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 $F = \text{Id}$ and $R = \varepsilon I$

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

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 5.1. 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, given a prior $P(\theta)$,

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

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},

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

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.

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

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:

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

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

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

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

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

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

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

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

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:

$$q_t(x,z) = P(y_t \mid x,z,\theta^*)P(x,z \mid \theta^*) \tag{16}$$
$$= P(y_t \mid x,R^*)P(x \mid m_z^*,Q_z^*)P(z \mid \pi^*) \tag{17}$$
$$= \mathcal{N}(y_t \mid x,R^*)\,\mathcal{N}(x \mid m_z^*,Q_z^*)\,\pi_z^* \tag{18}$$

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),

$$q_t(x,z) = \pi_z^*\,P(y_t|z,\theta^*)\,\mathcal{N}(x \mid \mu_{z,t}^*,\Sigma_{z,t}^*) \tag{19}$$
$$\Sigma_{z,t}^* = R^*\left(R^*+Q_z^*\right)^{-1}Q_z \tag{20}$$
$$\mu_{z,t}^* = \Sigma_{z,t}^*\left(R^{*-1}y_t + Q_z^{*-1}m_z\right) \tag{21}$$

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 $P(x_t,z_t|y_t,\theta^*)$:

$$\int_x P(x,z_t \mid y_t,\theta^*) = P(z_t \mid y_t,\theta^*) = \frac{P(y_t|z_t,\theta^*)P(z_t \mid \theta^*)}{P(y_t|\theta^*)} = \frac{P(y_t|z_t,\theta^*)\pi_z^*}{P(y_t|\theta^*)} \tag{22}$$
$$\int_x P(x,z_t \mid y_t,\theta^*) = \int_x \frac{q_t(x,z_t)}{P(y_t|\theta^*)} = \int_x \frac{K\pi_z^*\,\mathcal{N}(x \mid \mu_{z,t}^*,\Sigma_{z,t}^*)}{P(y_t|\theta^*)} = \frac{K\pi_z^*}{P(y_t|\theta^*)} \tag{23}$$

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

$$P(y_t|z,\theta^*) = \left[(2\pi)^{-M}\left(|R^*|\,|Q_z^*|\right)^{-1}\left|\Sigma_{z,t}^*\right|\right]^{1/2} \times \tag{24}$$
$$\exp\left\{-\frac{1}{2}\left(y_t^T R^{*-1}y_t \;+\; m_z^{*T}Q_z^{*-1}m_z^* \;-\; \mu_{z,t}^{*\,T}\Sigma_{z,t}^{*\,-1}\mu_{z,t}^*\right)\right\} \tag{25}$$

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 $C = A\left(A+B\right)^{-1}B$ and $c = CA^{-1}a + CB^{-1}b$. Moreover, if $Z_\Sigma = (2\pi)^{-D/2}\left|\Sigma\right|^{-1/2}$ is the usual Gaussian normalization factor for covariance matrix Σ, then equality is achieved using the following proportionality constant:*

$$N_aN_b = \frac{Z_aZ_b}{Z_c}\exp\left\{-\frac{1}{2}\left(a^TA^{-1}a \;+\; b^TB^{-1}b \;-\; c^TC^{-1}c\right)\right\}N_c. \tag{26}$$

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 $E_a = (x-a)^TA^{-1}(x-a)$. Then the exponents of N_aN_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.,

$$(E_a+E_b) - E_c = a^TA^{-1}a + b^TB^{-1}b + c^TC^{-1}c \tag{27}$$

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

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:

$$\pi_z^* P(y_t|z, \theta^*) \int_x \mathcal{N}(x \mid \mu_{z,t}^*, \Sigma_{z,t}^*) \Big[- \frac{D}{2} \log(2\pi) \tag{29}$$

$$- \frac{1}{2} \log |R| \tag{30}$$

$$- \frac{1}{2} (y_t - x)^T R^{-1} (y_t - x) \Big] \tag{31}$$

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:*

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

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 $\Sigma_z^* \to \mathbf{0}$.

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

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

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

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

3 Mix and match fitting framework

To enable modular combination of pose space distributions and morphs, we derive EM in a more general setting.

Pose space models will in general support a discrete state z whose distribution is allowed to vary by subject in terms of parameters π_n, and a continuous state distributed conditionally on z in terms of parameters ψ. Morph models will support a subject-wise affine transformation from pose space to keypoint space specified by parameters ϕ_n, which generate observed data up to measurement noise with covariance R. The total parameter set available to the pose space model $\theta = (\pi, \psi, R)$, and θ, ϕ together parameterize the full model.

$$z_n^t \sim G_{\pi_n} \qquad\qquad x_n^t \sim F_{\psi|z_n^t} \qquad\qquad y_n^t \sim \mathcal{N}(C_n(\phi) x_n^t + d_n(\phi), R) \tag{44}$$

Each morph model must simply compute the affine transformation specified by C_n, d_n, while most of the heavily lifting is left to the pose space models. The full models therefore function on an

extended parameter set $\theta_C = (\theta, C(\phi), d(\phi))$. When parameters for only one subject are needed, we will denote this as $\theta_{C,n} = (\pi_n, \psi, R, C_n(\phi), d_n(\phi))$. In addition, we consider a log-prior defined separately for parameters of the pose and morph models $\log P(\theta) = L_{\text{pose}}(\pi, \phi, R) + L_{\text{morph}}(\phi)$.

Each pose space model is responsible for computing the objective function that the M-step seeks to maximize, given the collection of linear transformations output by the morph model, in terms of an auxiliary distribution over the latents, $q_t(x, z)$:

$$J(\theta; C(\phi)) = \sum_{t,n} \sum_{z} \int_x q_t(x, z) \log P(y_n^t, x, z \mid \theta_C)\, dx \tag{45}$$

The theoretical guarantees of EM are achieved by taking the latent distribution to be the likelihood of latents given current parameter values θ^*, ϕ^*, and is usually calculated using Bayes' rule:

$$q_t(x, z) = P(x, z \mid y, \theta_C^*) = \frac{P(y, x \mid z, \theta^*, \phi^*) P(z \mid \theta^*)}{P(y \mid \theta_C^*)} \tag{46}$$

The main challenge in computing this objective function is the integral over x, which must be handled analytically. We begin by calculating the joint probability $P(y, x \mid z, \theta_{C,n}^*)$ for a particular subject and timepoint in terms of the linear transform provided by the morph model and the continuous pose state distribution F:

$$P(y, x \mid z, \theta_{C,n}^*) = P(y \mid x, \theta_{C,n}) P(x \mid z, \theta^*) \tag{47}$$

$$= \mathcal{N}(y \mid C_n(\phi^*)\, x + d_n(\phi^*), R^*) F_{\psi^* \mid z}(x) \tag{48}$$

Our central task for each pose space model will be to rewrite this probability as a constant in x times a probability distribution $s(x)$ for which expectations of log probability terms arising from $\log P(y_n^t, x, z \mid \theta_C)$ are known,

$$P(y, x \mid z, \theta_{C,n}^*) = K_{y,z,\theta_{C,n}^*} \cdot s(x;\, y, z, \theta_{C,n}^*) \tag{49}$$

We further simplify the job of the pose space model by recognizing that $P(y \mid \theta^*)$ in the objective function is obtained for free via the discrete state probabilities $P(z \mid \theta^*)$ and the constant $K_{y,z,\theta_{C,n}^*}$. Marginalizing q_t over x reveals that the constants K are themselves probabilities:

$$P(y \mid z, \theta_{C,n}^*) = \int_x P(y, x \mid z, \theta_{C,n}^*) = \int_x K_{y,z,\theta_{C,n}^*} \cdot s(x;\, y, z, \theta_{C,n}^*) = K_{y,z,\theta_{C,n}^*} \tag{50}$$

Therefore the probability of an observation under estimated parameters $P(y \mid \theta_{C,n})$ may be written in terms of quantities either readily available or already calculated by the pose space model:

$$P(y \mid \theta_{C,n}^*) = \sum_z P(y \mid z, \theta_{C,n}^*) P(z \mid \theta_n^*) = \sum_z K_{y,z,\theta_{C,n}^*} G_{\pi_n}(z) \tag{51}$$

Putting together equations 45, 46, 49, and 51, we may write the objective function as it is to be calculated by the pose model:

$$J(\theta; C(\phi)) = \sum_{t,n} \sum_{z} \frac{K_{y_n^t,z,\theta_{C,n}^*} G_{\pi_n}(z)}{\sum_{z'} K_{y_n^t,z',\theta_{C,n}^*} G_{\pi_n}(z')} \int_x s(x;\, y_n^t, z, \theta_{C,n}^*) \log P(y_n^t, x, z \mid \theta_C)\, dx \tag{52}$$

To evaluate the requirement that s be a distribution analytically integrable against log probability terms appearing in the objective function requires an enumeration of such terms. We may expand

the log probability to arrive at a set of terms, all of which will arise from the pose space model

$$\log P(y_n^t, x, z \mid \theta_C) = \log P(y_n^t \mid x, \theta_C) + \log P(x \mid z, \theta) + \log P(z \mid \theta) \tag{53}$$

$$= \log \mathcal{N}(y_n^t \mid C_n(\phi)\, x + d_n(\phi), R) + \log F_{\psi|z}(x) + \log G_{\pi_n}(z) \tag{54}$$

The final term does not vary in x and therefore does not need to be integrated against s. This requirement thus amounts to integrability of a normal and F log probabilities against s. The normal term will always need to be rewritten as PDF in x

$$\log \mathcal{N}(y \mid C_n\, x + d_n, R) = \log \left(\frac{Z_R}{Z_{C_n^{-1} R C_n^{T\,-1}}} \mathcal{N}\left(x \mid C_n^{-1}\,(y - d_n)\,,\; C_n^{-1} R\, C_n^{T\,-1}\right)\right) \tag{55}$$

$$= \log Z_R - \frac{1}{2} \left|\left|x - C_n^{-1}(y - d_n)\right|\right|^2_{C_n^{-1} R\, C_n^{T\,-1}} \tag{56}$$

where the dependence of C_n on ϕ is dropped for readability. The result is an unnormalized Gaussian log probability to be integrated against $s(x)$ and a normalizer term that may be computed independent of the morph model. We now prove the statement in which we inverted the affine-transformed Gaussian PDF:

Proof. For vectors $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^N$ and matrices $A \in \mathbb{R}^{N \times M}, B \in \mathbb{R}^{N \times N}$, we may expand a normal PDF as a normalizer times a squared Mahalanobis distance:

$$\mathcal{N}(y \mid Ax, B) = Z_B \exp\left\{-\frac{1}{2} ||y - Ax||^2_B\right\} \tag{57}$$

$$= Z_B \exp\left\{-\frac{1}{2}(y - Ax)^T B^{-1}(y - Ax)\right\} \tag{58}$$

$$= Z_B \exp\left\{-\frac{1}{2}(A^{-1}y - x)^T A^T B^{-1} A(A^{-1}y - x)\right\} \tag{59}$$

$$= Z_B \exp\left\{-\frac{1}{2} \left|\left|A^{-1}y - x\right|\right|^2_{A^{-1} B A^{-1\,T}}\right\} \tag{60}$$

$$= \frac{Z_B}{Z_{A^{-1} B A^{-1\,T}}} \mathcal{N}\left(x \mid A^{-1}y,\; A^{-1} B A^{-1\,T}\right) \tag{61}$$

Additionally, for x, y, z in $\mathbb{R}^N$ and $B \in \mathbb{R}^{N \times N}$, $\mathcal{N}(x - z \mid y, B) = \mathcal{N}(x \mid y + z, B)$. Applying these two identities yields Eq. 56. □

4 Morph models

4.1 Scalar morph

Before introducing spatially nonlinear morphs, we begin by working with a simple uniform scale around an affine offset. Letting the morph model be parameterized by scalars α_n and offsets $\mu_n \in \mathbb{R}^N$, we take $C_n(\phi) = e^{\alpha_n} I_{KD}$, and $d_n(\phi) = \mu_n$. Using this scalar morph the latent pose space dimension cannot compress keypoint space, so $M = KD$.

Priors As will be standard for most morph models, we define a prior to remove ambiguity both in the offsets μ_n and the scale parameters α_n. In particular, if $\bar{\mu}(\phi)$, $\bar{\alpha}(\phi)$ are the average offset $\mathbb{E}_n\left[\mu_n(\phi)\right]$ and log-scale, respectively, then the prior on ϕ for the scalar morph is

$$L_{\text{morph}}(\phi) = \log \mathcal{N}\left(\bar{\mu}(\phi) \mid 0, I_{KD}\right) + \log \mathcal{N}\left(\bar{\alpha}(\phi) \mid 0, I_1\right) \tag{62}$$

$$= \frac{1}{2} ||\bar{\mu}(\phi)||_2^2 + \frac{1}{2}\bar{\alpha}(\phi)^2 \tag{63}$$

Initialization We initialize offsets to the subject-wise mean in keypoint space $\mu_n = \mathbb{E}_t[y_n^t]$ and log-scale factors to the MLE of standard deviation for a spherical Gaussian centered at μ_n given the subject's keypoint data, $\alpha_n = \frac{1}{2}\log \mathbb{E}_t[||y_n^t - \mu_n||_2^2]$

4.2 Affine modal rotation

The first (dare I say ...only?) non-rigid morph we will consider is an affine linear transform which rotates a limited number of dimensions around a centroid. We now define this morph model and derive the required terms to implement it in the mix-and-match fitting framework.

Let L be an integer d.o.f. hyperparameter giving the number of axes to be rotated. In this case, we define the morph model in terms of the following parameters

- $\bar{\alpha} \in \mathbb{R}$ - log of the subject-wise uniform scale factor $\alpha = e^{\bar{\alpha}}$ to be applied to all dimensions of pose space
- $U \in \mathbb{R}^{KD \times L}$ - matrix whose columns give the population-wide dimensions to be morphed. *The "PC modal rotation" morph is the special case where the columns of U are the first L principal components and considered a hyperparameter instead of being learned.*
- $\hat{U}_n \in \mathbb{R}^{KD \times L}$ - matrix whose columns, $\hat{u_{n,l}}$, give subject-wise updates to the morph dimensions.
- $\mu_n \in \mathbb{R}^{KD}$ - center of the affine transform, so that $d_n(\phi) = \mu_n$.

Let $\bar{U}$ be the projection on to the orthogonal complement of U, which will be used as a pass-through for those dimensions of pose space that should be effected only by uniform scaling. Using these parameters, the modal rotation morph is the affine map given by the matrix $C_n(\phi)$ and the vector $d_n(\phi)$:

$$C_n(\phi) = \alpha\left[\bar{U} + \left(U + \hat{U}_n\right)U^+\right] \qquad \text{with } \phi = (U, \bar{\alpha}_n, \hat{U}_n, \mu_n,\ n \in 1...N) \tag{64}$$

$$d(\phi_n) = \mu_n \qquad \alpha_n = \exp\left(\bar{\alpha}_n\right) \tag{65}$$

$$\hat{u}_{n,l} = (\hat{U}_n)_{:,l} \sim \mathcal{N}(0, v^2 I_{KD}) \tag{66}$$

To understand the action of the morph, consider the $L = 2$ case in which the columns of U, which we call u_1 and u_2, are orthonormal. Let $\hat{u}_1, \hat{u}_2$ be the updates to the morph dimensions for subject n. Then the action of $C(\phi_n)$ on $\text{span}(u_1, u_2)$ - the subspace spanned by the morph dimensions - is a perturbation of the symmetric matrix with SVD $U(\alpha I_{KD})U^T$, namely

$$C(\phi_n) : x \in \text{span(U)} \mapsto \begin{bmatrix} | & | \\ u_1 + \hat{u}_1 & u_2 + \hat{u}_2 \\ | & | \end{bmatrix} \begin{bmatrix} \alpha & \\ & \alpha \end{bmatrix} \begin{bmatrix} - & u_1^T & - \\ - & u_2^T & - \end{bmatrix} x \tag{67}$$

The action of $C(\phi_n)$ on the orthogonal complement of U should then be the identity. Removing the requirement of orthogonal morph dimensions, and considering vectors outside the span of those morph dimensions, we have the general expression:

$$C(\phi_n) = \alpha\left[\bar{U} + \left(U + \hat{U}_n\right)\text{Diag}\left(s_n\right)U^+\right] \tag{68}$$

with U^+ being the Moore-Penrose pseudoinverse of U.

Priors We inherit the normal priors on the average offset and log-scale to remove ambiguity in μ_n and α_n as in Eq. 63. In addition, we include a normal prior on the mode adjustments, the

columns of the matrix $\hat{U}_n$, which we call $u_{n,l}$, according to a scale hyperparameter υ

$$L_{\text{morph}}(\phi) = \log \mathcal{N}\left(\bar{\mu}(\phi) \mid 0, I_{KD}\right) + \log \mathcal{N}\left(\bar{\alpha}(\phi) \mid 0, I_1\right) + \sum_{n,l=1}^{N,L} \log \mathcal{N}(u_{n,l}, 0, \upsilon^2 I_{KD}) \tag{69}$$

$$= \frac{1}{2}\left|\left|\bar{\mu}(\phi)\right|\right|_2^2 + \frac{1}{2}\bar{\alpha}(\phi)^2 + \frac{1}{2\upsilon^{2KD}} \sum_{n,l=1}^{N,L} ||u_{n,l}||_2^2 \tag{70}$$

5 Pose space models

5.1 Gaussian mixture

In the Gaussian mixture model, our discrete pose state is categorical with weights subject to a heirarchical Dirichlet prior with hyperparameters $\hat{\gamma}, \hat{\beta}$, and our continuous pose state is normal, with mean and covariance conditional on the discrete state.

$$\begin{aligned}
G_{\pi_n} = \text{Cat}(\beta_n) \qquad & \text{with } \pi_n = ((\hat{\beta}_n, \hat{\gamma}_n) \in \mathbb{R}^L \times \mathbb{R}^L,\ n \in 1...N) \\
& \gamma_n = \text{softmax}(\hat{\gamma}_n),\ \gamma_n \sim \text{Dir}(\hat{\gamma}) \\
& \beta_n = \text{softmax}(\hat{\beta}_n),\ \beta_n \sim \text{Dir}(\hat{\beta}\,\gamma_n) \\
F_{\psi|z} = \mathcal{N}(m_z, Q_z) \qquad & \text{with } \psi = ((m_z, Q_z) \in \mathbb{R}^M \times \mathbb{R}^{M\times M},\ z \in 1...L)
\end{aligned}$$

Our task is to rewrite the product of F and a normal PDF as in Eq. 48, as a constant K times a distribution s, as in Eq. 49. Since F is normal here, the product is itself proportional to a normal PDF, and the following proposition specifies the proportionality constant.

Proposition 5.1. *The product normal PDFs evaluated at a point, $\mathcal{N}(x \mid a, A)$ and $\mathcal{N}(x \mid b, B)$, is proportional to $\mathcal{N}(x \mid c, C)$ with and $C = A\left(A+B\right)^{-1}B$ and $c = CA^{-1}a + CB^{-1}b$. Moreover, if $Z_\Sigma = (2\pi)^{-D/2}\left|\Sigma\right|^{-1/2}$ is the usual Gaussian normalization factor for covariance matrix Σ, then equality is achieved using the following proportionality constant:*

$$\mathcal{N}(x \mid a, A)\mathcal{N}(x \mid b, B) = \frac{Z_A Z_B}{Z_C} \exp\left\{-\frac{1}{2}\left(a^T A^{-1} a \,+\, b^T B^{-1} b \,-\, c^T C^{-1} c\right)\right\} \mathcal{N}(x \mid c, C). \tag{71}$$

Proof. The proportionality result is standard, so we leave that proof to the reader and use the result to derive our proportionality constant. Let $||x-a||_A^2 = (x-a)A^{-1}(x-a)^T$ be the squared Mahalanobis distance between x and a under covariance A, so that $\mathcal{N}(x \mid a, A) = Z_A \exp(-\frac{1}{2}||x-a||_A^2)$. The normalizing constant may therefore be written

$$\frac{\mathcal{N}(x \mid a, A)\mathcal{N}(x \mid b, B)}{\mathcal{N}(x \mid c, C)} = \frac{Z_A Z_B}{Z_C} \exp\left\{-\frac{1}{2}\left(||x-a||_A^2 + ||x-b||_B^2 - ||x-c||_C^2\right)\right\} \tag{72}$$

Expanding the Mahalanobis distances, we see that the remaining terms are precisely those which to not vary in x (which is natural given the proportionality result) yielding the desired result. □

The terms $K_{y_n^t, z, \theta_{C,n}}$ and $s(x;\, y_n^t, z, \theta^*)$ for the Gaussian mixture pose space model may be calculated using this proposition by transforming the normal term in Eq. 48 as in 56. The means and covariances in the language of the proposition are then $a = C_n^{-1}(y + d_n)$, $A = C_n^{-1} R\, C_n^{T^{-1}}$, $b = m_z$, and $B = Q_z$. The extra two Gaussian normalizer terms introduced in Eq. 56 may be absorbed into K as well, leaving s to be a normal PDF and the normalizer S to be unity.

It remains to be seen that s can be analytically integrated against the requisite log probability terms: the quadratic form from Eq. 56 and $\log F$, which is also a quadratic form here. Do calculate these integrals, we will use the following proposition.

Proposition 5.2. *The expectation of a quadratic form in a normal variable is the sum of a Mahalanobis distance and a trace:*

$$\mathbb{E}_{x\sim\mathcal{N}(a,B)}\left[(x-c)^T D^{-1}(x-c)\right] = ||a-c||_D^2 + \operatorname{Tr}\left[BD^{-1}\right] \tag{73}$$

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 $(x-c)(x-c)^T$. We may now enumerate the log probability terms that are to be calculated by the Gaussian mixture pose space model:

- The expectation in s of the quadratic form from Eq. 56 is given by the proposition, scaled by $-\frac{1}{2}$
- The expectation in s of $\log F$ splits into two terms. First, we have the expectation of the quadratic form $-\frac{1}{2}\,||x-m_z||_{Q_z}^2$, which is again given by the scaled result of the composition. Additionally, there is a term $\log Z_{Q_z} = -\frac{1}{2}\log|Q_z|$ which is constant in x and so is its own expectation.

Priors We impose a heirarchical Dirichlet prior on the component weights π_n, but we do not constrain the parameters of the component distributions m_z, Q_z. Also, to remove ambiguity in the logits $\hat{\beta}_n, \hat{\gamma}_n$ we add normal priors on their means. This results in the following log-prior for the Gaussian mixture pose model:

$$L_{\text{morph}}(\pi_n) = \sum_n \log \operatorname{Dir}(\beta_n \mid \hat{\beta}\ \gamma_n) + \sum_n \log \operatorname{Dir}(\gamma_n \mid \hat{\gamma}) \tag{74}$$

$$+ \log \mathcal{N}\left(\mathbb{E}_n[\beta] \mid 0, I_1\right) + \mathcal{N}\left(\mathbb{E}_n[\gamma] \mid 0, I_1\right) \tag{75}$$

Initialization We initialize the Gaussian mixture pose space model based on a standard Gaussian mixture model (GMM) fit to pose space data from a reference subject $\hat{n}$. Specifically, given parameters for a morph model, ϕ, we form the dataset of pose space points $\{C_{\hat{n}}(\phi)y_n^t + d_{\hat{n}}(\phi)\}_t$ and fit a GMM in L components, yielding means $\hat{m}_z$, covariances $\hat{Q}_z$ and cluster weights $\hat{\pi}_z$. From these we may construct initialization parameters π and ψ as

- Component means and covariances of the pose space model are directly inherited from the reference-subject GMM. That is, $m_z = \hat{m}_z$ and $Q_z = \hat{Q}_z$.
- The heirarchical Dirichlet prior is initialized so that $\beta_{n,z} = \hat{\pi}_z$. We take the component weight logits as $\hat{\beta}_n = \log\hat{\pi}_z$ and take $\hat{\gamma} = \log\hat{\pi}_z$ so that the initial β_n's are the mean of their generating distribution, i.e. $\mathbb{E}[\operatorname{Dir}(\hat{\beta}\gamma)] = \beta_n$.

6 Optimizing the mixture of linear Gaussians

We assume the following generative model:

$$z_t^i \sim \operatorname{Cat}(\pi^i) \qquad \pi^i \sim \operatorname{Dir}(\alpha) \tag{76}$$

$$x_t^i \sim \mathcal{N}(m_{z_t^i}, Q_{z_t^i}) \qquad m_n, Q_n \sim \text{Normal-Inverse-Wishart} \tag{77}$$

$$y_t^i \sim \mathcal{N}(F^i x_t^i + d^i, I_{KD}) \qquad d^i, F^i \sim \text{Matrix-Normal} \tag{78}$$

where i indexes animals, t indexes frames, $x_t^i, y_t^i \in \mathbb{R}^{KD}$, $z_t^i \in \{1, ..., N\}$, 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 $q(z_t^i), q(x_t^i)$. 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 $q(x, z) = \prod_{t,i} q(z_t^i) q(x_t^i)$. 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:

$$\log q^*(z_t^i) = \mathbb{E}_{q(x_t^i)} \log P(y_t^i, x_t^i, z_t^i|\theta) + \text{const.} \tag{79}$$

$$= \int_{x_t^i} \mathcal{N}(x_t^i \mid \mu, \Sigma) \log \mathcal{N}(x_t^i \mid m_{z_t^i}, Q_{z_t^i}) + \log \pi_{z_t^i}^i + \text{const.} \tag{80}$$

$$= \mathcal{N}(\mu \mid m_{z_t}^i, Q_{z_t}^i) - \text{tr}[Q_{z_t^i}^{-1} \Sigma] + \log \pi_{z_t^i}^i + \text{const.} \tag{81}$$