Appendices to Combining Functional Data Registration and Factor Analysis

Cecilia Earls
Cornell University
and
Giles Hooker
Cornell University

February 3, 2016

APPENDIX A

Below, in detail, are the specifications for the hierarchical Bayesian registration and factor analysis model discussed in this paper. The first section includes the basic model for functional data registration and factor analysis. Section A.2 describes the MCMC sampling scheme for this model.

A.1 Factor Analysis

The initial assumption of this model is that we are interested in registering and possibly grouping functional data, $X_i(t), i = 1, \ldots, N$. The registered functions, $X_i(h_i(t))$, $i = 1 \ldots N$, are assumed to be characterized almost completely by a linear combination of two factors, $f_1(t)$ and $f_2(t)$.

Furthermore, we assume the unregistered functions, $X_i(t)$, $i = 1, \ldots, N$, are observed at time points, $\mathbf{t} = (t_1 \ldots t_p)'$, such that the data distribution of each observation, $\mathbf{X}_i$, of the ith unregistered function evaluated over $\mathbf{t}$ has the following property:

$$\mathbf{X}_i \mid \mathbf{w}_i, z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 \quad \propto \quad \mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2$$

where

$$\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 \;\sim\; N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}). \tag{1}$$

The joint distribution of the observed data and all unknown parameters is proportional to the data distribution and following priors:

$$\begin{aligned}
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j}(t_k - t_{k-1})e^{w_i(t_{k-1})}, \quad i = 1,\ldots,N, \quad j = 1,\ldots,p, \\
\mathbf{w}_i &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\mathbf{\Sigma} + \lambda_w^{-1}\mathbf{P}_w)\mathbb{1}\{t_1 + \sum_{k=2}^{p}(t_k - t_{k-1})e^{w_i(t_{k-1})} = t_p\}, \qquad (2) \\
&\quad i = 1,\ldots,N, \\
\mathbf{\Sigma} &= \mathbf{P}_1 + \mathbf{P}_2, \qquad (3) \\
\mathbf{P}_w &= \mathbf{P}_2, \\
z_{0i} \mid \sigma^2_{z0} &\sim N(0, \sigma^2_{z0}), \quad i = 1,\ldots,(N-1), \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i}, \\
\sigma^2_{z0} &\sim IG(a,b), \\
z_{1i} \mid \sigma^2_{z1} &\sim N(1, \sigma^2_{z1}), \quad i = 1,\ldots,N, \\
\sigma^2_{z1} &\sim IG(a,b), \\
z_{2i} \mid \sigma^2_{z2} &\sim N(0, \sigma^2_{z2}), \quad i = 1,\ldots,N, \\
\sigma^2_{z2} &\sim IG(a,b), \\
\mathbf{f}_1 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f), \\
\mathbf{f}_2 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f), \\
\mathbf{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2, \\
\eta_f &\sim G(c,d) \text{ and} \\
\lambda_f &\sim G(c,d).
\end{aligned}$$

The matrix $\mathbf{\Sigma}$ is a fixed matrix designed to penalize variation in any direction from the corresponding mean of the distribution in which it is utilized. It is composed of two matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$, such that $\mathbf{\Sigma} = \mathbf{P}_1 + \mathbf{P}_2$. The matrix $\mathbf{P}_1$ penalizes variation from the mean in constant and linear directions, and $\mathbf{P}_2$ penalizes variation from the mean in directions of curvature.

The following derivation provides the basis for the particular form of the covariance matrix, $\mathbf{\Sigma}$, utilized in our models. The derivations below are based on a more general discussion of functional penalties found in Ramsay and Silverman [2005].

As an initial step, consider a given function, $Y_i(t)$ as a sum of its linear and non-linear components. Under this assumption, each function, Y_i, $i = 1, \ldots, N$, can be represented as $Y_i = \xi_i(t) + \nu_i(t)$, where $\xi_i(t) = \sum_{k=1}^{2} c_{ik}\xi_{ik}(t)$ is the constant and linear terms of the function Y_i and $\nu_i(t) \in kerB$ where B is the constraint operator such that $BY_i = [Y_i(0), Y_i'(0)]'$.

Let L be the linear operator such that $LY_i(t) = Y_i''(t)$ and
$kerL \cap kerB = \emptyset$.

Then, $|| \ Y_i \ ||^2 = \eta(BY_i)'(BY_i) + \lambda \int (LY_i)^2(t)dt$ defines a penalty on Y_i such that η penalizes variation in constant and linear directions and λ penalizes curvature in Y_i.

Let $\mathbf{L}$ and $\mathbf{B}$ be matrix representations of the operators L and B that define a penalty on the finite approximations to the functions $Y_i, i = 1, \ldots, N$. Then,

$$\eta(BY_i)'(BY_i) + \lambda \int (LY_i)^2(t)dt \approx \eta \mathbf{Y_i'B'BY_i} + \lambda \mathbf{Y_i'L'LY_i} \tag{4}$$

Notice that (4) can be reexpressed as $\mathbf{Y}_i'(\eta \mathbf{B'B}\mathbf{Y}_i + \lambda \mathbf{L'L})\mathbf{Y}_i = \mathbf{Y}_i'(\eta \mathbf{P}_1^- + \lambda \mathbf{P}_2^-)\mathbf{Y}_i$ where here we set $\mathbf{P}_1^- = \mathbf{B'B}$ and $\mathbf{P}_2^- = \mathbf{L'L}$. The expression, $\mathbf{Y}_i'(\eta \mathbf{P}_1^- + \lambda \mathbf{P}_2^-)\mathbf{Y}_i$, takes a form proportional to the power of the exponential term for a multivariate Gaussian probability density function on $\mathbf{Y}_i$. Seen in this context, large values of this term correspond to less likely values of $\mathbf{Y}_i$. It follows that defining the precision matrix of a Gaussian distribution on $\mathbf{Y}_i$ to $\eta \mathbf{P}_1^- + \lambda \mathbf{P}_2^-$ allows us to define two types of penalties on the approximated functions. The first is a penalty on any deviation from the mean function. To impose this penalty, we set $\eta = \lambda$ where $\mathbf{P}_1^-$ penalizes constant and linear deviations and $\mathbf{P}_2^-$ penalizes deviations in all other directions. When we would like to use this type of penalty in our model, we set the precision matrix for the Gaussian distribution proportional to $\mathbf{\Sigma}^{-1} = \mathbf{P}_1^- + \mathbf{P}_2^-$. The other penalty of interest is a penalty on the roughness of $\mathbf{Y}_i$. Here we will again use a precision matrix of the form $\eta \mathbf{P}_1^- + \lambda \mathbf{P}_2^-$. However to use this precision matrix to penalize roughness, λ needs to be set greater than η so that the penalty on function curvature is much larger than the penalty on constant and linear functions. All of the precision matrices used in the multivariate Gaussian priors for this model are intended to include one or both of these types of penalties.

A.2 MCMC Sampling

Using these assumptions, the following full conditional distributions are derived to run a MCMC sampler. Note, this list will not include an exact full conditional for the base functions as their priors are not conjugate. The base functions are sampled via a Metropolis step.

$$
\begin{aligned}
\mathbf{w}_i \mid rest &\propto f(\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}) f(\mathbf{w}_i), \text{ the product of (15) and (16) above} \\
\mathbf{f}_1 \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}_1|rest}, \boldsymbol{\Sigma}_{\mathbf{f}_1|rest}) \\
\boldsymbol{\Sigma}_{\mathbf{f}_1|rest} &= (\sum_{i=1}^{N} z_{1i}^2(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_f^{-1})^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}_1|rest} &= \boldsymbol{\Sigma}_{\mathbf{f}_1|rest}\Big[(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} z_{1i}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1} + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2)\Big)\Big] \\
\mathbf{f}_2 \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}_2|rest}, \boldsymbol{\Sigma}_{\mathbf{f}_2|rest}) \\
\boldsymbol{\Sigma}_{\mathbf{f}_2|rest} &= \Big(\sum_{i=1}^{N} z_{2i}^2\Big(\frac{\gamma_2^2}{\gamma_1+\gamma_2}\Big)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_f^{-1}\Big)^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}_2|rest} &= \boldsymbol{\Sigma}_{\mathbf{f}_2|rest}\Big[\gamma_2\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} z_{2i}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1)\Big)\Big] \\
z_{0i} \mid rest &\sim N(\mu_{z_{0i}|rest}, \sigma^2_{z_{0i}|rest}) \\
\sigma^2_{z_{0i}|rest} &= (\sigma_{z_0}^{-2} + 2 * \mathbf{1}_p'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p)^{-1} \\
\mu_{z_{0i}|rest} &= \sigma^2_{z_{0i}|rest}\Big(\mathbf{X_i}(\mathbf{h_i}) - \mathbf{X_N}(\mathbf{h_N}) + (z_{1N}-z_{1i})\mathbf{f}_1 + \Big(\frac{\gamma_2}{\gamma_1+\gamma_2}\Big)(z_{2N}-z_{2i})\mathbf{f}_2 - \\
&\qquad \sum_{j=1}^{N-1} z_{0j}\mathbb{1}\{j\neq i\}\mathbf{1}_p\Big)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p \\
\sigma^2_{z_0} \mid rest &\sim IG(a + (N-1)/2, b + 1/2\sum_{i=1}^{N-1} z_{0i}^2) \\
z_{1i} \mid rest &\sim N(\mu_{z_{1i}|rest}, \sigma^2_{z_{1i}|rest}) \\
\sigma^2_{z_{1i}|rest} &= (\sigma_{z_1}^{-2} + \mathbf{f}_2'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{f}_2)^{-1} \\
\mu_{z_{1i}|rest} &= \sigma^2_{z_{1i}|rest}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2)\Big)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{f}_1 \\
\sigma^2_{z_1} \mid rest &\sim IG(a + N/2, b + 1/2\sum_{i=1}^{N} z_{1i}^2) \\
z_{2i} \mid rest &\sim N(\mu_{z_{2i}|rest}, \sigma^2_{z_{2i}|rest}) \\
\sigma^2_{z_{2i}|rest} &= (\sigma_{z_2}^{-2} + \mathbf{f}_2'\frac{\gamma_2^2}{\gamma_1+\gamma_2}\boldsymbol{\Sigma}^{-1}\mathbf{f}_2)^{-1} \\
\mu_{z_{2i}|rest} &= \sigma^2_{z_{2i}|rest}\gamma_2\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1)\Big)'\boldsymbol{\Sigma}^{-1}\mathbf{f}_2 \\
\sigma^2_{z_2} \mid rest &\sim IG(a + N/2, b + 1/2\sum_{i=1}^{N} z_{2i}^2) \\
\eta_f \mid rest &\sim G(c+2, d + \frac{1}{2}tr\big((\mathbf{f}_1\mathbf{f}_1' + \mathbf{f}_2\mathbf{f}_2')\mathbf{P}_1^-\big)) \\
\lambda_f \mid rest &\sim G(c+(p-2), d + \frac{1}{2}tr\big((\mathbf{f}_1\mathbf{f}_1' + \mathbf{f}_2\mathbf{f}_2')\mathbf{P}_2^-\big))
\end{aligned}
$$

APPENDIX B

B.1 Adapted Variational Bayes

The variational Bayes procedure described here is based on the variational methods proposed by Omerod and Wand [2010] and Bishop [2006]. Their proposed method optimizes a lower bound of the marginal likelihood which results in finding an approximate joint posterior density that has the smallest Kullback-Leibler (KL) distance, Kullback and Leibler [1951], from the true joint posterior density.

In minimizing the KL distance between the approximate and true posterior distribution, parameters are updated by an optimization method that requires an approximate posterior density that not only factors but factors into components of known parametric forms. Suppose, $q(\boldsymbol{\theta})$ is the approximated posterior joint distribution. Then for some partition of $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\}$, $q(\boldsymbol{\theta}) = \prod_{k=1}^{d} q_k(\boldsymbol{\theta}_k)$, where each distribution q_k is of a known parametric form.

In our model, the Gaussian process priors for the base functions, $w_i(t)$, $i = 1, \ldots, N$, are not conditionally conjugate to the likelihood function. Therefore, the traditional variational Bayes optimization method does not apply directly since $q_k(\mathbf{w}_i)$, $i = 1, \ldots, N$ are not known parametric distributions. Thus, we propose an adapted variational Bayes algorithm.

After initializing all parameters, in each iteration, the adapted variational Bayes algorithm performs two steps. In the first step, the 'likelihood' as a function of the base functions is maximized. For this 'likelihood', all other parameters are fixed at their current values. The second step uses a traditional variational Bayes iterative scheme to update all other parameters. Specifically, assuming $\boldsymbol{\theta}_k = \mathbf{w}_k$, for $k = 1 \ldots N$, so that, $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\}$, the adapted variational Bayes algorithm is as follows:

1. Initialize $\boldsymbol{\theta}$

2. For each iteration, m, and each k, $k = 1, \ldots, N$, update the estimate for $\mathbf{w}_k$ so that $\mathbf{w}_k^{(m)} = \sup_{\mathbf{w}_k} q_k(\mathbf{w}_k \mid \boldsymbol{\theta}_j^{(m-1)}, j = (N+1), \ldots, d)$

3. For each iteration, m, and each k, $k = (N+1), \ldots, d$, update q_k so that $q_k^{(m)} \propto exp[E_{(\boldsymbol{\theta}_{-k})}(log\ f(\boldsymbol{\theta}_k \mid rest)]$, where the expectation is taken with respect to the distributions $q_j^{(m-1)}(\boldsymbol{\theta}_j)$, $j = 1, \ldots, d$, $j \neq k$

4. Repeat steps (2) and (3) until the desired convergence criterion is met

This algorithm is guaranteed to converge. However, convergence is not guaranteed to a global maximum, and in practice it is sometimes necessary to adjust the registration and warping penalties as the functions become registered. An unregistered function that requires a substantial amount of warping can cause convergence to a local maximum due to the small penalty on warping. The flexibility in warping allowed by this small penalty can cause the function to deform rather than register. This can be remedied in two ways. The first option might be to perform a simple initial warping for this function that prevents the optimization from falling into a local mode. The second option is to adjust the registration and warping parameters over time. Initially a stronger warping penalty is employed to prevent function deformation. Then, as the functions register, the warping penalty can be reduced to allow for a more complete registration. When initializing an MCMC sampler, the final penalties on warping and registration from the adapted variational Bayes algorithm should be used. For further information on the convergence properties of the adapted variational Bayes algorithm and an analysis of how well adapted variational Bayes estimates correspond to MCMC estimates, see Earls and Hooker [2016].

Below are the approximate posterior distributions, $q_k(\boldsymbol{\theta}_k)$, $k = (N+1), \ldots, d$, for the adapted variational Bayes estimation procedure for the registration and factor analysis model. Note, the subscripts on the q distributions has been omitted. For a more thorough discussion and illustration of how the optimal q distributions are derived see Goldsmith et. al. [2011].

$$
\begin{aligned}
q(\mathbf{f}_1) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f}_1)}, \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}) \\
q(\mathbf{f}_2) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f}_2)}, \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}) \\
q(z_{0i}) &\sim N(\mu_{q(z_{0i})}, \sigma^2_{q(z_{0i})}) \\
q(\sigma^2_{z_0}) &\sim IG(a_{q(\sigma^2_{z_0})}, b_{q(\sigma^2_{z_0})}) \\
q(z_{1i}) &\sim N(\mu_{q(z_{1i})}, \sigma^2_{q(z_{1i})}) \\
q(\sigma^2_{z_1}) &\sim IG(a_{q(\sigma^2_{z_1})}, b_{q(\sigma^2_{z_1})}) \\
q(z_{2i}) &\sim N(\mu_{q(z_{2i})}, \sigma^2_{q(z_{2i})}) \\
q(\sigma^2_{z_2}) &\sim IG(a_{q(\sigma^2_{z_2})}, b_{q(\sigma^2_{z_2})}) \\
q(\eta_f) &\sim G(c_{q(\eta_f)}, d_{q(\eta_f)}) \\
q(\lambda_f) &\sim G(c_{q(\lambda_f)}, d_{q(\lambda_f)})
\end{aligned}
$$

The approximate joint posterior distribution of all parameters except the base functions is

$$
q(\boldsymbol{\theta}) = \prod_{k=(N+1)}^{d} q_k(\boldsymbol{\theta}_k) = q(\mathbf{f}_1)q(\mathbf{f}_2)q(\sigma^2_{z_0})q(\sigma^2_{z_1})q(\sigma^2_{z_2})q(\eta_f)q(\lambda_f) \prod_{i=1}^{(N-1)} q(z_{0i}) \prod_{i=1}^{N} q(z_{1i})q(z_{2i}) \qquad (5)
$$

As the q densities are all of known distributional forms, updating these densities is equivalent to updating their parameters. For each iteration, the following parameters are updated for the q densities found in (5). Here we have ordered these updates so that a formal convergence criterion to be calculated. Details on this convergence criterion can be found in Appendix B.2. However as an alternative to using the convergence criterion, in practice it may be more practical to instead monitor changes in the unknown parameters estimates from iteration to iteration and terminate the algorithm when these changes are below a certain threshold.

$$\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} = \Big[\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})}\mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})}\mathbf{P}_2^-\Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f}_1)} = \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\Big[\sum_{i=1}^{N}\mu_{q(z_{1i})}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)})\big)\Big]$$

$$\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} = \Big[\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})\frac{\gamma_2^2}{\gamma_1+\gamma_2}\boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})}\mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})}\mathbf{P}_2^-\Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f}_2)} = \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}\gamma_2\boldsymbol{\Sigma}^{-1}\Big[\sum_{i=1}^{N}\mu_{q(z_{2i})}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)})\big)\Big]$$

$$\sigma^2_{q(z_{0i})} = (\mu_{q(\sigma_{z_0}^{-2})} + 2\mathbf{1}_p'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p)^{-1}$$

$$\mu_{q(z_{0i})} = \Big[\sigma^2_{q(z_{0i})}\big(\mathbf{X}_i(\mathbf{h}_i)' - \mathbf{X}_N(\mathbf{h}_N)' + (\mu_{q(z_{1N})} - \mu_{q(z_{1i})})\mu'_{q(\mathbf{f}_1)} + \frac{\gamma_2}{\gamma_1+\gamma_2}(\mu_{q(z_{2N})} - \mu_{q(z_{2i})})\mu'_{q(\mathbf{f}_2)}\big) -$$
$$\sigma^2_{q(z_{0i})}\big(\sum_{j=1}^{N-1}\mu_{q(z_{0j})}\mathbb{1}\{i\neq j\}\mathbf{1}_p'\big)\Big](\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p$$

$$\sigma^2_{q(z_{1i})} = (\mu_{q(\sigma_{z_1}^{-2})} + tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)})(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{1i})} = \sigma^2_{q(z_{1i})}\Big(\mu'_{q(\mathbf{f}_1)}(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)})\big)\Big)$$

$$\sigma^2_{q(z_{2i})} = (\mu_{q(\sigma_{z_2}^{-2})} + \frac{\gamma_2^2}{\gamma_1+\gamma_2}tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)})\boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{2i})} = \sigma^2_{q(z_{2i})}\Big(\mu'_{q(\mathbf{f}_2)}\gamma_2\boldsymbol{\Sigma}^{-1}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)})\big)\Big)$$

$$d_{q(\eta_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_1^-(\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)} + \Sigma_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)}))$$

$$d_{q(\lambda_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_2^-(\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)} + \Sigma_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)}))$$

$$b_{q(\sigma^2_{z_0})} = b + 1/2\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})$$

$$b_{q(\sigma^2_{z_1})} = b + 1/2\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})$$

$$b_{q(\sigma^2_{z_2})} = b + 1/2\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})$$

B.2 Convergence Criterion

The adapted variational Bayes algorithm is run until changes in $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w},\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big]$ are below a certain threshhold. This value can be computed in each iteration as follows.

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ (f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}}) f(\boldsymbol{\theta}_{-\mathbf{w}})) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}}) + log\ f(\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_1) - log\ q(\mathbf{f}_1)\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_2) - log\ q(\mathbf{f}_2)\big] \\
&\quad + \sum_{i=1}^{(N-1)} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(z_{0i}) - log\ q(z_{0i})\big] \\
&\quad + \sum_{i=1}^{N} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(z_{1i}) - log\ q(z_{1i})\big] \\
&\quad + \sum_{i=1}^{N} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(z_{2i}) - log\ q(z_{2i})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_2}) - log\ q(\sigma^2_{z_2})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\eta_f) - log\ q(\eta_f)\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\lambda_f) - log\ q(\lambda_f)\big]
\end{aligned}
$$

Now looking at each piece individually,

$$
\begin{aligned}
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w}\mid\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
= \;& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N}\big(log[(2\pi)^{-p/2}\mid(\gamma_1+\gamma_2)^{-1}\boldsymbol{\Sigma}\mid^{-1/2}]\big)\Big] \\
& +\ E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N}-\frac{1}{2}[(\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i)-2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}(z_{0i}\mathbf{1}_p+z_{1i}\mathbf{f}_1+\frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)+ \\
& (z_{0i}\mathbf{1}_p+z_{1i}\mathbf{f}_1+\frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}(z_{0i}\mathbf{1}_p+z_{1i}\mathbf{f}_1+\frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)]\Big] \\
= \;& \sum_{i=1}^{N}\big(log[(2\pi)^{-p/2}\mid(\gamma_1+\gamma_2)^{-1}\boldsymbol{\Sigma}\mid^{-1/2}]\big) \\
& +\Big[\sum_{i=1}^{N}-\frac{1}{2}\Big(\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i)- \\
& 2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mu_{q(z_{0i})}\mathbf{1}_p-2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)}- \\
& 2\mathbf{X}_i(\mathbf{h}_i)'\gamma_2\boldsymbol{\Sigma}^{-1}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)}+ \\
& (\sigma^2_{q(z_{1i})}+\mu^2_{q(z_{1i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)}+\boldsymbol{\mu}_{q(\mathbf{f}_1)}\boldsymbol{\mu}'_{q(\mathbf{f}_1)})(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1})+ \\
& (\sigma^2_{q(z_{2i})}+\mu^2_{q(z_{2i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}+\boldsymbol{\mu}_{q(\mathbf{f}_2)}\boldsymbol{\mu}'_{q(\mathbf{f}_2)})\frac{\gamma_2^2}{(\gamma_1+\gamma_2)}\boldsymbol{\Sigma}^{-1})+ \\
& 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}'_p(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_1)}+2\mu_{q(z_{1i})}\mu_{q(z_{2i})}\boldsymbol{\mu}'_{q(\mathbf{f}_1)}\gamma_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}+ \\
& 2\mu_{q(z_{0i})}\mu_{q(z_{2i})}\mathbf{1}'_p\gamma_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big)\Big]- \\
& \Big[\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})}+\mu^2_{q(z_{0i})})+\frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\mu_{q(z_{0i})}\mu_{q(z_{0j})}\mathbb{1}\{j\neq i\}\Big]\mathbf{1}'_p(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p
\end{aligned}
$$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_1)-log\ q(\mathbf{f}_1)\big] \;=\; & E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{p}{2}log\ 2\pi+\frac{1}{2}log\ \mid\eta_f\mathbf{P}_1^-+\lambda_f\mathbf{P}_2^-\mid\Big]- \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}_1\mathbf{f}_1'(\eta_f\mathbf{P}_1^-+\lambda_f\mathbf{P}_2^-)]\Big]+ \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2}log\ 2\pi+\frac{1}{2}log\ \mid\boldsymbol{\Sigma}_{q(\mathbf{f}_1)}\mid\Big]+ \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}tr(\mathbf{f}_1\mathbf{f}_1'\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)})-\mathbf{f}_1'\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)}\boldsymbol{\mu}_{q(\mathbf{f}_1)}\Big]+ \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}\boldsymbol{\mu}'_{q(\mathbf{f}_1)}\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)}\boldsymbol{\mu}_{q(\mathbf{f}_1)}\Big] \\
= \;& C+\frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[2log\ \eta_f\big]+\frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(p-2)log\ \lambda_f\big]- \\
& \frac{1}{2}tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)}+\boldsymbol{\mu}_{q(\mathbf{f}_1)}\boldsymbol{\mu}'_{q(\mathbf{f}_1)})(\mu_{q(\eta_f)}\mathbf{P}_1^-+\mu_{q(\lambda_f)}\mathbf{P}_2^-)\Big)- \\
& \frac{1}{2}log\ \mid\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)}\mid+\frac{p}{2}
\end{aligned}
$$

where C is a constant that does not change from one iteration to the next. Similarly,

$$\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_2) - log\ q(\mathbf{f}_2)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^- \mid \Big] - \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}_2\mathbf{f}_2'(\eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^-)]\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_2)} \mid \Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2} tr(\mathbf{f}_2\mathbf{f}_2'\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}) - \mathbf{f}_2'\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}\boldsymbol{\mu}_{q(\mathbf{f}_2)}'\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big] \\
&= C + \frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[2log\ \eta_f\big] + \frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(p-2)log\ \lambda_f\big] - \\
&\quad \frac{1}{2} tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \boldsymbol{\mu}_{q(\mathbf{f}_2)}\boldsymbol{\mu}_{q(\mathbf{f}_1)}')(\mu_{q(\eta_f)}\mathbf{P}_1^- + \mu_{q(\lambda_f)}\mathbf{P}_2^-)\Big) - \\
&\quad \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1} \mid + \frac{p}{2}
\end{aligned}$$

where C is a constant that does not change from one iteration to the next. For $\mathbf{z}_0 = (z_{01}, \ldots, z_{0(N-1)})'$

$$\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N-1}{2} log\ 2\pi - \frac{N-1}{2} log\ \sigma_{z_0}^2 - \sum_{i=1}^{N-1} -\frac{1}{2\sigma_{z_0}^2} z_{0i}^2 + \\
&\quad \frac{N-1}{2} log\ 2\pi + \frac{N-1}{2} log\ \sigma_{q(z_{0i})}^2 + \\
&\quad \sum_{i=1}^{N-1} \frac{1}{2\sigma_{q(z_{0i})}^2}(z_{0i} - \mu_{q(z_{0i})})^2\Big] \qquad (6) \\
&= \frac{N-1}{2} log\ \sigma_{q(z_{0i})}^2 - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N-1}{2} log\ \sigma_{z_0}^2\Big] - \qquad (7) \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma_{z_0}^2})}\Big(\sum_{i=1}^{N-1}(\sigma_{q(z_{0i})}^2 + \mu_{q(z_{0i})}^2)\Big) + \frac{N-1}{2}
\end{aligned}$$

For $\mathbf{z}_1 = (z_{11}, \ldots, z_{1N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_1) - log\ q(\mathbf{z}_1)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_1} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_1}} z^2_{1i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_{1i})} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{1i})}} (z_{1i} - \mu_{q(z_{1i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{1i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_1}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_1}})}\Big(\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\mathbf{z}_2 = (z_{21}, \ldots, z_{2N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_2) - log\ q(\mathbf{z}_2)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_2} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_2}} z^2_{2i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_{2i})} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{2i})}} (z_{2i} - \mu_{q(z_{2i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{2i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_2}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_2}})}\Big(\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\sigma^2_{z_0}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1) log\ \sigma^2_{z_0} - b\frac{1}{\sigma^2_{z_0}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + (a_{q(\sigma^2_{z_0})} + 1) log\ \sigma^2_{z_0} + \\
&\quad b_{q(\sigma^2_{z_0})}\frac{1}{\sigma^2_{z_0}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[-(a+1) log\ \sigma^2_{z_0}\big] - b\mu_{q(\frac{1}{\sigma^2_{z_0}})} - log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(a_{q(\sigma^2_{z_0})} + 1) log\ \sigma^2_{z_0}\big] + b_{q(\sigma^2_{z_0})}\mu_{q(\frac{1}{\sigma^2_{z_0}})}
\end{aligned}
$$

For $\sigma^2_{z_1}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_1} - b\frac{1}{\sigma^2_{z_1}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + (a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1} + \\
&\quad b_{q(\sigma^2_{z_1})}\frac{1}{\sigma^2_{z_1}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[-(a+1)log\ \sigma^2_{z_1}\big] - b\mu_{q(\frac{1}{\sigma^2_{z_1}})} - log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1}\big] + b_{q(\sigma^2_{z_1})}\mu_{q(\frac{1}{\sigma^2_{z_1}})}
\end{aligned}
$$

For $\sigma^2_{z_2}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_2}) - log\ q(\sigma^2_{z_2})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_2} - b\frac{1}{\sigma^2_{z_2}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_2})}}_{q(\sigma^2_{z_2})}}{\Gamma(a_{q(\sigma^2_{z_2})})} + (a_{q(\sigma^2_{z_2})}+1)log\ \sigma^2_{z_2} + \\
&\quad b_{q(\sigma^2_{z_2})}\frac{1}{\sigma^2_{z_2}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[-(a+1)log\ \sigma^2_{z_2}\big] - b\mu_{q(\frac{1}{\sigma^2_{z_2}})} - log\ \frac{b^{a_{q(\sigma^2_{z_2})}}_{q(\sigma^2_{z_2})}}{\Gamma(a_{q(\sigma^2_{z_2})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(a_{q(\sigma^2_{z_2})}+1)log\ \sigma^2_{z_2}\big] + b_{q(\sigma^2_{z_2})}\mu_{q(\frac{1}{\sigma^2_{z_2}})}
\end{aligned}
$$

For η_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\eta_f) - log\ q(\eta_f)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \eta_f - d\eta_f - \\
&\quad log\ \frac{d^{c_{q(\eta_f)}}_{q(\eta_f)}}{\Gamma(c_{q(\eta_f)})} - c\ log\ \eta_f + d_{q(\eta_f)}\eta_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d^{c_{q(\eta_X)}}_{q(\eta_f)}}{\Gamma(c_{q(\eta_f)})} - 2E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ \eta_f\big] - d\mu_{q(\eta_f)} + \\
&\quad d_{q(\eta_f)}\mu_{q(\eta_f)}
\end{aligned}
$$

For λ_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\lambda_f) - log\ q(\lambda_f)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \lambda_f - d\lambda_f - \\
&\quad log\ \frac{d_{q(\lambda_f)}^{c_{q(\lambda_f)}}}{\Gamma(c_{q(\lambda_f)})} - \Big(\frac{p-2}{2} + c - 1\Big)\ log\ \lambda_f + d_{q(\lambda_f)}\lambda_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d_{q(\lambda_f)}^{c_{q(\lambda_f)}}}{\Gamma(c_{q(\lambda_f)})} - (p-2)E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ \lambda_f\big] - d\mu_{q(\lambda_f)} + \\
&\quad d_{q(\lambda_f)}\mu_{q(\lambda_f)}
\end{aligned}
$$

The expression for $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big]$ can be simplified much further by combining terms that cancel out. However, in some cases the ability to cancel terms depends on the order of the updates. For instance, in the expression, $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma_{z_0}^2) - log\ q(\sigma_{z_0}^2)\big]$, the terms $-b\mu_{q(\frac{1}{\sigma_{z_0}^2})}$ and $b_{q(\sigma_{z_0}^2)}\mu_{q(\frac{1}{\sigma_{z_0}^2})}$ cancel with $-\frac{1}{2}\mu_{q(\frac{1}{\sigma_{z_0}^2})}\Big(\sum_{i=1}^{N-1}(\sigma_{q(z_{0i})}^2 + \mu_{q(z_{0i})}^2)\Big)$ from $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)\big]$ as long as the parameters of $q(\mathbf{z}_0)$ are updated before $b_{q(\sigma_{z_0}^2)}$. For convenience, we have taken account the ordering necessary to compute the convergence criterion in the updates given above. Additionally, note all components in this expression that do not change from one iteration to the next can be ignored.

APPENDIX C

For comparison purposes, the remaining 88 cycles of the juggling dataset were split into a partition of 4 subsets of 22 functions. Each subset was registered using the proposed registration and factor analysis model. Figure 1, below, illustrates the consistency of our registration model through a combined plot of all estimated registered functions from all 5 subsets of the original data. It can be seen in this plot that the registration is consistent throughout the 5 subsets. Figures 2 and 3 show the similarity between the estimated factors resulting from the 4 additional runs. These estimated factors all contain similar features to the estimated factors from the original data set found in Figure 7 of the main text.

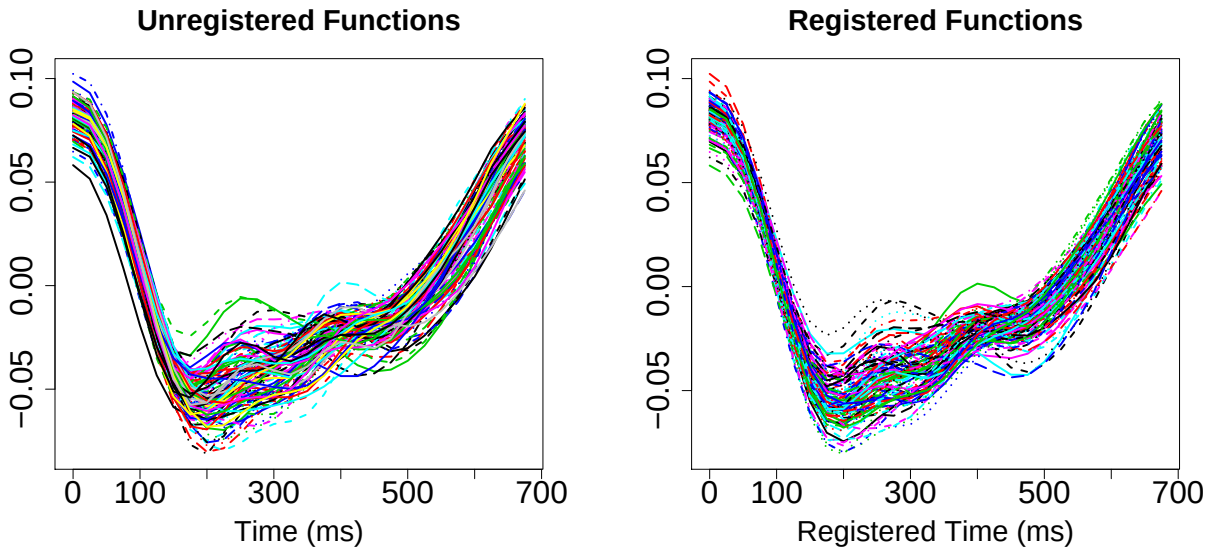

Figure 1: Consistency in registration of the entire dataset. **Left** Unregistered functions **Right** Registered functions

References

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Earls, C., and Hooker, G. (2016). Variational Bayes for Functional Data Registration, Smoothing, and Prediction. *in review.*

Goldsmith, J., Wand, M.P., and Crainiceanu, C.(2011). Functional regression via variational Bayes. *Electronic Journal of Statistics* **5**, 572.

Kullback, S., and Leibler, D.(1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79-86.

Omerod, J., and Wand, M. (2010). Explaining variational approximations. *The American Statistician* **64**, 140-153.

Ramsay, J., and Silverman, B. (2005). *Functional Data Analysis.* Springer, New York.

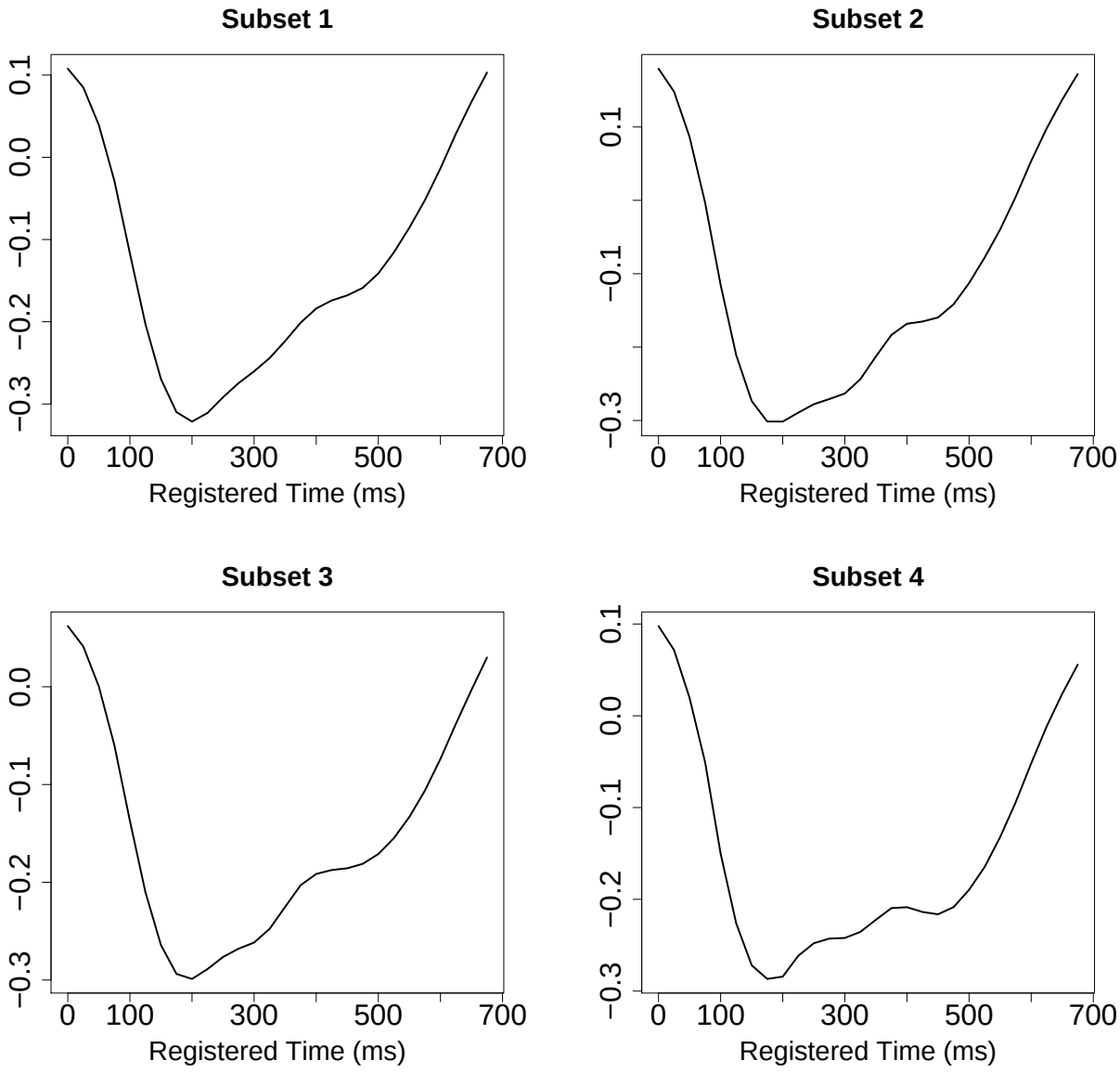

Figure 2: The estimated first factor for each of the 4 additional subsets.

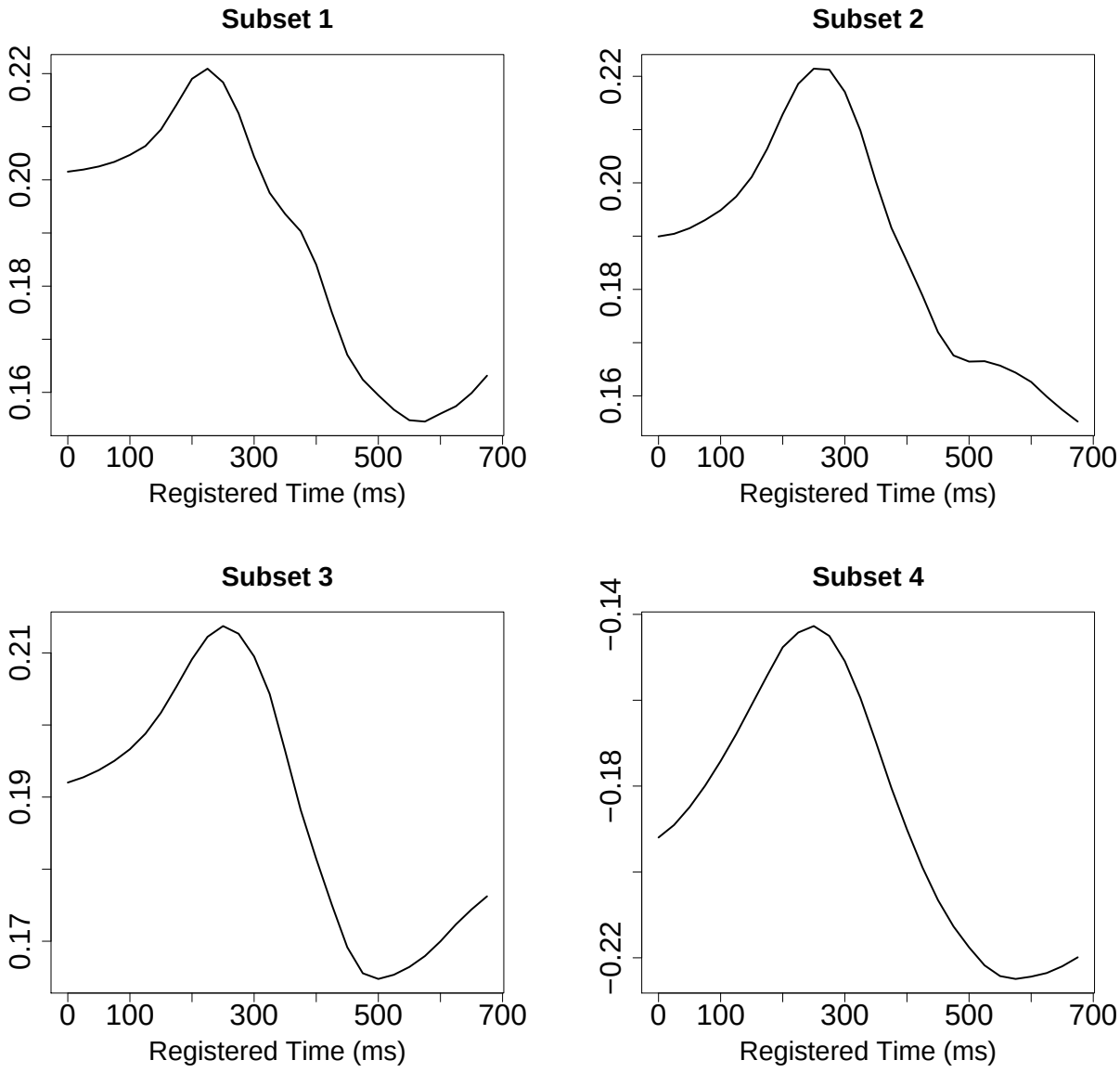

Figure 3: The estimated second factor for each of the 4 additional subsets.