Size norm model ideas

Caleb Weinreb and Kai Fox

May 2023

1 Modeling framework

Suppose we have pose data $\{y_t\}_{t=1}^T$ and $\{y'_t\}_{t=1}^{T'}$ for a pair of animals, where $y_t, y'_t \in \mathbb{R}^{KD}$, represent the positions of K keypoints in D dimensions. For now, we will assume that the keypoints are in egocentric coordinates, meaning each animal is always centered and pointing in the same direction. Let's also assume the data have been rescaled to account for gross size differences, so that all remaining differences are subtleties of body shape. Our goal is to define a canonical (low-dimensional) pose space and learn how the data from each animal map onto it. Put in terms of generative modeling, we wish to explain each animal's observed pose y_t as the (noisy) realization of some latent pose state $x_t \in \mathbb{R}^M$, where the space of latent states is shared between animals. This can be formalized as follows:

$$y_t = F(x_t) + \xi \text{ where } \xi \sim \mathcal{N}(0, R) \text{ and } x_t \sim P_x \tag{1}$$

$$y'_t = F'(x'_t) + \xi' \text{ where } \xi' \sim \mathcal{N}(0, R') \text{ and } x'_t \sim P'_x \tag{2}$$

where F, F' are respective functions mapping from the latent space to each animal's pose space, and P_x, P'_x are distributions over latent states. Ideally, we want F and F' to capture morphological differences between the two animals, and P_x, P'_x to capture differences in the frequency of behaviors. We also want to make sure that F and F' are as similar as possible, i.e. to avoid arbitrary rotations of the latent space for one animal compared to the other. Here are some ideas for how to achieve these goals:

- **Shared Gaussian pose distribution:** We could start with a simple shared distribution over the latent space $P_x = P'_x = \mathcal{N}(0, I_M)$.
- **Gaussian mixture pose distribution:** Another option is to model P_x and P'_x as Gaussian mixtures, where the mixture components are shared between animals, but the mixture weights are allowed to vary. This would naturally capture differences in the frequency of behaviors between animals.
- **Affine pose mappings:** We probably want to assume that F and F' are affine, i.e. that $F(z) = Az + b$ for some $A \in \mathbb{R}^{KD \times M}$ and $b \in \mathbb{R}^{KD}$.
- **Ensuring similar mappings:** There are a few ways to make sure that F and F' are similar. One is to model their parameters as additive perturbations of a common mapping, e.g. $F = (A + \Delta A, b + \Delta b)$, $F' = (A + \Delta A', b + \Delta b')$ where A, b can vary broadly but $\Delta A, \Delta b, \Delta A', \Delta b'$ have a tight prior. Another option is to perturb the common mapping multiplicatively.

A core simplifying assumption of all of our pose space models P_z will be independence at each time step. In this way our model is a deepening of keypoint-moseq's affine transform from pose space to keypoint space $y \sim \mathcal{N}(Cx + d, R)$, where we vary the transform per-mouse and investigate other transforms, while ignoring all dynamical character of behavior.

1.1 Pared-down Gaussian mixture model

Before deriving EM steps for our full desired context, we begin with a pared-down model for only one animal, where we take $F = \mathrm{Id}$ and $R = \varepsilon I$

$$z_t \sim \mathrm{Cat}(\pi) \qquad x_t \sim \mathcal{N}(m_{z_t}, Q_{z_t}) \qquad y_t \sim \mathcal{N}(x_t, \varepsilon I) \tag{3}$$

A couple of notes:

- Until we migrate to a Bayesian formulation, we do not have parameter priors like those specified in keypoint-moseq or in Section 2.1. The EM updates will be simpler to work with, so we're starting in this prior-less regime. We can then derive updates incorporating priors as in Scott Linderman's Stats 305C lecture notes (github) or Murphy 11.4.2.8 [cite]
- Because of the symmetry between ε and Q_k, will fix ε and treat it as a hyperparameter. Once the morph functions are not identity, we will impose separate priors on uniform keypoint error R and the scale factors of animals, which will break this symmetry. For forward-compatibility, we will work without the assumption $R = \varepsilon I$, however in implementing this model the parameter will be fixed.
- In practice, we will reparameterize the model to avoid constrained optimization by taking the mixture weights π to be the softmax of an unconstrained cluster weight vector $\bar{\pi}$.

Treating z_t, x_t as latent variables, $\theta = (\pi, m_k, Q_k)$ as a parameter vector, and ε as a hyperparameter, this model can be fit using expectation maximization, as described in Section 1.2

1.2 Expectation maximization

In expectation maximization, jointly optimize the log likelihood $\ell(\theta)$ of a set of model parameters θ given observations y. In particular, we do so while taking the expectation over unknown values of some latent variables — z and x in our case — according to an auxiliary distribution $q(x, z|y, \theta^*) \propto P(x, z|y, \theta^*)$ for a current parameter estimate θ^*. The theoretical basis for EM (see Murphy 11.4.7 [cite]) is that taking this expectation produces a lower bound for the likelihood $\ell(\theta)$, namely

$$\ell(\theta) \geq A(\theta, \theta^*) := \mathbb{E}_{q(x,z|y,\theta^*)} \log P(y, x, z|\theta) \tag{4}$$

The expectation maximization algorithm splits the problem of iteratively optimizing ℓ into two blocks: in the E-step we will calculate q using our current parameter estimates θ^*, so that in the M-step we can calculate $\arg\max_\theta A(\theta, \theta^*)$ to arrive at new parameter estimates θ^*_{new},

$$\theta^*_{\text{new}} := \arg\max\nolimits_\theta \mathbb{E}_{q(x,z|y,\theta^*)} \log P(y, x, z|\theta) \tag{5}$$

The key theoretical guarantee is that the new parameter estimates monotonically increase in likelihood, that is $\ell(\theta^*_{\text{new}}) \geq \ell(\theta^*)$.

E step: The core simplifying assumption from Section 1.1 of independence across time allows us to write q as a product distribution of i.i.d. latent variables x, z at each time point, $q_t(x, z) \propto P(x_t = x, z_t = z \mid y_t, \theta^*)$.

Our first move in calculating q_t will be to apply Bayes' rule and drop the denominator, since it does not vary in θ and therefore cannot affect computation of the arg max in Equation 5:

$$P(x_t = x, z_t = z \mid y_t, \theta^*) = \frac{P(y_t \mid x_t = x, z_t = z, \theta^*) P(x_t = x, z_t = z \mid \theta^*)}{P(y_t \mid \theta^*)} \tag{6}$$

$$q_t(x, z) = P(y_t \mid x_t = x, z_t = z, \theta^*) P(x_t = x, z_t = z \mid \theta^*) \tag{7}$$

We will then aim to compute this expression for q_t analytically for each model.

NOTE: We *cannot* drop the normalizing factor $P(y_t \mid \theta^*)$, since it weights terms in the sum over t. This error is fixed in the `SingleMouseGMM` code as of this commit.

M step: For simple models, maximization of A can be achieved analytically as well, but in general to improve model iteration speed we will handle the M-step using gradient ascent methods. The goal of our M-steps in this document will therefore be to calculate functional forms of A that are compatible with numerical optimization packages.

Rewriting Eq. 5 using timepoint independence and expanding the expectation from Eq. 5, we arrive at the formula that will be the basis for deriving each model's M-step.

$$\theta^*_{\text{new}} := \arg\max_\theta \sum_t \sum_z \int q_t(x,z)\,\log P(y_t, x, z|\theta)\,dx \tag{8}$$

The main requirement to enable numerical computation and differentiation of θ^*_{new} will be to evaluate the integral in x.

1.3 EM algorithm for the pared-down mixture model

In this section we present the results necessary to implement EM for the pared down Gaussian mixture model of Section 1.1. For details, see Section 2. The results of the E-step are implicit in the formulation of the objective function, but we will use the following constants that arise during the E-step:

$$\Sigma^*_{z,t} = R^*\left(R^* + Q^*_z\right)^{-1} Q_z \tag{9}$$

$$\mu^*_{z,t} = \Sigma^*_{z,t}\left({R^*}^{-1} y_t + {Q^*_z}^{-1} m_z\right) \tag{10}$$

$$P(y_t|z,\theta^*) = \left[(2\pi)^{-M}\left(|R^*|\,|Q^*_z|\right)^{-1}\left|\Sigma^*_{z,t}\right|\right]^{1/2} \times \tag{11}$$

$$\exp\left\{-\frac{1}{2}\left(y_t^T {R^*}^{-1} y_t \;+\; {m^*_z}^T {Q^*_z}^{-1} m^*_z \;-\; {\mu^*_{z,t}}^T {\Sigma^*_{z,t}}^{-1} \mu^*_{z,t}\right)\right\} \tag{12}$$

which arise as Equations 20, 21, and 24 respectively in the derivation. We then proceed in the M-step to numerically maximize the objective function

$$J(\theta) = -\sum_t \sum_z \frac{\pi^*_z\, P(y_t \mid z)}{2}\Bigg(\log|R| + d^2_M(\mu^*_{z,t}, y_t;\, R) + \mathrm{Tr}\left[\Sigma^*_{z,t} R^{-1}\right] \tag{13}$$

$$+\log|Q_z| + d^2_M(\mu^*_{z,t}, m_z;\, Q_z) + \mathrm{Tr}\left[\Sigma^*_{z,t} Q_z^{-1}\right] \tag{14}$$

$$-\,2\bar{\pi}_k + 2\log\sum e^{\bar{\pi}}\Bigg) \tag{15}$$

whose terms are derived in Equations 33, 34, and 35.

2 Derivation of EM for the pared-down mixture model

In this section we expand upon parameter inference procedure outlined in Section 1.2 for the pared down Gaussian mixture model of 1.3.

E step: In the "expectation" step, we begin from Equation 7 to derive the auxiliary distribution $q_t(x, z)$. The first action will be to remove unnecessary conditional terms and apply the assumed distributions of the model:

$$
\begin{aligned}
q_t(x,z) &= P(y_t \mid x, z, \theta^*) P(x, z \mid \theta^*) && (16)\\
&= P(y_t \mid x, R^*) P(x \mid m^*_z, Q^*_z) P(z \mid \pi^*) && (17)\\
&= \mathcal{N}(y_t \mid x, R^*)\,\mathcal{N}(x \mid m^*_z, Q^*_z)\,\pi^*_z && (18)
\end{aligned}
$$

We now move to write the functional form we will use for our auxiliary distribution q, which should be an unnormalized distribution equal to the numerator of equation 6. In general, the product of normal PDFs is proportional to another normal PDF (we will show why the proportionality constant is $P(y_t|z_t, \theta^*)$ momentarily),

$$
\begin{aligned}
q_t(x,z) &= \pi^*_z\, P(y_t|z,\theta^*)\,\mathcal{N}(x \mid \mu^*_{z,t}, \Sigma^*_{z,t}) && (19)\\
\Sigma^*_{z,t} &= R^*\,(R^* + Q^*_z)^{-1}\,Q_z && (20)\\
\mu^*_{z,t} &= \Sigma^*_{z,t}\left(R^{*-1} y_t + Q^{*-1}_z m_z\right) && (21)
\end{aligned}
$$

To understand where our proportionality constant $P(y_t|z,\theta^*)$ arises from, we can marginalize over x using two different forms of the latent posterior $P(x_t, z_t|y_t, \theta^*)$:

$$
\begin{aligned}
\int_x P(x, z_t \mid y_t, \theta^*) &= P(z_t \mid y_t, \theta^*) = \frac{P(y_t|z_t,\theta^*)P(z_t \mid \theta^*)}{P(y_t|\theta^*)} = \frac{P(y_t|z_t,\theta^*)\pi^*_z}{P(y_t|\theta^*)} && (22)\\
\int_x P(x, z_t \mid y_t, \theta^*) &= \int_x \frac{q_t(x,z_t)}{P(y_t|\theta^*)} = \int_x \frac{K\pi^*_z\,\mathcal{N}(x \mid \mu^*_{z,t}, \Sigma^*_{z,t})}{P(y_t|\theta^*)} = \frac{K\pi^*_z}{P(y_t|\theta^*)} && (23)
\end{aligned}
$$

In computations, it will be useful to have a more explicit form of this proportionality constant,

$$
\begin{aligned}
P(y_t|z,\theta^*) = &\left[(2\pi)^{-M}\,(|R^*|\,|Q^*_z|)^{-1}\,\left|\Sigma^*_{z,t}\right|\right]^{1/2} \times && (24)\\
&\exp\left\{-\frac{1}{2}\left(y_t^T R^{*-1} y_t \;+\; m^{*T}_z Q^{*-1}_z m^*_z \;-\; \mu^*_{z,t}{}^T \Sigma^*_{z,t}{}^{-1} \mu^*_{z,t}\right)\right\} && (25)
\end{aligned}
$$

which we derive in an abstract setting to simplify notation:

Proposition 2.1. *The product normal PDFs evaluated at a point, $N_a = \mathcal{N}(x \mid a, A)$ and $N_b = \mathcal{N}(x \mid b, B)$, is proportional to $N_c = \mathcal{N}(x \mid c, C)$ with and $C = A\,(A+B)^{-1}\,B$ and $c = CA^{-1}a + CB^{-1}b$. Moreover, if $Z_\Sigma = (2\pi)^{-D/2}\,|\Sigma|^{-1/2}$ is the usual Gaussian normalization factor for covariance matrix Σ, then equality is achieved using the following proportionality constant:*

$$
N_a N_b = \frac{Z_a Z_b}{Z_c} \exp\left\{-\frac{1}{2}\left(a^T A^{-1} a \;+\; b^T B^{-1} b \;-\; c^T C^{-1} c\right)\right\} N_c. \tag{26}
$$

Proof. The proportionality result is standard, so we leave that proof to the reader and use the result to derive our proportionality constant. Let E_i be the exponent in the normal PDF N_i, namely $E_a = (x-a)^T A^{-1}(x-a)$. Then the exponents of $N_a N_b$, $E_a + E_b$, and the exponent of N_c, E_c only differ in those terms which are constant in x, i.e.,

$$
(E_a + E_b) - E_c = a^T A^{-1} a + b^T B^{-1} b + c^T C^{-1} c \tag{27}
$$

Finally using the Gaussian PDF normalization constant $Z_\Sigma = (2\pi)^{-D/2}\,|\Sigma|^{-1/2}$, we can write the constant K such that $N_a N_b = K N_c$ as $\frac{Z_a Z_b}{Z_c} \exp\left\{-\frac{1}{2}\left(E_a + E_b - E_c\right)\right\}$. □

M step: In our "maximization" step, we make $A(\theta, \theta^*)$ numerically computable to enable gradient ascent optimization. In particular, we must evaluate the integral w.r.t. x that appears in Equation 8. These derivations closely follow the standard ones for expectation maximization of a Gaussian mixture model. Expanding the joint probability in equation 8, we arrive at three terms that to integrate against q_t:

$$\arg\max_\theta \sum_t \sum_z \int_x q_t(x, z) \left[\log P(y_t \mid x, R) + \log P(x \mid m_z, Q_z) + \log P(z \mid \pi)\right] \tag{28}$$

Note that we may optimize many of the parameters separately, since each $\log P$ is constant in all but a few of them. In particular, we may find an analytical optimum for π, which requires constrained optimization, and then continue with our numerical optimization procedure for other variables, meaning that we do not need to evaluate the integral on the third term.

Integration against $\log P(y_t \mid x, R)$. We aim to evaluate $\int_x q_t(x, z) \log P(y_t \mid x, R)$ for a given z, t. By applying the model assumption that y_t is normally distributed with parameters x, R, the log probability may be expanded as:

$$\begin{aligned}\pi_z^*\, P(y_t|z, \theta^*) \int_x \mathcal{N}(x \mid \mu_{z,t}^*, \Sigma_{z,t}^*) \Big[&- \frac{D}{2} \log(2\pi) && (29)\\ &- \frac{1}{2} \log |R| && (30)\\ &- \frac{1}{2}(y_t - x)^T R^{-1} (y_t - x) \Big] && (31)\end{aligned}$$

The first term (Eq. 29) is constant in θ and therefore may be dropped. The second (Eq. 30) is constant in x, so integration against a normal PDF in x is identity. For the third term (Eq. 31), we apply a general formula for integration of a quadratic form against a normal PDF:

Proposition 2.2. *The expectation of a quadratic form in a normal variable is the sum of a Mahalanobis distance and a trace:*

$$\mathbb{E}_{x \sim \mathcal{N}(a,B)} \left[(x - c)^T D^{-1} (x - c)\right] = d_M^2(a, c;\, D^{-1}) + \operatorname{Tr}\left[B D^{-1}\right] \tag{32}$$

The proof is left to the reader, but is achieved by wrapping the whole expectation in a trace and cycling its arguments to arrive at an outer product $(x - c)(x - c)^T$. Combining the observations above, we arrive at

$$\int_x q_t(x, z) \log P(y_t \mid x, R) = -\frac{1}{2} \pi_z^*\, P(y_t|z, \theta^*) \left(\log |R| + d_M^2(\mu_{z,t}^*, y_t;\, R) + \operatorname{Tr}\left[\Sigma_{z,t}^* R^{-1}\right]\right) \tag{33}$$

Integration against $\log P(x \mid m_z, Q_z)$. Next we calculate $\int_x q_t(x, z) \log P(x \mid m_z, Q_z)$ for a given z, t. The procedure is the same as for $P(y_t \mid x, R)$ but for a normal with parameters m_z, Q_z, and results in:

$$\int_x q_t(x, z) P(x \mid m_z, Q_z) = -\frac{1}{2} \pi_z^*\, P(y_t|z, \theta^*) \left(\log |Q_z| + d_M^2(\mu_{z,t}^*, m_z;\, Q_z) + \operatorname{Tr}\left[\Sigma_{z,t}^* Q_z^{-1}\right]\right) \tag{34}$$

Integration against $\log P(z \mid \pi$. Finally, we calculate $\int_x q_t(x, z) \log P(z \mid \pi)$ for a given z, t. Because the log probability does not depend on x, the integral marginalizes the normal PDF in $q_t(x, x)$, and we arrive at $\pi^*_z\, P(y_t|z,\theta^*)\, \log \pi_k$. Substituting then the logits vector $\bar{\pi}$ results in the form amenable to unconstrained optimization:

$$\int_x q_t(x,z) \log P(z \mid \pi) = \pi^*_z\, P(y_t|z,\theta^*) \left(\bar{\pi}_k - \log \sum_i e^{\bar{\pi}_i} \right) \tag{35}$$

2.1 Zero-noise limit in a single mixture component

NOTE: This section uses a $q_t(x,z)$ that includes the normalizing factor $P(y_t \mid \theta^*)$, combined with conditional terms from the application of Bayes' rule to result the normalizer $P(z \mid y_t, \theta^*)$. This normalizer will be further detailed in a future commit.

For debugging, we seek a simple and highly interpretable case, to which end we explore the $\varepsilon \to 0$ case with $N = 1$. Here,

$$\Sigma^*_{z,t} = R^* \left(R^* + Q^*_z\right)^{-1} Q_z \to \mathbf{0} \tag{36}$$

$$\mu^*_{z,t} = \Sigma^*_{z,t} \left({R^*}^{-1} y_t + {Q^*_z}^{-1} m_z \right) \tag{37}$$

$$= Q^*_z (R^* + Q^*_z)^{-1} y_t + R^*_z (R^* + Q^*_z)^{-1} m^*_z \to y_t \tag{38}$$

$$q_t(x,z) = P(z \mid y_t, \theta^*)\, \mathcal{N}(x;\, \mu^*_{z,t}, \Sigma^*_z) \to \pi^*_z P(z \mid y_t, \theta^*)\, \delta_{y_t}(x) \tag{39}$$

We are then able to recapitulate the objective function 15 up to the $P(y_t \mid x, R)$ term, which we drop here for ease since it is both infinite and fixed in the optimized variables, and $\log 2\pi$ which is constant. Note that $\mathrm{Tr}\left[\Sigma^*_z Q_z^{-1}\right]$ does not appear since $\Sigma^*_z \to \mathbf{0}$.

$$\arg\max_\theta A(\theta, \theta^*) = \sum_{z,t} \int_x q_t(x,z) \log P(y_t, x, z|\theta) \tag{40}$$

$$= \sum_{z,t} \int_x \delta_{y_t}(x) P(z \mid y_t, \theta^*) \big[\log P(y_t \mid x, R) + \log P(x \mid m_z, Q_z) + \log P(z \mid \theta) \big] \tag{41}$$

$$= \sum_{z,t} P(z \mid y_t, \theta^*) \left[\log \mathcal{N}(y_t \mid m_z, Q_z) + \log \pi_z \right] \tag{42}$$

$$= \sum_{z,t} P(z \mid y_t, \theta^*) \left[-\frac{1}{2} \log |Q_z| - \frac{1}{2} d^2_M(y_t, m_z;\, Q_z) + \log \pi_z \right] \tag{43}$$

3 Optimizing the mixture of linear Gaussians

We assume the following generative model:

$$z^i_t \sim \mathrm{Cat}(\pi^i) \qquad\qquad \pi^i \sim \mathrm{Dir}(\alpha) \tag{44}$$

$$x^i_t \sim \mathcal{N}(m_{z^i_t}, Q_{z^i_t}) \qquad\qquad m_n, Q_n \sim \text{Normal-Inverse-Wishart} \tag{45}$$

$$y^i_t \sim \mathcal{N}(F^i x^i_t + d^i, I_{KD}) \qquad\qquad d^i, F^i \sim \text{Matrix-Normal} \tag{46}$$

where i indexes animals, t indexes frames, $x^i_t, y^i_t \in \mathbb{R}^{KD}$, $z^i_t \in \{1, ..., N\}$, and $\sigma^2 \in \mathbb{R}^+$ is fixed. The parameters can be optimized using variation mean field EM, in which the posterior over latent variables is approximated by the product of factors $q(z^i_t), q(x^i_t)$. During the M-step, we find the parameters $\theta = (m, Q, d, F, \pi)$ that maximize $\mathbb{E}_{q(x,z)} \log P(y, x, z|\theta)$, where $q(x,z) = \prod_{t,i} q(z^i_t) q(x^i_t)$.

During the E-step, we iteratively update the factors $q(z_t^i), q(x_t^i)$ using coordinate ascent, with one or more passes through the data per iteration. The updates are given by:

$$\begin{aligned}
\log q^*(z_t^i) &= \mathbb{E}_{q(x_t^i)} \log P(y_t^i, x_t^i, z_t^i | \theta) + \text{const.} \qquad (47)\\
&= \int_{x_t^i} \mathcal{N}(x_t^i \mid \mu, \Sigma) \log \mathcal{N}(x_t^i \mid m_{z_t^i}, Q_{z_t^i}) + \log \pi^i_{z_t^i} + \text{const.} \qquad (48)\\
&= \mathcal{N}(\mu \mid m^i_{z_t}, Q^i_{z_t}) - \mathrm{tr}[Q^{-1}_{z_t^i} \Sigma] + \log \pi^i_{z_t^i} + \text{const.} \qquad (49)
\end{aligned}$$