Supporting Material

ICON: an adaptation of infinite HMMs for time traces with drift

I. Sgouralis and S. Pressé

In this Supplement we provide: (i) a description of the source code and GUI implementing ICON as describe in the main text, Sect. S1; (ii) additional time series analyses demonstrating the effects of the drift function $f(t)$, Sect. S2; and (iii) a complete summary of the involved models throughout the main text, Sect. S3.

S1 ICON implementation

The implementation, graphical user interface (GUI), and source code of ICON is similar to the implementation of the plain iHMM described in the companion perspectives article. Here we highlight only the main differences. For further details and an in depth presentation of the offered capabilities we refer to the documentation found in the perspectives.

The present implementation comes in two versions: *single trace ICON*, and *double trace ICON*. These versions are described in the Methods section of the main text and can be used to analyze single molecule data from one or two traces, respectively.

Additionally to the hyperparameters associated with the emission model described in the perspectives article, both ICON versions require M, the total number of interpolation nodes, to be specified by the user.

In the *single trace ICON* the time series to be analyzed must be provided as an 1D array with one value per time level, while in the *double trace ICON* the time series to be analyzed must be provided as a 2D array with two values per time level.

Single trace ICON exports (i) one drift time series, (ii) one emission trajectory, (iii) one state transition matrix, and (iv) the corresponding individual samples. *Double trace ICON* exports (i) two drift time series, (ii) two emission trajectories, (iii) one state transition matrix, and (iv) the corresponding individual samples.

S2 Additional analyses

Here we demonstrate the effects of the various choices of the interpolation $f(t)$, see Methods section in the main text. Specifically, we consider the following choices: basis functions and number of nodes for the interpolation.

Figure S1 shows the resulting state and drift trajectories for two different choices of bases functions: (i) cubic splines (as in the main text) and (ii) linear polynomials. As can be seen, the resulting estimates are minimally affected by this choices.

Figure **??** shows the resulting state and drift trajectories for three different numbers of nodes: (i) $M = 5$, (ii) $M = 10$ (as in the main text), and (iii) $M = 20$. As can be seen, the resulting estimates are also minimally affected by this choices.

S3 Summary of the models used

Here we present a detailed list of the statistical models presented in this study. For definitions and notation we refer to Section 2 of the main text.

S3.1 Hidden Markov Model (HMM)

$$s_0 = \sigma_0 \tag{1}$$

$$s_n \big| s_{n-1} \sim Cat(\tilde{\pi}_{s_{n-1}}), \qquad n = 1, 2, \dots, N \tag{2}$$

$$x_n \big| s_n \sim F_{s_n}, \qquad n = 1, 2, \dots, N \tag{3}$$

where $Cat(\tilde{\pi}_{s_{n-1}})$ is the categorical probability distribution with parameters gathered in the probability vector $\tilde{\pi}_{s_{n-1}} = (\pi_{s_{n-1}\to\sigma_1}, \pi_{s_{n-1}\to\sigma_2}, \dots)$.

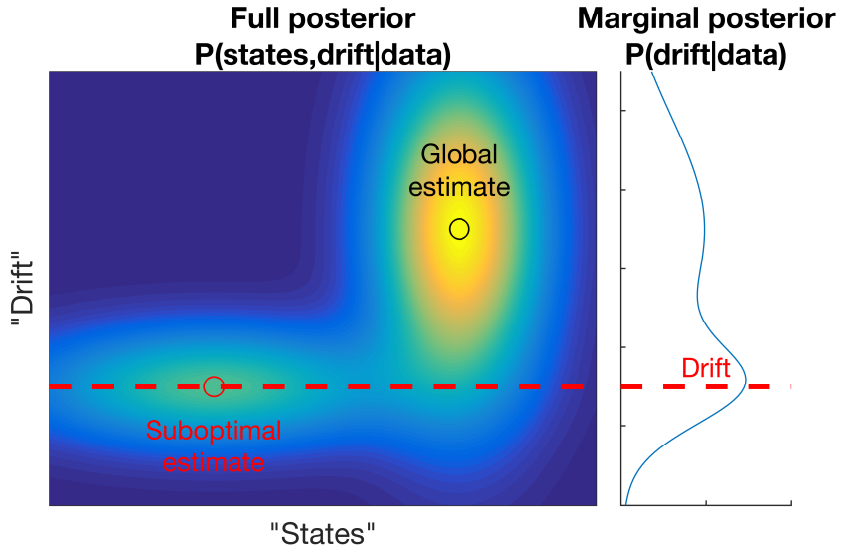

Figure S1: State (left) and drift (right) trajectories estimated using cubic splines and linear polynomials for the interpolation function $f(t)$.

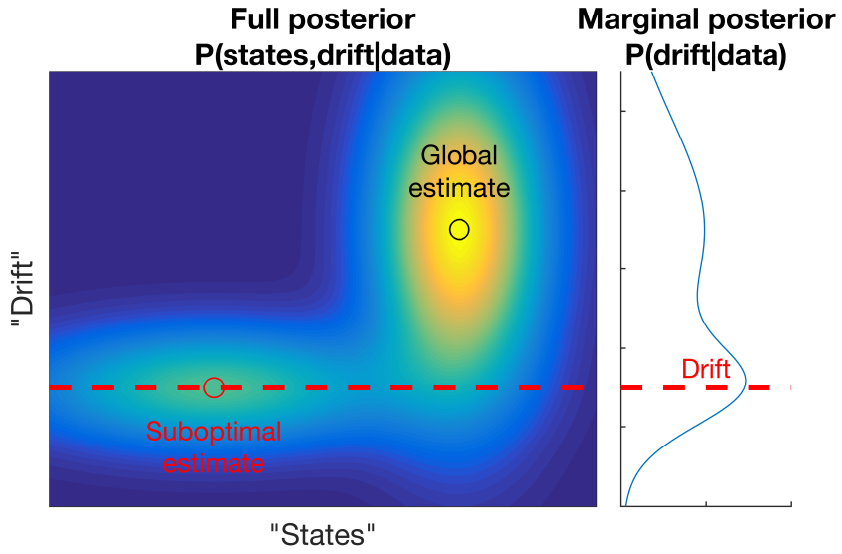

Figure S2: State (left) and drift (right) trajectories estimated using $M = 5$, 10 and 20 in the interpolation function $f(t)$.

S3.2 infinite Hidden Markov Model (iHMM)

$$\tilde{\beta} \sim GEM(\gamma) \tag{4}$$
$$\tilde{\pi}_{\sigma_k} \big| \tilde{\beta} \sim DP(\alpha, \tilde{\beta}), \qquad k = 0, 1, 2, \dots \tag{5}$$
$$\phi_{\sigma_k} \sim H, \qquad k = 1, 2, \dots \tag{6}$$

$$s_0 = \sigma_0 \tag{7}$$
$$s_n \big| s_{n-1}, \tilde{\pi}_{s_{n-1}} \sim Cat(\tilde{\pi}_{s_{n-1}}), \qquad n = 1, 2, \dots, N \tag{8}$$
$$x_n \big| s_n, \phi_{s_n} \sim F_{s_n}, \qquad n = 1, 2, \dots, N \tag{9}$$

where $GEM(\gamma)$ is the Griffiths-Engen-McCloskey "stick-breaking" process, $DP(\alpha, \tilde{\beta})$ is the Dirichlet distribution and $\tilde{\beta} = (\beta_{\sigma_1}, \beta_{\sigma_2}, \dots)$ is the base measure.

S3.3 Single trace ICON

$$\tilde{\beta} \sim GEM(\gamma) \tag{10}$$
$$\tilde{\pi}_{\sigma_k} \big| \tilde{\beta} \sim DP(\alpha, \tilde{\beta}), \qquad k = 0, 1, 2, \dots \tag{11}$$
$$\phi_{\sigma_k} \sim H, \qquad k = 1, 2, \dots \tag{12}$$

$$s_0 = \sigma_0 \tag{13}$$
$$s_n \big| s_{n-1}, \tilde{\pi}_{s_{n-1}} \sim Cat(\tilde{\pi}_{s_{n-1}}), \qquad n = 1, 2, \dots, N \tag{14}$$
$$x_n \big| s_n, \phi_{s_n} \sim F_{s_n}, \qquad n = 1, 2, \dots, N \tag{15}$$

$$h^* \sim \mathcal{N}(\mu^*, \tau^*) \tag{16}$$
$$w^* \sim \mathcal{G}(a^*, b^*) \tag{17}$$
$$y^*_m \big| h^*, w^* \sim \mathcal{N}(h^*, w^*), \qquad m = 1, 2, \dots, M \tag{18}$$
$$v^* \big| y^*_1, y^*_2, \dots, y^*_M \sim \delta \tag{19}$$

$$y_n = f(y^*_1, y^*_2, \dots, y^*_M; t_n), \qquad n = 1, 2, \dots, N \tag{20}$$
$$z_n = x_n + y_n, \qquad n = 1, 2, \dots, N \tag{21}$$

where δ is the Dirac delta, $\mathcal{N}(\mu^*, \tau^*)$ is the normal probability distribution with mean value μ^* and precision τ^*, $\mathcal{G}(a^*, b^*)$ is the gamma probability distribution with shape a^* and scale b^*, and $f(y^*_1, y^*_2, \dots, y^*_M; t)$ is the interpolant of the nodes (t^*_m, y^*_m) evaluated at t.

A graphical model summarizing this model is shown in Fig. S3.

S3.4 Double trace ICON

$$\tilde{\beta} \sim GEM(\gamma) \tag{22}$$
$$\tilde{\pi}_{\sigma_k}\big|\tilde{\beta} \sim DP(\alpha, \tilde{\beta}), \qquad k = 0, 1, 2, \ldots \tag{23}$$
$$\phi^1_{\sigma_k} \sim H^1, \qquad k = 1, 2, \ldots \tag{24}$$
$$\phi^2_{\sigma_k} \sim H^2, \qquad k = 1, 2, \ldots \tag{25}$$

$$s_0 = \sigma_0 \tag{26}$$
$$s_n\big|s_{n-1}, \tilde{\pi}_{s_{n-1}} \sim Cat(\tilde{\pi}_{s_{n-1}}), \qquad n = 1, 2, \ldots, N \tag{27}$$
$$x^1_n\big|s_n, \phi^1_{s_n} \sim F^1_{s_n}, \qquad n = 1, 2, \ldots, N \tag{28}$$
$$x^2_n\big|s_n, \phi^2_{s_n} \sim F^2_{s_n}, \qquad n = 1, 2, \ldots, N \tag{29}$$

$$h^{*1} \sim \mathcal{N}(\mu^{*1}, \tau^{*1}) \tag{30}$$
$$h^{*2} \sim \mathcal{N}(\mu^{*2}, \tau^{*2}) \tag{31}$$
$$w^{*1} \sim \mathcal{G}(a^{*1}, b^{*1}) \tag{32}$$
$$w^{*2} \sim \mathcal{G}(a^{*2}, b^{*2}) \tag{33}$$
$$y^{*1}_m\big|h^{*1}, w^{*1} \sim \mathcal{N}(h^{*1}, w^{*1}), \qquad m = 1, 2, \ldots, M^1 \tag{34}$$
$$y^{*2}_m\big|h^{*2}, w^{*2} \sim \mathcal{N}(h^{*2}, w^{*2}), \qquad m = 1, 2, \ldots, M^2 \tag{35}$$
$$v^*\big|y^{*1}_1, y^{*1}_2, \ldots, y^{*2}_{M^1} \sim \delta \tag{36}$$
$$v^{*2}\big|y^{*2}_1, y^{*2}_2, \ldots, y^{*2}_{M^2} \sim \delta \tag{37}$$

$$y^1_n = f(y^{*1}_1, y^{*1}_2, \ldots, y^{*1}_{M^1}; t_n), \qquad n = 1, 2, \ldots, N \tag{38}$$
$$y^2_n = f(y^{*2}_1, y^{*2}_2, \ldots, y^{*2}_{M^2}; t_n), \qquad n = 1, 2, \ldots, N \tag{39}$$
$$z^1_n = x^1_n + y^1_n, \qquad n = 1, 2, \ldots, N \tag{40}$$
$$z^2_n = x^2_n + y^2_n, \qquad n = 1, 2, \ldots, N \tag{41}$$

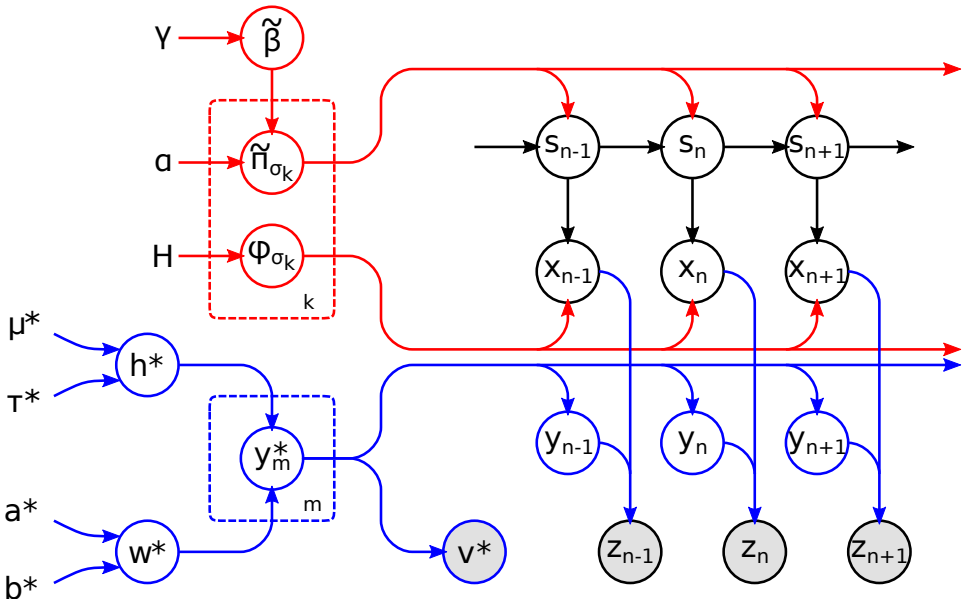

Figure S3: **Graphical representation of single trace ICON.** The hidden Markov model is highlighted with **black,** the prior formulating the infinite hidden Markov model is highlighted with **red,** and the drift representation is highlighted with **blue.**