S1 Text. Mathematical details

Derivation of ELBO and Inference

1 Evidence Lower Bound (ELBO)

Here we derive the evidence lower bound (ELBO) used as a variational objective by our inference algorithm. That is, we want to calculate

$$L \equiv \mathbb{E}_q \left[ \log \frac{p(\Theta | N)}{q(\Theta)} \right] = \mathbb{E}_q [\log p(\Theta | N)] + \mathcal{H}[q(\Theta)]$$

From [1], this can be written

$$L = \mathbb{E}_q(\pi) \left[ \log \frac{p(\pi)}{q(\pi)} \right] + \mathbb{E}_q(A) \left[ \log \frac{p(A)}{q(A)} \right] + \mathbb{E}_q(\lambda) \left[ \log \frac{p(\lambda)}{q(\lambda)} \right] + \mathbb{E}_q(\gamma) \left[ \log \frac{p(\gamma)}{q(\gamma)} \right]$$

For the first two terms, updates are standard and covered in [1]. The rest we do piece-by-piece below:

1.1 Log evidence

We would like to calculate $\mathbb{E}_q \left[ \log \frac{p(N, z | x, \Theta)}{q(z)} \right]$. To do this, we make use of expectations calculated via the posteriors returned from the forward-backward algorithm

$$\xi_t \equiv p(z_t | N, \Theta) \quad \Xi_{t,ij} \equiv p(z_{t+1} = j, z_t = i | N, \Theta) \quad \log Z_t = \log p(N_t+1 | N_t, \Theta)$$

Here, we have suppressed the latent feature index $k$ and abuse notation by writing the observation index as $t$, but in the case of multiple observations at a given time, we pool across units and presentations: $N_t \equiv \sum_{m,t(m)=t} N_m$. From this, we can write

$$\mathbb{E}_q \left[ \log \frac{p(N, z | x, \Theta)}{q(z)} \right] = \sum_{mkr} \left[ N_m \left( \log \theta_m + \log \lambda_{0u(m)} + \xi_{t(m)k} \log \lambda_{zuk} + x_{t(m)r} \log \lambda_{xu(m)r} \right) \right]$$

$$- \sum_{m} \theta_m \mathbb{E}_q \left[ \Lambda_{t(m)u(m)} \right] + \sum_{tk} \left[ \text{tr} \left( \Xi_{tk} \frac{\log A_k}{\xi_{tk}} \right) + \xi_{tk} \frac{\log \pi_k}{\xi_{tk}} \right] + \text{constant} \quad (4)$$

where for notational simplicity, we have replaced expectations with overlines: $\bar{x} \equiv \mathbb{E}[x]$. In what follows, we will drop the irrelevant constant. For $\log y$, where $y \in \{\theta, \lambda_0, \lambda_z, \lambda_x\}$, the assumption $q(y) = \text{Ga}(\alpha, \beta)$ gives

$$\log y = \psi(\alpha) - \log \beta$$

with $\psi(x)$ the digamma function. Likewise, the expectation $\bar{y}$ is straightforward. For the expectation of the rate, we have

$$\mathbb{E}_q \left[ \lambda_{0u} \prod_k (\lambda_{zuk})^{z_{tk}} \prod_r (\lambda_{xur})^{x_{tr}} \right] = \frac{\alpha_{0u}}{\beta_{0u}} \prod_k \left( 1 - \xi_{tk} + \xi_{tk} \frac{\alpha_{zuk}}{\beta_{zuk}} \right) \prod_r \frac{1}{\beta_{xur}^{x_{tr}}} \frac{\Gamma(\alpha_{xur} + x_{tr})}{\Gamma(\alpha_{xur})} \quad (6)$$
However, for $\alpha \gg x$, we have $\Gamma(\alpha + x)/\Gamma(\alpha) \approx \alpha^x$, so that we can write

$$
\mathbb{E}_q[\Lambda_{tu}] = H_{0u} F_{tu} G_{tu} \tag{7}
$$

with $G_{tu} \approx \prod_r (\alpha_{xur}/\beta_{xur})^{x_{tr}}$. In addition, it will later be useful to have the expectation over all except a particular feature $k$ or $r$, for which we define

$$
F_{tk} = \prod_{k' \neq k} \left(1 - \xi_{tk'} + \xi_{tk} \frac{\alpha_{zuk'}}{\beta_{zuk'}} \right) \tag{8}
$$

$$
G_{tur} = \prod_{r' \neq r} \left(\frac{\alpha_{xur'} \beta_{xur'}}{\beta_{xur'}} \right)^{x_{tr}} \tag{9}
$$

Finally, we want the entropy of the variational posterior over $z$, $\mathbb{E}_q[- \log q(z)]$. We can write this in the form

$$
- \sum_{tk} \left[ \xi_{tk}^T \eta_{tk} + \text{tr} \left( A_k^T \tilde{A}_k \right) + \xi_{tk}^T \tilde{\pi}_k - \log Z_{tk} \right] \tag{10}
$$

with $(\eta, \tilde{A}, \tilde{\pi})$ the parameters of the variational posterior corresponding to the emission, transition, and initial state probabilities of the Markov chain (interpreted as matrices) and $Z$ the normalization constant. From [1], we have that variational updates should give

$$
\tilde{A}_k \leftarrow \frac{\log A_k}{\lambda_{zu}} \tag{11}
$$

$$
\tilde{\pi}_k \leftarrow \frac{\log \pi_k}{\theta} \tag{12}
$$

while the effective emission probabilities in the “on” ($z = 1$) state of the HMM are

$$
\eta_{tk} \leftarrow \delta_{zk,1} \sum_{m:t(m)=t} N_{m} \log \lambda_{zu(m)} - \sum_{m:t(m)=t} \bar{D}_{m} H_{0u(m)} F_{tku(m)} G_{tu(m)} \tag{13}
$$

Given these update rules, we can then alternate between calculating $(\eta, \tilde{A}, \tilde{\pi})$, performing forward-backward to get $(\xi, \Xi, \log Z)$ and recalculating $(\eta, \tilde{A}, \tilde{\pi})$.

1.2 Overdispersion, firing rate effects

Both the case of $p(\theta)$ and $p(\lambda)$ are straightforward. If we ignore subscripts and write $p(y) = \text{Ga}(a, b)$, $q(y) = \text{Ga}(\alpha, \beta)$, then

$$
\mathbb{E}_q \left[ \log \frac{p(y)}{q(y)} \right] = (\bar{\pi} - 1) \overline{\log y} + \overline{\log \theta} + \mathcal{H}[q(y)] \tag{14}
$$

where again, $\overline{\log y}$, $\overline{\log \theta}$ and $\mathcal{H}[q(y)]$ are straightforward properties of the Gamma distribution. Expectations of the prior parameters are listed in Table [1]. Note in the last line that there is no expectation, since we have not assumed a hierarchy over firing rate effects for the covariates, $x$.

1.3 Hyperparameters

As shown in the main text, the hyperparameters $c$ and $d$, given gamma priors, have conjugate gamma posteriors, so that their contribution to the evidence lower bound, $\mathbb{E}_q \left[ \log \frac{p(c)}{q(c)} \right]$ is a sum of terms of the form

$$
(a_c - 1) \overline{\log c} + b_c \overline{c} + \mathcal{H}[q(c)] + (a_d - 1) \overline{\log d} + b_d \overline{d} + \mathcal{H}[q(d)] \tag{15}
$$
Table 1. Expectations of prior parameters for overdispersion and firing rates

| Variable | $E_q[a]$ | $E_q[b]$ |
|----------|----------|----------|
| $\theta$ | $\tau_0$ | $c_0d_0$ |
| $\lambda_0$ | $\tau_z$ | $c_zd_z$ |
| $\lambda_z$ | $a_x$ | $b_x$ |

In other words, these are straightforward gamma expectations, functions of the prior parameters $a$ and $b$ for each variable and the corresponding posterior parameters $\alpha$ and $\beta$. Similarly, the overdispersion terms are exactly the same with the substitutions $c \rightarrow s$, $d \rightarrow 1$.

As we will see below, the expectations under the variational posterior of $s$, $c$, and $d$ are themselves straightforward to calculate.

1.4 Autocorrelated noise

In the main text, we assumed the $\theta_m$ to be uncorrelated. However, it is possible to model temporal autocorrelation among the $\theta_m$ when observations correspond to the same neuron at successive time points. More specifically, let us replace the observation index $m$ by $\tau$, the experimental clock time. We then write a particular observation as corresponding to a stimulus time $t(\tau)$ and a set of units $u(\tau)$, which, for simplicity, we will assume fixed and simply write as $u^T$. To model the autocorrelation of noise across successive times, we then write

$$N_m \sim \text{Pois}(\Lambda_{t(\tau),u} \theta_{\tau u})$$

Assuming $\theta_{\tau u} = \phi_{\tau u} \theta_{\tau -1,u}$  \hspace{1cm} (16)

If $\phi_{0u} = \theta_{0u}$, then we have $\theta_{\tau u} = \prod_{\tau' \leq \tau} \phi_{\tau' u}$. In essence, this is a log-autoregressive process in which the innovations are not necessarily normally distributed. In fact, if we further assume that $p(\phi_{\tau u}) = \text{Ga}(s_u, r_u)$ and $q(\phi_{\tau u}) = \text{Ga}(\omega_{\tau u}, \zeta_{\tau u})$, we can once again make use of conjugate updates.

Given these assumptions, we need to make the following modifications for the third and fifth terms in the evidence lower bound of Eq (4):

$$\sum_{mkr} N_m \xi(t(m)) k \log \lambda_{zuk} \to \sum_{mkr} N_m \xi(t(m)) k \log \lambda_{zuk} + \sum_{\tau u} \log \phi_{\tau u} \sum_{\tau' \geq \tau} N_{\tau' u}$$

$$- \sum_{m} \theta_{\tau u} E_q[\Lambda_{t(m)u(m)}] \to \sum_{\tau u} H_{0u} F_{\tau ku} G_{\tau u} \prod_{\tau' \leq \tau} \frac{\omega_{\tau' u}}{\zeta_{\tau' u}}$$

Thus the effective emission probabilities in State 1 now become:

$$\eta_{tk} \to \delta_{ztk,1} \sum_{m} N_m \log \lambda_{zuk} - \sum_{u; t(\tau) = t} H_{0u} F_{\tau ku} G_{\tau u} \prod_{\tau' \leq \tau} \frac{\omega_{\tau' u}}{\zeta_{\tau' u}}$$

1.5 Latent states, semi-Markov dynamics

We model each of the latent states $z_{tk}$ as an independent Markov process for each feature $k$. That is, each $k$ indexes an independent Markov chain with initial state probability $\pi_k \sim \text{Dir}(\alpha_k)$ and transition matrix $A_k \sim \text{Dir}(\alpha_A)$. For the semi-Markov case, we assume that the dwell times in each state are distributed independently for

---

1. The generalization to partially overlapping neurons and stimuli is straightforward but complex and notationally cumbersome.
each chain according to an integer-valued, truncated lognormal distribution with support on the integers $1 \ldots D$:

$$p_k(d|z = j) = \text{Log-Normal}(d|m_{jk}, s_{jk}^2)/W_{jk}$$  \hspace{1cm} (20)

$$W_{jk} = \sum_{d=1}^{D} \text{Log-Normal}(d|m_{jk}, s_{jk}^2)$$  \hspace{1cm} (21)

Note that we have allowed the dwell time distribution to depend on both the feature $k$ and the state of the Markov chain $j$. In addition, we put independent Normal-Gamma priors on the mean $(m_{kj})$ and precision $(\tau_{kj} \equiv s_{jk}^{-2})$ parameters of the distribution: $(m, \tau) \sim \text{NG}(\mu, \lambda, \alpha, \beta)$.

In this case, we additionally need to perform inference on the parameters $(m, \tau)$ of the dwell time distributions for each state. In the case of continuous dwell times, our model in Equation (20) would have $W = 1$ and be conjugate to the Normal-Gamma prior on $(m, \tau)$, but the restriction to discrete dwell times requires us to again lower bound the variational objective:

$$E_q[-\log W_{jk}] = E_q[-\log \left(\sum_{d=1}^{D} p(d|j)\right)] \geq -\log \sum_{d=1}^{D} E_q[p(d|j)]$$  \hspace{1cm} (22)

This correction for truncation must then be added to $E_q[p(z|\Theta)]$. For inference in the semi-Markov case, we use an extension of the forward-backward algorithm, as well as the algorithms in [3,4], at the expense of computational complexity $O(SDT)$ ($S = 2$) per latent state, to calculate $q(z_k)$. For the $4SK$ hyperparameters of the Normal-Gamma distribution, we perform an explicit BFGS optimization on the $4S$ parameters of each chain on each iteration (detailed in Subsection 2.3).

The new ELBO for the semi-Markov model thus adds to the terms involving $\Xi$ and $q_0$ in (4) an additional piece:

$$\sum_{djk} C(d, j, k) E_q \left[ \log p_k(d|j) - \log \left(\sum_{d=1}^{D} p_k(d|j)\right) \right] \geq \sum_{djk} C(d, j, k) \left[ E_q[\log p_k(d|j)] - \log \sum_{d=1}^{D} E_q[p_k(d|j)] \right]$$  \hspace{1cm} (23)

where as noted above, the second term inside the expectation arises from the need to normalize $p(d|j)$ over discrete times and the inequality follows from Jensen’s Inequality. Here, $d$ is the duration variable, $j$ labels states of the chain (here 0 and 1), and $k$, as elsewhere, labels chains [2,4]. $C$ is defined for each state and chain as the probability of a dwell time $d$ in state $j$ conditioned on the event of just having transitioned to $j$.

Thus the objective to be optimized is

$$\sum_{djk} C(d, j, k) \left[ E_q[\log p_k(d|j)] - \log \sum_{d=1}^{D} E_q[p_k(d|j)] \right] + E_q \left[ \log p(m, \tau) \right]$$  \hspace{1cm} (25)

\footnote{Thus $C$ is equivalent to $\mathcal{D}_{\nu/\gamma}$ in [2].}

\footnote{We also note that, just as the emission, transition, and initial state probabilities have their counterparts in variational parameters ($\eta, A, \pi$) in $q(z)$, so $C$ is matched by a term $-\sum_{djk} C(d, j, k)\nu_{djk}$ in $E_q[-\log q]$. And in analogy with the HMM case, variation with respect to $\nu$ gives

$$\nu_{djk} = E_q[\log p_k(d|j)] - \log \sum_{d=1}^{D} E_q[p_k(d|j)],$$  \hspace{1cm} (24)

implying that there are cancellations between $\log p(N, z|\Theta)$ and $\log q(z)$ in $\mathcal{L}$ and care must be taken when calculating it [1].}
We focus on each of these and their updates below.

#### 1.5.1 Contributions to the ELBO

For the case of \( p(d|m, s^2) \) Log-Normal and \((m, \tau = s^{-2})\) Normal-Gamma, the terms \( \mathbb{E}_q[\log p(d|j)] \) involve only routine expectations of natural parameters of the Normal-Gamma, and similarly for the expected log prior and entropy in the last term. Writing the delay distribution in the exponential family form with base measure \( h \) and natural parameter \( \eta \)

\[
p(d|z = i) = h(d) \exp(\eta_i \cdot T(d) - A(\eta_i)) \tag{26}
\]

yields for the lognormal distribution \( \log d \sim \mathcal{N}(m, s^2) \)

\[
h(d) = \frac{1}{\sqrt{2\pi d}} \tag{27}
\]

\[
T(d) = \left[ \frac{\log d}{(\log d)^2} \right] \tag{28}
\]

\[
\eta = \left[ -\frac{m}{2s^2} \right] \tag{29}
\]

\[
A(\eta) = \frac{m^2}{2s^2} + \log s \tag{30}
\]

If we then put a Normal-Gamma posterior on \((m, \tau)\) with parameters \((\mu, \lambda, \alpha, \beta)\), we can easily calculate

\[
\mathbb{E}_q[\eta] = \mathbb{E}_q \left[ \frac{m \tau}{\frac{1}{\tau}} \right] = \left[ \frac{\mu \alpha}{\beta} \right] \tag{31}
\]

\[
\mathbb{E}_q[A(\eta)] = \mathbb{E}_q \left[ \frac{1}{2} m^2 \tau - \frac{1}{2} \log \tau \right] \tag{32}
\]

\[
= \frac{1}{2} \mathbb{E}_\tau \left[ \left( \mu^2 + \frac{1}{\lambda \tau} \right) \tau - \frac{1}{2} (\psi(\alpha) - \log \beta) \right] \tag{33}
\]

\[
= \frac{1}{2} \left( \mu^2 \alpha + \frac{1}{\lambda} \right) - \frac{1}{2} (\psi(\alpha) - \log \beta) \tag{34}
\]

The relevant piece of the evidence lower bound for updating \((\mu, \lambda, \alpha, \beta)\) is

\[
\sum_{i, d} C(d, i) [\log h(d) + \mathbb{E}_q[\eta_i] \cdot T(d) - \mathbb{E}_q[A(\eta_i)] + \mathbb{E}_q \left[ \log \frac{p(m, \tau)}{q(m, \tau)} \right]] \tag{35}
\]

again, with \( q(m, \tau) = \text{Normal-Gamma}(\mu, \lambda, \alpha, \beta) \). We can thus write

\[
\mathbb{E}_q[\log p(d|j)] = -\frac{1}{2} \log 2\pi C_0 - C_1 + \mathbb{E}_q[m \tau]C_1 - \frac{1}{2} \mathbb{E}_q[\tau]C_2 - \frac{1}{2} C_0 \mathbb{E}_q[m^2 \tau] + \frac{1}{2} C_0 \mathbb{E}_q[\log \tau]
\]

\[
= G + F \cdot T(\alpha, \beta, \lambda, \mu) \tag{36}
\]
where

\[ G \equiv -\frac{1}{2} \log 2\pi C_0 - C_1 \quad (37) \]

\[ F \cdot T = \begin{bmatrix} C_0 \quad -C_2 \quad C_1 \quad -C_0 \end{bmatrix} \cdot \mathbb{E}_q \begin{bmatrix} \log \tau \\ \tau m \tau \\ m^2 \tau \end{bmatrix} \quad (38) \]

\[ C_0 \equiv \sum_{id} C_{id} \quad (39) \]

\[ C_1 \equiv \sum_{id} C_{id} \log d \quad (40) \]

\[ C_2 \equiv \sum_{id} C_{id} (\log d)^2 \quad (41) \]

For calculating the final term in (25), we can make use of conjugacy (since \( p(m, \tau | \alpha_0, \beta_0, \lambda_0, \mu_0) \) is itself Normal-Gamma) and the fact that (36) only depends linearly on the expected sufficient statistics of \( q \) to write the contribution of \( \mathbb{E}_q \log p(m, \tau) \) as an update to \( F \):

\[ F' = \begin{bmatrix} C_0^2 + \alpha_0 \\ -C_2^2 - \beta_0 - \lambda_0 \mu_0^2 \\ C_1 + \lambda_0 \mu_0^2 \\ -C_0^2 - \lambda_0^2 \end{bmatrix} \quad (42) \]

Moreover, we can make use of the standard entropy formula for the Normal-Gamma distribution:

\[ H[q] = \mathbb{E}_q[-\log q] = \mathbb{E}_q[-\log q(\tau) - \log q(m|\tau)] = H_g(\alpha, \beta) + \mathbb{E}_q \left[ -\log \sqrt{\frac{\lambda \tau}{2\pi}} + \frac{\lambda \tau}{2} (m - \mu)^2 \right] = \alpha - \log \beta + \log \Gamma(\alpha) + (1 - \alpha) \psi(\alpha) - \frac{1}{2} \log \frac{\lambda}{2\pi} - \frac{1}{2} (\psi(\alpha) - \log \beta) + \frac{1}{2} \quad (43) \]

Only slightly more complicated is the term arising from the normalization constant, which in the standard \((\mu, \lambda, \alpha, \beta)\) parameterization of the Normal-Gamma takes the form

\[ \mathbb{E}_q[p(d|m, \tau)] = \int d\tau dm \ p(d|m, \tau) q(m, \tau) = \frac{1}{\sqrt{2\pi d}} \sqrt{\frac{\lambda}{1 + \lambda}} \ \Gamma(\alpha + 1/2) \ \beta^{-\frac{1}{2}} \quad (44) \]

with

\[ \beta \equiv 1 + \frac{1}{2\beta} \frac{\lambda}{1 + \lambda} (\log d - \mu)^2 \quad (45) \]

This can be derived either by performing the integral directly or by noting that the result should be proportional to the posterior (Normal-Gamma, by conjugacy) of \((m, \tau)\) after having observed a single data point, \( \log d \).

Taken together, (30), (36) (with \( F \rightarrow F' \)), (13), and (44) allow (25) to be calculated directly as a function of the variational parameters \((\alpha, \beta, \lambda, \mu)\) of \( q(m, \tau) \). Were it not for the normalization constant arising from the restriction of discrete \( d \), the updates resulting from optimizing this objective would reduce by conjugacy to simple parameter updates. With that additional term, we are forced to use a brute-force BFGS optimization step instead (see 2.3 below).
1.5.2 Forward-Backward Algorithm for HSMM: Notation

Here, we follow Yu and Kobayashi, mutatis mutandis. Define:

\[ \psi_t(m) = p(y_t|z_t = m) \]  
observation probability

\[ \alpha_{t|z}(m, d) = p(z_t = m, d_t = d|y_{1:t}) \]  
condition on data

\[ \gamma_{t|z}(m) = \sum_d \alpha_{t|z}(m, d) \]  
marginalize out \( d \)

\[ \psi^*_t(m) = \frac{\alpha_{t|z}(m, d)}{\alpha_{t|z-1}(m, d)} = \frac{\psi_t(m)}{p(y_t|y_{1:t-1})} \]  
4. Calculate

\[ r_t^{-1} = p(y_t|y_{1:t-1}) = \sum_{m,d} \alpha_{t|z-1}(m, d)\psi_t(m) \]

\[ = \sum_m \gamma_{t|z-1}(m)\psi_t(m) \]

\[ Z = p(y_{1:T}) = \left( \prod_{t=1}^T r_t \right)^{-1} \]  
probability of data

\[ D_{t|z}(m, d) = p(z_t = m, d_t = d, d_{t-1} = 1|y_{1:t}) \]  
\( \rightarrow (m, d) \)

\[ T_{t|z}(n, m) = p(z_t = n, z_{t-1} = m, d_{t-1} = 1|y_{1:t}) \]  
\( m \rightarrow n \)

\[ E_t(m) = p(z_t = m, d_t = 1|y_{1:t}) = \alpha_{t|z-1}(m, 1)\psi^*_t(m) \]  
m ends at \( t \)

\[ S_t(m) = p(d_t = 1, z_{t+1} = m|y_{1:t}) = \sum_n A(m, n)E_t(n) \]  
m starts at \( t + 1 \)

and for the backward pieces

\[ \beta_t(m, d) = \frac{p(z_t = m, d_t = d|y_{1:T})}{p(z_t = m, d_t = d|y_{1:t-1})} = \frac{p(y_{t:T}|z_t = m, d_t = d)}{p(y_{t:T}|y_{1:t-1})} \]

\[ E^*_t(m) = \frac{p(y_{t:T}|z_t = m, d_{t-1} = 1)}{p(y_{t:T}|y_{1:t-1})} = \sum_d p_m(d)\beta_t(m, d) \]

\[ S^*_t(m) = \frac{p(y_{t:T}|z_{t-1} = m, d_{t-1} = 1)}{p(y_{t:T}|y_{1:t-1})} = \sum_n E^*_t(n)A(n, m) \]

These quantities then allow us to write the posterior probabilities

\[ \alpha_{t|T}(m, d) = \beta_t(m, d)\alpha_{t|z-1}(m, d) \]  
posterior prob of \( (m, d) \)

\[ T_{t|T}(n, m) = E^*_t(n)A(n, m)E_{t-1}(m) \]  
posterior prob of \( m \rightarrow n \)

\[ D_{t|T}(m, d) = \beta_t(m, d)p_m(d)S_{t-1}(m) \]  
posterior prob of entering \( (m, d) \) at \( t \)

1.5.3 Forward

From Yu and Kobayashi, we have the following forward pass:

1. \( \alpha_{1|0}(m, d) = \pi_m p_m(d) \) for time 1; otherwise, use

\[ \alpha_{t|z-1}(m, d) = S_{t-1}(m)p_m(d) + \psi^*_{t-1}(m)\alpha_{t-1|z-2}(m, d + 1) \quad (46) \]

2. Use \( \alpha_{t|z-1} \) and \( \psi_t \) to get \( r_t \); together, these give \( \psi^*_t \). Store \( r_t \)

3. Calculate and save \( E_t \).

4. Calculate \( S_t \).

5. Loop.
1.5.4 Backward

Again following Yu and Kobayashi:

1. Initialize $\beta_T(m,d) = \psi_T^*(m) = \psi_T(m)r_T$
2. $\psi_t^*(m) = \psi_t(m)r_t$
3. Calculate $\beta_t(m,d)$ by
   $\beta_t(m,d) = \begin{cases} S_{t+1}^*(m)\psi_t^*(m), & d = 1 \\ \beta_{t+1}(m,d-1)\psi_t^*(m), & d > 1 \end{cases}$ \hspace{1cm} (47)

where, again,

- $S_{t+1}^*(m) = \sum_n E_{t+1}^*(n)A(n,m)$ \hspace{1cm} (48)
- $E_{t+1}^*(n) = \sum_d \beta_{t+1}(n,d)p_n(d)$ \hspace{1cm} (49)

4. Use the above equation to calculate $E_t^*$ and $S_t^*$. Save $E_t^*$.

1.5.5 Sufficient Statistics

Finally, we can calculate and return:

- $\log Z = -\sum_{t=1}^{T} \log r_t$ \hspace{1cm} (50)
- $\xi_t(m) = \sum_d \alpha_{t+1}(m,d) = \gamma_{t|T}(m) = \sum_d \alpha_{t|t-1}(m,d)\beta_t(m,d)$ \hspace{1cm} (51)
- $\Xi_{t+1,t}(n,m) = \Theta_{t+1|T}(n,m) = E_{t+1}^*(n)A(n,m)E_t(m)$ \hspace{1cm} (52)
- $C_t(m,d) = D_{t|T}(m,d) = \begin{cases} \pi_m p_m(d)\beta_t(m,d), & t = 1 \\ S_{t-1}(m)p_m(d)\beta_t(m,d), & t > 1 \end{cases}$ \hspace{1cm} (53)

Note that, because $\Xi_{t+1,t}$ is the joint absolute probability of the transition (i.e., not conditioned on $d_t = 1$) it is not normalized when summing over $m$ and $n$. The same is true for $C$ when summing over $m$ and $d$.

2 Inference

2.1 Conjugate updates

For updates on the overdispersion, firing rate, and hyperparameter variables, we have the simple conjugate update rules depicted in Table 2. These can be derived either from free-form variational arguments, or exponential family rules, but are in any case straightforward [5]. If we assume a $\text{Ga}(a,b)$ prior and $\text{Ga}(\alpha,\beta)$ variational posterior, all of these have a simple algebraic update in terms of sufficient statistics.

Here, we overload notation to write $N_{tu} = \sum_m \delta_{u(m),u}\delta_{t(m),t}N_m$, $\theta_u = \sum_m \delta_{u(m),u}\theta_m$, and make use of $H, G$, and $F$ as defined in Eq (7) - (9). Indeed, our implementation caches $F$ and $G$ for a substantial speedup (at the cost of additional memory requirements). For autocorrelated noise, the update rules are for the $j$-th item in the series of autocorrelation.
We also supply the optimizer with the gradient:

\[ \nabla_{\beta} \mathcal{L}_c = \alpha_{xur} - b_{xur}e^{-\epsilon_{xur}} - \sum_{u, m; u(m) = u} x_{t(m)ur} \bar{b}_m H_{0u(m)} F_{t(m)u(m)} e^{-\sum_{r \neq j} \epsilon_{mr} x_{t(m)ur(r)}} \]

2.2 Non-conjugate updates

We employ explicit optimization steps for two updates in our iterative algorithm. In each case, we employ an off-the-shelf optimization routine, though more efficient alternatives are likely possible.

2.2.1 Covariate firing rate effects

Because we do not restrict the covariates \( x(t) \) to be binary, Equation 6 no longer yields a conditional log probability for \( \lambda_x \) in the exponential family. This leaves us with a transcendental equation to solve for \( \beta_{xr} \). However, from Table 2, we see that \( \alpha_x \gtrsim \sum_r x_{tr} \), allowing us to approximate \( G_t \approx \prod_r (\alpha_x / \beta_{xr})^{x_{tr}} \) as we have done above. Moreover, since the sum \( \sum_r G_t \sim TG \), we expect \( \alpha_x / \beta_{x} \approx 1 \) at the optimum for most reasonable datasets. We thus reparameterize \( \beta_{xur} = \alpha_{xur}e^{-\epsilon_{xur}} \) and write the relevant \( \epsilon \)-dependent piece of the objective function \( \mathcal{L} \) as

\[ \mathcal{L}_c = \sum_{ur} [a_{xur} \epsilon_{ur} - b_{xur}e^{-\epsilon_{xur}}] - \sum_m \bar{b}_m H_{0u(m)} F_{t(m)u(m)} e^{-\sum_r \epsilon_{ur} x_{t(m)ur(m)}} \]

We also supply the optimizer with the gradient:

\[ \nabla_{\beta} \mathcal{L}_c = \alpha_{xur} - b_{xur}e^{-\epsilon_{xur}} - \sum_{m; u(m) = u} x_{t(m)ur} \bar{b}_m H_{0u(m)} F_{t(m)u(m)} e^{-\sum_r \epsilon_{mr} x_{t(m)ur}} \]

where we sum only over observations with \( u(m) = u \). On each iterate, we then optimize \( \mathcal{L}_c \), initializing \( \beta_x \) to the just-updated value of \( \alpha_x \). In addition, we do not update firing rate effects for covariates separately, but optimize \( \beta_{xur} \) for all \( r \) together. Again, more efficient schemes are no doubt possible.

2.3 Semi-Markov duration distribution

If the dwell time distributions of states in the semi-Markov model were truly continuous, the parameters \( (m, x^2) \) of a lognormal distribution \( p(dj) \) would have a Normal-Gamma conjugate prior, and updates would be closed-form. However, the requirement that the durations be integers and that there exist a maximal duration, \( D \), over which these probability mass functions must normalize results in an extra term in \( \mathbb{E}_q[\log p] \) arising
from the normalization constant. In this case, we must explicitly optimize for the parameters of the Normal-Gamma prior on \((m, s^2)\). However, unlike the case of 2.2.1, these parameters are only updated for one latent feature at a time. Since the Normal-Gamma distribution has only 4 parameters and the number of states of the latent variable is \(S = 2\), this requires updating \(4SK\) parameters, but only taken \(4S = 8\) at a time, a much more manageable task.

3 Experiments

Code for all algorithms and analyses is available at [https://github.com/pearsonlab/spiketopics](https://github.com/pearsonlab/spiketopics).
References

1. Beal MJ. Variational algorithms for approximate Bayesian inference. University of London; 2003.

2. Yu SZ, Kobayashi H. Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model. Signal Processing, IEEE Transactions on. 2006;54(5):1947–1951.

3. Mitchell CD, Jamieson LH. Modeling Duration in a Hidden Markov Model with the Exponential Family. In: Acoustics, Speech, and Signal Processing; 1993.

4. Mitchell C, Harper M, Jamieson L. On the complexity of explicit duration HMM’s. IEEE Trans Audio Speech Lang Processing. 1995;3(3):213–217.

5. Blei DM, Jordan MI. Variational inference for Dirichlet process mixtures. Bayesian Anal. 2006;1(1):121–143.