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. Our goal is to define a canonical (low-dimensional) pose space that does not reflect body shape, and learn how to the data from each animal map onto it. Put in terms of generative modeling, we wish to explain each animal n observed pose at time t, y_n^t, as the (noisy) realization of some latent pose state $x_n^t \in \mathbb{R}^M$, where the space of latent states is shared between animals.

$$y_1^t = f_1(x_1^t) + \xi \text{ where } \xi \sim \mathcal{N}(0, R) \text{ and } x_1^t \sim P_1 \tag{1}$$

$$y_2^t = f_2(x_2^t) + \xi \text{ where } \xi \sim \mathcal{N}(0, R) \text{ and } x_2^t \sim P_2 \tag{2}$$

where f_1, f_2 are respective *morph* functions mapping from the latent space to each animal's pose space, and P_1, P_2 are distributions over latent pose states. Ideally, we want f_1 and f_2 to capture morphological differences between the two animals, and P_1, P_2 to capture differences in the frequency of behaviors.

This formulation explicitly puts f and P in competition to explain variation in observed keypoint data as differences subject-wise body morphology or differences in occupancy of pose space. For example — an animal with higher density in observation space where keypoints on the back are far apart could be due to a larger body or due to an upregulation of exploratory behavior such as rearing.

We want to make sure that f_1 and f_2 are as similar as possible, i.e. to avoid mapping similar regions of latent pose space onto completely different bodily realizations for different animals or "collapsing" truly different observed postures onto the same latent poses. However, we also wish to keep P_1, P_2 as similar as possible to avoid trivial explanations that fail to unify variation across animals in the latent space. These desires will be encoded both in structural constraints on f and P such as the number of degrees of freedom in their parameterizations, and in hyperparameters on their parameter priors. Later work will specify a battery of controls to interrogate the correct amount of flexibility to afford f and P respectively for a given dataset.

We take the approach of formulating general results with minimal specifications on f and P, and developing a set of responsibilities that a particular *pose space model* P or *morph model* f must perform. This modular framework allows us to examine the ways that implicit assumptions in the structure of each model cause them to trade off explanation of the observed keypoint distributions. We will however work within a limited set of constraints on the kinds of models we consider.

- *Affine morphs.* We assume that morphs may be written $f(x) = Cx + d$ for a matrix $C \in \mathbb{R}^{KD \times M}$ and a vector $d \in \mathbb{R}^{KD}$. Note that linearity in this sense does not entail simple scaling and shearing in D-dimensional keypoint space, but should rather be thought of similarly to

the linearity of PCA in that certain directions or *modes* in latent pose space will correspond to certain postures in keypoint space. The specification of a different morph for each animal allows us to define the way those modes are realized on each body.

- *Time independence.* A core simplifying assumption of all of our pose space models F_n will be that x^t for a given animal are i.i.d. over timesteps. 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.
- *Phoneme variation.* We will consider pose space models that are common across animals, conditional upon an additional discrete latent variable z specifying pose "phonemes" (*ponemes*, *postu-nemes*?), much as the discrete variable in keypoint-moseq describes dynamical syllables. We allow the distribution of the discrete latent z to vary per-animal, such that the flexibility of the pose space distribution P is in defining the frequency of occupying a set of phonemes that are shared (up to body morphology) across the population.

With these structural assumptions, we may more formalize our generative modeling framework in more detail.

Pose space: A pose space model in our framework is a joint distribution over x and z for a single timepoint (extended i.i.d. to sequences). We parameterize such a model by γ defining per-animal distributions G_n of the discrete phoneme latent z, and parameters ψ defining the distribution F of latent pose states x, conditional on z. In this way a pose space model is given by a tuple of PDFs (F, G_n) that together specify

$$P_n(x, z) = F(x \mid z, \psi)\, G_n(z \mid \gamma) \tag{3}$$

Morph: A morph model in our framework is a per-animal affine map from latent pose space to noisy observations in keypoint space. We parameterize this map by a tuple ϕ which defines a matrix C_n and a vector d_n for subject n. A pose space model is therefore given by a tuple of functions (C_n, d_n) mapping ϕ to $\mathbb{R}^{KD \times M}$ and $\mathbb{R}^{KD}$ respectively, that together specify

$$f_n(x_n^t) = C_n(\phi)\, x_n^t + d_n(\phi) \tag{4}$$

The total parameter set for the combined model, including the observation covariance R, is therefore $\theta = (\gamma, \psi, \theta, R)$. The full model may therefore be written

$$z_n^t \sim G_n(\cdot \mid \gamma) \qquad x_n^t \sim F(\cdot \mid z_n^t, \psi) \qquad y_n^t \sim \mathcal{N}(C_n(\phi)\, x_n^t + d_n(\phi), R) \tag{5}$$

As we will see below, 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. In addition, we consider a log-prior defined separately for parameters of the pose and morph models $\log P(\theta) = L_{\text{pose}}(\gamma, \psi, R) + L_{\text{morph}}(\phi)$.

1.1 Expectation maximization

We fit our models using expectation maximization (EM). That is, we learn model parameters θ that maximize the log likelihood $\ell(\theta) = \log P(y, x, z \mid \theta)$ of all observed keypoints y and latent pose states (x, z) under particular set of model parameters θ. In particular, we do so by taking the expectation of the log likelihood over the unknown values the latent pose states (x, z) according to an auxiliary distribution $q(x, z|y, \theta^*) \propto P(x, z|y, \theta^*)$ defined by 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 $P(\theta)$,

$$\ell(\theta) \geq A(\theta, \theta^*) := \mathbb{E}_{q(x,z|y,\theta^*)} \log P(y, x, z|\theta) + \log P(\theta) \tag{6}$$

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

$$\theta^*_{\text{new}} := \arg\max_\theta \mathbb{E}_{q(x,z|y,\theta^*)} \log P(y, x, z|\theta) + \log P(\theta) \tag{7}$$

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

The core simplifying assumption from Section 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. We write $q_n^t(x, z) = P(x_n^t = x, z_n^t = z \mid y_n^t, \theta^*)$. Rewriting Eq. 7 using this timepoint independence and expanding the expectation as an integral, we arrive at the formula that will be the basis for deriving each model's parameter updates:

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

For simple models, maximization of A can be achieved analytically, 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. In particular, we must handle integrals of the log probability $\log P(y_t, x, z \mid \theta)$ against $q_n^t(x, z)$ analytically. This means also that our E-step will not take the standard form of computing probabilities q_n^t for each value of the latent state (x, z), but instead will involve precomputing values for the integral that are independent of θ.

1.2 Responsibilities of morph and pose space models in EM

The majority of the calculations in both the E- and M-steps can be performed in terms of the affine map (C_n, d_n) and therefore may be posed as responsibilities of the pose space model, given the affine map produced by the morph model as input. The only responsibilities of the morph model are to calculate these matrices and vectors, evaluate priors and initializations for the parameters, and to provide convenience functions such as finding maximum likelihood estimates of latents given observations and parameters.

The requirements of the pose space model on the other hand are to calculate priors, initializations, and convenience functions as the morph model does, but also to evaluate the objective function A of the maximization routine in Eq. 8. We now expand upon the framework for analytically integrating our problematic integral. In doing so we enumerate both the necessary arithmetic results to derive for the M-step of each pose space model, and the computations to be performed in their E-steps.

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

$$q_n^t(x, z) = P(x, z \mid y_n^t, \theta^*) = \frac{P(y_n^t, x \mid z, \theta^*) P(z \mid \theta^*)}{P(y_n^t \mid \theta^*)} \tag{9}$$

For integration in x, this fraction breaks down as a normalizer times $P(y_n^t, x \mid z, \theta^*)$. As shown below, that probability the product the normal PDF around $f(x)$ giving the probability of the noisy observation y_n^t and the probability of the pose state x given by F. It will be the central task of the M-step to rewrite that product in turn as a constant in x times an auxiliary probability distribution s to be integrated against the data likelihood $\log P(y_n^t, x, z \mid \theta)$:

$$P(y_n^t, x \mid z, \theta^*) = P(y_n^t \mid x, \theta^*) P(x \mid z, \theta^*) \tag{10}$$

$$= \mathcal{N}(y_n^t \mid C_n(\phi^*)\, x + d_n(\phi^*), R^*) F(x \mid z, \psi^*) \tag{11}$$

$$:= K_{y_n^t, z, \theta^*} \cdot s(x;\, y_n^t, z, \theta^*) \tag{12}$$

We expand upon the requirement that $s(x)$ be a PDF in which expectations of the log probability terms arising from the data likelihood are known below in Note that for brevity we write s, K as functions of θ, however by Eq. 11 they are functions only of $(\psi^*, R^*, C_n(\phi^*), d_n(\phi^*))$ and therefore may be evaluated by the morph model operating only on the affine transform output by an arbitrary pose space model.

We find that these constants K and expectations in s, together with the finite array of probabilities for the discrete latents $G_n(z \mid \gamma)$, are sufficient to express the objective function $A(\theta; \theta^*)$. The remaining term in the numerator of q_n^t in Eq. 9, $P(z \mid \theta^*)$, is simply $G(z \mid \gamma^*)$. Additionally, marginalizing Eq. 12 over x shows that the constants K are in fact the probabilities $P(y_n^t \mid z, \theta^*)$, so the denominator $P(y_n^t \mid \theta^*)$ of q_n^t in Eq. 9 may be expanded by the law of total probability as $\sum_z K_{y_n^t, z, \theta^*} G(z \mid \gamma^*)$. The objective function as defined in Eq. 6 may therefore be written

$$A(\theta; \theta^*) = \sum_{t,n} \sum_z \frac{K_{y_n^t, z, \theta^*} G(z \mid \gamma^*)}{\sum_{z'} K_{y_n^t, z', \theta^*} G(z' \mid \gamma^*)} \, \mathbb{E}_{s(x;\, y_n^t, z, \theta^*)} \left[\log P(y_n^t, x, z \mid \theta) \right] + \log P(\theta) \tag{13}$$

Using Eq. 13, we may state explicitly the arithmetic work to be done in deriving a pose space model and the computations that a pose space model must perform during an EM iteration in our modeling framework:

Derivation: A pose space model must define (F, G), and rewrite the product of F with a Gaussian probability as a constant K times a PDF $s(x)$, as in Eq. 12. Closed form integrals of s against the two terms outlined in Sec. 1.3 must be derived.

E step: A pose space model must, during the E-step, calculate the constants $K_{y_n^t, z, \theta^*}$ as a function of the estimated parameters θ^* resulting in an $\mathbb{R}^{|y| \times |z|}$ matrix, where $|y|$ is the number of observed samples and $|z|$ is the cardinality of the discrete latent "phoneme" space. The function s will be defined parametrically, and during the E-step a pose space model must also calculate the $\mathbb{R}^{|y| \times |z|}$ matrix of parameters as a function of θ^*.

M step: A pose space model will use the derived closed form expectations as functions of θ and s (parameterized as the results of the E-step) alongside the constants K calculated in the E-step to compute the objective function Eq. 13 during the M-step. General model-agnostic methods will perform gradient-based maximization based on this output.

1.3 Requirements for the auxiliary distribution of a pose model

It is crucial for each pose space model to formulate an auxiliary PDF s for which the expectations in the rewritten objective function (Eq. 13) have a closed form expression. To elaborate on this constraint, we expand the log probability of the data using the structure of our generative model

(Eq. 5):

$$\log P(y_n^t, x, z \mid \theta) = \log P(y_n^t \mid x, \theta) + \log P(x \mid z, \theta) + \log P(z \mid \theta) \tag{14}$$

$$= \log \mathcal{N}(y_n^t \mid C_n(\phi)\, x + d_n(\phi), R) + \log F(x \mid z, \psi) + \log G(z \mid \gamma) \tag{15}$$

Of these, only the first two are functions of x and thereby impose a requirement on $s(x)$ of having a known closed form expectation.

It will in general be necessary when finding the closed form expectations to recognize that the first term may be written as a normal PDF in x by transforming the normal distribution of y around $f(x)$ to a scaled normal distribution of x around $f^{-1}(y)$. Let $||a - b||_A$ is the Mahalanobis distance of a, b according to covariance A. Then, dropping the dependence of C_n, d_n on ϕ for readability,

$$\log \mathcal{N}(y \mid C_n\, x + d_n, R) = \log \left(\frac{Z_R}{Z_{C_n^{-1} R C_n^{T\,-1}}} \mathcal{N}\left(x \mid C_n^{-1}\,(y - d_n)\,,\; C_n^{-1} R\, C_n^{T\,-1}\right)\right) \tag{16}$$

$$= \log Z_R - \frac{1}{2} \left|\left|x - C_n^{-1}(y - d_n)\right|\right|^2_{C_n^{-1} R\, C_n^{T\,-1}} \tag{17}$$

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. For completeness, we now explicate a proof general inversion of an 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:

$$\mathcal{N}(y \mid Ax, B) = Z_B \exp\left\{-\frac{1}{2} ||y - Ax||^2_B\right\} \tag{18}$$

$$= Z_B \exp\left\{-\frac{1}{2}(y - Ax)^T B^{-1} (y - Ax)\right\} \tag{19}$$

$$= Z_B \exp\left\{-\frac{1}{2}(A^{-1}y - x)^T A^T B^{-1} A (A^{-1}y - x)\right\} \tag{20}$$

$$= Z_B \exp\left\{-\frac{1}{2} \left|\left|A^{-1}y - x\right|\right|^2_{A^{-1} B A^{-1\,T}}\right\} \tag{21}$$

$$= \frac{Z_B}{Z_{A^{-1} B A^{-1\,T}}} \mathcal{N}\left(x \mid A^{-1}y,\; A^{-1} B A^{-1^T}\right) \tag{22}$$

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. 17. □

To summarize, during the derivation of a pose space model, expectations in the random variable $x \sim s$ of the following terms must be derived as closed-form expressions. Note that while s is defined in terms of estimated parameters θ^*, the following terms are functions of test parameters θ during gradient based optimization of A.

- The Gaussian log-PDF centered at morphed latent poses $f(x) = C_n x + d_n$ with observation covariance R evaluated at observations y_n^t. This PDF breaks down as in Eq. 17 into a quadratic form and a Gaussian normalizer $\log Z_R$ that is constant in x.
- The log-PDF of the pose space distribution $F(x \mid z, \psi)$ at a latent pose x under given discrete pose state z and test parameters ψ.

2 Morph models

2.1 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 $\mathbb{E}_n[\mu_n(\phi)]$ and log-scale, respectively, then the prior on ϕ for the scalar morph is

$$L_{\text{morph}}(\phi) = \log \mathcal{N}(\bar{\mu}(\phi) \mid 0, I_{KD}) + \log \mathcal{N}(\bar{\alpha}(\phi) \mid 0, I_1) \tag{23}$$

$$= -\frac{1}{2} ||\bar{\mu}(\phi)||_2^2 - \frac{1}{2}\bar{\alpha}(\phi)^2 \tag{24}$$

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, $\alpha_n = \frac{1}{2} \log \mathbb{E}_t[||y_n^t - \mu_n||_2^2]$

2.2 Affine modal alignment

The first (dare I say ...only?) spatially non-rigid morph we will consider is an affine linear transform which aligns a limited number of dimensions around a centroid. We first give the intuition for this model, then define it and specify priors and an initialization routine.

As in the scalar morph model, we make the assumption that all latent poses for a subject are related to their keypoints by the same offset and scaling, i.e. that there is a certain homologous centroid pose that varies per-animal and that each animal is characterized by a scale of variations from that centroid. In addition, we consider that there might be variations from the centroid that differ by subject but should also be considered homologous — for example large mice might turn more sharply to follow the same path, or that the limbs of older mice might take flatter trajectories when walking if their bodies are held lower to the ground. We therefore allow variation of a certain number, L, of posture *modes* to be mapped differently from latent pose space to each animal's body. We may represent a pose as a magnitude of variation from the centroid along each of these dimensions (as is done in PCA) along with an additional $(KD - L)$-dimensional component describing the remaining variation, which we leave unaltered when mapping from pose space to keypoint space.

Let L be an integer d.o.f. hyperparameter giving the number of pose modes to be aligned. 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 modes in pose space that may be morphed differently per-animal. *In the special case where the columns of U are the first L principal components and are considered a hyperparameter instead of being learned, we call this morph a "PC modal alignment".*
- $\hat{U}_n \in \mathbb{R}^{KD \times L}$ - 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 alignment morph is the affine map given by the matrix $C_n(\phi)$ and the vector $d_n(\phi)$:

$$C_n(\phi) = \alpha_n \left[\bar{U} + \left(U + \hat{U}_n\right) U^+\right] \qquad \text{with } \phi = (U, \bar{\alpha}_n, \hat{U}_n, \mu_n,\ n \in 1...N) \tag{25}$$

$$d(\phi_n) = \mu_n \qquad \alpha_n = \exp\left(\bar{\alpha}_n\right) \tag{26}$$

$$\hat{u}_{n,l} = (\hat{U}_n)_{:,l} \sim \mathcal{N}(0, \upsilon^2 I_{KD}) \tag{27}$$

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 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

$$C(\phi_n) : x \in \text{span(U)} \mapsto \begin{bmatrix} | & | \\ u_1 + \hat{u}_1 & u_2 + \hat{u}_2 \\ | & | \end{bmatrix} \begin{bmatrix} \alpha & \\ & \alpha \end{bmatrix} \begin{bmatrix} - & u_1^T & - \\ - & u_2^T & - \end{bmatrix} x \tag{28}$$

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:

$$C(\phi_n) = \alpha \left[\bar{U} + \left(U + \hat{U}_n\right) \text{Diag}\left(s_n\right) U^+\right] \tag{29}$$

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. 24. In addition, we include a Gaussian prior on the mode adjustments, the columns of the matrix $\hat{U}_n$ which we call $u_{n,l}$, after scaling by the norm of the corresponding mode u_l, the l-th column of the matrix U, according to a scale hyperparameter υ. Finally, we remove symmetry in $||u_l||$ and arrive at the following morph parameter log likelihood function:

$$L_{\text{morph}}(\phi) = \log \mathcal{N}\left(\bar{\mu}(\phi) \mid 0, I_{KD}\right) + \log \mathcal{N}\left(\bar{\alpha}(\phi) \mid 0, I_1\right) \tag{30}$$

$$+ \sum_{n,l=1}^{N,L} \log \mathcal{N}(u_{n,l}, 0, \upsilon^2 I_{KD}) + \sum_{l=1}^{L} \log \mathcal{N}(\log ||u_l||, 0, 1) \tag{31}$$

$$= -\frac{1}{2} ||\bar{\mu}(\phi)||_2^2 - \frac{1}{2}\bar{\alpha}(\phi)^2 - \frac{1}{2\upsilon^{2KD}} \sum_{n,l=1}^{N,L} ||u_{n,l}||_2^2 \, ||u_l||_2^{-2} - \frac{1}{2} \sum_{l=0}^{1} \log^2 ||u_l|| \tag{32}$$

3 Pose space models

3.1 Gaussian mixture

In the Gaussian mixture model, our discrete latent "phoneme" state is categorical over L values (a hyperparameter) which we term components as in a standard GMM. The discrete distribution weights subject to a heirarchical Dirichlet prior with hyperparameters for usage uniformity across components, $\hat{\beta}$, component usage uniformity across animals $\hat{\pi}$, and a saturation level for logits $\pi_{\max}$. Our continuous pose state is normal, with mean and covariance conditional on the discrete

state.

$$
\begin{aligned}
G\left(z \mid \gamma\right) &= \mathrm{Cat}(\pi_n) && \text{with } \gamma = (\bar{\beta} \in \mathbb{R}^L, \bar{\pi}_n \in \mathbb{R}^L,\ n \in 1...N) \\
& && \quad \beta = \mathrm{softmax}(\bar{\beta}),\ \beta \sim \mathrm{Dir}(\hat{\beta}) \\
& && \quad \pi_n = \mathrm{softmax}(\pi_{\max}\tanh(\bar{\pi}_n/\pi_{\max})),\ \pi_n \sim \mathrm{Dir}(\hat{\pi}\ \gamma) \\
F(x \mid z, \psi) &= \mathcal{N}(m_z, Q_z) && \text{with } \psi = ((m_z, Q_z) \in \mathbb{R}^M \times \mathbb{R}^{M\times M},\ z \in 1...L)
\end{aligned}
$$

3.1.1 Auxiliary PDF and constants

Our first task is to rewrite the product of F and a normal PDF in y_n^t, from Eq. 11, as a constant K times a PDF $s(x)$. Since F is itself Gaussian in this case, we may apply the same transformation as in 17 to arrive at a product of two Gaussians in x, which is proportional to another normal PDF in x. The following proposition specifies the proportionality constant.

Proposition 3.1. *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:*

$$
\mathcal{N}(x \mid a, A)\mathcal{N}(x \mid b, B) = \frac{Z_A Z_B}{Z_C}\exp\left\{-\frac{1}{2}\left(a^T A^{-1}a\ +\ b^T B^{-1}b\ -\ c^T C^{-1}c\right)\right\}\mathcal{N}(x \mid c, C). \tag{33}
$$

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

$$
\frac{\mathcal{N}(x \mid a, A)\mathcal{N}(x \mid b, B)}{\mathcal{N}(x \mid c, C)} = \frac{Z_A Z_B}{Z_C}\exp\left\{-\frac{1}{2}\left(||x-a||_A^2 + ||x-b||_B^2 - ||x-c||_C^2\right)\right\} \tag{34}
$$

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 Proposition 3.1 by inverting the affine transform of x in Eq. 11 and applying it to y_n^t, as in 17. The mean and covariances of the constituent normals are then given in the language of the proposition by $a = C_n^{-1}(y + d_n)$, $A = C_n^{-1}R\,C_n^{T^{-1}}$, $b = m_z$, and $B = Q_z$, with dependence of C_n, d_n on ϕ^* suppressed for readability, leaving s to be the normal PDF $N(x \mid c, C)$ and K to be the proportionality constant in Eq. 33.

3.1.2 Expectations of log probabilities

It remains to be seen that s can be analytically integrated against the requisite log probability terms: the quadratic form from Eq. 17 and $\log F$, which is also a quadratic form here. To calculate these integrals, we will use the following proposition.

Proposition 3.2. *The expectation of a quadratic form in a normal variable is the sum of a Mahalanobis distance and a trace:*

$$
\mathbb{E}_{x\sim\mathcal{N}(a,B)}\left[(x-c)^T D^{-1}(x-c)\right] = ||a-c||_D^2 + \mathrm{Tr}\left[BD^{-1}\right] \tag{35}
$$

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. 17 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 $-\frac{1}{2}\left|\left|x-m_z\right|\right|^2_{Q_z}$, which is again given by the scaled result of the composition. Additionally, there is a term $\log Z_{Q_z} \propto -\frac{1}{2}\log|Q_z|$ 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. This results in the following log-prior for the Gaussian mixture pose model:

$$L_{\text{morph}}(\gamma) = \sum_n \log \text{Dir}(\pi_n \mid \hat{\pi}\ \beta) + \log \text{Dir}(\beta \mid \hat{\beta}/L) \tag{36}$$

Note that the hyperparameters $\hat{\beta}$, $\hat{\pi}$ define the variation tolerable in the expected weight of each component and the variation tolerable from that mean respectively. In particular, $\mathbb{E}[\beta]$ is the discrete uniform over L choices, with $\text{Var}[\beta_l]$ inversely proportional to $\hat{\beta}+1$. Moreover, $\mathbb{E}[\pi_n] = \beta$ with $\text{Var}[\pi_{n,l}]$ inversely proportional to $\hat{\pi}+1$.

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 $\{C_{\hat{n}}(\phi)y^t_n + d_{\hat{n}}(\phi)\}_t$ 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 follows:

- 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 $\pi_{n,z} = \hat{\pi}_z$ and so that the initial π_n's are the mean of their generating distribution, i.e. we take the component weight logits as $\bar{\pi}_{n,z} = \log \hat{\pi}_z$, and we take $\bar{\beta}_z = \log \hat{\pi}_z$ so that $\mathbb{E}[\text{Dir}(\hat{\pi}\beta)] = \pi_n$.