0.1 Joint model over frequency estimates and regression with a subdivision specific child effect

Below summarises my first attempt at a joint model over both frequency estimation and regression that takes into account structure at the level of the child within and across subdivisions of the data. It is basically the joint model that the extended frequency model plus would-be regression step approximates, but with dependence between haplotypes. That is, each child has a subdivision specific child effect. Ie, it doesn't capture the fact that a child with relatively low immunity in 2008 is likely to have relatively low immunity in 2009.

Due to the dependence between both $\pi_{1r}, \dots \pi_{Kr}$ and $\pi_{k1}, \dots \pi_{kR}$, instead of modelling $\boldsymbol{\pi}_{\text{child}}$ as a realisation from a distribution centred about $\boldsymbol{\pi}_{\text{population}} \in \mathbb{S}_R$, $\boldsymbol{\pi}_{\text{child}}$ ought to be modelled jointly for $k = 1, \dots, K$ as a realisation, $\boldsymbol{\pi}_{\text{child}} \in (0,1)^{K \times R}$, from a matrix variate distribution centred about $\boldsymbol{\pi}_{\text{population}} \in (0,1)^{K \times R}$, where both $\boldsymbol{\pi}_{\text{child}}$ and $\boldsymbol{\pi}_{\text{population}}$ are a $K \times R$ matrices of dependent haplotype frequencies. I therefore need a distribution over matrices in which the columns sum to one (each row is a realisation from a dirichlet), and the rows are correlated. Unfortunately, I can't model $\boldsymbol{\pi}_{\text{child } k} \in (0,1)^R \sim \mathcal{D}\text{irichlet}(\boldsymbol{\pi}_{\text{child}} \in (0,1)^R)$ and $\boldsymbol{\pi}_{\text{child}} \in (0,1)^R \sim \mathcal{D}\text{irichlet}(\boldsymbol{\pi}_{\text{population}} \in (0,1)^R))$, since $\boldsymbol{\pi}_{\text{population}}$ needs to be a $K \times R$ in order to do regression with correlation across $K \times R$. I therefore assumed $\pi_{\text{child } k}$ were independent given $\pi_{\text{population } k}$. The correlation between $k = 1, \dots, K$ is accounted for by $\boldsymbol{V}_k$ (scale matrix among over rows) in the regression. This is basically the joint model that the extended model plus regression using the approach

$$
\begin{aligned}
\rho(\boldsymbol{\pi}_{\text{poppulation}}, \boldsymbol{\pi}_{\text{child}}, \boldsymbol{a}, \boldsymbol{m}, \boldsymbol{B}, \boldsymbol{V} \mid \boldsymbol{y}) \propto & \prod_{\text{child}=1}^{\#\text{children}} \left\{ \prod_{i=1}^{I_{\text{child}_k}} \left\{ \prod_{j=1}^{J} \{\rho\left(\boldsymbol{y}_{ijk \text{ child}} | \boldsymbol{a}_{ik \text{ child}}\right)\} \right.\right. \\
& \times \rho\left(\boldsymbol{a}_{ik \text{ child}} | m_{ik \text{ child}}, \boldsymbol{\pi}_{k \text{ child}}\right) \rho\left(m_{ik \text{ child}}\right)\} \\
& \left.\left. \times \prod_{k=1}^{K} \{\rho\left(\boldsymbol{\pi}_{k \text{ child}} \mid \boldsymbol{\pi}_{k \text{ population}}\right)\} \right\} \right. \\
& \times \rho\left(\boldsymbol{\theta}_{\text{population}} \mid \boldsymbol{B}, \boldsymbol{V}_k, \boldsymbol{V}_r\right) \left| \det\left(\left\{ \frac{\partial \boldsymbol{\pi}_{\text{population}}}{\partial \boldsymbol{\theta}_{\text{population}}} \right\}^{-1} \right) \right| \\
& \times \rho(\boldsymbol{B}) \, \rho(\boldsymbol{V}_k) \rho(\boldsymbol{V}_r) \qquad (1)
\end{aligned}
$$

where $\rho\left(\boldsymbol{y}_{ijk \text{ child}} | \boldsymbol{a}_{ik \text{ child}}\right)$, $\rho\left(\boldsymbol{a}_{ik \text{ child}} | m_{ik \text{ child}}, \boldsymbol{\pi}_{k \text{ child}}\right)$, and $\rho\left(m_{ik \text{ child}}\right)$ are equivalent to previous definitions and

$$
\begin{aligned}
\rho\left(\boldsymbol{\pi}_{\text{child } k} \mid \boldsymbol{\pi}_{\text{population } k}\right) &= \mathcal{D}\text{irichlet}(\boldsymbol{\pi}_{\text{population } k}) && (2) \\
\rho\left(\boldsymbol{\theta}_{\text{population}}\right) &= \mathcal{MN}_{KR-1}(\boldsymbol{XB}, \boldsymbol{V}_k, \boldsymbol{V}_r) && (3) \\
\rho(\boldsymbol{B}) &= \mathcal{MN}_{PR-1}(\boldsymbol{M}, \boldsymbol{U}_p, \boldsymbol{U}_r), \text{say, and} && (4) \\
\rho(\boldsymbol{V}_k) &= \mathcal{W}_K^{-1}(\boldsymbol{\Psi_k}, \nu_k) && (5) \\
\rho(\boldsymbol{V}_r) &= \mathcal{W}_{R-1}^{-1}(\boldsymbol{\Psi_r}, \nu_r) && (6)
\end{aligned}
$$

Where the two inverse wishart prior distributions, $\mathcal{W}^{-1}(\cdot)$, each have two parameters, a scale matrix ($\boldsymbol{\Psi_k} \in \mathbb{R}_0^{+K \times K}$ and $\boldsymbol{\Psi_r} \in \mathbb{R}_0^{+R-1 \times R-1}$), that either encodes the prior knowledge about the correlation across studies ($\boldsymbol{\Psi_k}$), or the correlation

across haplotypes ($\boldsymbol{\Psi_r}$), and a tuning parameter, (ν_k and ν_r) that specifies strength of the prior belief.

0.2 A note regarding child-level frequencies based on the Ugandan data.

Note that, when there was no child effect, $\pi_{\text{child}\,r}$, were generally well aligned (there were some peculiar results - for example *pfcrt*-76, AL, 2012 - perhaps due to poor mixing, but I didn't look at the trace for every child in every partition). When there was a child effect, $\pi_{\text{child}\,r}$ variation was high. For the most part all children had the same major and minor allele. For one study, however (*pfmdr*-86, AL, 2010) had different children appeared to have different major and minor alleles. This is probably due to the random sample of 10 children used to generate the plot (which differs to the 16 children used for the m_i and $\boldsymbol{a}_i$ plots). For example, for *pfmdr*-86, AL, 2010, only 34 of 200 samples were wild type. If a given child have multiple wild type infections (which is compatible with the hypothesis of inter-individual immunity if the episodes are further apart than the AL half life within 2010), they might not experience much shrinkage, hence have different major/minor alleles to the overall mean. To test this I would need to investigate the major/minor relationship in light of the infections per child (the mutatednness, the number of episodes and the duration between them). I'm not going to do this, as the goal in this study was to concentrate on $\boldsymbol{\pi}_{\text{population}}$, treating $\boldsymbol{\pi}_{\text{child}}$ as a nuisance parameter. In any case, there were too few data per child to report on $\boldsymbol{\pi}_{\text{child}}$ in a meaningful manner (many children suffered only one episode per partition).

0.3 A note regarding spurious precision under the exchangeability assumption.

It seems unlikely that spurious precision with have an effect on the frequency trends (recall that, the frequency trends based on point estimates (Figure X) are practically identical to the trends based on the full posterior densities (Figure X), despite the difference in uncertainty being 100%).

0.4 A note regarding a child effect on the MOI.

Note that, neither a population nor child estimate of the MOI ($m_{\text{population}}$ nor m_{child}, respectively) feature in the child effect model (equation (**??**)) because we opt to treat $m_{i\,\text{child}}$ as a nuisance parameter, included to aid estimation of $\boldsymbol{\pi}_{\text{population}}$. That said, human attributes such as acquired immunity are thought to impact the MOI distribution (REF). We could allow different children to have different MOI propensities (replace $\rho\left(m_{i\,\text{child}}\right)$ with $\rho\left(m_{i\,\text{child}} \mid m_{\text{child}}\right)$, say, and let $m_{\text{child}} \sim \mathcal{G}\text{eometric}(\lambda^{-1}, 0.05)$, say); however, there is a scarcity of information in the data to support MOI estimates (at best, we can discern demonstrably multiclonal infections), so we opt not to include m_{child}.

0.5 Future work

0.5.1 Regression

Need to read about how to incorporate correlations due to time. Also how to implement this model. p.480(Mendeley) p.380 Bayesian Data Analysis onwards. How does this impact the choice of conjugate priors? Need to read into it. $\boldsymbol{\beta}_r|\boldsymbol{V}_r$ can no-longer be equivalent to $\mathcal{N}_P(\boldsymbol{\mu}_0, \sigma^2\boldsymbol{V}_0)$ since $\boldsymbol{V}_r$ is $K \times K$ and the prior is a P variate normal distribution, but can use a inverse wishart prior on $\boldsymbol{V}_r$

0.5.2 Joint model

To capture dependence between haplotypes and across subdivisions of the data at the level of the child the following model (figure **??**) was proposed,

$$\begin{aligned}\rho(\boldsymbol{B}, \boldsymbol{V}, \boldsymbol{\mu}_{\text{children}}, \boldsymbol{\theta}_{\text{children}\,1:K}, \boldsymbol{a}, \boldsymbol{m} \mid \boldsymbol{y}) \propto \prod_{\text{child}=1}^{291} \Bigg\{ \prod_{k=1}^{K} \Bigg\{ \prod_{i=1}^{I_{\text{child}_k}} \Bigg\{ \prod_{j=1}^{J} \{\rho\left(\boldsymbol{y}_{ijk\ \text{child}} | \boldsymbol{a}_{ik\ \text{child}}\right)\} \\ \times \rho\left(\boldsymbol{a}_{ik\ \text{child}} | m_{ik\ \text{child}}, f(\boldsymbol{\theta}_{k\ \text{child}})\right) \rho\left(m_{ik\ \text{child}}\right)\} \\ \times \rho\left(\boldsymbol{\theta}_{\text{child}_k} \mid \boldsymbol{B}, \boldsymbol{V}, \boldsymbol{\mu}_{\text{child}}\right)\} \\ \times \rho(\boldsymbol{\mu}_{\text{child}})\} \\ \times\ \rho(\boldsymbol{B})\rho(\boldsymbol{V}) \end{aligned} \tag{7}$$

- $\boldsymbol{B}$ is a $P \times R-1$ matrix of regression coefficients;
- $\boldsymbol{V}$ is a $R-1 \times R-1$ covariance matrix;
- $\boldsymbol{\mu}_{\text{children}}$ denotes the collection of $\boldsymbol{\mu}_{\text{child}}$, where each $\boldsymbol{\mu}_{\text{child}}$ denotes a $1 \times R-1$ random effect vector associated with a specific child, for child $= 1, \ldots, 291$;
- $\boldsymbol{\theta}_{\text{children}\,1:K}$ denotes the collection of $\boldsymbol{\theta}_{\text{child}\,k}$, where each $\boldsymbol{\theta}_{\text{child}\,k}$ denotes a $1 \times R-1$ vector of haplotype frequencies mapped onto the real line, for child $= 1, \ldots, 291$ and $k = 1, \ldots, K$;

and where

- $\rho\left(\boldsymbol{y}_{ijk\ \text{child}} | \boldsymbol{a}_{ik\ \text{child}}\right)$, $\rho\left(\boldsymbol{a}_{ik\ \text{child}} | m_{ik\ \text{child}}, \boldsymbol{\pi}_{k\ \text{child}}\right)$, $\rho\left(m_{ik\ \text{child}}\right)$ are equivalent to the definitions for $\rho\left(\boldsymbol{y}_{ij} | \boldsymbol{a}_i\right)$, $\rho\left(\boldsymbol{a}_i | m_i, \boldsymbol{\pi}\right)$, $\rho\left(m_i\right)$ (equations X, X and X)
- $\rho\left(\boldsymbol{\theta}_{\text{child}_k} \mid \boldsymbol{B}, \boldsymbol{V}, \boldsymbol{\mu}_{\text{child}}\right) = \mathcal{N}_{R-1}(\boldsymbol{x}_k \cdot \boldsymbol{B} + \boldsymbol{\mu}_{\text{child}}, \boldsymbol{V})$
- $\rho(\boldsymbol{B}) = \mathcal{MN}_{PR-1}(\boldsymbol{M}, \boldsymbol{U}, (\boldsymbol{X}^T\boldsymbol{X})^{-1})$
- $\rho(\boldsymbol{V}) = \mathcal{W}_{R-1}^{-1}(\boldsymbol{\Psi}, v)$
- $\rho\left(\boldsymbol{\mu}_{\text{child}}\right) \sim \mathcal{N}_{R-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

This model assumes the child effect is fixed year upon year, or across different durations since last treatment, and that $\boldsymbol{\theta}_{\text{child}\ k}$ are conditionally independent given $\boldsymbol{B}, \boldsymbol{V}$ and $\boldsymbol{\theta}_{\text{child}}$, where the non-diagonal components of $\boldsymbol{V}$ capture dependence between haplotypes, not subdivisions. Hence, the only dependence between subdivisions is due $\boldsymbol{\theta}_{\text{child}}$. The key parameter of interest is $\boldsymbol{B}$ (since $\boldsymbol{B}$ describes the trends over time or duration since last treatment).