rue, and Lindström (2011) developed a method to approximate Matérn fields on a triangulated mesh, transforming the finite-dimensional approximation into a Gaussian Markov random field (GMRF). This approach enables sparse computations and has gained popularity through the `R-INLA` (Lindgren and Rue, 2015) and `inlabru` (Bachl et al., 2019) `R` packages, the latter tailored for distributional regression models with latent Gaussian fields.

Integrating such fields into HMMs can be beneficial in two distinct scenarios. First, the state-dependent observation process may be perturbed by measurement error or a smooth, possibly quasi-periodic, trend (McClintock, King, et al., 2012; Esquivel et al., 2025). Such signal-plus-noise decompositions are the classical domain of state-space and unobserved-components models, estimated by Kalman or particle filters (Durban and Koopman, 2012). The distinguishing feature here is that the underlying signal is itself state-switching. Second, in more standard mixed model settings — such as those considered in `inlabru` for regression — a component of the HMM (e.g., transition probabilities between states or a state-dependent parameter) can be related to a latent (spatial) Gaussian field. Fast frequentist likelihood-based inference for both scenarios is desirable but challenging due to the high-dimensional integration required to evaluate the likelihood function. The Template Model Builder framework and its associated `R` packages, `TMB` (Kristensen et al., 2016) and `RTMB` (Kristensen, 2026), address this challenge by leveraging automatic differentiation to perform the Laplace approximation of the high-dimensional integral. However, the approach requires computing the Hessian matrix of the log-likelihood with respect to the latent variables, which is only feasible in high-dimensional settings if the Hessian is a sparse matrix. Unfortunately, such

a sparsity pattern is not present in HMMs, necessitating a tailored approximation of the HMM log-likelihood function.

In this paper, we propose such an approximation by modifying the forward algorithm to compute the HMM log-likelihood. This modification ensures local time-dependence only and consequently yields a sparse second derivative with respect to latent variables spaced out over time. Combined with the SPDE approach by Lindgren, Rue, and Lindström (2011), this enables the integration of Gaussian fields into HMMs and facilitates rapid, numerically efficient inference that scales well to high dimensions. We demonstrate the practical utility and performance of our approach through two case studies: (1) modeling stellar flares, where brightness measurements are perturbed by quasi-periodic oscillations in star brightness, and (2) movement ecology, where a spatial field influences the probability of transitions between behaviors.

2 Methods

Before we detail the novel inference procedures outlined above, we shall briefly introduce the two model components required throughout this paper.

2.1 Hidden Markov models

A basic hidden Markov model comprises two stochastic processes: an observed process $\{Y_t\}$ and a hidden process $\{S_t\}$. The unobserved process is an N-state Markov chain on the state space $\{1, \ldots, N\}$ that satisfies the Markov property

$$\Pr(S_t = s_t \mid S_{t-1} = s_{t-1}, \ldots, S_1 = s_1) = \Pr(S_t = s_t \mid S_{t-1} = s_{t-1}).$$

Due to this Markovian nature, the state process is fully specified by its initial distribution $\boldsymbol{\delta}^{(1)} = \big(\Pr(S_1 = 1), \ldots, \Pr(S_1 = N)\big)$ and one-step transition probabilities $\gamma_{ij}^{(t)} = \Pr(S_t =$

$j \mid S_{t-1} = i$) which are commonly summarized in a transition probability matrix (t.p.m.)

$$\mathbf{\Gamma}^{(t)} = (\gamma_{ij}^{(t)})_{i,j=1,\dots,N}.$$

The time index t is included here because in many applications covariates $\boldsymbol{z}_t$, $t = 1, \dots, T$, affect the probability of transitioning from one state to another. If at time $t-1$, the process is in state i, the categorical distribution of states for time t is given by $(\gamma_{i1}^{(t)}, \dots, \gamma_{iN}^{(t)})$. The covariate influence on this categorical distribution is usually modeled using a multinomial logistic regression, achieved by applying the inverse multinomial logistic link function (softmax) to linear predictors for each row of the t.p.m.,

$$\gamma_{ij}^{(t)} = \frac{\exp(\eta_{ij}^{(t)})}{\sum_{k=1}^{N} \exp(\eta_{ik}^{(t)})},$$

setting $\eta_{ii}^{(t)} = 0$ (reference category). The linear predictors $\eta_{ij}^{(t)}$ may include any type of parametric relationship (e.g. linear, quadratic, interaction), simple random effects or, as will be considered here, *random fields*.

Given a realization of the hidden state S_t, the distribution of Y_t is independent of all other observations and hidden states. More formally, if f denotes a general density,

$$f(y_t \mid s_1, \dots, s_T, y_1, \dots, y_{t-1}, y_{t+1}, \dots, y_T) = f(y_t \mid s_t).$$

As S_t only takes on finitely many values, to simplify notation we denote the corresponding densities as $f_j(y_t) = f(y_t \mid S_t = j)$, $j = 1, \dots, N$. If y_t is multivariate, a common assumption — as we make in Section 3.2 — is that its components are conditionally independent given the current state. Consequently, in this case $f_j(y_t)$ is the product of univariate densities.

2.2 Gaussian processes with sparse precision

Gaussian processes (GPs) provide a flexible framework for function modeling by specifying their distribution through a mean function, m, and a covariance kernel, $\mathcal{K}$. The defining

characteristic of a *Gaussian* process u is that for any finite collection of evaluation points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$, each in $\mathbb{R}^d$, the corresponding function values follow a multivariate *Gaussian* distribution with mean vector and covariance matrix determined by m and $\mathcal{K}$, respectively. More formally,

$$\Big(u(\boldsymbol{r}_1), \ldots, u(\boldsymbol{r}_n)\Big)^\top \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}), \tag{1}$$

where $\boldsymbol{m} = \Big(m(\boldsymbol{r}_1), \ldots, m(\boldsymbol{r}_n)\Big)^\top$, and $\boldsymbol{K}_{ij} = \mathcal{K}(\boldsymbol{r}_i, \boldsymbol{r}_j)$. In spatial and spatio-temporal settings, this framework is commonly referred to as a Gaussian random field (GRF). A major practical limitation of GP models is their computational cost. For n evaluation points, likelihood evaluation and inference typically require $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ storage due to dense covariance matrices.

However, Lindgren, Rue, and Lindström (2011) showed that Gaussian fields with Matérn covariance can be represented as the weak solution to the stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\alpha/2} u(\boldsymbol{r}) = \mathcal{W}(\boldsymbol{r}), \tag{2}$$

where Δ denotes the Laplacian operator, $\mathcal{W}(\boldsymbol{r})$ is Gaussian white noise, $\kappa > 0$ controls the spatial range, and $\boldsymbol{r} \in \mathbb{R}^d$. The parameter α determines the smoothness of the field and is usually fixed as it is poorly identified. For spatial dimension d, the corresponding Matérn smoothness parameter is given by $\nu = \alpha - d/2$.

Approximating the weak solution of (2) using a finite element basis representation,

$$u(\boldsymbol{r}) \approx \sum_{i=1}^{l} \psi_i(\boldsymbol{r})\, w_i, \tag{3}$$

where $\{\psi_i\}_{i=1,\ldots,l}$ are locally supported basis functions defined on a triangulated mesh, yields a finite-dimensional Gaussian approximation of the Matérn field. In practice, the triangulation is obtained as a Delaunay mesh over the spatial domain, refined so that the basis functions are denser where finer spatial resolution is needed (Hjelle and Dæhlen, 2006;

Lindgren and Rue, 2015). Due to the local support of the basis functions and the locality of the differential operator, the resulting system matrices are sparse. Consequently, the coefficient vector $\boldsymbol{w} = (w_1, \ldots, w_l)^\top$ follows a Gaussian Markov random field (GMRF) with sparse precision matrix. For example, for dimension $d = 2$ and $\alpha = 2$, we obtain the precision matrix

$$\boldsymbol{Q}_{\tau,\kappa} = \tau^2\big(\kappa^4\boldsymbol{C} + 2\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}\big), \tag{4}$$

where $\tau > 0$ is an overall precision parameter that scales the driving white noise and thereby determines the marginal variance of the resulting field. The $l \times l$ matrices $\boldsymbol{C}$ and $\boldsymbol{G}$ have entries

$$C_{ij} = \int \psi_i(\boldsymbol{r})\psi_j(\boldsymbol{r})\, d\boldsymbol{r}, \quad \text{and} \quad G_{ij} = \int \nabla\psi_i(\boldsymbol{r})^\top \nabla\psi_j(\boldsymbol{r})\, d\boldsymbol{r},$$

respectively. Using piecewise linear basis functions yields sparse matrices $\boldsymbol{C}$ and $\boldsymbol{G}$. In practice, $\boldsymbol{C}$ is often replaced by its lumped (diagonal) approximation such that $\boldsymbol{C}^{-1}$ is sparse (see Bolin and Lindgren, 2009 for justification).

The stochastic weights $\boldsymbol{w}$ then have multivariate Gaussian density

$$(2\pi)^{-l/2}|\boldsymbol{Q}_{\tau,\kappa}|^{1/2} \exp\Big(-\tfrac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^\top \boldsymbol{Q}_{\tau,\kappa}(\boldsymbol{w}-\boldsymbol{\mu})\Big), \tag{5}$$

where $\boldsymbol{\mu}$ is determined by the mean function m and $|\cdot|$ denotes the determinant. Evaluating this density is computationally efficient because the quadratic form only involves the non-zero entries of $\boldsymbol{Q}_{\tau,\kappa}$, and sparse Cholesky factorization can be used to compute the log-determinant efficiently.

2.3 Example model formulations

Before turning to inference, we shall highlight two examples in which combining hidden Markov models (HMMs) with Gaussian fields or Gaussian processes is natural and beneficial.

Example 1 Suppose the state-switching process of interest is not observed directly but

is contaminated by an additive noise component. A natural model formulation for the contaminated observations $y_1, \ldots, y_T$ is then

$$Y_t = X_t + u_t, \qquad t = 1, \ldots, T,$$

where the process $\{X_t\}$ follows an HMM, with each X_t's distribution depending on a discrete-valued, unobserved state S_t. The process of interest is $\{X_t\}$ — and how it is driven by $\{S_t\}$ — but we only observe its noisy version $\{Y_t\}$. The noise component u_t is independent of the state S_t and may simply be mean-zero white noise. However, it may also exhibit temporal structure, for example $u_t = u(t)$ with u following a Gaussian process with zero mean. The latter allows for smooth trends or temporally correlated disturbances that cannot be captured by independent errors. We consider such a model in the first case study.

Example 2 In spatial settings, suppose we may observe not only (error-free) $y_1, \ldots, y_T$ but also corresponding spatial locations $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_T$, each in $\mathbb{R}^2$. The observations $y_1, \ldots, y_T$ may then follow a standard state-dependent model,

$$Y_t \mid \{S_t = j\} \sim f_j(\cdot),$$

where $\{f_j\}_{j=1,\ldots,N}$ are parametric state-dependent densities. While the observation model remains parametric, the state process itself may depend on spatial location. For example, transition probabilities $\gamma_{ij}^{(t)}$ can be linked to linear predictors of the form

$$\eta_{ij}^{(t)} = \boldsymbol{\beta}_{ij}^\top \boldsymbol{z}_t + u_{ij}(\boldsymbol{r}_t), \qquad i \neq j,\ \eta_{ii}^{(t)} = 0,$$

where $\boldsymbol{z}_t$ denotes observed covariates and each $u_{ij}(\boldsymbol{r}_t)$ is a spatial Gaussian field evaluated at location $\boldsymbol{r}_t$. This formulation allows behavioral switching dynamics to vary smoothly across space. In the second case study, we apply such a model to animal movement data.

More generally, since the SPDE representation of the field is formally just another smoother (Miller et al., 2020), it may enter the model at any level — additively on the observations,

on the state-dependent parameters (as in Markov-switching generalized additive models, Langrock, Kneib, Glennie, et al., 2017), or on the transition probabilities — with inference proceeding in exactly the same way in each case.

2.4 Inference via marginal likelihood using (R)TMB

Conducting frequentist inference, i.e. maximum likelihood estimation, in any of the aforementioned models requires access to the likelihood function of the observations under the particular model of interest. The main challenge is that, although it is usually straightforward to compute the joint likelihood of the data and the latent variables, denoted by $\boldsymbol{x}$ in the following, the latter must be integrated out to obtain the marginal likelihood of the data. Precisely, letting f denote a general density and using subscripts to denote parameter dependence, we require the marginal likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \int_{\mathbb{R}^l} f_\theta(\boldsymbol{y} \mid \boldsymbol{x}) f_\theta(\boldsymbol{x}) \, d\boldsymbol{x} \tag{6}$$

of observations $\boldsymbol{y} = (y_1, \ldots, y_T)^\top$. The vector $\boldsymbol{\theta} \in \mathbb{R}^p$ parameterizes the t.p.m., the state-dependent distributions and the distribution of the Gaussian field. Equation (6) covers two distinct cases, corresponding to the example model formulations from Section 2.3 and case studies from Section 3.

1. In Example 1, $f_\theta(\boldsymbol{x})$ gives the HMM likelihood of a latent process $\boldsymbol{x} = (x_1, \ldots, x_T)^\top$ that is then perturbed by an additive Gaussian field to yield the observed series $y_1, \ldots, y_T$, with one basis function placed at each time point so that $l = T$. The term $f_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ is then the GMRF density (5), evaluated at the residual $\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{x}$, corresponding to a measurement model of the observations conditional on the process.

2. In Example 2, $f_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ is the HMM likelihood of observations $y_1, \ldots, y_T$ conditional on a latent Gaussian field, which is discretized according to (3). Here the latent variable is the vector of field weights itself, $\boldsymbol{x} = \boldsymbol{w}$, so that $f_\theta(\boldsymbol{x})$ is the GMRF

density (5), evaluated at $\boldsymbol{x} \in \mathbb{R}^l$. This conditioning may act through the transition probabilities, as in our second case study, or through the state-dependent densities.

Due to the structure of the HMM likelihood, the high-dimensional integral in Equation (6) is analytically intractable. A powerful tool for approximate frequentist inference is given by the so-called *Laplace approximation* (Erkanli, 1994; Van der Vaart, 2000). Omitting the dependence on the data $\boldsymbol{y}$ for notational simplicity and letting

$$g(\boldsymbol{x}, \boldsymbol{\theta}) = -\log f_\theta(\boldsymbol{y} \mid \boldsymbol{x}) - \log f_\theta(\boldsymbol{x}), \tag{7}$$

we can rewrite the marginal likelihood in Equation (6) as

$$\mathcal{L}(\boldsymbol{\theta}) = \int_{\mathbb{R}^l} \exp\Big(-g(\boldsymbol{x}, \boldsymbol{\theta})\Big)\, d\boldsymbol{x}. \tag{8}$$

For fixed $\boldsymbol{\theta}$ the Laplace approximation now proceeds by performing a second-order Taylor approximation of g about its minimum $\hat{\boldsymbol{x}}_\theta = \underset{\boldsymbol{x}}{\operatorname{argmin}}\ g(\boldsymbol{x}, \boldsymbol{\theta})$, yielding the quadratic approximation

$$g^*(\boldsymbol{x}, \boldsymbol{\theta}) = g(\hat{\boldsymbol{x}}_\theta, \boldsymbol{\theta}) + \tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta)^\top \boldsymbol{H}_\theta (\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta),$$

where $\boldsymbol{H}_\theta = g''_{xx}(\hat{\boldsymbol{x}}_\theta, \boldsymbol{\theta})$ is the Hessian matrix of g w.r.t. $\boldsymbol{x}$. Plugging g^* back into (8), yields

$$\mathcal{L}^*(\boldsymbol{\theta}) = \exp\Big(-g(\hat{\boldsymbol{x}}_\theta, \boldsymbol{\theta})\Big) \int_{\mathbb{R}^l} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta)^\top \boldsymbol{H}_\theta (\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta)\Big)\, d\boldsymbol{x},$$

where the latter part is a multivariate Gaussian integral with closed-form solution. Solving the integral, applying the logarithm and multiplying by -1, we arrive at the approximate marginal negative log-likelihood function

$$-\log \mathcal{L}^*(\boldsymbol{\theta}) = g(\hat{\boldsymbol{x}}_\theta, \boldsymbol{\theta}) - \tfrac{l}{2}\log(2\pi) + \tfrac{1}{2}\log|\boldsymbol{H}_\theta|. \tag{9}$$

This procedure is conveniently automated through the Template Model Builder framework and its associated `R` package `TMB` (Kristensen et al., 2016) and has recently become even more attractive through the `R` interface provided by the novel `RTMB` package (Kristensen,

2026). With the latter, the user simply needs to implement and supply the joint negative log-likelihood g as an `R` function. Using reverse-mode automatic differentiation for the derivatives of g, the package then returns (9) along with its gradient w.r.t. $\boldsymbol{\theta}$. Each call to the objective function (or its gradient) then triggers an inner Newton optimization to determine $\hat{\boldsymbol{x}}_\theta$, required for the Laplace approximation. Fitting the model then amounts to numerically minimizing (9) using any suitable gradient based numerical minimization routine, yielding a nested numerical optimization procedure. Due to the nature of the Laplace approximation, after the optimization predicted latent variables $\hat{\boldsymbol{x}}_{\hat{\theta}}$, corresponding to the posterior mode, are immediately available. Furthermore, Template Model Builder provides an estimate of the joint precision matrix of the parameters and latent variables (Bates and Watts, 1988). Using this matrix, we can sample from the approximate Gaussian distribution and use the acquired samples to quantify uncertainty. The Laplace approximation is exact when g is quadratic in $\boldsymbol{x}$ and remains accurate when the integrand is dominated by a single, sharply peaked mode. Away from this regime its accuracy is not guaranteed and should be checked in practice (Ogden, 2021). To this end we implement an importance-sampling diagnostic that quantifies the departure of the integrand from the Gaussian approximation, described in Supplementary Material S2.

A crucial practical consideration is the structure of $\boldsymbol{H}_\theta$. If the dimensionality l of $\boldsymbol{x}$ is large, computation of $\boldsymbol{H}_\theta$ becomes prohibitive in general. Only if many entries of $\boldsymbol{H}_\theta$ are exactly zero, i.e. it is a *sparse* matrix, computations become feasible and fast. An entry H_{ij} is zero if $\partial^2 g / \partial x_i \partial x_j = 0$ for all $\boldsymbol{x}$ and all $\boldsymbol{\theta}$. This requirement makes the SPDE approach outlined in Section 2.2 particularly attractive. The term corresponding to the GMRF log-likelihood will be the logarithm of (5) and hence take the form

$$-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{Q}_{\tau,\kappa}(\boldsymbol{x}-\boldsymbol{\mu}) + \text{const.}$$

The second derivative of this summand w.r.t. $\boldsymbol{x}$ (or $\boldsymbol{\mu}$) is simply $-\boldsymbol{Q}_{\tau,\kappa}$, a sparse matrix

by construction. If such a pattern is present, it is automatically detected by `(R)TMB` and exploited for all required computations. Specifically, only non-zero cross-derivatives are computed and the linear algebra required to compute (9) is performed using a sparse Cholesky factor.

In either of the two cases from Section 2.3, the remaining summand in (7) corresponds to the HMM log-likelihood. Evaluating this term requires marginalization over the latent state sequence $s_1, \ldots, s_T$ which, as we shall see, results in a dense Hessian matrix. Naively trying to apply the Laplace approximation is hence doomed to fail as evaluating $\boldsymbol{H}_\theta$ and performing computations with it is prohibitive. This marginalization is also what places these models outside the range of existing tools like `INLA`, which assume the observations to be conditionally independent given a Gaussian field; we borrow the SPDE construction from that literature purely as a sparse representation of the field. To remedy the resulting dense Hessian, in the next section we develop an approximation to the algorithm required to evaluate the HMM log-likelihood.

2.5 A banded forward algorithm

Given our two examples, we either need to compute the HMM likelihood of a latent process $x_1, \ldots, x_T$, or of observations $y_1, \ldots, y_T$ conditional on a random field $\boldsymbol{x} \in \mathbb{R}^l$. To simplify notation, in this section we focus on the first case and use the term "observations" for $x_1, \ldots, x_T$. Note, however, that this does not restrict generality. The main objective is to construct a log-likelihood function that yields zero cross-derivatives w.r.t. any two x_t and $x_{t'}$ entering log-likelihood summands at time t and t' if $|t - t'|$ is sufficiently large.

Computing the likelihood of $x_1, \ldots, x_T$ under an HMM requires marginalizing over all possible unobserved states $s_1, \ldots, s_T$ that could have generated this particular state-dependent sequence. The task of summing out the states is made feasible through the so-called forward

algorithm (Zucchini et al., 2016). Defining the *forward variables*

$$\alpha_t(j) = f(x_1, \ldots, x_t, s_t = j), \quad \boldsymbol{\alpha}_t = \Big(\alpha_t(1), \ldots, \alpha_t(N)\Big),$$

and the diagonal matrix $\boldsymbol{P}(x_t) = \text{diag}\Big(f_1(x_t), \ldots, f_N(x_t)\Big)$, we have

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}^{(1)}\boldsymbol{P}(x_1), \quad \text{and} \quad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)$$

by element-wise application of the law of total probability. Running the recursion from $t = 1$ to $t = T$ and building the sum of the entries of the last forward variable yields the likelihood of interest. However, this approach can suffer from numerical under- or overflow, and thus it is common to consider the following scaling strategy (Lystig and Hughes, 2002). Let $\mathbf{1} = (1, \ldots, 1)$ and define the *scaled* forward variables

$$\boldsymbol{\phi}_t = \frac{\boldsymbol{\alpha}_t}{\boldsymbol{\alpha}_t\mathbf{1}^\top},$$

and thus $\phi_t(j) = f(s_t = j \mid x_1, \ldots, x_t)$ by the definition of $\alpha_t(j)$. Now it can be shown that

$$\boldsymbol{\phi}_t = \frac{\boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)}{\boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top}, \quad \text{and} \quad \boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top = f(x_t \mid x_1, \ldots, x_{t-1}).$$

Observing that $\boldsymbol{\phi}_t$, $t = 1, \ldots, T$, can thus also be calculated recursively, this yields an alternative way to calculate the likelihood based on the product rule

$$f(x_1, \ldots, x_T) = \underbrace{f(x_1)}_{\boldsymbol{\delta}^{(1)}\boldsymbol{P}(x_1)\mathbf{1}^\top} \prod_{t=2}^{T} \underbrace{f(x_t \mid x_1, \ldots, x_{t-1})}_{\boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top},$$

and taking logarithms, we obtain a summation over all time points:

$$\log f(x_1, \ldots, x_T) = \log \underbrace{f(x_1)}_{\boldsymbol{\delta}^{(1)}\boldsymbol{P}(x_1)\mathbf{1}^\top} + \sum_{t=2}^{T} \log \underbrace{f(x_t \mid x_1, \ldots, x_{t-1})}_{\boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top}. \tag{10}$$

The algorithm for computing (10) then involves a recursion over time where at each time point we a) update the scaled forward variable based on the previous one and b) add the corresponding log-likelihood contribution to a running summation. The key observation here is that each term in the summation in (10) is conditional on the entire process history

Algorithm 1 Forward Recursion for log-likelihood and scaled forward variable

Require: Observations $x_1, \ldots, x_T$, transition matrices $\boldsymbol{\Gamma}^{(t)}$, state-dependent densities $f_j(x_t)$, initial state vector $\boldsymbol{\rho}$

Ensure: Log-likelihood l and last scaled forward variable $\boldsymbol{\phi}$

1: Initialize scaled forward variable:

$$\boldsymbol{\phi} \leftarrow \frac{\boldsymbol{\rho}\boldsymbol{P}(x_1)}{\boldsymbol{\rho}\boldsymbol{P}(x_1)\mathbf{1}^\top}$$

2: Initialize log-likelihood:

$$l \leftarrow \log(\boldsymbol{\rho}\boldsymbol{P}(x_1)\mathbf{1}^\top)$$

3: **for** $t = 2$ **to** T **do**

4: Update scaled forward variable:

$$\boldsymbol{\phi} \leftarrow \frac{\boldsymbol{\phi}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)}{\boldsymbol{\phi}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top}$$

5: Update log-likelihood:

$$l \leftarrow l + \log(\boldsymbol{\phi}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top)$$

6: **end for**

7: **return** $(l, \boldsymbol{\phi})$

through the scaled forward variable $\boldsymbol{\phi}_{t-1}$. Consequently, the state-dependent process of a hidden Markov model is not Markovian. Hence, the Hessian of (10) with respect to $\boldsymbol{x}$ is a dense matrix.

This second point extends more generally to derivatives taken with respect to parameters that enter only one or a small number of summands in (10). Such a situation arises in the second example, where we consider the HMM likelihood of an observed sequence $y_1, \ldots, y_T$, conditional on the GMRF $\boldsymbol{x} = (x_1, \ldots, x_l)^\top$ (recall $\boldsymbol{x} = \boldsymbol{w}$ from Section 2.4). Each weight is associated with a basis function of compact support. Indeed, at any spatial location $\boldsymbol{r}$, at most three basis functions $\psi_i(\boldsymbol{r})$ are non-zero, giving an index set $I(\boldsymbol{r})$ with at most three entries. The summand in the log-likelihood for time t can then be written as

$$\log\Big(\boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}(x_{i_1}, x_{i_2}, x_{i_3})\boldsymbol{P}(y_t)\mathbf{1}^\top\Big), \quad i_1, i_2, i_3 \in I(\boldsymbol{r}_t),$$

only depending on a small subset of the GMRF weights. The same argument applies

unchanged if the field enters the state-dependent densities rather than the transition matrix. However, the forward variable $\boldsymbol{\phi}_{t-1}$ carries information from all previous time points. Without further modification, derivatives with respect to field weights appearing at different times can interact through the recursion, producing a dense Hessian unless modifications are applied.

To achieve independence of terms sufficiently far apart in time, we split the sum in (10) into B non-overlapping blocks of equal length k. We refer to k as the *bandwidth* parameter. To simplify notation, we assume that $T = Bk$ and define $t_b = (b-1)k$. The log-likelihood can then be written as

$$\log f(x_1, \ldots, x_T) = \log f(x_1, \ldots, x_k) + \sum_{b=2}^{B} \sum_{t=t_b+1}^{bk} \log \underbrace{f(x_t \mid x_1, \ldots, x_{t-1})}_{\boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}\boldsymbol{P}(x_t)\mathbf{1}^\top}.$$

The first block, $\log f(x_1, \ldots, x_k)$, can be evaluated exactly, with the forward recursion initialized using the initial distribution $\boldsymbol{\delta}^{(1)}$, yielding the scaled forward variable $\boldsymbol{\phi}_k$. This quantity can then also be used to evaluate the log-likelihood contribution of the second block exactly.

The key difference arises when evaluating the (approximate) log-likelihood contributions for blocks $b \geq 3$. Usually, each corresponding recursion needs to be initialized with the scaled forward variable $\boldsymbol{\phi}_{t_b}$ which depends on $x_1, \ldots, x_{t_b}$. However, to achieve independence of $x_1, \ldots, x_{t_{b-1}}$, we construct an approximation $\tilde{\boldsymbol{\phi}}^k_{t_b}$. To do so, we define a fixed probability vector $\boldsymbol{\rho}$, e.g. $\boldsymbol{\rho} = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$, and propagate it through the forward recursion over the *previous* block only. That is, at time t_b, we replace the exact scaled forward variable at the start of the block, $\boldsymbol{\phi}_{t_b}$, by

$$\tilde{\boldsymbol{\phi}}^k_{t_b} = \frac{\boldsymbol{\rho} \prod_{t=t_{b-1}+1}^{t_b} \boldsymbol{\Gamma}^{(t)} \boldsymbol{P}(x_t)}{\boldsymbol{\rho} \prod_{t=t_{b-1}+1}^{t_b} \boldsymbol{\Gamma}^{(t)} \boldsymbol{P}(x_t) \mathbf{1}^\top},$$

with entries $\tilde{\phi}^k_{t_b}(j) = f(s_{t_b} = j \mid x_{t_{b-1}+1}, \ldots, x_{t_b})$. In doing so, for each block we *truncate* the conditioning on the process history and hence obtain the approximate log-likelihood

function

$$\log f(x_1, \ldots, x_T) \approx \log f(x_1, \ldots, x_k) + \sum_{b=2}^{B} \sum_{t=t_b+1}^{bk} \log \underbrace{f(x_t \mid x_{t_{b-1}+1}, \ldots, x_{t-1})}_{\tilde{\boldsymbol{\phi}}_{t-1} \boldsymbol{\Gamma}^{(t)} \boldsymbol{P}(x_t) \mathbf{1}^\top}. \tag{11}$$

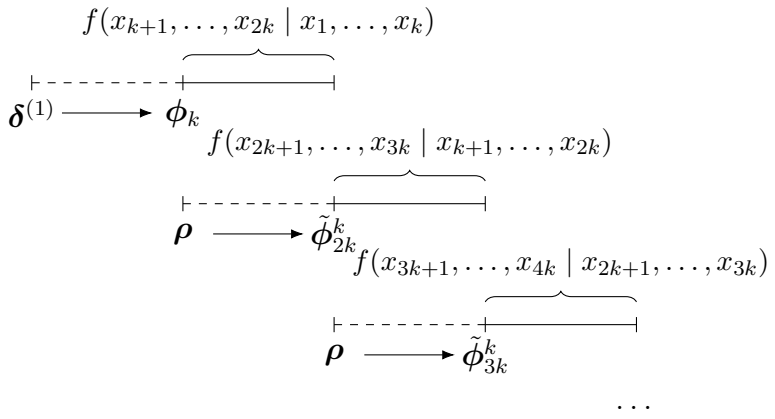


Figure 1: Visual demonstration of the approximation employed for computing the banded forward algorithm. Each block is used twice: 1) for computing its likelihood contribution and 2) for constructing an approximate scaled forward variable for the next block's likelihood contribution.

The precise algorithms for the forward recursion and the banded forward algorithm are given in Algorithms 1 and 2, respectively. Visual intuition is given by Figure 1. As each block is used *twice* to run the forward recursion, the computational cost of the algorithm is twice that of the cost of the standard forward algorithm. We verified this in a brief experiment on the data from the first case study: for a range of bandwidths k, we timed the banded forward algorithm, averaging over 100 runs (see Figure S7 of Supplementary Material S5). Beyond the low-bandwidth regime ($k \leq 5$), computation times are almost exactly double those of the standard forward algorithm. In high-dimensional random-effect settings, this is however greatly compensated by the fact that it generates a banded Hessian matrix as shown in Figure 2.

Motivation for this local approximation is given by the fact that Markov chains forget exponentially fast about their initial condition (Le Gland and Mevel, 2000). Hence, for a

Algorithm 2 Banded Forward Algorithm for Approximate log-likelihood

Require: Observations $x_1, \ldots, x_T$, transition matrices $\mathbf{\Gamma}^{(t)}$, state-dependent densities $f_j(x_t)$, initial distribution $\boldsymbol{\delta}^{(1)}$, bandwidth k, fixed initialization vector $\boldsymbol{\rho}$

Ensure: Approximate log-likelihood l

$B \leftarrow T/k$ ▷ Assume T is a multiple of k for simplicity

Compute exact log-likelihood and final scaled forward variable of first block:

$$(l, \boldsymbol{\phi}) \leftarrow \text{ForwardRecursion}(x_1, \ldots, x_k, \mathbf{\Gamma}^{(t)}, f_j, \boldsymbol{\delta}^{(1)})$$

Compute exact log-likelihood for second block

$$l \leftarrow l + \text{ForwardRecursion}(x_{k+1}, \ldots, x_{2k}, \mathbf{\Gamma}^{(t)}, f_j, \boldsymbol{\phi})$$

for $b = 3$ **to** B **do**

  Compute approximate scaled forward variable from previous block:

$$\tilde{\boldsymbol{\phi}} \leftarrow \text{ForwardRecursion}(x_{(b-2)k+1}, \ldots, x_{(b-1)k}, \mathbf{\Gamma}^{(t)}, f_j, \boldsymbol{\rho})$$

  Update log-likelihood for current block:

$$l \leftarrow l + \text{ForwardRecursion}(x_{(b-1)k+1}, \ldots, x_{bk}, \mathbf{\Gamma}^{(t)}, f_j, \tilde{\boldsymbol{\phi}})$$

end for

return l

sufficiently large lag k, we expect $\tilde{\boldsymbol{\phi}}_t^k$ to be indistinguishable from $\boldsymbol{\phi}_t$ regarding numerical precision, and thus will be the conditional likelihood contribution. Formal theoretical justification for the algorithm is given by the following theorem.

Theorem 1 (Geometric decay of the blockwise-likelihood approximation error)**.** *Let $\ell_T(\boldsymbol{\theta}, \boldsymbol{x})$ denote the exact log-likelihood given in (10) and $\tilde{\ell}_T^{\mathrm{b},k}(\boldsymbol{\theta}, \boldsymbol{x})$ its bandwidth-k blockwise approximation given in (11). Consider either of the following cases:*

(1) ***Latent-process case:*** *$\ell_T(\boldsymbol{\theta}, \boldsymbol{x}) = \log f(x_1, \ldots, x_T)$ is the log-likelihood of a latent process $\boldsymbol{x} = (x_1, \ldots, x_T)^\top$, as in Example 1.*

(2) ***Observed-data case:*** *$\ell_T(\boldsymbol{\theta}, \boldsymbol{x}) = \log f(y_1, \ldots, y_T \mid \boldsymbol{x})$ is the conditional log-likelihood of observations $y_1, \ldots, y_T$ given a latent field, discretized and represented by the finite-dimensional $\boldsymbol{x} = (x_1, \ldots, x_l)^\top$, as in Example 2.*

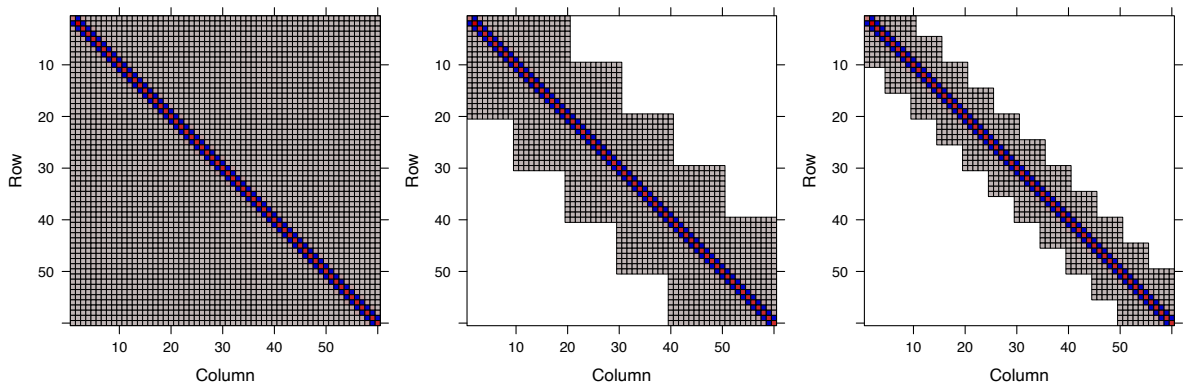


Figure 2: Comparison of the Hessian matrix w.r.t. observation sequence $x_1, \ldots, x_{60}$ when the log-likelihood is calculated using the regular forward algorithm (left), or using the banded forward algorithm with bandwidth $k = 10$ (center) and $k = 5$ (right).

Gray corresponds to values that are very close to zero in magnitude while white corresponds to exact zeros.

Let z_t denote the argument of the t-th state-dependent density, i.e. $z_t = x_t$ in Case 1 and $z_t = y_t$ in Case 2. Assume that $\boldsymbol{\Theta} \times \mathcal{X}$ is compact, that the state-dependent densities satisfy

$$0 < m \leq f_j(z_t) \leq M < \infty \qquad \textit{for all } j,\ t \textit{ and all } (\boldsymbol{\theta}, \boldsymbol{x}) \in \boldsymbol{\Theta} \times \mathcal{X}$$

(in Case 2 a condition on the observed sequence $y_1, \ldots, y_T$), and that the underlying Markov chain is uniformly ergodic in the sense of Assumption 3 and Remark S3 in Supplementary Material S1. Then there exist constants $C > 0$ and $\rho \in (0, 1)$, independent of k and T, such that

$$\sup_{(\boldsymbol{\theta}, \boldsymbol{x}) \in \boldsymbol{\Theta} \times \mathcal{X}} \left| \ell_T(\boldsymbol{\theta}, \boldsymbol{x}) - \tilde{\ell}_T^{\mathrm{b},k}(\boldsymbol{\theta}, \boldsymbol{x}) \right| \leq TC\rho^k.$$

For the proof, see Supplementary Material S1. Note that the ergodicity assumption permits structurally forbidden transitions, provided some power of $\boldsymbol{\Gamma}$ is strictly positive. This covers the stellar-flare model of Section 3.1, where $\boldsymbol{\Gamma}$ has zero entries but $\boldsymbol{\Gamma}^2$ does not. While Theorem 1 yields a theoretical guarantee that the approximation error can be made arbitrarily small by choosing a sufficiently large k, in practice, the constants C and ρ

can vary substantially with the characteristics of the data at hand. For good practical performance — balancing approximation accuracy and numerical efficiency — a suitable bandwidth should therefore always be tuned to the specific application. We study the practical implications of varying k and overall estimation performance in two simulation experiments, using a temporal and a spatial field, in Supplementary Material S3. Both indicate accurate recovery of the field and all model parameters, with close to nominal coverage of confidence intervals, and results are mostly insensitive to the bandwidth beyond $k = 2$. A moderate bandwidth thus generally suffices, though the most efficient choice remains application-specific.

In practice, we expect values around $k = 15$ to be useful initial guesses, while values below 10 should almost always be avoided. For highly persistent processes — i.e. when the diagonal entries of the transition probability matrix are close to one (e.g. > 0.95) — a larger bandwidth is typically required. The reason is that persistence corresponds to a small spectral gap and hence slow mixing, so the chain forgets its initial condition slowly and the recursion must be unrolled over more steps. The required bandwidth is not governed by the transition dynamics alone, however: as is apparent from the proof of Theorem 1, the forgetting rate also depends on the data through the state-dependent densities. A more practical strategy for selecting k is therefore to profile the banded HMM log-likelihood against k: evaluating it at fixed initial parameter values for a sequence of increasing k, and selecting the smallest k beyond which the value no longer changes appreciably. This provides a sensible starting point for the first model fit without requiring the model to be re-fitted. After the initial fit, a complementary sanity check is recommended: analysts should increase the bandwidth by, e.g., 5, and verify that the estimated parameters remain equal (to reasonable precision). If they do, this indicates that the approximation is sufficiently accurate.

2.6 Practical implementation

We use the automatic Laplace approximation implemented in the `RTMB` package (Kristensen, 2026), choosing `RTMB` over `TMB` simply because defining custom likelihood functions in `R` is usually more convenient than writing the `C++` code necessary for `TMB`.

The approximate forward algorithm from Section 2.5 is implemented in the `LaMa R` package (Fischer, 2026), a modular framework for building and fitting latent Markovian models (Mews, Koslik, et al., 2025). For the core function `forward()` that runs the forward algorithm, the user simply needs to pass a `bw` argument, corresponding to the bandwidth parameter k.

For the SPDE approach, we rely on the `fmesher R` package (Lindgren, 2026) to construct the required triangulated mesh and finite element matrices. Efficient computation of the GP likelihood is possible through `RTMB`'s `dgmrf()` function which evaluates a multivariate Gaussian density parameterized in terms of a sparse precision matrix, and thus works seamlessly with the sparse matrices provided by `fmesher`.

Penalized-complexity priors (Simpson et al., 2017) on the Matérn parameters τ and κ could in principle be added as penalties on the log-likelihood, but calibrating them is challenging without knowing the scale of the latent field a priori, and the diagnostic of Supplementary Material S2 suggests they are not required in our examples.

3 Case studies

The following two case studies demonstrate the practical usage of the proposed method for the model formulations presented in Section 2.3. All code and data for full reproducibility of the analyses are available at https://github.com/janolefi/HMMs_GFs.

3.1 Detecting stellar flares in photometric data

We revisit the analysis of Esquivel et al. (2025), focussing on stellar flare detection in the brightness measurements of M dwarf TIC 031381302 observed by TESS at a 2-minute cadence. Our goal is to demonstrate the computational advantages of our frequentist estimation approach over the dynamic Hamiltonian Monte Carlo algorithm used in their analysis. Stellar flares are sudden bursts of energy emitted by stars, primarily caused by magnetic reconnection events. Such events are crucial for understanding stellar magnetic activity, rotation, and the radiation environment of orbiting exoplanets. Hence, detecting and characterizing flares in photometric time series data is essential for studying stellar behavior.

Flares are characterized by a rapid increase in brightness followed by a gradual decay. Detection is however complicated by the fact that this is usually superimposed on quasi-periodic oscillations in the star's brightness. Traditional methods for flare detection, such as sigma-clipping, often struggle to distinguish low-energy flares from these oscillations, leading to biased energy estimates. To address these challenges, Esquivel et al. (2025) introduced a 3-state HMM, modeling the star's light curve as transitioning between "quiet", "firing", and "decaying" states, superimposed with an additive quasi-periodic trend. We retain the core structure of their model, where the observed light curve $\{Y_t\}$ is decomposed into a flaring channel $\{X_t\}$ and a quasi-periodic trend $\{u_t\}$ as

$$Y_t = X_t + u_t, \quad t = 1, \ldots, T.$$

We model u_t as a Gaussian process with scalar mean μ and, slightly deviating from Esquivel et al. (2025), with a Matérn GP via the SPDE approach. We place the non-zero mean on the trend rather than on $\{X_t\}$, which we keep mean-zero (as specified below); since the two components are additive, a constant location can be absorbed into either, so this is equivalent to the formulation in Section 2.3. In order to mirror the rotation kernel

employed by Esquivel et al. (2025), we use the extension to *oscillating* covariance functions, introduced in Section 3.3 of Lindgren, Rue, and Lindström (2011). This specification results in a precision matrix of the form

$$\boldsymbol{Q}_{\tau,\kappa,\omega} = \tau^2\Big(\kappa^4\boldsymbol{C} + 2\cos(\pi\omega)\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}\Big)$$

with matrices $\boldsymbol{C}$ and $\boldsymbol{G}$ as in Section 2.2 and oscillation parameter $\omega \in (0, 1)$. The corresponding oscillating covariance function is given in Supplementary Material S4. For this GP with 1D domain, mesh construction is simple. Using the `fmesher` package, we construct degree one B-spline basis functions (tent functions) with knots at the observed time points and compute the finite element matrices.

Again following Esquivel et al. (2025), for the states "quiet" (Q), "firing" (F), and "decaying" (D), we assume the (partially autoregressive) state-dependent model

$$\begin{aligned}
X_t \mid \{S_t = Q\} &\sim \mathcal{N}(0, \sigma^2),\\
X_t \mid \{S_t = F\}, X_{t-1} &\sim \mathcal{N}(X_{t-1}, \sigma^2) + \mathrm{Exp}(\lambda),\\
X_t \mid \{S_t = D\}, X_{t-1} &\sim \mathcal{N}(rX_{t-1}, \sigma^2)
\end{aligned}$$

The state process model is a homogeneous Markov chain, where transitions between quiet (Q) and decaying (D) and transitions between firing (F) and quiet (Q) are forbidden. Transitions from decaying (D) to firing (F) are permitted, with such a behavior being called a *compound flare*. Hence, the model has transition matrix

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{QQ} & \gamma_{QF} & 0 \\ 0 & \gamma_{FF} & \gamma_{FD} \\ \gamma_{DQ} & \gamma_{DF} & \gamma_{DD} \end{pmatrix},$$

with all non-zero entries being estimated freely. Note that, while $\boldsymbol{\Gamma}$ has zero entries, Theorem 1 still applies because all entries of $\boldsymbol{\Gamma}^2$ are strictly positive, i.e. the Markov chain is ergodic with index of primitivity 2.

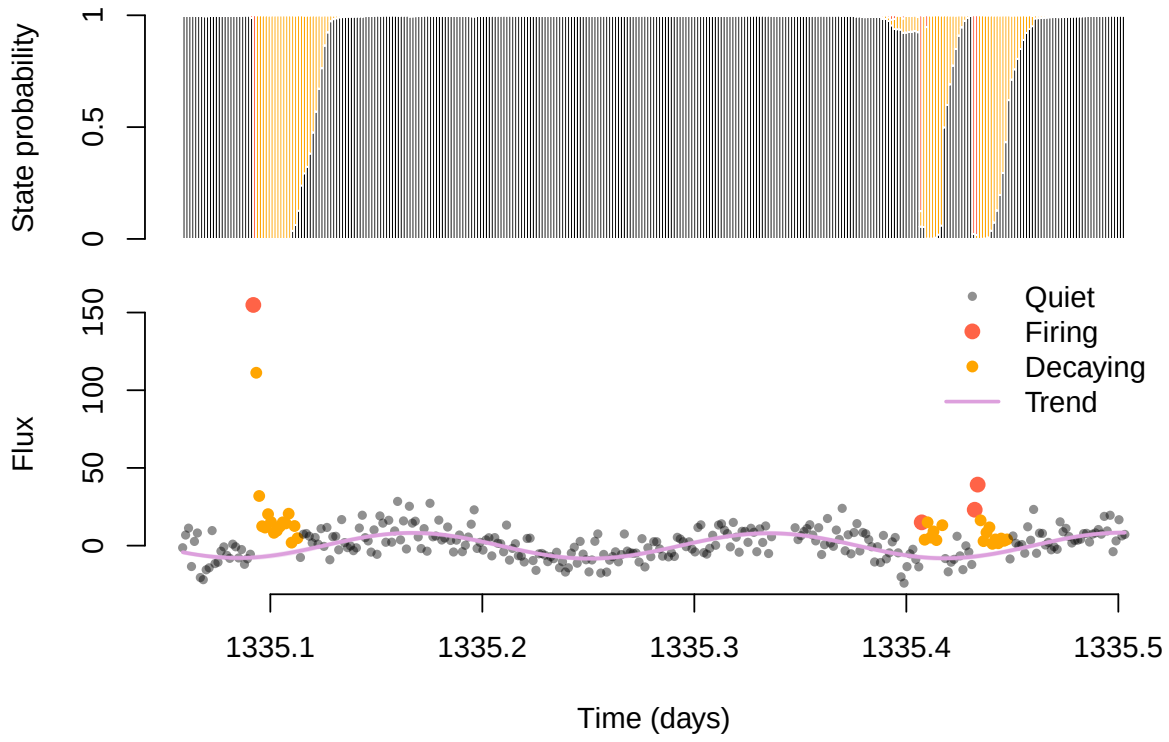


Figure 3: Top panel: locally decoded state probabilities ($\Pr(S_t = i \mid \hat{\boldsymbol{x}})$), stacked from bottom to top for the quiet, decaying, and firing states, for a subsection of the brightness time series. Bottom panel: corresponding brightness values, with symbol size indicating the Viterbi-decoded (most-likely) state (see legend). The continuous line is the estimated quasi-periodic trend, i.e. the posterior mode of the Gaussian process, shown with an approximate 95% credible interval that is barely visible due to the high estimation precision.

We implemented a custom likelihood function corresponding to the model above via the `LaMa` `R` package. We split the time series in two because of a long sequence of missing values in the middle. Finally, the unobserved true process is integrated out using the automatic Laplace approximation of the `RTMB` package. Note that in our implementation we integrate over the GMRF weights $\boldsymbol{w}$ rather than the process $\boldsymbol{x}$, which, being the linear transformation $\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{x}$ of the integration variable, gives an identical result and is the more intuitive way to implement the model. Any missing observations are, owing to the autoregressive state-dependent distributions, replaced by latent variables that are integrated out alongside the trend.

The model is fitted by numerically optimizing the Laplace-approximated marginal likelihood function using `R`'s `nlminb()` (Gay, 1990). Fitting the model to the full remaining data set of two time series and a total of $T = 19{,}259$ observations with a bandwidth parameter of $k = 30$, which we chose based on the likelihood profile shown in Figure S11 of Supplementary Material S5, took about 10 minutes on a single core of an Apple M2 chip with 16 GB of memory. This is considerably faster than the Hamiltonian Monte Carlo approach of Esquivel et al. (2025), who report 1–4 hours per 2,000-observation chunk on comparable hardware (two cores of an M1 MacBook Pro with 16 GB of memory); the difference reflects the linear scalability of the SPDE-based approach, which fits the entire series at once rather than in chunks. Their model additionally uses a different covariance kernel, so the two are not strictly identical. For this fit the importance-sampling diagnostic of Supplementary Material S2 is inconclusive, as the high latent dimension renders the check unreliable; on a subset of 5000 observations, where it is reliable, the Laplace approximation is confirmed to be accurate.

For the time series of M dwarf TIC 031381302, the model flagged 18 flare events, with an average duration of 8.7 observations including the decaying phase. We find that the model

is able to detect flare events while simultaneously controlling for the quasi-periodic trend in oscillation (cf. Figure 3). In the period analyzed by Esquivel et al. (2025), the two models agree: both flag 11 flare events with the same average flare duration of 10.3 observations, and near-identical state-switching dynamics — our maximum-likelihood estimate of the quiet-to-firing transition probability is 0.0018, against their posterior median of 0.0017. In order to detect flaring events, we performed state decoding at the predicted random effects $\hat{\boldsymbol{x}}$. The Viterbi-decoded (Viterbi, 1967) sequence shown in the bottom panel of Figure 3 is thus the maximizer of $\Pr(S_1 = j_1, \ldots, S_T = j_T \mid \hat{x}_1, \ldots, \hat{x}_T)$, and the stacked probabilities in the top panel are the local decoding probabilities $\Pr(S_t = j \mid \hat{x}_1, \ldots, \hat{x}_T)$. Specifically, note that the second flare in Figure 3 would be very difficult to detect using a simpler model without a quasi-periodic trend, as the brightness value of the observations decoded as "flaring" is lower than that of previous observations that are decoded as "quiet". The full time series with detected flares is shown in Figure S12 in Supplementary Material S5.

We do not aim to provide a formal performance comparison with the approach of Esquivel et al. (2025), nor to claim that our specification achieves the same level of physical validity. Their work carefully investigated model adequacy and detection performance using injection experiments, which is beyond the scope of this case study. Instead, our objective is to illustrate that a closely related modeling framework can be fitted using a Laplace-approximated marginal likelihood combined with an SPDE-based GP representation, resulting in a substantially reduced computational load. The example highlights the potential of this frequentist inference strategy as a scalable and practically attractive alternative.

3.2 Lion movement in Central Kalahari

In our second case study, we consider GPS movement data of lions (*Panthera leo*) in the Central Kalahari, Botswana, publicly available through the Movebank Data Repository

(MacFarlane, 2014). Data were collected from December 2008 to March 2012 and originally comprised 241,858 observations. Tracks from sensors with long, systematically missing periods not missing at random (e.g., no data between 8 a.m. and 3 p.m.) were excluded. The remaining near-hourly data were regularized to exact hourly positions by aligning observations within 15 minutes of the full hour to the hour and discarding all others. This resulted in a final dataset of $T = 53,453$ hourly positions. The GPS measurements were converted into *step lengths* (km) and *turning angles* (radians) as it is commonly done when analysing horizontal movement data (Langrock, King, et al., 2012).

We specified the state-dependent distributions for step lengths and turning angles as

$$\text{step}_t \mid \{S_t = j\} \sim \text{Gamma}(\mu_j, \sigma_j), \quad \text{angle}_t \mid \{S_t = j\} \sim \text{wrapped Cauchy}(\nu_j, \rho_j),$$

for states $j = 1, \ldots, N$. After some initial exploration we decided that a 2-state model — distinguishing between predominantly inactive and more active behavior — was the most appropriate. Although a 3-state model further subdivided the movement state, the overlap between the state-dependent distributions was substantial, and the estimated transition probability matrix exhibited little persistence. This indicates that, at an hourly temporal resolution, different types of active movement could not be reliably disentangled. We modeled the state-switching dynamics of the latent 2-state Markov chain as

$$\text{logit}(\gamma_{12}^{(t)}) = \beta_0^{(12)} + h_{12}(\text{hour}_t), \quad \text{logit}(\gamma_{21}^{(t)}) = \beta_0^{(21)} + h_{21}(\text{hour}_t) + u(\boldsymbol{r}_t),$$

where $\boldsymbol{r}_t$ is the 2D location in the study area at time t and u is a Matérn field. The periodic function in both equations is

$$h_{ij}(\text{hour}) = \sum_{k=1}^{3} \beta_{1k}^{(ij)} \sin\left(\frac{2\pi k \,\text{hour}}{24}\right) + \sum_{k=1}^{3} \beta_{2k}^{(ij)} \cos\left(\frac{2\pi k \,\text{hour}}{24}\right).$$

Only the linear predictor for the transition from state 2 (active) to state 1 (resting) includes the spatial field, as state 1 is characterized by virtually no movement. Consequently, a

switch to state 2 occurs very near to the location where the animal was resting and depends primarily on the time spent there and the time of day.

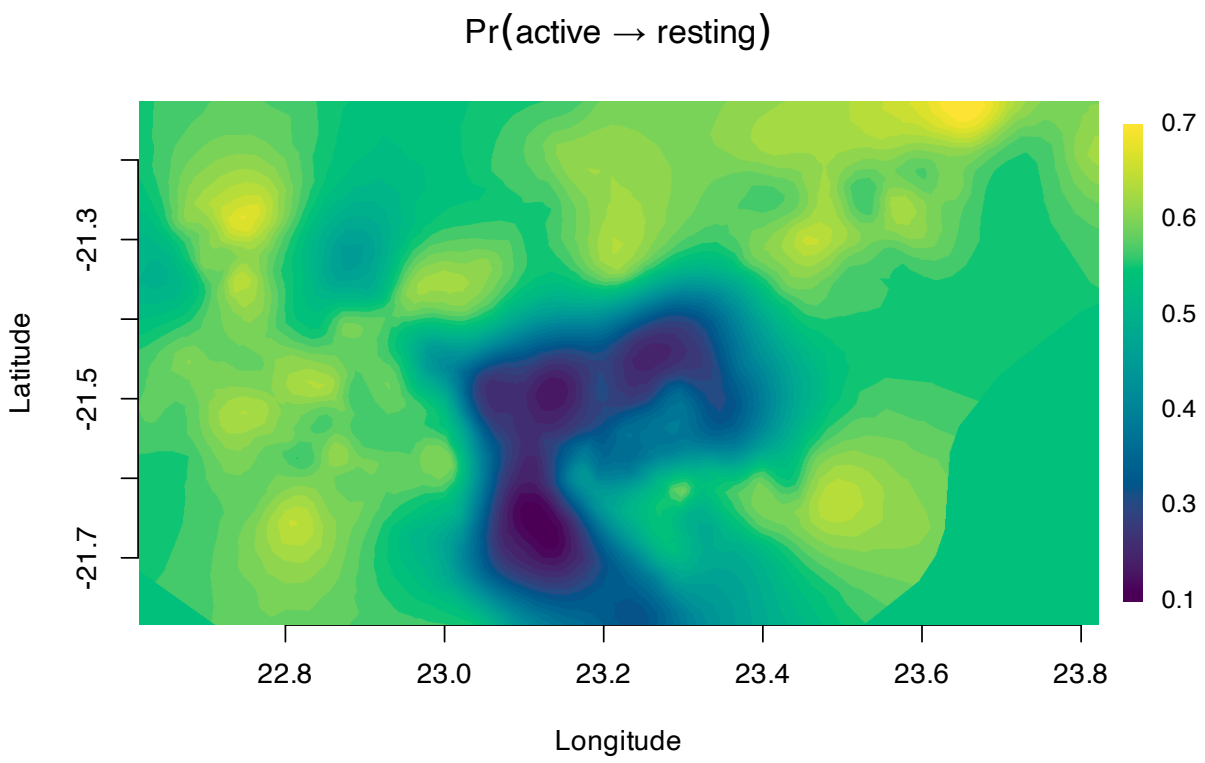


Figure 4: Transition probability from active to resting, based on the posterior mode of the spatial field fitted to the lion data. Darker areas indicate a low probability of transitioning into the resting state, while lighter areas indicate a high probability.

Again using `fmesher`, we approximated the spatial field by constructing a triangulated mesh with 2517 nodes, leading to a GMRF approximation with $l = 2517$ weights and precision matrix as given in Equation (4). The mesh is shown in Figure S13 of Supplementary Material S5. Missing positions were linearly interpolated so that the spatial field can be evaluated at every time point; missing step lengths and turning angles require no such treatment, as they are handled directly by the HMM likelihood (Zucchini et al., 2016). Fitting the model with a bandwidth parameter of $k = 15$, chosen based on the likelihood profile shown in Figure S14 of Supplementary Material S5, took about 6 minutes on a single

core of an Apple M2 chip with 16 GB of memory. Additionally, we also fitted a homogeneous model and models including only the random field or only the periodic functions h_{ij}. For the fully parametric models, fitting times were negligible in comparison to the models including a spatial field. Fitting the purely spatial model took about 4 minutes. The importance-sampling diagnostic of Supplementary Material S2 confirms that the Laplace approximation is accurate for this fit.

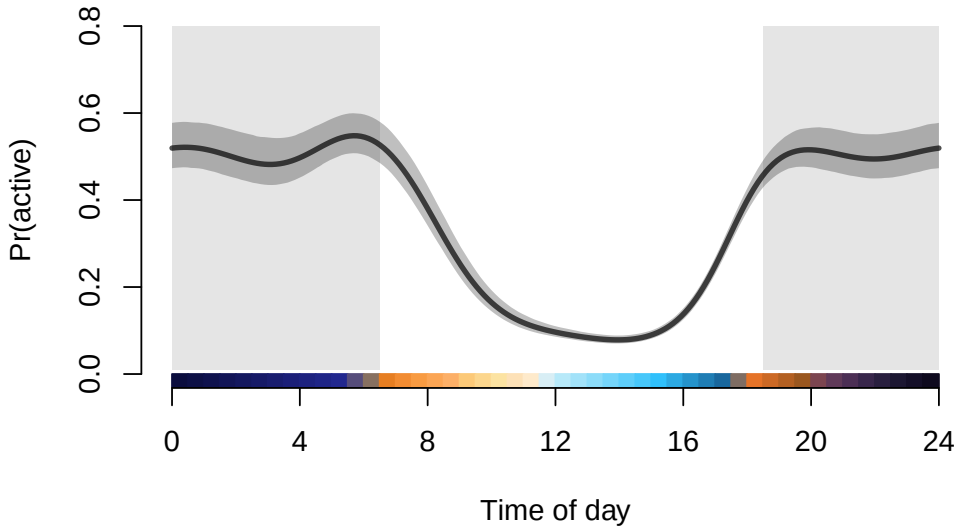


Figure 5: Probability of being active as a function of the time of day, obtained based on the periodically stationary state distribution. The gray band corresponds to pointwise 95% confidence intervals based on the approximate multivariate normal distribution of the MLE.

A comparable model can be fitted with `hmmTMB` (Michelot, 2025) through `mgcv`'s Gaussian-process smoother (`bs = "gp"`; Wood, 2023), but the SPDE approach remains preferable on two counts. Following Kammann and Wand (2003), the `gp` smoother fixes the Matérn range a priori, forcing an overly smooth field, whereas the SPDE construction estimates both the marginal variance and the range and thus recovers localized structure supported by the data. It is also far cheaper: relying on `mgcv`'s dense matrices, even a fit with 70

basis functions — arguably too few to resolve the finer features in Figure 4 — took roughly 1.4 hours for the purely spatial model, against the 4 minutes reported above. To investigate scaling performance of our approach, we refitted the spatial model on a finer mesh with $l = 5220$ basis functions, which resulted in an increased computation time of about 6.5 minutes, against the 4 minutes reported above.

The estimated state-dependent distributions are shown in Figure S15 of Supplementary Material S5. We find that the animals are very unlikely to transition into the resting state within a distinct region in the center of the study area (cf. Figure 4). In other words, once active in this area, they tend to remain active rather than settle. A comparison with satellite imagery did not reveal any obvious landscape features explaining this pattern. This spatial structure may therefore reflect more subtle ecological drivers, such as prey distribution, social interactions, or unobserved habitat characteristics, and warrants further ecological investigation.

Isolating the effect of time of day via the periodically stationary distribution (Koslik, Feldmann, et al., 2025) shows a clear diel pattern (cf. Figure 5). Activity is highest during the night and early morning hours, and lowest during the daytime, which is consistent with the predominantly nocturnal behavior of lions in arid environments.

4 Discussion

We consider combining hidden Markov models (HMMs) and latent Gaussian fields in a unified modeling framework, where, depending on the application, the GF component may enter at a different level of the model hierarchy. Recent advances in automatic differentiation and automatic Laplace approximation have, in principle, made such models computationally feasible. In particular, these tools integrate naturally with the SPDE approach, which provides Gaussian Markov random field approximations to Gaussian fields and consequently

yields sparse precision matrices. This sparsity is essential for scalable inference when the latent Gaussian component is high-dimensional. A central computational obstacle specific to HMMs is that marginalization over the discrete latent states — required to evaluate the HMM log-likelihood — induces global temporal dependence. Consequently, the Hessian matrix with respect to the latent variables is dense, undermining the efficiency gains of the automatic Laplace approximation as implemented in (R)TMB.

To address this issue, we developed a modified forward algorithm for evaluating the HMM log-likelihood. The key feature of this construction is that it generates a sparse Hessian, thereby restoring compatibility with SPDE-based GMRF representations and automatic Laplace approximation tools. Taken together, these three ingredients allow rapid and scalable inference in models that integrate discrete state switching with richly parameterized Gaussian fields. While we exploit this sparsity within (R)TMB, the construction is agnostic to the fitting method and not necessarily frequentist: any procedure forming a Gaussian approximation to the full conditional of $\boldsymbol{x}$ benefits in the same way — including, for example, INLA if the construction were to be implemented there.

We illustrated the practical performance of the approach in two case studies. The first concerned stellar flare detection, where a state-switching flaring process was masked by quasi-periodic oscillations in stellar brightness modeled via a Gaussian process. The second analyzed movement data from African lions in the Central Kalahari, where behavioral switching was modulated by a latent spatial field entering the HMM likelihood. In both cases, the proposed framework enabled flexible modeling while retaining remarkable computational efficiency.

The approach rests on the Laplace approximation, whose accuracy warrants comment. It is exact only when the joint negative log-likelihood g is quadratic in the latent variables — which holds for the Gaussian-field component but not for the HMM likelihood. In practice

it is accurate when the conditional posterior of $\boldsymbol{x}$ is unimodal and sharply peaked, but may deteriorate when it is multimodal or strongly skewed, for instance under weak identifiability or low posterior curvature. Rigorously assessing the approximation error in high-dimensional settings remains an active area of research (Ogden, 2021). We discuss the issue further, including a simple diagnostic, in Supplementary Material S2.

Our construction is structurally related to locality-based filtering in high-dimensional factorial hidden Markov models (Rimella and Whiteley, 2022), where a reach parameter analogous to our bandwidth controls a sparse approximate forward pass. There the locality is over a factor graph of a high-dimensional discrete state, whereas here it is over the temporal memory of the chain, with the SPDE representation inducing an additional spatial graph of conditional dependencies among the field weights.

Several extensions of the approach appear promising. For instance, the SPDE approach has already been generalized to latent fields that are non-stationary (Lindgren, Rue, and Lindström, 2011), and non-Gaussian or spatio-temporal processes are also possible (Lindgren, Bolin, et al., 2022). Furthermore, the SPDE framework can be applied to non-Euclidean manifolds (Lindgren, Rue, and Lindström, 2011), such as curved surfaces or even networks. Exploring these richer settings as building blocks inside HMMs — and, more broadly, a unified treatment of locality-based forward approximations spanning discrete factorial states and continuous latent fields — is an interesting avenue for future research.

5 Acknowledgments

The author thanks Roland Langrock and Nancy Heckman for helpful comments on an earlier version of this manuscript, Kasper Kristensen for answering various questions related to `RTMB`, and the reviewers for their very helpful comments, in particular the useful suggestion regarding the importance-sampling check of the Laplace approximation.

The author received no specific funding for this work.

6 Disclosure statement

The author declares that there are no conflicts of interest.

7 Data Availability Statement

All data and code to reproduce the simulation experiments and case studies are available at https://github.com/janolefi/HMMs_GFs.

References

Ammann, Katharina, Timo Adam, and Jan-Ole Koslik (2026). "Non-homogeneous Markov-switching generalised additive models for location, scale, and shape". In: *arXiv preprint arXiv:2601.03760*.

Amoros, Ruben et al. (2019). "A continuous-time hidden Markov model for cancer surveillance using serum biomarkers with application to hepatocellular carcinoma". In: *Metron* 77.2, pp. 67–86.

Bachl, Fabian E, Finn Lindgren, David L Borchers, and Janine B Illian (2019). "inlabru: an R package for Bayesian spatial modelling from ecological survey data". In: *Methods in Ecology and Evolution* 10.6, pp. 760–766.

Bates, Douglas M and Donald G Watts (1988). *Nonlinear Regression Analysis and its Applications*. Vol. 2. Wiley New York.

Bolin, David and Finn Lindgren (2009). "Wavelet Markov models as efficient alternatives to tapering and convolution fields". In: *Preprints in Mathematical Sciences* 2009.13.

Chen, Sida and Bärbel Finkenstädt (2023). "Bayesian spline-based hidden Markov models with applications to actimetry data and sleep analysis". In: *Journal of the American Statistical Association*, pp. 1–11.

Cressie, Noel (2015). *Statistics for Spatial Data*. John Wiley & Sons.

Durban, James and Siem Jan Koopman (2012). *Time Series Analysis by State Space Methods*. 2nd. Oxford University Press.

Erkanli, Alaattin (1994). "Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space". In: *Journal of the American Statistical Association* 89.425, pp. 250–258.

Esquivel, J Arturo et al. (2025). "Detecting stellar flares in photometric data using hidden Markov models". In: *The Astrophysical Journal* 979.2, p. 141.

Feldmann, Carlina C, Sina Mews, Angelica Coculla, Ralf Stanewsky, and Roland Langrock (2023). "Flexible modelling of diel and other periodic variation in hidden Markov models". In: *Journal of Statistical Theory and Practice* 17.3, p. 45.

Fischer, Jan-Ole (2026). *LaMa: Fast Numerical Maximum Likelihood Estimation for Latent Markov Models*. R package version 2.1.3. URL: https://CRAN.R-project.org/package=LaMa.

Foreman-Mackey, Daniel, Eric Agol, Sivaram Ambikasaran, and Ruth Angus (2017). "Fast and scalable Gaussian process modeling with applications to astronomical time series". In: *The Astronomical Journal* 154.6, p. 220.

Gay, David M (1990). "Usage summary for selected optimisation routines". In: *Computing Science Technical Report* 153.153, pp. 1–21.

Hjelle, Øyvind and Morten Dæhlen (2006). *Triangulations and Applications*. Springer Science & Business Media.

Kammann, EE and Matthew P Wand (2003). "Geoadditive models". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 52.1, pp. 1–18.

Koslik, Jan-Ole (2024). "Efficient smoothness selection for nonparametric Markov-switching models via quasi restricted maximum likelihood". In: *arXiv preprint arXiv:2411.11498*.

– (2025). "Tensor-product interactions in Markov-switching models". In: *arXiv preprint arXiv:2507.01555*.

Koslik, Jan-Ole, Fanny Dupont, et al. (2025). "Flexible unimodal density estimation in hidden Markov models". In: *arXiv preprint arXiv:2511.17071*.

Koslik, Jan-Ole, Carlina C Feldmann, Sina Mews, Rouven Michels, and Roland Langrock (2025). "Inference on the state process of periodically inhomogeneous hidden Markov models for animal behavior". In: *The Annals of Applied Statistics* 19.4, pp. 2724–2737.

Kristensen, Kasper (2026). *RTMB: 'R' Bindings for TMB'*. R package version 1.8. URL: https://CRAN.R-project.org/package=RTMB.

Kristensen, Kasper, Anders Nielsen, Casper W Berg, Hans Skaug, and Bradley M Bell (2016). "TMB: automatic differentiation and Laplace approximation". In: *Journal of Statistical Software* 70, pp. 1–21.

Langrock, Roland, Ruth King, et al. (2012). "Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions". In: *Ecology* 93.11, pp. 2336–2342.

Langrock, Roland, Thomas Kneib, Richard Glennie, and Théo Michelot (2017). "Markov-switching generalized additive models". In: *Statistics and Computing* 27, pp. 259–270.

Langrock, Roland, Thomas Kneib, Alexander Sohn, and Stacy L DeRuiter (2015). "Nonparametric inference in hidden Markov models using P-splines". In: *Biometrics* 71.2, pp. 520–528.

Le Gland, Francçois and Laurent Mevel (2000). "Exponential forgetting and geometric ergodicity in hidden Markov models". In: *Mathematics of Control, Signals and Systems* 13.1, pp. 63–93.

Lindgren, Finn (2026). *fmesher: Triangle Meshes and Related Geometry Tools*. R package version 0.8.0. URL: https://CRAN.R-project.org/package=fmesher.

Lindgren, Finn, David Bolin, and Håvard Rue (2022). "The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running". In: *Spatial Statistics* 50, p. 100599.

Lindgren, Finn and Håvard Rue (2015). "Bayesian spatial modelling with R-INLA". In: *Journal of Statistical Software* 63, pp. 1–25.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73.4, pp. 423–498.

Liu, Xinyi, Dimitris Margaritis, and Peiming Wang (2012). "Stock market volatility and equity returns: Evidence from a two-state Markov-switching model with regressors". In: *Journal of Empirical Finance* 19.4, pp. 483–496.

Lystig, Theodore C and James P Hughes (2002). "Exact computation of the observed information matrix for hidden Markov models". In: *Journal of Computational and Graphical Statistics* 11.3, pp. 678–689.

MacFarlane, K. (2014). "The ecology and management of Kalahari lions in a conflict area in central Botswana". Dissertation. Canberra (Australia): Australian National University.

McClintock, Brett T, Ruth King, et al. (2012). "A general discrete-time modeling framework for animal movement using multistate random walks". In: *Ecological Monographs* 82.3, pp. 335–349.

McClintock, Brett T, Roland Langrock, et al. (2020). "Uncovering ecological state dynamics with hidden Markov models". In: *Ecology Letters* 23.12, pp. 1878–1903.

Mews, Sina, Jan-Ole Koslik, and Roland Langrock (2025). "How to build your latent Markov model: The role of time and space". In: *Statistical Modelling*. in press.

Mews, Sina, Roland Langrock, Ruth King, and Nicola Quick (2022). "Multistate capture–recapture models for irregularly sampled data". In: *The Annals of Applied Statistics* 16.2, pp. 982–998.

Michelot, Théo (2025). “hmmTMB: Hidden Markov models with flexible covariate effects in R”. In: *Journal of Statistical Software* 114, pp. 1–45.

Michels, Rouven, Robert Bajons, and Jan-Ole Fischer (2026). “Integrating unsupervised and supervised learning for the prediction of defensive schemes in American football”. In: *arXiv preprint arXiv:2602.10784*.

Michels, Rouven and Roland Langrock (2025). “Nonparametric estimation of bivariate hidden Markov models using tensor-product B-splines”. In: *Statistical Modelling* 0.0, p. 1471082X251335431.

Miller, David L, Richard Glennie, and Andrew E Seaton (2020). “Understanding the stochastic partial differential equation approach to smoothing”. In: *Journal of Agricultural, Biological and Environmental Statistics* 25.1, pp. 1–16.

O’Hagan, Anthony (1978). “Curve fitting and optimal design for prediction”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 40.1, pp. 1–24.

Ogden, Helen (2021). “On the error in Laplace approximations of high-dimensional integrals”. In: *Stat* 10.1, e380.

Ötting, Marius, Roland Langrock, and Antonello Maruotti (2023). “A copula-based multivariate hidden Markov model for modelling momentum in football”. In: *AStA Advances in Statistical Analysis* 107.1, pp. 9–27.

Rimella, Lorenzo and Nick Whiteley (2022). “Exploiting locality in high-dimensional factorial hidden Markov models”. In: *Journal of Machine Learning Research* 23.4, pp. 1–34.

Simpson, Daniel, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye (2017). “Penalising model component complexity: a principled, practical approach to constructing priors”. In: *Statistical Science* 32.1, pp. 1–28.

Soper, Braden C, Mari Nygård, Ghaleb Abdulla, Rui Meng, and Jan F Nygård (2020). “A hidden Markov model for population-level cervical cancer screening data”. In: *Statistics in Medicine* 39.25, pp. 3569–3590.

Van der Vaart, Aad W (2000). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13.2, pp. 260–269.

Wood, Simon N (2023). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.9-1. URL: https://CRAN.R-project.org/package=mgcv.

Zhang, Mengqi, Xin Jiang, Zehua Fang, Yue Zeng, and Ke Xu (2019). "High-order hidden Markov model for trend prediction in financial time series". In: *Physica A: Statistical Mechanics and its Applications* 517, pp. 1–12.

Zucchini, Walter, Iain L MacDonald, and Roland Langrock (2016). *Hidden Markov Models for Time Series: An Introduction using R*. 2nd. Chapman and Hall/ CRC press.