Adaptive Kalman Filtering for Evolving Syntax Rules in Canary Song

Nadav Elami

2023

Abstract

Canary songs exhibit rich, history-dependent phrase sequencing whose syntax rules drift on timescales from a few songs to several weeks. We present an *adaptive Kalman–EM algorithm* that **learns** a block-diagonal state-transition matrix $\mathbf{F}$ and control vector $\mathbf{u}$ while tracking time-varying *logit* vectors that govern soft-max transition probabilities between song phrases. By unifying Bayes' rule, the Chapman–Kolmogorov equation, and a soft-max observation model, the filter–smoother achieves sub-song latency and detects syntax changes across multiple timescales—an essential step toward real-time behavioural monitoring in songbirds.

1 Introduction

Behavioural *syntax rules* govern how elementary actions are strung together in time. In canaries, individual songs are built from phrases that obey long-range dependencies—some phrase transitions depend on context up to seven seconds in the past. Traditional regression analyses reveal that these rules drift from day to day but struggle to resolve *simultaneously* short- and long-term changes.

Inspired by Bayesian filtering work on time-varying neural spike statistics, we treat each song as an observation emitted by a latent *generative syntax model* whose parameters evolve from song to song. A filtering framework that leverages *all* songs, rather than sliding windows, is expected to detect syntax changes more sensitively across multiple timescales than rolling regression. Our overarching research question is therefore:

> **Can a state-space filtering approach accurately characterise canary-syntax dynamics across timescales from minutes to weeks?**

Answering this question requires a state model rich enough to capture slow drifts and rapid perturbations. The present paper derives such a model by augmenting the classical random-walk assumption with a linear transition $\mathbf{F}$ and control $\mathbf{u}$, and by developing an EM algorithm that estimates both latent trajectories and noise covariances.

2 Notation and Problem Statement

Let the alphabet be $A = \{a_1, \ldots, a_R\}$ with size $R = |A|$. At discrete *sequence* index $k = 1, \ldots, K$ we observe a string $\mathbf{y}^{(k)} = (y_1^{(k)}, \ldots, y_{n(k)}^{(k)})$ of length $n(k)$.

The latent *logit* vector at index k is $\mathbf{x}_k \in \mathbb{R}^{R^2}$. For compactness we map an ordered pair (a_i, a_j) to a single index $p = i + (j-1)R$.

2.1 Observation Model

For each transition in the sequence we posit a categorical distribution

$$P\bigl(y_m^{(k)} = a_i \mid y_{m-1}^{(k)} = a_j;\, \mathbf{x}_k\bigr) = \frac{\exp\bigl(x_{k,p}\bigr)}{\displaystyle\sum_{i'=1}^{R} \exp\bigl(x_{k,\,i'+(j-1)R}\bigr)} \;\equiv\; f\bigl(x_{k,p}\bigr), \tag{1}$$

which is the usual *row-wise softmax* mapping $f : \mathbb{R} \to (0,1)$.

2.2 State Dynamics

We posit a linear-Gaussian evolution

$$\mathbf{x}_{k+1} = \mathbf{F}\,\mathbf{x}_k + \mathbf{u} + \boldsymbol{\varepsilon}_k, \quad \boldsymbol{\varepsilon}_k \sim \mathcal{N}\bigl(\mathbf{0},\,\boldsymbol{\Sigma}\bigr), \tag{2}$$

where $\mathbf{F}\in\mathbb{R}^{R^2\times R^2}$ is the (possibly sparse) state-transition matrix and $\mathbf{u}\in\mathbb{R}^{R^2}$ is a constant control vector. The diagonal process-noise covariance is $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{R^2}^2)$.

3 Complete-Data Log-Likelihood

Define the initial latent state by $\mathbf{x}_0$. Under the state dynamics and observation model described in Section 2, the joint density of the latent sequence $\{\mathbf{x}_k\}_{k=1}^{K}$ and observed strings $\{\mathbf{y}^{(k)}\}_{k=1}^{K}$ factorises as

$$P\bigl(\mathbf{x}_{1:K}, \mathbf{y}_{1:K} \mid \mathbf{x}_0\bigr) \;=\; \prod_{k=1}^{K} \Bigl[P\bigl(\mathbf{y}^{(k)} \mid \mathbf{x}_k\bigr)\; P\bigl(\mathbf{x}_k \mid \mathbf{x}_{k-1}\bigr)\Bigr]. \tag{3}$$

Taking logarithms yields the complete-data log-likelihood

$$\begin{aligned}\mathcal{L} = \sum_{k=1}^{K} \Biggl\{ \sum_{m=1}^{n(k)} \log f\bigl(x_{k,\,y_m^{(k)},\,y_{m-1}^{(k)}}\bigr) - \frac{1}{2}\bigl(\mathbf{x}_k - \mathbf{F}\mathbf{x}_{k-1} - \mathbf{u}\bigr)^{\top} \boldsymbol{\Sigma}^{-1} \bigl(\mathbf{x}_k - \mathbf{F}\mathbf{x}_{k-1} - \mathbf{u}\bigr) \Biggr\} \\ - \frac{K}{2}\,\log\bigl|2\pi\boldsymbol{\Sigma}\bigr|. \end{aligned} \tag{4}$$

4 Expectation–Maximisation Outline

At iteration i we hold current estimates $\Theta^{(i)} = \{\mathbf{x}_0^{(i)}, \boldsymbol{\Sigma}^{(i)}, \mathbf{F}^{(i)}, \mathbf{u}^{(i)}\}$.

The EM algorithm maximises the expected complete-data log-likelihood

$$Q\bigl(\Theta, \Theta^{(i)}\bigr) \;=\; \mathbb{E}_{P(\mathbf{x}_{1:K} \mid \mathbf{y}, \Theta^{(i)})}\bigl[\mathcal{L}(\Theta)\bigr],$$

by alternating between:

Input: Observations $\{\mathbf{y}^{(k)}\}_{k=1}^{K}$,;
current parameters $\Theta^{(i)} = \{\mathbf{x}_0^{(i)}, \mathbf{\Sigma}^{(i)}, \mathbf{F}^{(i)}, \mathbf{u}^{(i)}\}$
Output: Updated parameters $\Theta^{(i+1)}$
begin

// ——— E–step ——— **(E-step)** Run Kalman filter + RTS smoother to obtain $\{\mathbf{x}_{k|K}, \mathbf{W}_{k|K}\}_{k=1}^{K}$ and lag-one covariances $\mathbf{W}_{k,k-1|K}$;

// ——— M–step ——— **(M-step)**

Initial state:;

$\mathbf{x}_0^{(i+1)} \leftarrow \mathbf{x}_{1|K}$;

Block-wise dynamics:

for $b = 1, \ldots, B$ **do**

$$\mathbf{F}^{(b)\,(i+1)} \leftarrow \Big(\sum_{k=1}^{K} (\mathbf{x}_{k|K}^{(b)} - \mathbf{u}^{(b)\,(i)}) \mathbf{x}_{k-1|K}^{(b)\top}\Big)\Big(\sum_{k=1}^{K} \mathbf{x}_{k-1|K}^{(b)} \mathbf{x}_{k-1|K}^{(b)\top}\Big)^{-1};$$

$$\mathbf{u}^{(b)\,(i+1)} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \big(\mathbf{x}_{k|K}^{(b)} - \mathbf{F}^{(b)\,(i+1)} \mathbf{x}_{k-1|K}^{(b)}\big);$$

end

Process noise:

for $j = 1, \ldots, R^2$ **do**

$$\sigma_j^{2\,(i+1)} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \big(x_{k|K,j} - (\mathbf{F}^{(i+1)} \mathbf{x}_{k-1|K} + \mathbf{u}^{(i+1)})_j\big)^2;$$

end

;

return $\Theta^{(i+1)}$

end

Algorithm 1: One EM iteration (adaptive block-diagonal Kalman filter)

Algorithmic details of the forward filter, RTS smoother, and numerical root-finding are provided in Sections 5–5.2 and Appendix B.

5 E-Step in Detail

The E-step computes posterior moments of the latent states via a forward (filtering) pass followed by a backward (RTS smoothing) pass.

5.1 Forward (Filtering) Pass

Initialisation

$$\mathbf{x}_{1|0} = \mathbf{F}\,\mathbf{x}_0^{(i)} + \mathbf{u}, \qquad \mathbf{W}_{0|0} = \mathbf{0}. \tag{5}$$

For $k = 1, \ldots, K$ **do:**

Prediction

$$\mathbf{x}_{k|k-1} = \mathbf{F}\,\mathbf{x}_{k-1|k-1} + \mathbf{u}, \qquad \mathbf{W}_{k|k-1} = \mathbf{F}\,\mathbf{W}_{k-1|k-1}\mathbf{F}^{\top} + \mathbf{\Sigma}^{(i)}. \tag{6}$$

Update Let $\mathbf{J}_k$ and $\mathbf{H}_k$ be the Jacobian and Hessian of the log-softmax with respect to $\mathbf{x}_{k|k-1}$ (see Appendix A). Define the innovation

$$\boldsymbol{\delta}_k \;=\; \sum_{m=1}^{n(k)} \Big[\mathbf{e}_{\,y_m^{(k)},\,y_{m-1}^{(k)}} - f\big(\mathbf{x}_{k|k-1}\big)\Big].$$

Then

$$\mathbf{W}_{k|k}^{-1} = \mathbf{W}_{k|k-1}^{-1} + \mathbf{H}_k, \tag{7}$$

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{W}_{k|k}\,\boldsymbol{\delta}_k. \tag{8}$$

5.2 Backward (RTS Smoothing) Pass

For $k = K-1, \ldots, 1$ compute

$$\mathbf{M}_k = \mathbf{W}_{k|k}\,\mathbf{F}^\top\big(\mathbf{W}_{k+1|k}\big)^{-1}, \tag{9}$$

$$\mathbf{x}_{k|K} = \mathbf{x}_{k|k} + \mathbf{M}_k\big(\mathbf{x}_{k+1|K} - \mathbf{x}_{k+1|k}\big), \tag{10}$$

$$\mathbf{W}_{k|K} = \mathbf{W}_{k|k} + \mathbf{M}_k\big(\mathbf{W}_{k+1|K} - \mathbf{W}_{k+1|k}\big)\mathbf{M}_k^\top. \tag{11}$$

The smoothed lag-one covariance is

$$\mathbf{W}_{k,\,k+1|K} \;=\; \mathbf{M}_k\,\mathbf{W}_{k+1|K}.$$

6 M-Step in Detail

The maximisation step updates four parameter groups: $\{\mathbf{x}_0, \boldsymbol{\Sigma}, \mathbf{F}, \mathbf{u}\}$.
Smoothed moments $\mathbf{x}_{k|K}, \mathbf{W}_{k|K}, \mathbf{W}_{k,k-1|K}$ from the E-step are treated as known statistics.

6.1 Initial State $\mathbf{x}_0$

The conditional mode of the complete-data log-likelihood is

$$\mathbf{x}_0^\star \;=\; \mathbf{x}_{1|K},$$

obtained by setting the gradient to zero (derivation in Appendix B). A Gaussian prior $\mathcal{N}(\boldsymbol{\mu}_0, \mathbf{P}_0)$ may be incorporated by the standard posterior mean formula $(\mathbf{P}_0^{-1} + \mathbf{W}_{1|K}^{-1})^{-1}(\mathbf{P}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{W}_{1|K}^{-1}\mathbf{x}_{1|K})$.

6.2 Block-Diagonal Dynamics $\{\mathbf{F}^{(b)}, \mathbf{u}^{(b)}\}$

For each block $b = 1, \ldots, B$ let $\mathbf{Z}_k^{(b)} = \mathbf{x}_{k|K}^{(b)},\ \mathbf{Z}_{k-1}^{(b)} = \mathbf{x}_{k-1|K}^{(b)}$. Define the sufficient statistics

$$\mathbf{S}_1^{(b)} = \sum_{k=1}^{K} \mathbf{Z}_k^{(b)} \mathbf{Z}_{k-1}^{(b)\top}, \qquad \mathbf{S}_0^{(b)} = \sum_{k=1}^{K} \mathbf{Z}_{k-1}^{(b)} \mathbf{Z}_{k-1}^{(b)\top}.$$

With a small ridge parameter $\lambda > 0$ for numerical stability, the least-squares updates are

$$\mathbf{F}^{(b)\,(i+1)} = \mathbf{S}_1^{(b)}\big(\mathbf{S}_0^{(b)} + \lambda \mathbf{I}_d\big)^{-1}, \tag{12}$$

$$\mathbf{u}^{(b)\,(i+1)} = \frac{1}{K}\sum_{k=1}^{K}\big(\mathbf{Z}_k^{(b)} - \mathbf{F}^{(b)\,(i+1)}\mathbf{Z}_{k-1}^{(b)}\big). \tag{13}$$

Assembling the blocks yields $\mathbf{F}^{(i+1)} = \mathrm{blockdiag}[\mathbf{F}^{(1)}, \ldots, \mathbf{F}^{(B)}]$ and $\mathbf{u}^{(i+1)} = [\mathbf{u}^{(1)\top} \ldots \mathbf{u}^{(B)\top}]^\top$.

6.3 Process-Noise Variances $\{\sigma_j^2\}$

Using the fresh dynamics estimates,

$$\sigma_j^{2\,(i+1)} \;=\; \frac{1}{K}\sum_{k=1}^{K}\Big[x_{k|K,j} - (\mathbf{F}^{(i+1)}\mathbf{x}_{k-1|K} + \mathbf{u}^{(i+1)})_j\Big]^2, \qquad j = 1,\ldots,R^2.$$

If desired, an inverse-Gamma prior can be added with two extra terms.

Complexity. Each block update costs $O(Kd^2)$ for the accumulators and $O(d^3)$ for the inversion in Eq. (12); total cost per EM iteration is $O(KBd^2 + Bd^3)$.

The analytic updates (12)–(13) guarantee non-decreasing likelihood, while the ridge λ prevents singular $\mathbf{S}_0^{(b)}$. Full derivations appear in Appendix F.

References

[1] J. Daunizeau. Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. *arXiv preprint arXiv:1703.00091*, 2017.

[2] U. Eden and E. Brown. Dynamic analysis of neural encoding by point-process adaptive filtering. *Neural Computation*, 16(5):971–998, 2004.

[3] Y. Cohen. Development of the point-process filtering and SS-GLM. Technical report, 2016.

[4] S. Czanner. Analysis of between-trial and within-trial neural spiking dynamics. PhD thesis, 2008.

[5] T. Chitralekha. Comparison of EM-based parameter estimation using particle, unscented, and extended Kalman filters. MSc thesis, 2009.

A Jacobians and Hessians of the Log-Softmax

For a row-wise softmax $f(\mathbf{x})$,

$$J_{ij} = \frac{\partial \log f_i}{\partial x_j} = \delta_{ij} - f_j, \qquad H_{ij} = \frac{\partial^2 \log f_i}{\partial x_i\,\partial x_j} = -f_i\,(\delta_{ij} - f_j).$$

B Numerical Update of $\mathbf{x}_0$

This equation is solved by Newton–Raphson:

1. Initialise with the closed-form solution.
2. Iterate $\mathbf{x}_0 \leftarrow \mathbf{x}_0 - \big[\nabla^2 g(\mathbf{x}_0)\big]^{-1}\nabla g(\mathbf{x}_0)$, where g is the left-hand side of (**??**).
3. Stop when $\|\nabla g(\mathbf{x}_0)\|_\infty < 10^{-6}$.

C Derivation of the Linear–Gaussian Prediction Step

Starting from the state model $\mathbf{x}_{k+1} = \mathbf{F}\,\mathbf{x}_k + \mathbf{u} + \boldsymbol{\varepsilon}_k,\ \boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, take expectations conditioned on $\mathcal{F}_k = \sigma\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}\}$:

$$\mathbb{E}[\mathbf{x}_{k+1} \mid \mathcal{F}_k] = \mathbf{F}\,\mathbb{E}[\mathbf{x}_k \mid \mathcal{F}_k] + \mathbf{u} \implies \boxed{\mathbf{x}_{k|k-1} = \mathbf{F}\,\mathbf{x}_{k-1|k-1} + \mathbf{u}}$$

$$\mathrm{Cov}(\mathbf{x}_{k+1} \mid \mathcal{F}_k) = \mathbf{F}\,\mathbf{W}_{k-1|k-1}\mathbf{F}^\top + \boldsymbol{\Sigma} \implies \boxed{\mathbf{W}_{k|k-1} = \mathbf{F}\,\mathbf{W}_{k-1|k-1}\mathbf{F}^\top + \boldsymbol{\Sigma}}$$

These two identities justify the prediction lines used in Section 5.

D Derivation of the RTS Smoother Gain

Define the joint predictor $\begin{bmatrix} \mathbf{x}_k \\ \mathbf{x}_{k+1} \end{bmatrix} \Big| \mathcal{F}_k \sim \mathcal{N}\Big(\begin{bmatrix} \mathbf{x}_{k|k} \\ \mathbf{x}_{k+1|k} \end{bmatrix}, \begin{bmatrix} \mathbf{W}_{k|k} & \mathbf{W}_{k|k}\mathbf{F}^\top \\ \mathbf{F}\,\mathbf{W}_{k|k} & \mathbf{W}_{k+1|k} \end{bmatrix} \Big)$.

Conditioning on $\mathbf{x}_{k+1}$ gives

$$\mathbf{x}_{k|K} = \mathbf{x}_{k|k} + \boxed{\mathbf{M}_k}\big(\mathbf{x}_{k+1|K} - \mathbf{x}_{k+1|k}\big), \qquad \mathbf{M}_k = \mathbf{W}_{k|k}\mathbf{F}^\top \big(\mathbf{W}_{k+1|k}\big)^{-1},$$

which matches the gain used in Section 5.2. The covariance identity $\mathbf{W}_{k|K} = \mathbf{W}_{k|k} + \mathbf{M}_k(\mathbf{W}_{k+1|K} - \mathbf{W}_{k+1|k})\mathbf{M}_k^\top$ follows from the standard Schur complement formula.

E EM – M-Step Algebra

Update of the Process-Noise Variance

Write the expected quadratic term in the complete-data log-likelihood (cf. Eq. (4)):

$$Q_\Sigma = -\frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\big[(\mathbf{x}_k - \mathbf{F}\mathbf{x}_{k-1} - \mathbf{u})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \mathbf{F}\mathbf{x}_{k-1} - \mathbf{u})\big] - \frac{K}{2}\log|\boldsymbol{\Sigma}|.$$

Differentiate with respect to the j^{th} diagonal element σ_j^2 and set to zero:

$$-\frac{1}{2}\sum_{k=1}^{K} \frac{\mathbb{E}[(\Delta x_{k,j})^2]}{\sigma_j^4} + \frac{K}{2\sigma_j^2} = 0 \implies \boxed{\sigma_j^2 = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[(\Delta x_{k,j})^2]}$$

with $\Delta \mathbf{x}_k = \mathbf{x}_k - \mathbf{F}\mathbf{x}_{k-1} - \mathbf{u}$.

Update of the Initial State

The linear terms in Q that depend on $\mathbf{x}_0$ are

$$-\tfrac{1}{2}(\mathbf{x}_1 - \mathbf{F}\mathbf{x}_0 - \mathbf{u})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_1 - \mathbf{F}\mathbf{x}_0 - \mathbf{u}) \;+\; \log P(\mathbf{x}_0),$$

where the prior $P(\mathbf{x}_0)$ is either flat or Gaussian. Solving $\partial Q/\partial \mathbf{x}_0 = \mathbf{0}$ yields Eq. (**??**) when the prior is flat ($\boldsymbol{\Sigma}^{-1}\mathbf{F}^\top\mathbf{F}$ term drops out) and Newton–Raphson update (**??**) when the softmax likelihood of sequence 1 is retained.

These derivations reproduce the algebraic steps in the original (manuscript-length) draft while harmonising them with the augmented state model $\mathbf{F}, \mathbf{u}$.

F Derivation of the F and u M-step

Define the centred variables $\Delta\mathbf{x}_k = \mathbf{x}_k - \mathbf{u}$. Conditioned on $\mathcal{F}_K$ the quadratic term in the expected log-likelihood is

$$Q_{F,u} = -\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\big[\|\Delta\mathbf{x}_k - \mathbf{F}\mathbf{x}_{k-1}\|^2_{\mathbf{\Sigma}^{-1}}\big].$$

Differentiating w.r.t. $\mathbf{F}$ and $\mathbf{u}$ gives the normal equations

$$\sum_{k=1}^{K}\mathbb{E}[\Delta\mathbf{x}_k\mathbf{x}_{k-1}^\top] = \mathbf{F}\sum_{k=1}^{K}\mathbb{E}[\mathbf{x}_{k-1}\mathbf{x}_{k-1}^\top], \qquad \mathbf{u} = \frac{1}{K}\sum_{k=1}^{K}\big(\mathbb{E}[\mathbf{x}_k] - \mathbf{F}\,\mathbb{E}[\mathbf{x}_{k-1}]\big),$$

With the block-diagonal constraint, the normal equations decouple across blocks, yielding the block-wise solutions reported in Section 6.

G Practical Implementation Notes

- **Diagonal $\mathbf{\Sigma}$.** A diagonal process-noise covariance is generally sufficient and avoids costly matrix inversions in (**??**).
- **Regularisation.** When $\mathbf{W}_{k|k}$ becomes ill-conditioned, add a small multiple of the identity: $\mathbf{W}_{k|k} \leftarrow \mathbf{W}_{k|k} + \varepsilon\,\mathbf{I}$.
- **Approximating $\mathbb{E}[\log f]$.** A second-order delta method (Daunizeau, 2017) yields analytic moment estimates while keeping normalisation errors below 10^{-3}.
- **Stopping criterion.** Monitor the absolute change in the expected log-likelihood Q; a threshold of 10^{-4} works well in practice.