Motivation

The previous model using a mixing matrix had bias/variance and identifiability issues because the parameters lacked constraints (other than the rows summing to 1). In an effort to fix this issue, we've now built a model with more structure and far fewer parameters.

Formulation

We are still using a negative binomial formulation but the expression for the mean has been modified to express mixing using a Gaussian kernel:

$$y_{ik} \sim \text{NegBin}(\omega_k \mu_{ik}, \omega_k) \qquad E(y_{ik}) = \mu_{ik} \qquad Var(y_{ik}) = \mu_{ik} + \frac{\mu_{ik}}{\omega_k}$$

Now let $a_i \in (-\infty, \infty)$ and $g_i \in \{M, F\}$ denote the age and gender of ego i, respectively, while we let g_k denote the gender of alter name k. Then we can derive the mean expression as follows:

$$\mu_{ik} = N_i p(k, g_k | a_i, g_i) = N_i \int_s p(k, s, g_k | a_i, g_i) ds = N_i \int_s p(k | s, g_k, a_i, g_i) p(s, g_k | a_i, g_i) ds$$

$$= N_i \int_s p(k | s, g_k) p(s | g_k, a_i, g_i) p(g_k | a_i, g_i) ds = N_i \int_s p(k | s, g_k) p(s | g_k, a_i, g_i) p(g_k | g_i) ds$$

$$= N_i p_{g_i g_k} \int_s p(k | s, g_k) \mathcal{K}_{g_i g_k}(a_i, s) ds$$

where the kernel is parametrized as:

$$\mathcal{K}_{g_i g_k}(a_i, a_j) = \frac{1}{\sqrt{2\pi \lambda_{g_i g_k}}} e^{-\frac{(a_i - a_j)^2}{2\lambda_{g_i g_k}}}$$

Simplification Options

1. If we assume $p(k|s, g_k) \sim \mathcal{N}(\mu_k, \sigma_k)$, we can simplify the integral in the expression of μ_{ik} using the product of unconstrained normals:

$$\int \mathcal{K}_{g_i g_k}(a_i, s) p(k | s, g_k) ds = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi \lambda_{g_i g_k}}} e^{-\frac{(a_i - s)^2}{2\lambda_{g_i g_k}}} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(s - \mu_k)^2}{2\sigma_k^2}} ds = \frac{e^{-\frac{(a_i - \mu_k)^2}{2(\lambda_{g_i g_k} + \sigma_k^2)}}}{\sqrt{2\pi(\lambda_{g_i g_k} + \sigma_k^2)}}$$

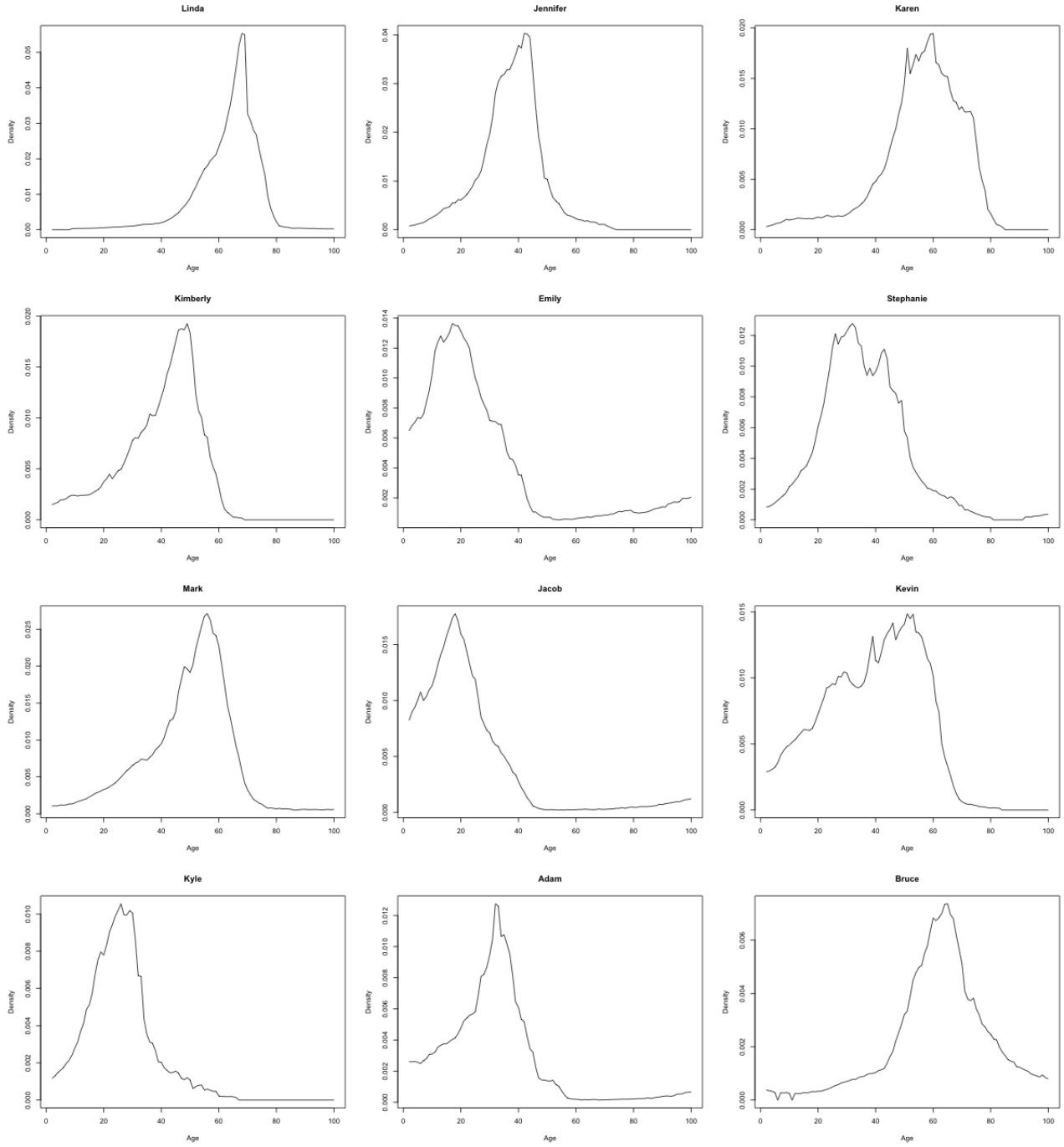
We can then reformulate the expectation as such:

$$\mu_{ik} = \frac{N_i p_{g_i g_k}}{\sqrt{2\pi(\lambda_{g_i g_k} + \sigma_k^2)}} e^{-\frac{(a_i - \mu_k)^2}{2(\lambda_{g_i g_k} + \sigma_k^2)}}$$

2. An alternative simplification is to refactorize $p(k|s, g_k) = \frac{p(s|k, g_k) p(k, g_k)}{p(s, g_k)} = \frac{p(s|k, g_k) p(k|g_k)}{p(s|g_k)}$ and then assume that $p(s|k, g_k) \sim \mathcal{N}(\mu_k, \sigma_k)$ and $p(s|g_k) \sim \mathcal{N}(\mu_{g_k}, \sigma_{g_k})$.

Normality Validity

To assess the validity of option 1, we present the empirical distributions of $p(k|s, g_k)$:



To assess the validity of option 2, we present the empirical distributions of $p(s|k, g_k)$ and $p(s|g_k)$: