1. 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.

2. Kernel Model 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.

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

Continuous Ages

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

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

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

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

where the kernel is parametrized as:

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

Discrete Ages

Alternatively, if we let $a_i \in \{0, \ldots, 100\}$, then we can derive the mean expression as follows:

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

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

$$= N_i p_{g_i g_k} \sum_s p(k | s, g_k) \frac{1}{C_{g_i g_k}} \mathcal{K}_{g_i g_k}(a_i, s)$$

The kernel is defined similar to the continuous case but with a discrete normalizing constant:

$$\mathcal{K}_{g_i g_k}(a_i, a_k) = e^{-\frac{(a_i - a_k)^2}{2\lambda_{g_i g_k}}} \qquad C_{g_i g_k} = \sum_s \mathcal{K}_{g_i g_k}(a_i, s)$$

3. Continuous Model Simplification

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:

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

4. Simulation

We simulate responses to questions about 12 names using estimated age means/variances (for each name) and simulated respondent degrees, name overdispersions, kernel lengthscales, and gender mixing rates in an attempt to determine whether our model can recover the parameters.

$$\log N_i \sim \mathcal{N}(6.2, 0.5) \qquad \frac{1}{\frac{1}{\omega_k} + 1} \sim Beta(10, 2)$$

Var	Linda	Jen.	Karen	Kim.	Emily	Steph.	Mark	Jacob	Kevin	Kyle	Adam	Bruce
μ_k	63.3	37.4	56.1	39.8	28.0	35.6	49.2	22.5	38.8	25.6	31.0	62.4
σ_k	10.4	10.6	13.9	13.0	22.9	14.3	14.9	18.0	16.2	10.8	16.2	16.7
ω_k	4.23	8.42	7.65	3.83	9.79	10.91	5.01	2.80	2.22	2.60	12.69	4.95

$$\lambda = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 225 & 144 \\ 100 & 256 \end{pmatrix}$$

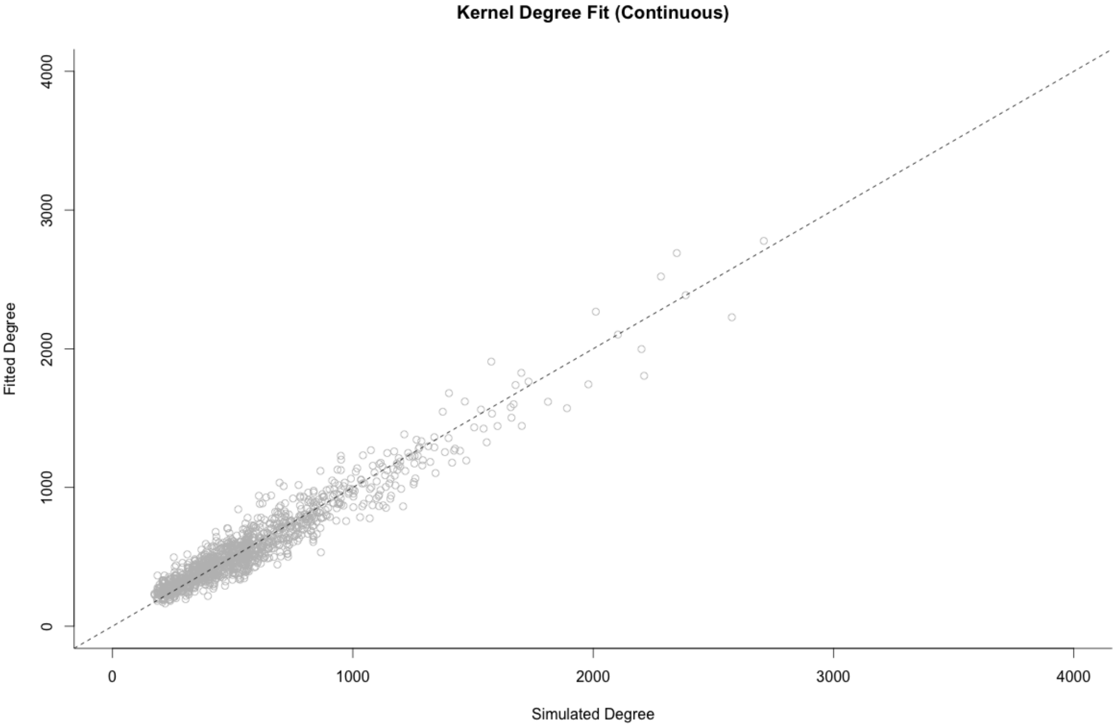
$$\rho = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.4 \\ 0.45 & 0.55 \end{pmatrix}$$

With these "true" parameter values, we then simulate two data sets: one from the continuous model and one from the discrete model. Our results outlining how well the continuous model recovers the parameters from each data set are then discussed in the next sections. In the final section, we discuss and interpret the results of fitting the continuous model on real data.

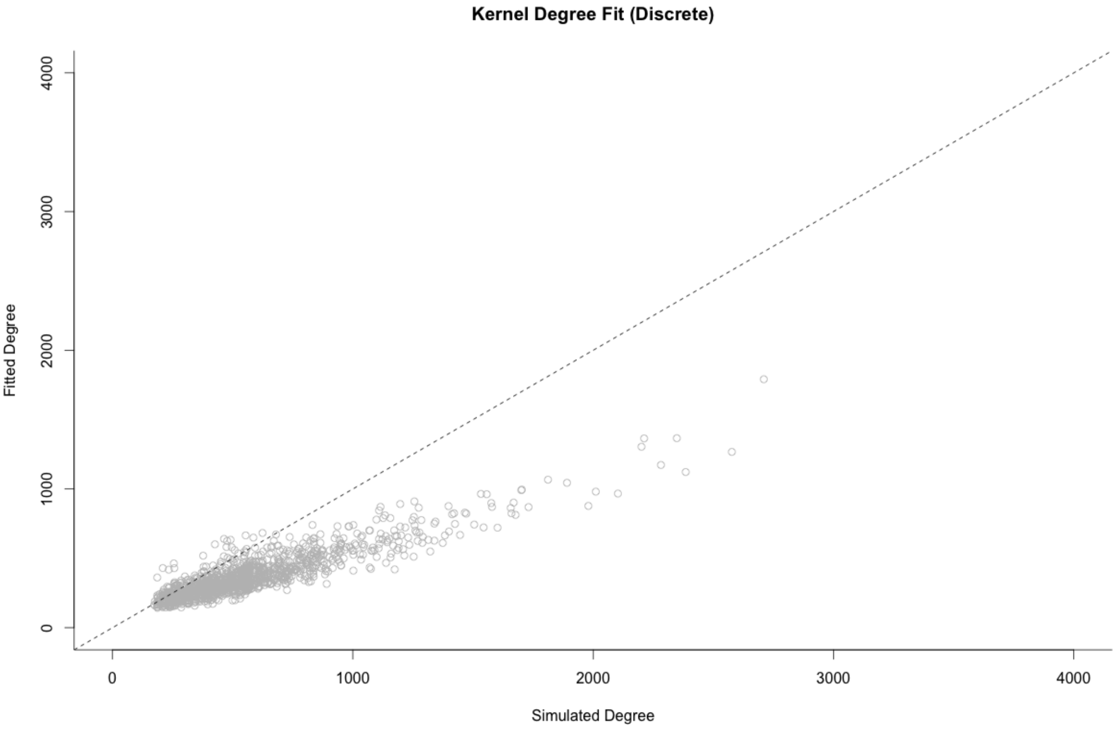
5. Simulation Results

Respondent Degrees (N_i)

The degrees are recovered well from the continuous model data, with a correlation of 0.97 between the simulated values and their posterior means.

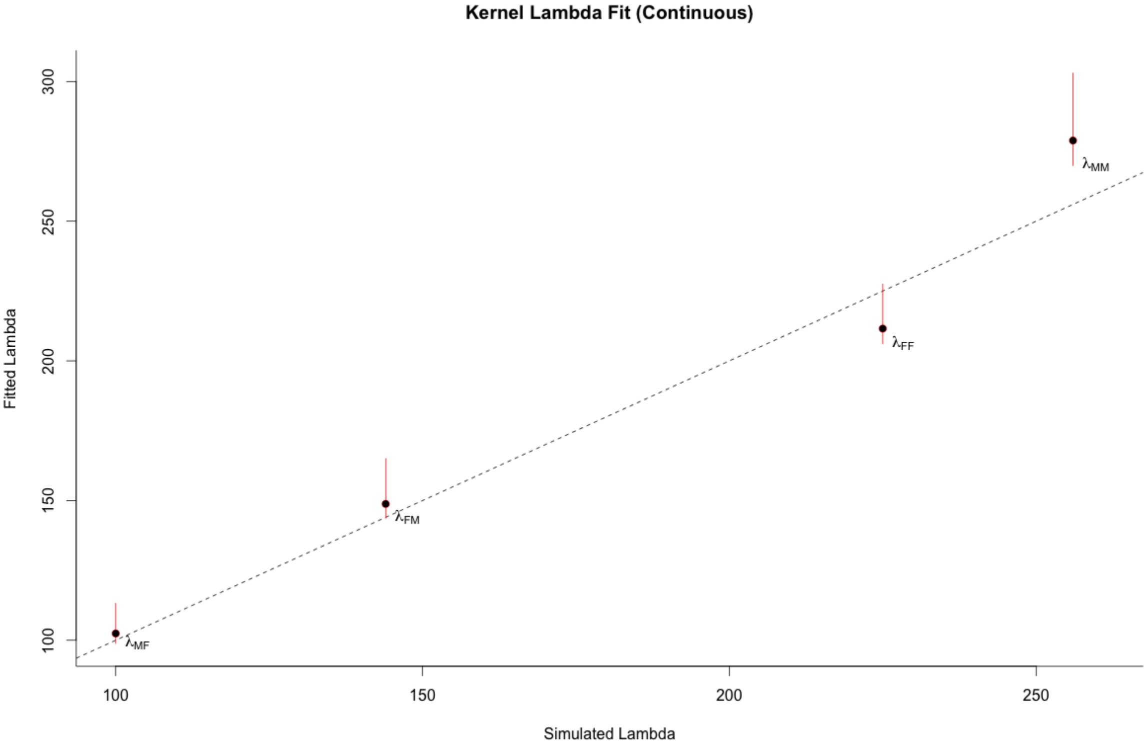


The degrees are largely underestimated for the discrete model data with a strangely linear offset.

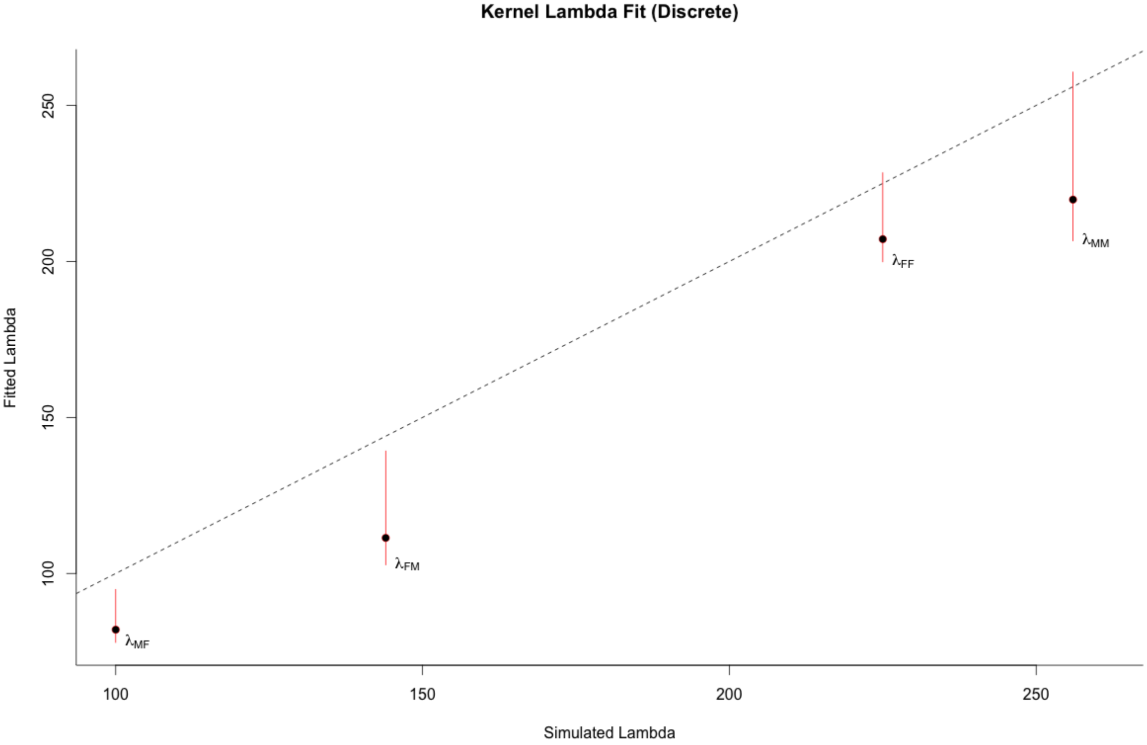


Kernel Lengthscale ($\lambda_{g_i g_j}$)

The continuous model does a great job of recovering the lambdas from the continuous model data.

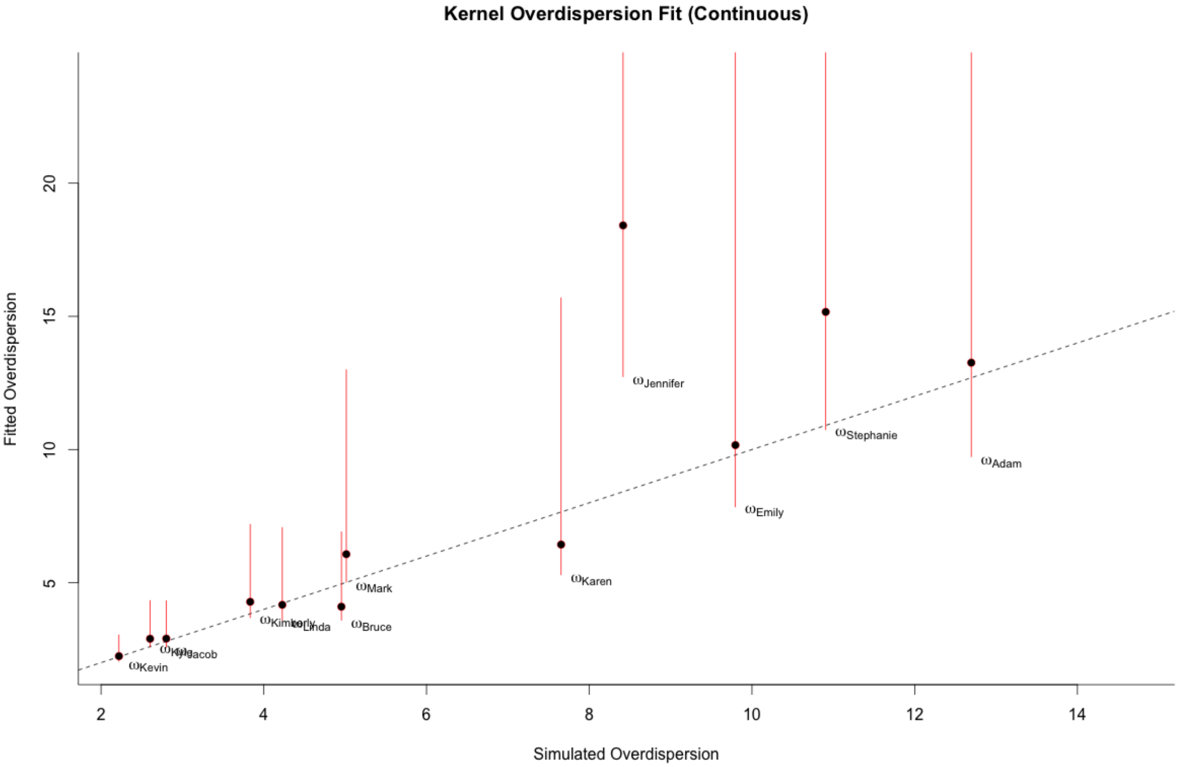


The continuous model underestimates the lambdas from the discrete model data.

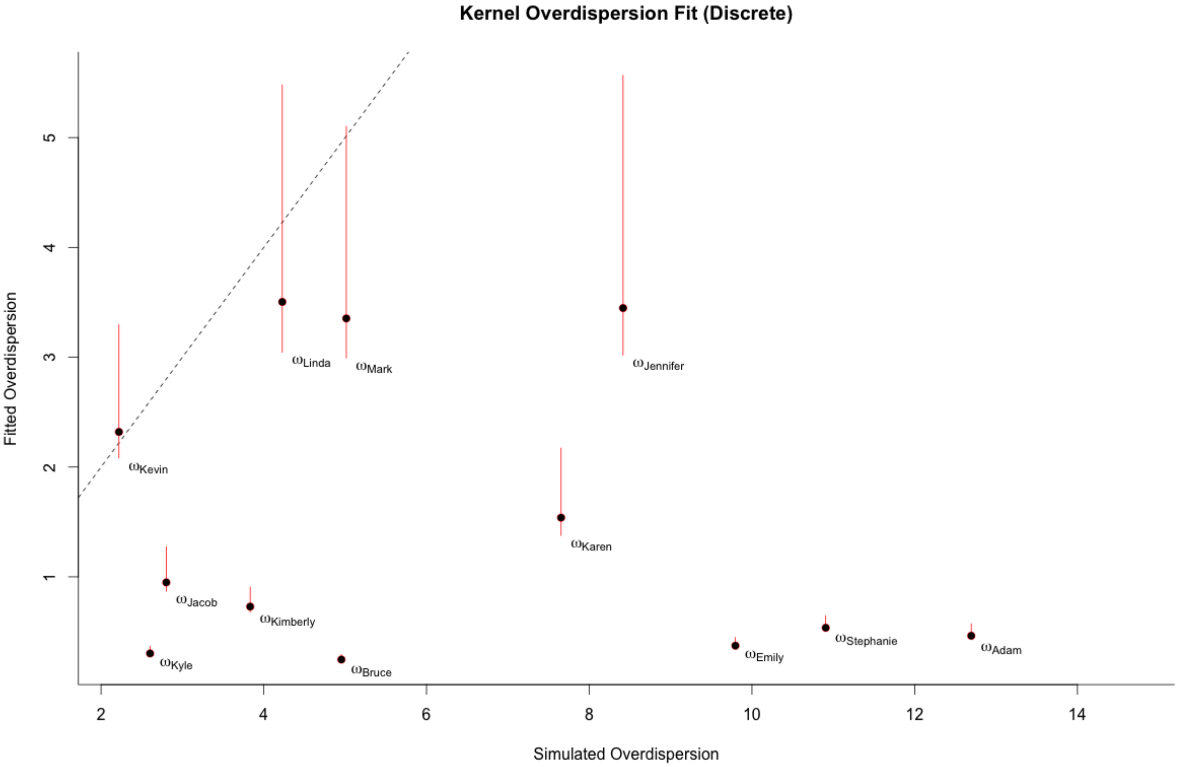


Name Overdispersions (ω_k)

The overdispersions are also mostly contained within the central 95% of their posterior distributions, but posterior uncertainty is quite large. Here the posterior median is preferred to the posterior mean.

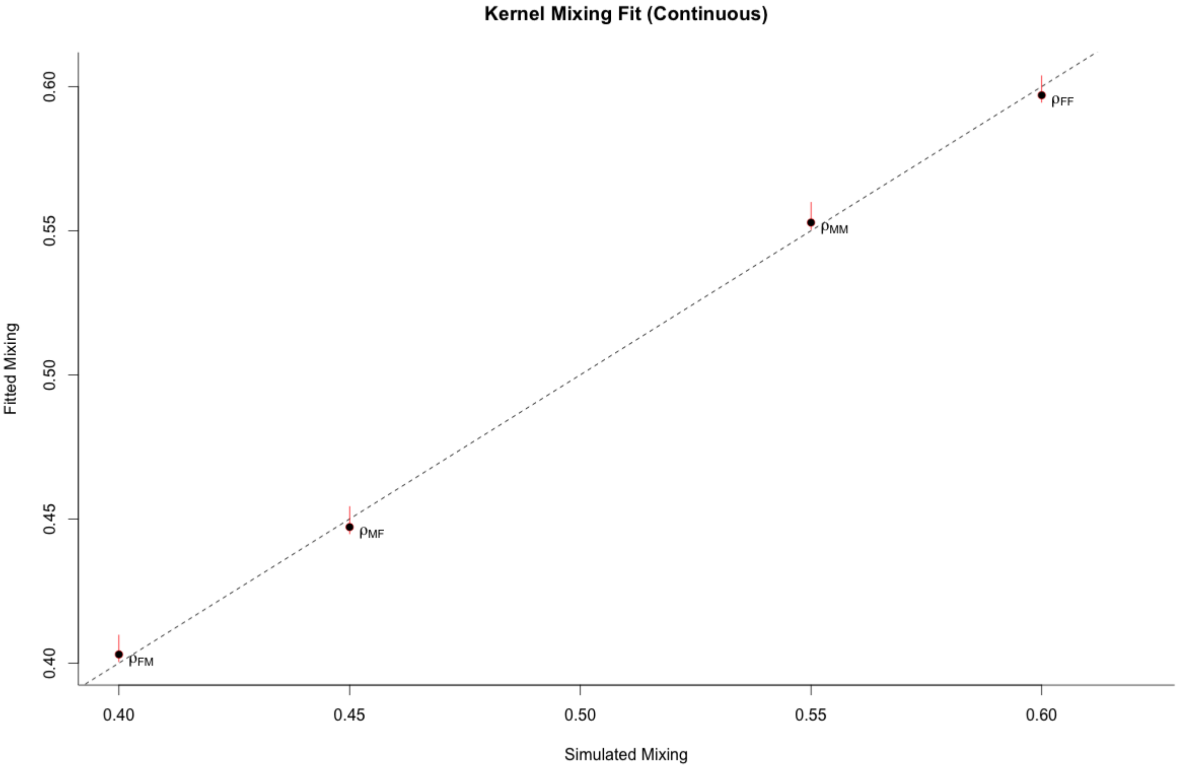


The overdispersions are poorly recovered from the discrete model data.

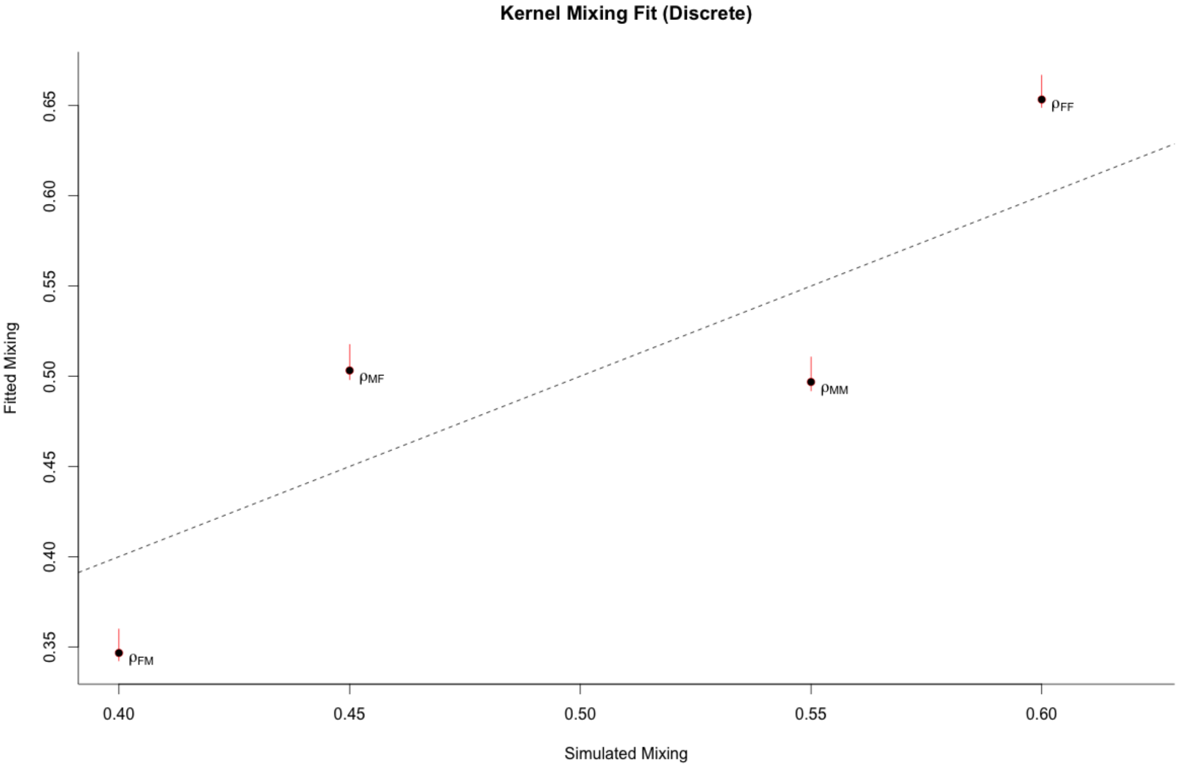


Gender Mixing Rates ($\rho_{g_i g_k}$)

Lastly, the gender mixing rates are recovered very well from the continuous model data.



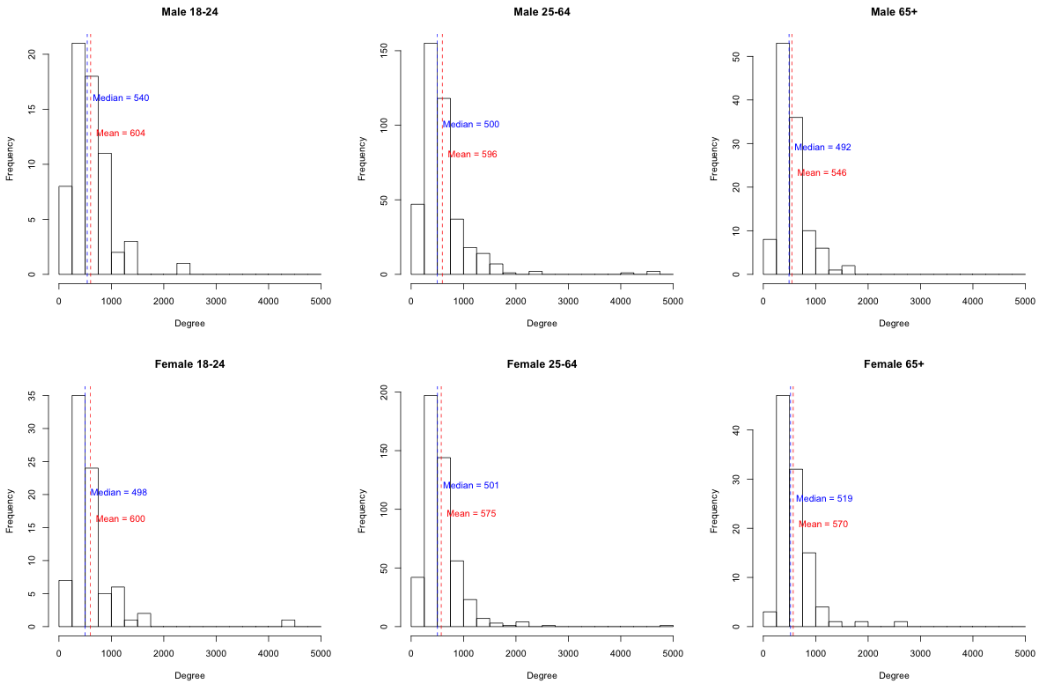
However, the mixing is poorly recovered from the discrete model data.



6. Omni Data Results

Respondent Degrees

The continuous model estimates imply decreasing network size by age for men, but possibly increasing by age for women.



Kernel Lengthscales

The length scales estimated from the actual data are quite large (compared to the ones we choose for the simulated data), implying very flat kernels. The important distinction here, then, is perhaps the relative size of the lengthscales.

$$\lambda_{BAYES} = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 5092 & 8334 \\ 4545 & 13710 \end{pmatrix}$$

Indeed, it seems that the female to female kernel is much tighter than the male to male kernel, implying that women tend to know a narrow age range of other women but men tend to know a wide age range of other men. Alternatively, the female to male kernel is much wider than the male to female kernel, implying that women know a wider age range of men while men known a very narrow age range of women.

Gender Mixing Rates

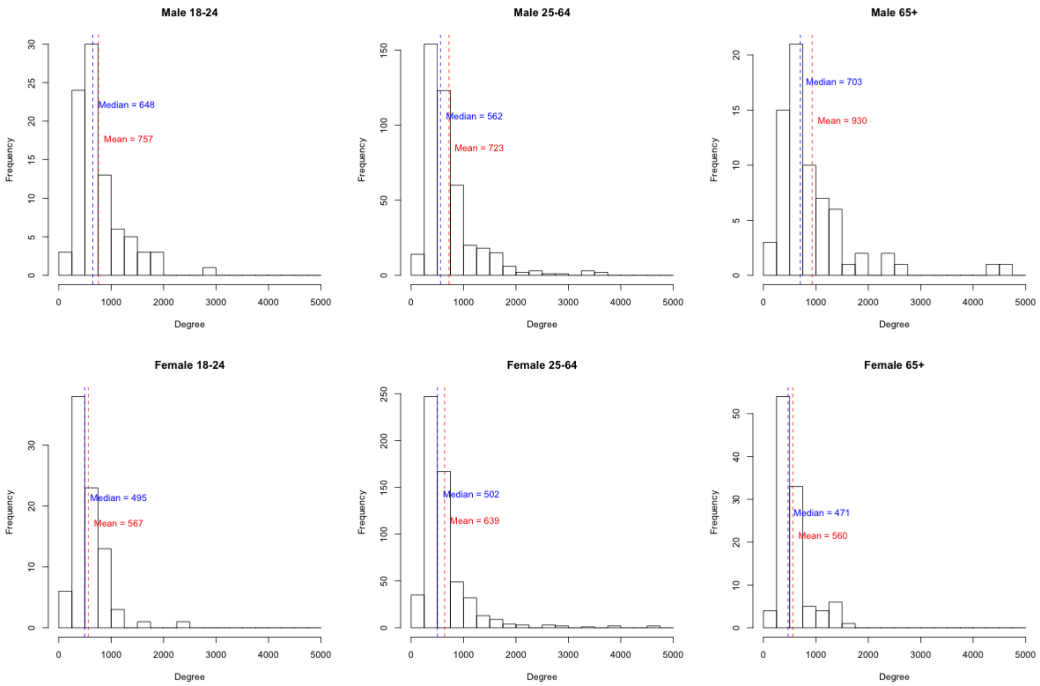
The gender mixing rates seem to correlate with the kernel lengthscales. Here we see that about 54% of a female's network is female, while 61% of a male's network is male. This also implies that females mix more with males than the other way around.

$$\rho_{BAYES} = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.54 & 0.46 \\ 0.39 & 0.61 \end{pmatrix}$$

7. McCarty Data Results

Respondent Degrees

The McCarty data imply network size is a minimum for middle aged men, while it is a maximum for middle aged women.



Kernel Lengthscales

The relative lengthscales for the McCarty data are somewhat similar to the relative lengthscales estimated from the Omni data.

$$\lambda_{BAYES} = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 136 & 506 \\ 334 & 772 \end{pmatrix}$$

Namely, we see once again that the female to female kernel is much tighter than the male to male kernel. Additionally, we see once again that the female to male kernel is wider than the male to female kernel.

Gender Mixing Rates

The gender mixing rates, however, are quite surprising, and don't correlate well with the kernel lengthscales.

$$\rho_{BAYES} = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.86 & 0.14 \\ 0.81 & 0.19 \end{pmatrix}$$