Supplementary Material for "Fast and scalable inference in hidden Markov models with Gaussian fields"

S1 Proof of Theorem 1

Notation and setup

Let $\ell_T(\boldsymbol{\theta}, \boldsymbol{x})$ denote the log-likelihood function based on T observations, given in Equation (10) of the main paper. This covers two cases:

- **Case 1:** $\ell_T(\boldsymbol{\theta}, \boldsymbol{x}) = \log f(x_1, \ldots, x_T)$ is the log-likelihood of latent variables $x_1, \ldots, x_T$, as in Example 1.
- **Case 2:** $\ell_T(\boldsymbol{\theta}, \boldsymbol{x}) = \log f(y_1, \ldots, y_T \mid \boldsymbol{x})$ is the conditional log-likelihood of observations $y_1, \ldots, y_T$ given a latent field which is discretized and represented by the finite-dimensional parameter $\boldsymbol{x} = (x_1, \ldots, x_l)^\top$, as in Example 2.

In both cases, let $\tilde{\ell}_T^{\mathrm{b},k}(\boldsymbol{\theta}, \boldsymbol{x})$ be the blockwise, bandwidth-k approximation given in Equation (11) of the main paper. From now on, we use the auxiliary variable z_t to denote x_t in Case 1 and y_t in Case 2, such that we can cover both cases at once. In Case 2, all densities are additionally conditioned on the latent field $\boldsymbol{x}$, which we suppress from the notation throughout. Using the auxiliary variable z_t, let $\boldsymbol{\phi}_t$ denote the exact scaled forward variable with entries $\phi_t(j) = \Pr(S_t = j \mid z_1, \ldots, z_t)$, and let $\tilde{\boldsymbol{\phi}}_t^k$ denote the truncated approximation based on the last k observations, with entries $\tilde{\phi}_t^k(j) = \Pr(S_t = j \mid z_{t-k+1}, \ldots, z_t)$. Now define the prediction filter $\boldsymbol{p}_t = \boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}^{(t)}$ and its approximation $\tilde{\boldsymbol{p}}_t^k = \tilde{\boldsymbol{\phi}}_{t-1}^k\boldsymbol{\Gamma}^{(t)}$. We analyze two approximations to ℓ_T. The *elementwise* approximation $\tilde{\ell}_T^k$ replaces each predictive filter $\boldsymbol{p}_t$ by the truncated filter $\tilde{\boldsymbol{p}}_t^k$ defined above, which conditions on exactly the last k observations $z_{t-k+1}, \ldots, z_t$. The *blockwise* approximation $\tilde{\ell}_T^{\mathrm{b},k}$ of Equation (11), which we ultimately use, instead re-anchors the filter once per block of length k, so that each likelihood contribution

conditions on *at least* k preceding observations: the first element of a block has lag k and the last has lag up to $2k-1$. Our main argument will use the forgetting bound of Le Gland and Mevel (2000), which is monotone in the lag — additional conditioning observations contribute further factors in $(0,1)$ to the product bound (see the proof of Theorem 1). Therefore, each blockwise filter error is bounded by the corresponding elementwise error at lag k. It therefore suffices to establish the geometric bound for the elementwise approximation; the same bound, with the same constants C and ρ, then holds for the blockwise approximation $\tilde{\ell}_T^{\mathrm{b},k}$. We prove the elementwise bound in the remainder of this section.

Assumptions

We make the following assumptions (see Section "Remarks on the assumptions" for more details)

1. **Compact parameter and latent space:** $\boldsymbol{\Theta} \times \mathcal{X}$ is a compact subset of $\mathbb{R}^p \times \mathbb{R}^l$.

2. **Bounded state-dependent densities:** For all states j, the emission densities satisfy

$$0 < m \leq f_j(z_t) \leq M < \infty,$$

   uniformly over $(\boldsymbol{\theta}, \boldsymbol{x}) \in \boldsymbol{\Theta} \times \mathcal{X}$, where $z_t = x_t$ in Case 1 and $z_t = y_t$ (for the fixed observations $t = 1, \ldots, T$) in Case 2.

3. **Uniform ergodicity of the Markov chain:** There exists a $q \geq 1$ and a constant $\varepsilon > 0$, the minimum positive single-step entry of the transition matrices, such that, for all t,

$$\left(\boldsymbol{\Gamma}^{(t)}\boldsymbol{\Gamma}^{(t+1)} \cdots \boldsymbol{\Gamma}^{(t+q-1)}\right)_{ij} \geq \varepsilon^q \quad \text{for all states } i, j.$$

   We call q the index of primitivity; see Remark S3.

Lemma S1 (Filter bound propagates to log-likelihood)**.** *Under Assumptions 1–3, we have for all T*

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x}) \in \boldsymbol{\Theta} \times \mathcal{X}} \left| \ell_T(\boldsymbol{\theta}, \boldsymbol{x}) - \tilde{\ell}_T^k(\boldsymbol{\theta}, \boldsymbol{x}) \right| \leq \sum_{t=1}^{T} \frac{M}{m} \sup_{(\boldsymbol{\theta},\boldsymbol{x})} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\|$$

where $\|\boldsymbol{u}\| = \sum_i |u_i|$ denotes the L_1-norm of a vector $\boldsymbol{u}$.

Proof. Recall that for $j = 1, \ldots, N$ we use the shorthand $f_j(z_t) = f(z_t \mid S_t = j)$ to denote the state-dependent densities. The conditional densities multiplied in the likelihood are then

$$L_t \coloneqq f(z_t \mid z_1, \ldots, z_{t-1}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \phi_{t-1}(i) \gamma_{ij}^{(t)} f_j(z_t) = \sum_{j=1}^{N} p_t(j) f_j(z_t),$$

by definition of $p_t(j)$. Denote with $\tilde{L}_t^k$ the same quantity with $\boldsymbol{\phi}_{t-1}$ replaced by $\tilde{\boldsymbol{\phi}}_{t-1}^k$ (equivalently, $\boldsymbol{p}_t$ by $\tilde{\boldsymbol{p}}_t^k$). Since $f_j(z_t) \leq M$ for all j and

$$L_t - \tilde{L}_t^k = \sum_{j=1}^{N} \Big(p_t(j) - \tilde{p}_t^k(j)\Big) f_j(z_t),$$

the error in the conditional densities is bounded by

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x})} |L_t - \tilde{L}_t^k| \leq M \sup_{(\boldsymbol{\theta},\boldsymbol{x})} \sum_{j=1}^{N} \Big|p_t(j) - \tilde{p}_t^k(j)\Big| = M \sup_{(\boldsymbol{\theta},\boldsymbol{x})} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\|.$$

By the mean value theorem for logarithms,

$$|\log L_t - \log \tilde{L}_t^k| = \frac{|L_t - \tilde{L}_t^k|}{\xi} \leq \frac{1}{m} |L_t - \tilde{L}_t^k|,$$

where $\xi \in (\min(L_t, \tilde{L}_t^k), \max(L_t, \tilde{L}_t^k)) \geq m$, since both L_t and $\tilde{L}_t^k$ are convex combinations of the $f_j(z_t) \geq m$. Combining with the previous bound,

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x})} |\log L_t - \log \tilde{L}_t^k| \leq \frac{M}{m} \sup_{(\boldsymbol{\theta},\boldsymbol{x})} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\|.$$

Summing over $t = 1, \ldots, T$ and applying the triangle inequality to $\ell_T = \sum_{t=1}^{T} \log L_t$ (and likewise for $\tilde{\ell}_T^k$), we obtain

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x})} |\ell_T(\boldsymbol{\theta}, \boldsymbol{x}) - \tilde{\ell}_T^k(\boldsymbol{\theta}, \boldsymbol{x})| \leq \sum_{t=1}^{T} \frac{M}{m} \sup_{(\boldsymbol{\theta},\boldsymbol{x})} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\|.$$

□

Proof of Theorem 1: Case 1. We first cover the result when $\ell_T(\boldsymbol{\theta}, \boldsymbol{x}) = \log f(x_1, \ldots, x_T)$ is a deterministic function of a parameter $\boldsymbol{\theta}$ and latent variables $\boldsymbol{x}$. We initially show that the approximation error decays geometrically in k, that is

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x}) \in \boldsymbol{\Theta} \times \mathcal{X}} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\| \leq C' \rho^k, \qquad \text{for all } t,$$

$C' > 0$ and $\rho \in (0, 1)$, relying on the work of Le Gland and Mevel (2000), in order to then apply Lemma S1. Under Assumptions 1–3, in particular the uniform minorization of Remark S3, the assumptions of Le Gland and Mevel (2000) (primitivity of the chain) are satisfied. The truncated filter $\tilde{\boldsymbol{p}}_t^k$ corresponds to a filter started at time $t-k$ from an arbitrary initial distribution and propagated forward to time t. In the notation of Le Gland and Mevel

(2000), this is equivalent to comparing filters on an interval $[m, n]$ with $m = t-k$ and $n = t$, so that the exponential forgetting rate depends only on the lag $n - m = k$. In our notation, their Theorem 2.1 states

$$\|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\| \leq 2 \prod_{\kappa=1}^{\lfloor (k+1)/q \rfloor} (1 - \varepsilon^q [\delta(x_{(t-k)+(\kappa-1)q+1}) \cdots \delta(x_{(t-k)+\kappa q-1})]^{-1}), \tag{S1}$$

where q is the index of primitivity of the chain (see Assumption 3) and we define

$$\delta(x) = \frac{\max_j f_j(x)}{\min_j f_j(x)}.$$

Since in our case $\boldsymbol{\Gamma}^{(t)}$ might be time-varying due to covariate-dependence, we define $\varepsilon = \min^+_{i,j,t} \gamma_{ij}^{(t)}$, the minimum positive entry over *all* transition matrices, where Le Gland and Mevel (2000) require only the single $\min_+ \gamma_{ij}$ of their time-homogeneous chain. This worst-case choice yields a uniform per-step minorization ε valid at every t, so their bound applies with ε in place of the homogeneous minimum, provided the support pattern is time-invariant so that ε composes to the q-step bound ε^q (Remark S3). The product over κ corresponds to successive blocks of length q, and Equation (S1) holds for $k \geq q-1$, so that at least one full block is present. Boundedness of the state-dependent densities implies $\delta(x) \leq \frac{M}{m}$ for any x, so the bracketed term satisfies $[\delta(\cdots)\cdots\delta(\cdots)]^{-1} \geq (\frac{m}{M})^{q-1}$ (there are $q-1$ factors per block). Each factor in the product is therefore bounded by

$$1 - \varepsilon^q [\delta(\cdots)\cdots\delta(\cdots)]^{-1} \leq 1 - \varepsilon^q \Big(\tfrac{m}{M}\Big)^{q-1} =: 1 - \beta, \qquad \beta = \varepsilon^q \Big(\tfrac{m}{M}\Big)^{q-1} \in (0, 1),$$

and hence

$$\|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\| \leq 2 \prod_{\kappa=1}^{\lfloor (k+1)/q \rfloor} \Big(1 - \beta\Big) = 2\,(1-\beta)^{\lfloor (k+1)/q \rfloor}.$$

By definition of the floor, $\left\lfloor \frac{k+1}{q} \right\rfloor \geq \frac{k+1-q}{q}$, and since $1 - \beta \in (0, 1)$,

$$(1-\beta)^{\lfloor (k+1)/q \rfloor} \leq (1-\beta)^{(k+1-q)/q} = (1-\beta)^{-(q-1)/q} \Big((1-\beta)^{1/q}\Big)^k.$$

Setting $\rho = (1-\beta)^{1/q} \in (0, 1)$ and $C' = 2\,(1-\beta)^{-(q-1)/q}$ gives

$$\|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\| \leq C' \rho^k.$$

Application of Lemma S1 then yields

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x})\in\boldsymbol{\Theta}\times\mathcal{X}} \left|\ell_T(\boldsymbol{\theta},\boldsymbol{x}) - \tilde{\ell}_T^k(\boldsymbol{\theta},\boldsymbol{x})\right| \leq \sum_{t=1}^{T} \frac{M}{m} \sup_{(\boldsymbol{\theta},\boldsymbol{x})} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\| \leq TC\rho^k,$$

where $C = \frac{M}{m}C' = 2\frac{M}{m}(1-\beta)^{-(q-1)/q}$. □

Proof of Theorem 1: Case 2. In Case 2 the observations $y_1, \ldots, y_T$ are fixed, finite values, so by Assumption 2 the realized state-dependent densities satisfy $0 < m \leq f_j(y_t) \leq M < \infty$ for all t, uniformly over $(\boldsymbol{\theta}, \boldsymbol{x}) \in \boldsymbol{\Theta} \times \mathcal{X}$. Consequently $\delta(y_t) \leq M/m$ for every t, and Theorem 2.1 of Le Gland and Mevel (2000) applies to the fixed sequence $y_1, \ldots, y_T$ exactly as in Case 1, with the conditioning on $\boldsymbol{x}$ left intact throughout. The filter bound

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x})\in\boldsymbol{\Theta}\times\mathcal{X}} \|\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t^k\| \leq C'\rho^k$$

therefore holds with the same constants $C' = 2(1-\beta)^{-(q-1)/q}$ and $\rho = (1-\beta)^{1/q}$ as in Case 1, and application of Lemma S1 gives

$$\sup_{(\boldsymbol{\theta},\boldsymbol{x})\in\boldsymbol{\Theta}\times\mathcal{X}} \left|\ell_T(\boldsymbol{\theta},\boldsymbol{x}) - \tilde{\ell}_T^k(\boldsymbol{\theta},\boldsymbol{x})\right| \leq TC\rho^k, \qquad C = \frac{M}{m}C'.$$

□

Combining both cases, the elementwise approximation $\tilde{\ell}_T^k$ satisfies

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

with $C = 2\frac{M}{m}(1-\beta)^{-(q-1)/q}$ and $\rho = (1-\beta)^{1/q}$. By the reduction given in the "Notation and setup" Subsection, each contribution of the blockwise approximation $\tilde{\ell}_T^{\mathrm{b},k}$ conditions on at least k observations, so the same bound holds for $\tilde{\ell}_T^{\mathrm{b},k}$ with identical constants, which proves the theorem.

Remarks on the assumptions

Remark S1 (Compactness of $\mathcal{X}$). Within the Laplace approximation, the latent field enters through the GMRF log-density $-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}$ (up to a constant), where the precision matrix $\boldsymbol{Q}$ is positive definite. The mode in $\boldsymbol{x}$ maximizes

$$\ell_T(\boldsymbol{\theta},\boldsymbol{x}) - \tfrac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}.$$

The upper bound on the state-dependent densities, M, requires only non-degenerate variance parameters and holds for all $\boldsymbol{x}$; hence $\ell_T \leq T \log M$ uniformly in $\boldsymbol{x}$. Since $\boldsymbol{Q}$ is positive definite, the quadratic penalty therefore dominates as $\|\boldsymbol{x}\| \to \infty$ and the objective tends to $-\infty$. The maximizer lies in a bounded region, so restricting $\mathcal{X}$ to a compact ball around $\hat{\boldsymbol{x}}_\theta$ entails no loss. The positive lower bound on the state-dependent densities, m, then holds on this compact set, as used in the proof.

Remark S2 (Boundedness of state-dependent densities). The two bounds in Assumption 2 are secured differently. The upper bound $f_j(z_t) \leq M$ requires only non-degenerate emission parameters — variances bounded below and all shape parameters in the compact $\boldsymbol{\Theta}$ — and holds for any argument. The lower bound $f_j(z_t) \geq m > 0$ requires z_t to lie in a bounded region, since these densities have no positive lower bound over an unbounded domain. This holds in Case 1 because $z_t = x_t$ with $\boldsymbol{x} \in \mathcal{X}$ compact, and in Case 2 because the $z_t = y_t$ are fixed, finite observations, so that $m = \min_{t,j} f_j(y_t) > 0$ unless some state assigns vanishing density to an observed point — a sign of misspecification rather than a defect of the approximation. In Case 2 the densities are additionally conditioned on $\boldsymbol{x} \in \mathcal{X}$, so all bounds are uniform in $(\boldsymbol{\theta}, \boldsymbol{x})$.

Remark S3 (Uniform ergodicity of the Markov chain). The required uniform minorization (Assumption 3) follows from the compactness of $\boldsymbol{\Theta} \times \mathcal{X}$ together with a condition on the *user-specified* zero-pattern of the transition matrices. If all entries of $\boldsymbol{\Gamma}^{(t)}$ are structurally positive and depend continuously on covariates and parameters in $\boldsymbol{\theta}$, or on random effects $\boldsymbol{x}$, all lying in the compact set $\boldsymbol{\Theta} \times \mathcal{X}$, then each entry attains a positive minimum, so $\gamma_{ij}^{(t)} \geq \varepsilon > 0$ for all i, j, t and the index of primitivity is $q = 1$. When some transitions are forbidden, the corresponding entries are fixed to zero, while the remaining entries might vary with covariates. Crucially, the resulting zero-pattern is the *same* at every time t, since which transitions are forbidden is a property of the model specification, not of the covariates. Whether an entry of the q-step product $(\boldsymbol{\Gamma}^{(t)} \cdots \boldsymbol{\Gamma}^{(t+q-1)})_{ij}$ is positive therefore depends only on this fixed pattern, not on t: it is positive if and only if the support graph admits a path of length q from i to j. We require the user to specify a zero-pattern whose support graph is *primitive*, i.e. irreducible and aperiodic, so that there exists a smallest integer $q \geq 1$ for which every such path exists. Denoting by $\varepsilon > 0$ the uniform lower bound on the structurally positive entries, each surviving length-q path then contributes at least ε^q, giving

$$\left(\boldsymbol{\Gamma}^{(t)} \cdots \boldsymbol{\Gamma}^{(t+q-1)}\right)_{ij} \geq \varepsilon^q > 0 \quad \text{for all } i, j, t,$$

which is Assumption 3 with index of primitivity q. Because the pattern is time-invariant, a single q is valid for all t, and the worst-case single-step bound ε composes correctly across

the q steps. A self-loop $\gamma_{ii}^{(t)} > 0$ at any state guarantees aperiodicity, so for the flare model of Section 3.1 of the main paper (which retains the diagonal entries) primitivity holds with $q = 2$. Conversely, the assumption is violated if the chain is reducible (e.g. an absorbing state) or periodic (e.g. a deterministic cycle with no self-loops), in which case no finite q yields a uniformly positive product and the forward approximation is not applicable.

S2 Quality of the Laplace approximation

The Laplace approximation replaces the conditional distribution of the latent variables $\boldsymbol{x}$ given the data with a Gaussian distribution centered at its mode (Tierney and Kadane, 1986). If the function g in the exponent of the posterior is exactly quadratic, the Laplace approximation is exact. This occurs, for example, when both the observation and process distributions are Gaussian, and the observation model depends linearly on $\boldsymbol{x}$ through its mean. In our model, this condition only holds for one of the two densities, while the other one corresponds to a non-Gaussian, non-linear HMM likelihood. Even in this case, the Laplace approximation provides a local Gaussian approximation around the posterior mode. Its accuracy depends on the posterior being dominated by this single mode. In practice, it often performs well, particularly when the posterior is concentrated around a single mode and the sample size is sufficient to constrain the distribution in a relatively small region of the high-dimensional space, i.e., when the posterior is sharply peaked near the mode, corresponding to relatively large eigenvalues of the Hessian $\boldsymbol{H}_\theta$ (Tierney and Kadane, 1986; Ogden, 2021).

Whether the approximation is adequate can be assessed through a simple importance-sampling check. For fixed parameters, write the importance weight

$$\tilde{\ell}(\boldsymbol{x}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{x})\, f(\boldsymbol{x})}{\mathcal{N}(\boldsymbol{x} \mid \hat{\boldsymbol{x}}, \boldsymbol{H}_\theta^{-1})},$$

where $\mathcal{N}(\cdot \mid \hat{\boldsymbol{x}}, \boldsymbol{H}_\theta^{-1})$ is the Gaussian density used by the Laplace approximation. The approximation itself equals $\tilde{\ell}(\hat{\boldsymbol{x}})$, whereas the expectation of $\tilde{\ell}(\boldsymbol{x})$ under $\boldsymbol{x} \sim \mathcal{N}(\hat{\boldsymbol{x}}, \boldsymbol{H}_\theta^{-1})$ equals the exact marginal likelihood; the two coincide precisely when g is quadratic in $\boldsymbol{x}$. Drawing $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n$ from this Gaussian, the spread of the log-weights $\log \tilde{\ell}(\tilde{\boldsymbol{x}}_i)$ therefore quantifies the departure of the integrand from the fitted Gaussian — a spread near zero indicating a near-exact approximation — while their average $\frac{1}{n}\sum_i \tilde{\ell}(\tilde{\boldsymbol{x}}_i)$ provides a bias-corrected estimate of the marginal likelihood. A caveat is that, like any importance-sampling estimate, this diagnostic degrades as the latent dimension grows, since the weights become increasingly heavy-tailed; its reliability in a given fit can itself be monitored, for example

through the Pareto-$\hat{k}$ statistic of the weights (Vehtari et al., 2024). With the function `laplace_check()`, we provide a general implementation for `RTMB` models in the `RTMBdist` package (Fischer, 2026).

We applied the check to both case studies. For the lion model ($l = 2517$) the approximation is accurate: the log-weights had a standard deviation of 0.39, a relative effective sample size of 85%, and an importance-sampling estimate of the log marginal likelihood within 0.33 of the Laplace value (a relative error of 1.9×10^{-5}). For the flare model at full resolution ($l \approx 19{,}000$) the diagnostic saturates: the weights are heavy-tailed (Pareto-$\hat{k} = 0.88 > 0.7$, Vehtari et al., 2024) and the importance-sampling estimate is therefore unreliable, so the check is inconclusive. This reflects the degradation of importance sampling in high dimension rather than a diagnosed inaccuracy — re-fitting the same model to a reduced series of 5000 observations ($l \approx 5000$), where the diagnostic remains reliable (effective sample size 66%), confirms an accurate approximation (log-weight SD 0.39; relative bias 1.5×10^{-5}). Assessing the Laplace approximation directly in such high-dimensional settings is itself an open problem (Ogden, 2021).

S3 Simulation experiments

To assess the performance of the proposed inference approach, we conducted two simulation experiments whose setups mirror the two case studies. All code for full reproducibility of the simulation experiments is available at https://github.com/janolefi/HMMs_GFs.

S3.1 Temporal trend

In our first experiment, we generate data from a 2-state HMM with homogeneous transition probability matrix

$$\mathbf{\Gamma} = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix},$$

corresponding to a Markov chain with relatively high temporal persistence. Conditional on the state at time t, we assume normal state-dependent densities

$$X_t \mid \{S_t = j\} \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad j = 1, 2,$$

with mean values $\mu_1 = -1$, $\mu_2 = 3$ and standard deviations $\sigma_1 = 1$, $\sigma_2 = 2$. In each simulation run, we then draw a quasi-periodic trend $u_t = u(t)$ from a Gaussian process with the oscillating covariance function given in Section S4, with SPDE parameters $\kappa = 0.14$ and

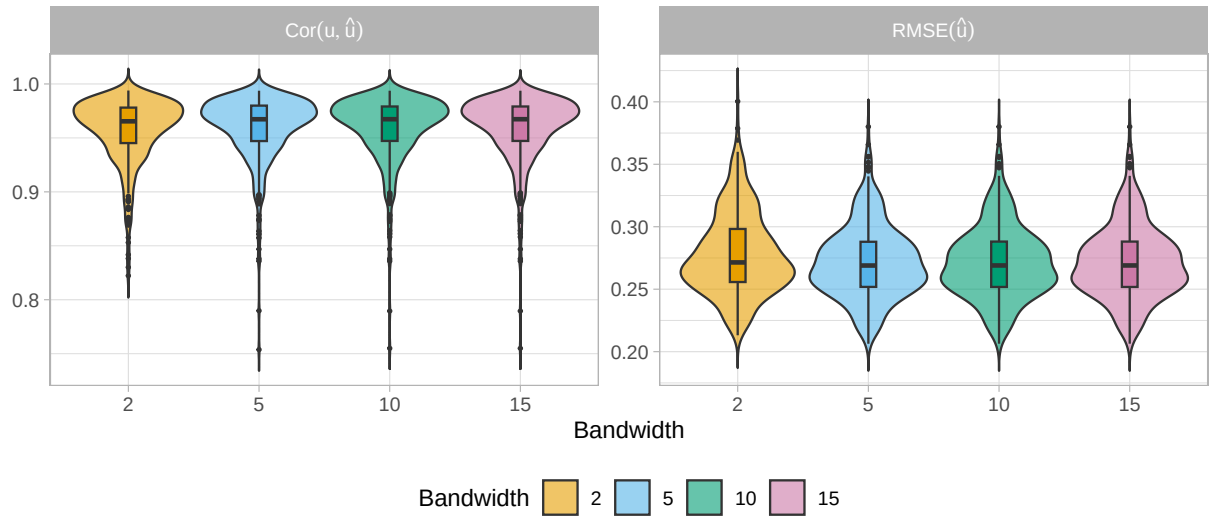


Figure S1: Violin plots and boxplots of the correlation between the estimated and true trend (left) and the root mean squared error between them (right), for bandwidth parameters $k = 2, 5, 10, 15$.

$\tau = 66.5$ and oscillation parameter $\omega = 0.99$, corresponding to a marginal standard deviation of about 1.1 and an oscillation period of roughly 44 time steps. The mean function is set to zero. We compute observations $Y_t = X_t + u_t$ and generate 500 data sets of size $T = 1000$. All model parameters are initialized at their true values and all field weights are initialized at zero, corresponding to $u_t = 0$ for all t. To each data set, we fit the corresponding HMM with bandwidth parameters $k \in \{2, 5, 10, 15\}$ to investigate potential biases due to a poor approximation for low bandwidth values.

Figure S8 shows the true and estimated trend for a single simulation run, indicating adequate recovery of the true field. As the predicted field $\hat{u}$ is pulled towards a zero mean by the GMRF density while the state-dependent means absorb the overall level, we summarize trend recovery through the correlation between $\hat{u}(t)$ and $u(t)$, which is invariant to this offset, and through the root mean squared error (RMSE). We find that correlation is high, with an average value around 0.96, and the RMSE averages around 0.27 for $k = 15$ (cf. Figure S1). Most influence of the bandwidth parameter k can be observed when comparing the distributions for $k = 2$ and $k = 5$. For $k \geq 5$, the distribution stabilizes and there are no visible differences anymore.

The estimates for the state-dependent means μ_1 and μ_2 and standard deviations σ_1 and σ_2 appear mostly unbiased (cf. Figure S9), with the same bandwidth dependence as the field recovery above: estimates for $k = 2$ are slightly biased while this bias vanishes for larger k. The field hyperparameters τ, κ and ω and the transition probabilities are likewise centered on

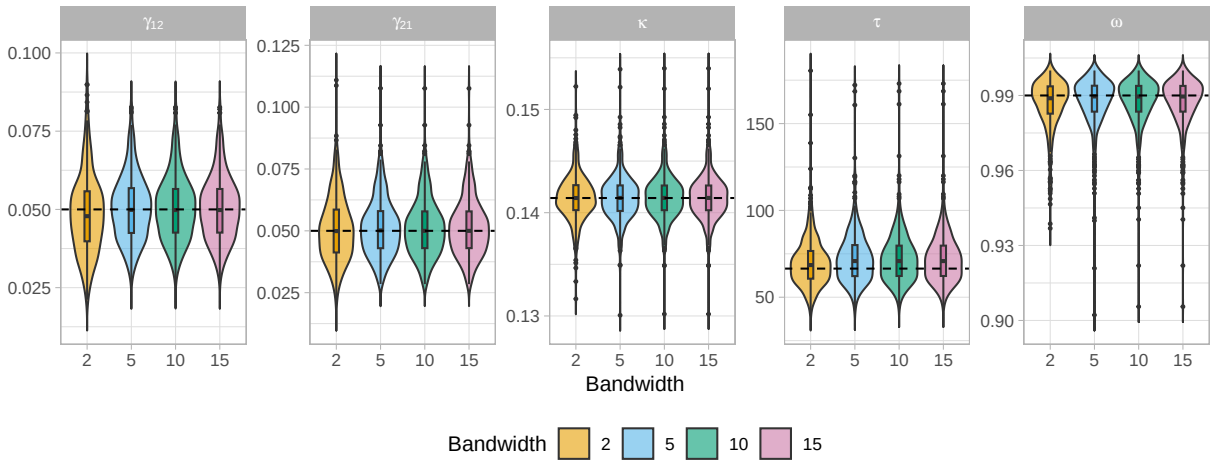


Figure S2: Violin plots and boxplots of the estimated transition probabilities γ_{12}, γ_{21} and field hyperparameters κ, τ, ω for bandwidth parameters $k = 2, 5, 10, 15$. Dashed lines mark the true values.

their true values (cf. Figure S2). The effect of the bandwidth parameter is barely noticeable, again only with visible differences between $k = 2$ and $k = 5$.

Finally, we assess the calibration of the confidence intervals obtained from the approximate joint posterior. For each run and each scalar parameter, we form a 95% interval from the Monte Carlo draws and record whether it contains the true value; for the trend, we record pointwise coverage averaged over t. Table S1 reports the resulting empirical coverage. For a bandwidth of $k = 15$, coverage is close to the nominal level for the state-dependent parameters, the marginal precision τ, and the transition probabilities, and marginally below nominal for the pointwise trend, as is typical for smoothers. The range parameter κ is slightly below nominal, covering at around 0.92. The oscillation parameter ω is likewise marginally below nominal, covering at around 0.93. At the smallest bandwidth $k = 2$, coverage of the state-dependent parameters is somewhat reduced but recovers for larger bandwidth values.

S3.2 Spatial field

In our second experiment, we generate data from a 2-state HMM where the transition probability matrix is driven by a spatial field. We simulate state-dependent gamma-distributed step lengths, reparameterized in terms of mean and standard deviation,

$$\text{step}_t \mid \{S_t = j\} \sim \text{Gamma}(\mu_j, \sigma_j), \quad j = 1, 2$$

with mean values $\mu_1 = 0.2$, $\mu_2 = 5$ and standard deviations $\sigma_1 = 0.5$, $\sigma_2 = 3$. At each time point, we also randomly draw a turning angle from a von Mises distribution, independently

Table S1: Empirical coverage of nominal 95% confidence intervals over 500 simulation runs, by parameter and bandwidth k. The last row reports pointwise coverage of the trend, averaged over time. The Monte Carlo standard error is approximately 0.01.

	$k = 2$	$k = 5$	$k = 10$	$k = 15$
μ_1	0.91	0.95	0.95	0.95
μ_2	0.93	0.93	0.94	0.94
σ_1	0.92	0.95	0.95	0.95
σ_2	0.94	0.94	0.95	0.95
γ_{12}	0.94	0.93	0.93	0.93
γ_{21}	0.93	0.96	0.96	0.96
τ	0.96	0.95	0.96	0.96
κ	0.90	0.92	0.92	0.92
ω	0.92	0.93	0.93	0.93
$u(t)$ (pointwise)	0.93	0.94	0.94	0.94

of the latent state. The mean is chosen as to imply a biased random walk towards the location $(0, 0)$ to prevent the simulated track from dispersing too much, with the concentration parameter being set to 0.3. We then compute Euclidean locations at each time point, initializing with the position $(0, 0)$ and calculating the subsequent position based on the sampled step length and direction. The transition probability γ_{12} is homogeneous while $\gamma_{21}^{(t)}$ depends on a spatial field. Specifically, we set

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

with $\beta_0^{(12)} = \beta_0^{(21)} = -2$. We fix the initial distribution of the chain at $\boldsymbol{\delta}^{(1)} = (0.5, 0.5)$ and do not estimate it as it is not the focus of this study and usually of minor practical relevance. The true spatial field u which we use for simulation is parametric with

$$u(\boldsymbol{r}) = u(x, y) = 2\sin(2\pi\tfrac{x}{40}) + 2\cos(2\pi\tfrac{y}{40}).$$

For time-series length $T \in \{5000, 10{,}000\}$, we simulate 200 data sets each. For each simulation run, we then fit the corresponding HMM with bandwidth parameters $k \in \{2, 5, 10, 15\}$ to investigate the practical effect of varying this parameter. For each run, we construct a GMRF approximation to the spatial field based on a triangulated mesh. However, we misspecify the model by additionally including a spatial field for the transition probability γ_{12} in order to inspect the model's capability of smoothing this field appropriately to zero. Figure S3 shows the true and estimated fields for a single simulation run, indicating adequate recovery of the true field.

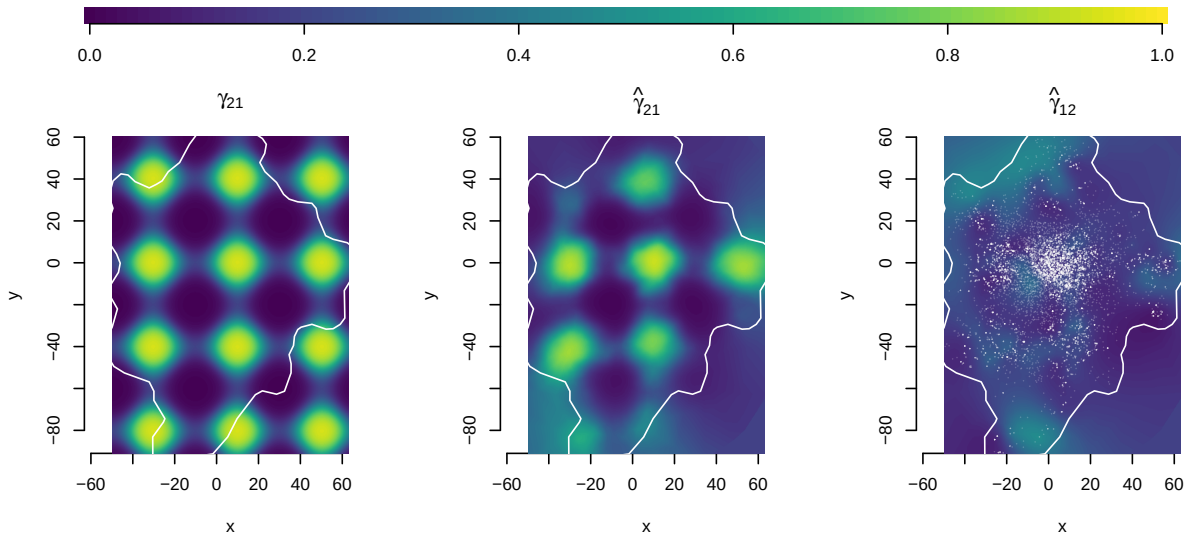


Figure S3: Estimated spatial field in one of the 200 simulation runs with 10,000 observations and bandwidth $k = 15$. True transition probability $\gamma_{21}(\boldsymbol{r})$ as a function of location (left), estimated transition probability $\hat{\gamma}_{21}(\boldsymbol{r})$ (middle) and estimated transition probability $\hat{\gamma}_{12}(\boldsymbol{r})$ (right). For the latter, the true function is completely flat, hence the ideal fit would be a flat surface. The white lines indicate the non-convex hull around the observed locations used for mesh construction in the GMRF approximation. The right panel additionally shows the visited locations as white dots.

As in the first experiment, the predicted field $\hat{u}$ is pulled towards a zero mean by the GMRF density while the true function u might have a non-zero mean inside the area that was visited, so we cannot directly compare the estimated intercept $\beta_0^{(21)}$ to its true value. Hence, we compute the mean of $\hat{\eta}_{21}^{(t)} - \eta_{21}^{(t)}$ over time, and similarly for $\eta_{12}^{(t)}$. We find a minor upward bias in the average difference for both linear predictors, i.e. the estimated models tend to be slightly less persistent than the true one on average for both sample sizes and all bandwidth parameters (cf. Figure S4). Overall, the bandwidth parameter seems to have even less influence on the estimates than in the first simulation experiment and for $k \geq 5$ all distributions are virtually identical.

To investigate interpolation performance, we compute the true and estimated field on a 512×512 grid in each run. As we do not expect to approximate the true field outside the area of observations, we restrict ourselves to those grid cells that lie inside the non-convex hull spanned by observed locations, which is computed anyway in each run for mesh construction (cf. Figure S3). Inside this region, the correlation between the estimate $\hat{\gamma}_{21}(\boldsymbol{r})$ and its true value $\gamma_{21}(\boldsymbol{r})$ is computed on the grid. Computing the corresponding quantity for $\hat{\gamma}_{12}(\boldsymbol{r})$ is not possible as the true field is constant and hence has zero variance. We find that correlation is

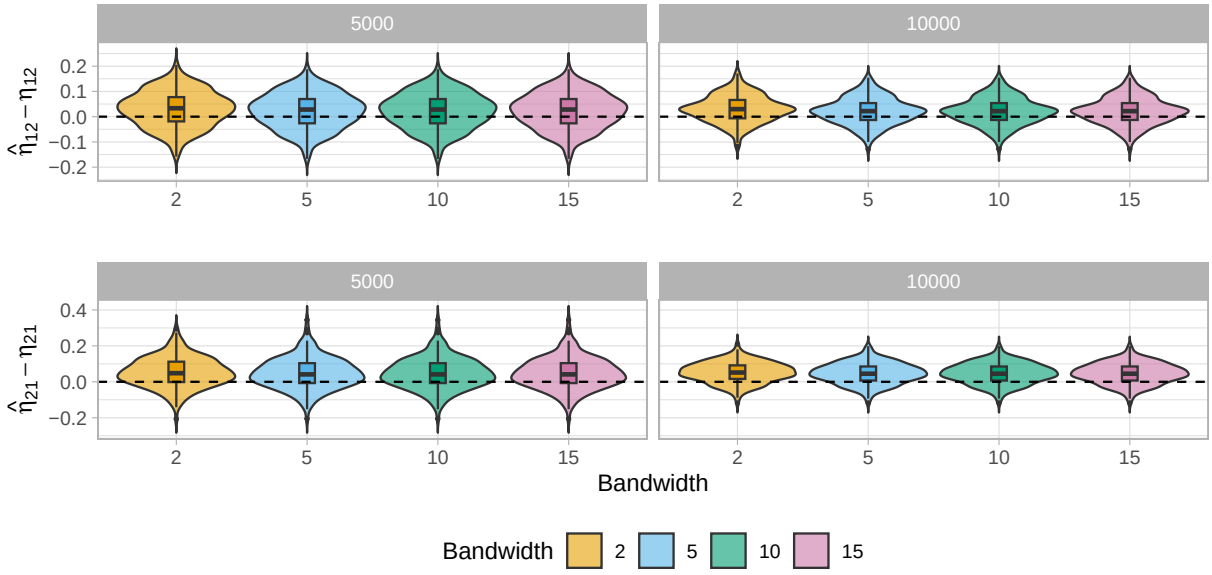


Figure S4: Violin plots and boxplots of the average difference $\frac{1}{T}\sum_{t=1}^{T}(\hat{\eta}_{ij}^{(t)} - \eta_{ij}^{(t)})$ between estimated and true linear predictors in sample, i.e. evaluated at the observed locations for sample sizes $T = 5000, 10{,}000$ and bandwidth parameters $k = 2, 5, 10, 15$.

high, with an average value slightly below 0.9 (cf. Figure S5). As expected, the distribution becomes more concentrated for the larger sample size of $T = 10{,}000$. Again, virtually no effect of the bandwidth parameter is visible.

Additionally, we compute the root mean squared errors (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i}\Big(\hat{\gamma}_{12}(\boldsymbol{r}_i) - \gamma_{12}\Big)^2}, \quad \text{and} \quad \sqrt{\frac{1}{n}\sum_{i}\Big(\hat{\gamma}_{21}(\boldsymbol{r}_i) - \gamma_{21}(\boldsymbol{r}_i)\Big)^2},$$

where n is the number of grid cells inside the area of interest. The RMSE between $\hat{\gamma}_{12}(\boldsymbol{r})$ and the constant field has an average value around 0.06, with the distribution stabilizing for $k \geq 5$ (cf. Figure S5). Performance for the true field is slightly worse, with an average RMSE around 0.12. Again deviation in the distribution of values is only visible between $k = 2$ and $k = 5$.

As in the first experiment, the estimates for the state-dependent means μ_1 and μ_2 and standard deviations σ_1 and σ_2 appear mostly unbiased (cf. Figure S10). As is to be expected, variance decreases with increasing sample size. The effect of the bandwidth parameter is barely noticeable, again only with visible differences between $k = 2$ and $k = 5$.

Finally, we assess the calibration of the confidence intervals obtained from the approximate joint posterior — in the same fashion as in the first experiment. For the two linear

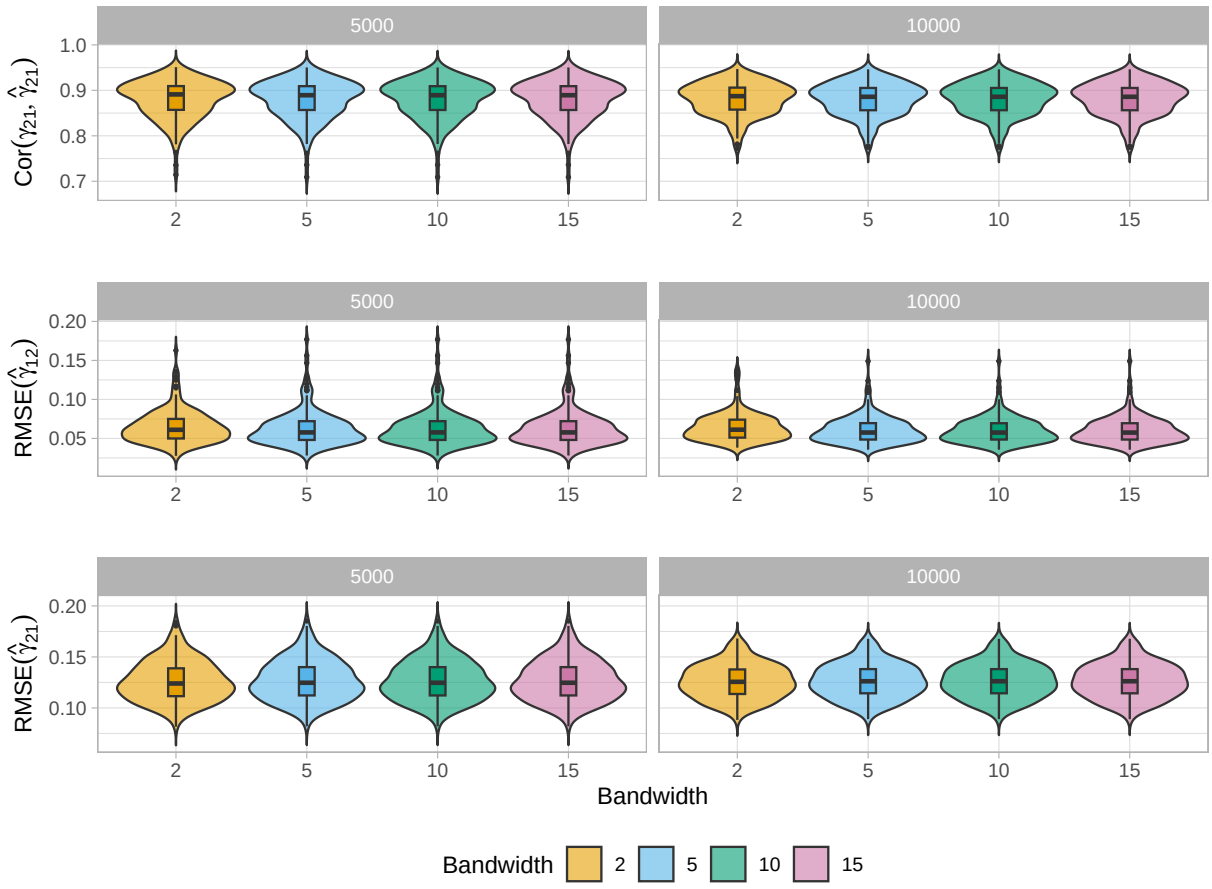


Figure S5: Violin plots and boxplots of correlation between true and estimated transition probability $\gamma_{21}(\boldsymbol{r})$ on a grid (inside the non-convex hull spanned by the observed locations) (top panel), and root mean squared error (RMSE) between estimated and true transition probabilities (bottom two panels). The correlation between $\hat{\gamma}_{12}(\boldsymbol{r})$ and the constant true field is undefined as the latter has zero variance.

Table S2: Empirical coverage of nominal 95% confidence intervals over 200 simulation runs, by parameter, sample size T, and bandwidth k. The rows η_{21} and η_{12} report pointwise coverage of the linear predictors at the observed locations; the remaining rows report coverage of the state-dependent parameters. The Monte Carlo standard error is approximately 0.015.

	$T = 5{,}000$				$T = 10{,}000$			
	$k = 2$	$k = 5$	$k = 10$	$k = 15$	$k = 2$	$k = 5$	$k = 10$	$k = 15$
μ_1	0.93	0.94	0.93	0.93	0.92	0.92	0.92	0.92
μ_2	0.97	0.97	0.96	0.96	0.96	0.95	0.95	0.95
σ_1	0.94	0.94	0.93	0.94	0.92	0.91	0.91	0.91
σ_2	0.93	0.94	0.94	0.93	0.95	0.95	0.95	0.95
η_{21} (pointwise)	0.95	0.95	0.95	0.94	0.96	0.96	0.96	0.96
η_{12} (pointwise)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

predictors, we record pointwise coverage of the true value, averaged over the *observed* locations. Table S2 reports the resulting empirical coverage. Coverage is close to the nominal 95% level for the field predictor η_{21} and for the state-dependent parameters, across both sample sizes and all bandwidths. For the misspecified predictor η_{12}, whose true value is constant, the intervals are conservative (around 0.99), reflecting that the corresponding field is smoothed towards its constant truth. In contrast to the first experiment, the bandwidth has a very small effect on coverage.

S4 GP with oscillating covariance function

For the stellar-flare case study and the temporal-trend simulation of Section S3.1, we use the Gaussian process detailed in Section 3.3 of Lindgren et al. (2011). This is based on a complex version of the SPDE given in Equation (2) of the main paper,

$$\Big(\kappa^2 \exp(i\pi\omega) - \Delta\Big)\Big(u_1(\boldsymbol{r}) + iu_2(\boldsymbol{r})\Big) = \mathcal{W}_1(\boldsymbol{r}) + i\mathcal{W}_2(\boldsymbol{r})$$

where $\omega \in (0, 1)$ is the oscillation parameter, and $\mathcal{W}_1(\boldsymbol{r})$ and $\mathcal{W}_2(\boldsymbol{r})$ are independent Gaussian white noise processes. Oscillation increases with the parameter ω. The real and imaginary parts, $u_1(\boldsymbol{r})$ and $u_2(\boldsymbol{r})$, are independent with identical spectra. Hence, the same oscillating covariance function

$$\mathcal{K}(v, w) = \frac{1}{2\sin(\pi\omega)\kappa^3} \exp\Big(-\kappa\cos\Big(\frac{\pi\omega}{2}\Big)|v - w|\Big) \sin\Big(\frac{\pi\omega}{2} + \kappa\sin\Big(\frac{\pi\omega}{2}\Big)|v - w|\Big)$$

applies to either component. Its simple form combines exponential decay with periodic oscillation. Figure S6 gives some intuition for the behavior of $\mathcal{K}$. The associated precision matrix has closed form and is given in the main paper (Section 3.1).

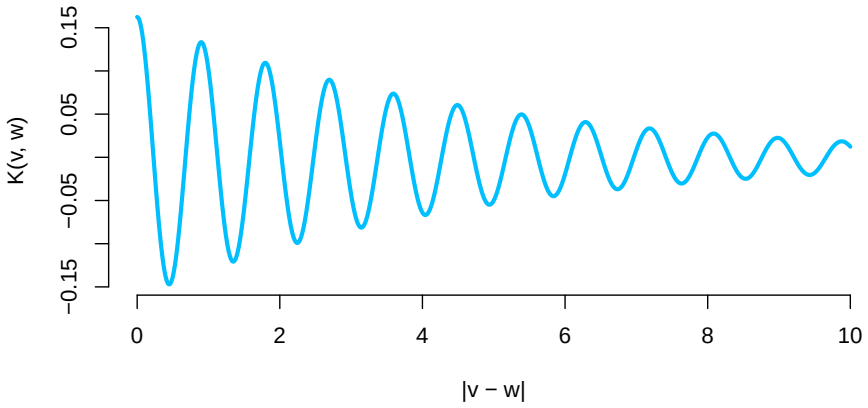


Figure S6: Oscillating covariance function $\mathcal{K}(v, w)$ for parameters $\omega = 0.98$ and $\kappa = 7$.

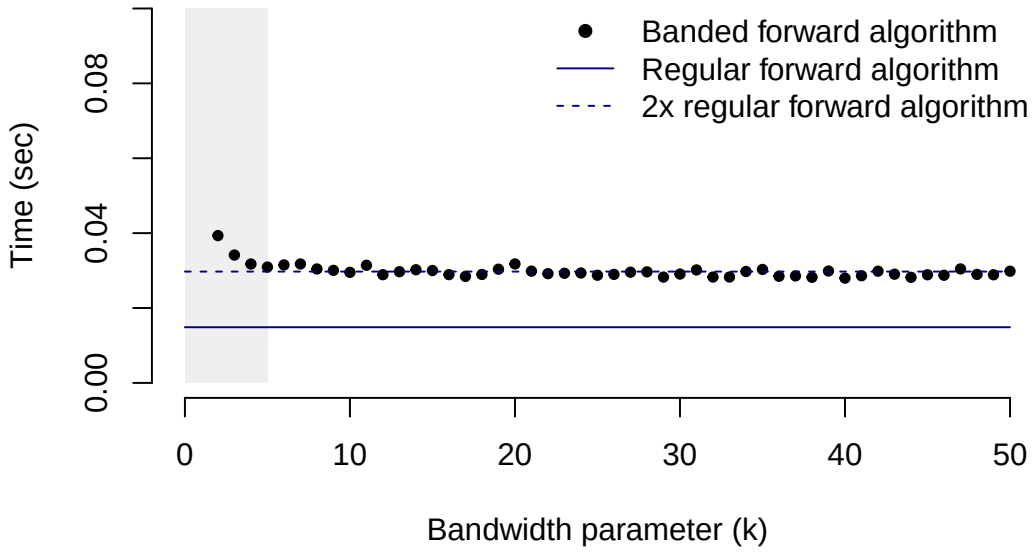


Figure S7: Average computation time of the banded forward algorithm as a function of the bandwidth k, using the data from the stellar-flare case study and averaged over 100 runs (black points). The horizontal blue line marks the average computation time of the standard forward algorithm. Beyond the low-bandwidth regime ($k \leq 5$), the banded algorithm takes almost exactly twice as long.

S5 Additional figures

References

Fischer, Jan-Ole (2026). *RTMBdist: Distributions Compatible with Automatic Differentiation by 'RTMB'*. R package version 1.0.5. URL: https://CRAN.R-project.org/package=RTMBdist.

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, 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.

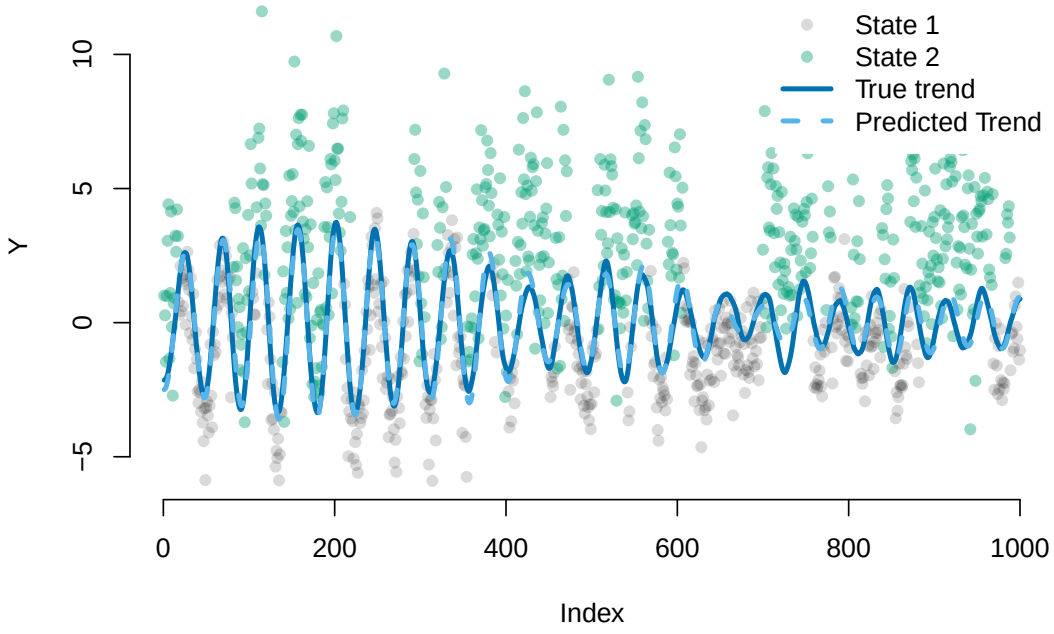


Figure S8: Estimated temporal trend in one of the 500 simulation runs of the first simulation experiment, with bandwidth $k = 15$. Points show the observations Y_t, distinguished by the decoded state; the solid line is the true trend $u(t)$ and the dashed line the point estimate $\hat{u}(t)$.

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

Tierney, Luke and Joseph B Kadane (1986). "Accurate approximations for posterior moments and marginal densities". In: *Journal of the American Statistical Association* 81.393, pp. 82–86.

Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry (2024). "Pareto smoothed importance sampling". In: *Journal of Machine Learning Research* 25.72, pp. 1–58.

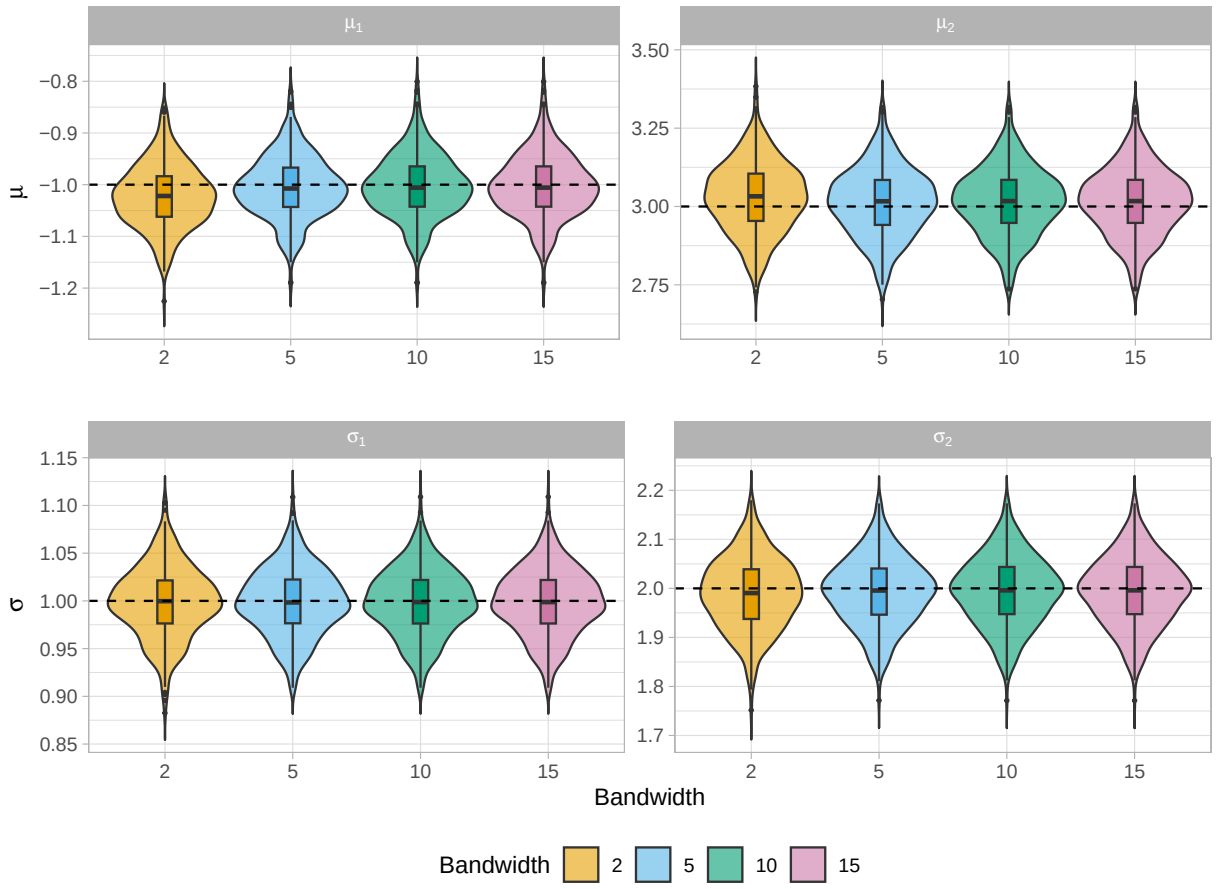


Figure S9: Violin plots and boxplots of the estimated state-dependent means μ_1, μ_2 and standard deviations σ_1, σ_2 in the first simulation experiment, for $T = 1000$ observations and bandwidth parameters $k = 2, 5, 10, 15$. Dashed lines mark the true values.

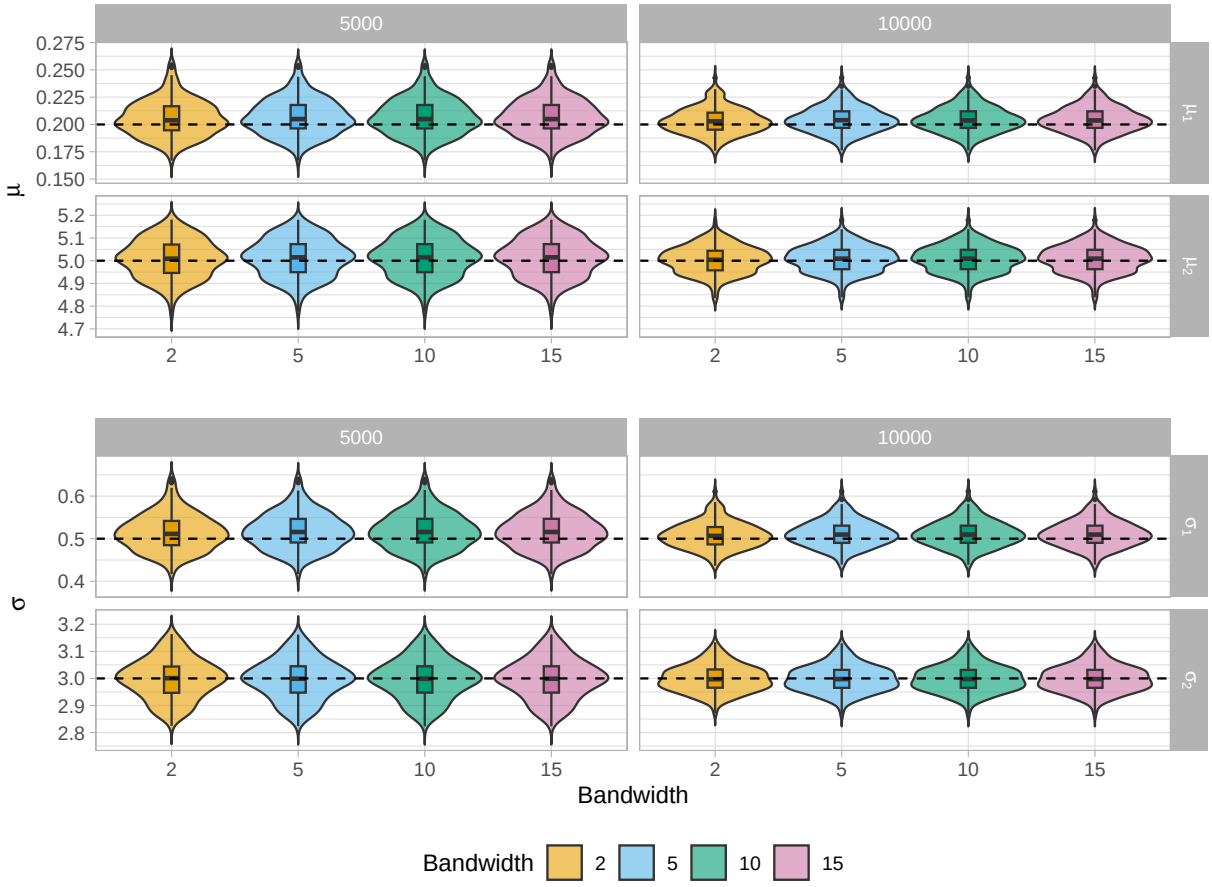


Figure S10: Violin plots and boxplots of the estimated state-dependent means μ_1, μ_2 and standard deviations σ_1, σ_2 in the second simulation experiment for $T = 5000, 10{,}000$ observations and bandwidth parameter $k = 2, 5, 10, 15$.

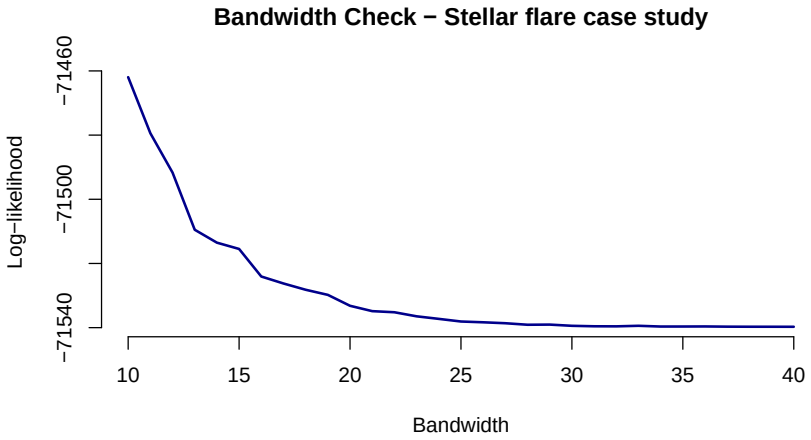


Figure S11: Approximate HMM log-likelihood, evaluated at the initial parameter values, as a function of the bandwidth k for the stellar-flare case study. The curve stabilizes beyond $k \approx 30$, indicating that the banded approximation has converged to the exact log-likelihood; we therefore set $k = 30$.

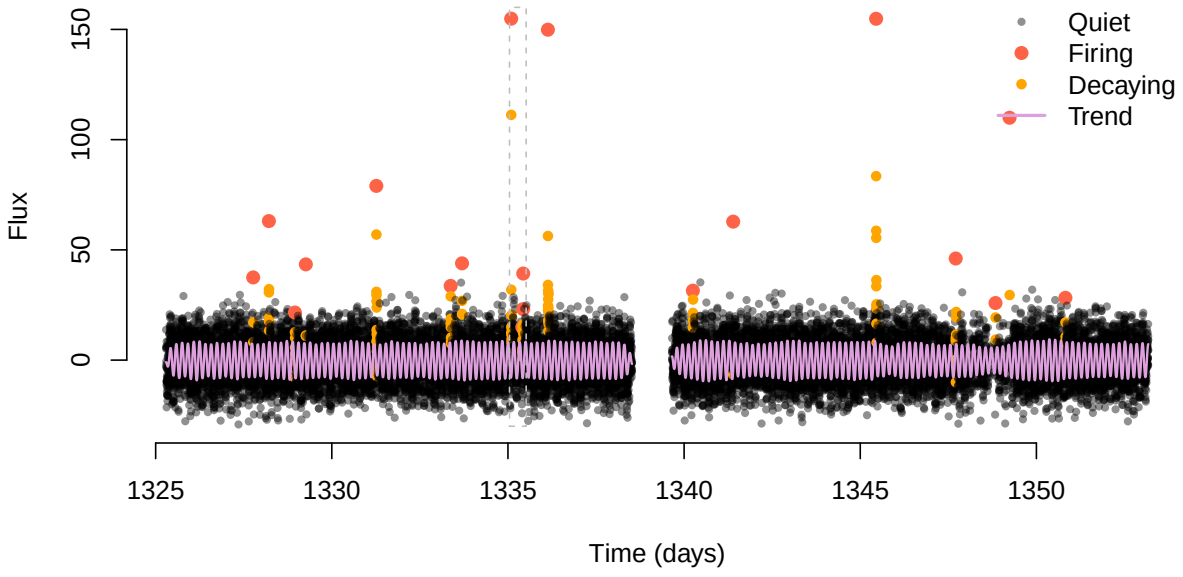


Figure S12: Complete time series of M dwarf TIC 031381302, colored according to the most likely decoded state. The dashed gray box shows the section of the time series depicted in the main paper.

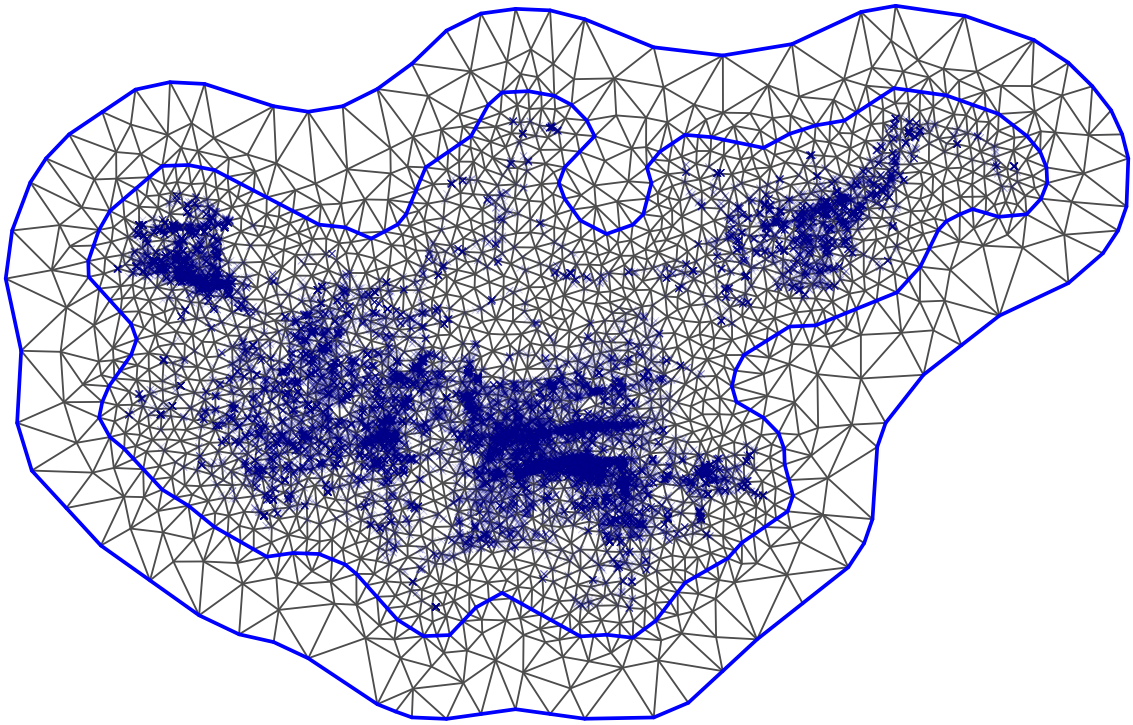


Figure S13: Triangulated mesh used to approximate the Gaussian field in the lion HMM by a GMRF. Observed locations are shown as dark blue crosses.

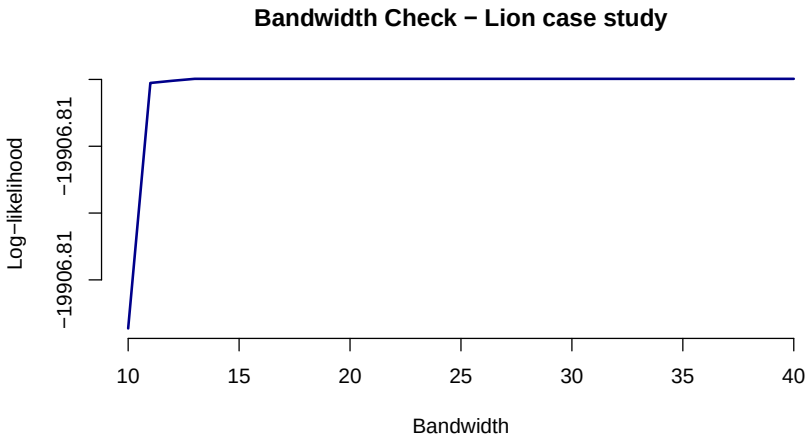


Figure S14: Approximate HMM log-likelihood, evaluated at the initial parameter values, as a function of the bandwidth k for the lion case study. The curve stabilizes beyond $k \approx 15$, indicating that the banded approximation has converged to the exact log-likelihood; we therefore set $k = 15$.

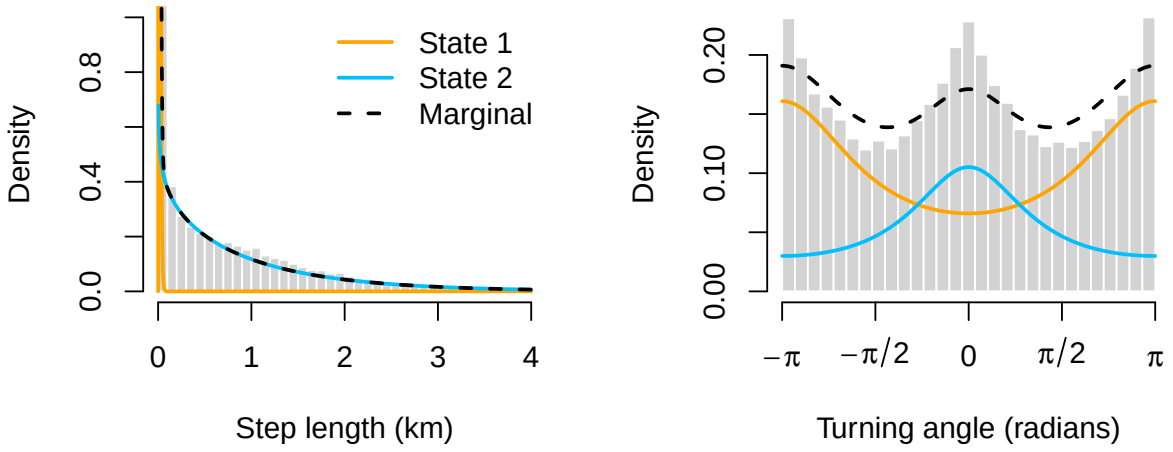


Figure S15: State-dependent distributions for the 2-state lion HMM, weighted by the approximate frequency of state occupancy (based on the decoded states).