MDI Conditional probability for the context similarity parameter

Stephen Coleman[1,*]

[1]MRC Biostatistics Unit, Cambridge, UK
[*]stephen.coleman@mrc-bsu.cam.ac.uk

ABSTRACT

The derivation of the conditional probability for the context similarity parameter ϕ_{12} between 2 different contexts. The final form is a mixture of Gamma distributions.

1 Context similarity parameter (ϕ_{12})

For multiple dataset integration (**MDI**) in the case of n observations in 2 datasets (also referred to as *contexts*):

$$p(\{c_{i1}, c_{i2}\}_{i=1}^n, v) \propto (1 + \phi_{12}\mathbb{I}(c_{i1} = c_{i2})) \prod_{k=1}^{2} \gamma_{c_{ik}k} \tag{1}$$

We assume priors of $\gamma_{1k}, \ldots, \gamma_{nk} \overset{i.i.d}{\sim} Gamma(\alpha_k/N, 1) \,\forall\, k \in \{1, 2\}$ where N is the smaller of the number of clusters in the two contexts. Similarly $\phi_{12} \sim Gamma(a, b)$.

From (1) we calculate the normalising constant Z, and find:

$$Z = \sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \left((1 + \phi_{12}\mathbb{I}(j_1 = j_2)) \prod_{k=1}^{2} \gamma_{j_k k} \right) \tag{2}$$

The joint density is hence:

$$p(\{c_{i1}, c_{i2}\}_{i=1}^n, v) = \frac{1}{Z} \prod_{i=1}^{n} \left((1 + \phi_{12}\mathbb{I}(c_{i1} = c_{i2})) \prod_{k=1}^{2} \gamma_{c_{ik}k} \right) \tag{3}$$

Intoduce a strategic latent variable v such that the form is:

$$p(\{c_{i1}, c_{i2}\}_{i=1}^n, v) = \frac{v^{n-1} \exp(-vZ)}{(n-1)!} \prod_{i=1}^{n} \left((1 + \phi_{12}\mathbb{I}(c_{i1} = c_{i2})) \prod_{k=1}^{2} \gamma_{c_{ik}k} \right) \tag{4}$$

Where Z remains as in (2).

1.1 Conditional likelihood

Consider the conditional probability of ϕ, then from (4) and expanding Z:

$$p(\phi_{12} | \{c_{i1}, c_{i2}\}_{i=1}^n, v) \propto \exp\left(-v \sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \left((1 + \phi_{12}\mathbb{I}(j_1 = j_2)) \prod_{k=1}^{2} \gamma_{j_k k} \right) \right) \prod_{i=1}^{n} \left((1 + \phi_{12}\mathbb{I}(c_{i1} = c_{i2})) \prod_{k=1}^{2} \gamma_{c_{ik}k} \right) \tag{5}$$

Now, consider the coefficients of the two occurences of ϕ_{12}:

$$a = \prod_{i=1}^{n} \left((1+\phi_{12}\mathbb{I}(c_{i1}=c_{i2})) \prod_{k=1}^{2} \gamma_{c_{ik}k} \right) \tag{6}$$

$$b = \sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \left((1+\phi_{12}\mathbb{I}(j_1=j_2)) \prod_{k=1}^{2} \gamma_{j_k k} \right) \tag{7}$$

Beginning with a from above:

$$\begin{aligned}
a &= \prod_{i=1}^{n} \gamma_{c_{i1}1}\gamma_{c_{i2}2}\,(1+\phi_{12}\mathbb{I}(c_{i1}=c_{i2})) && (8)\\
&= (1+\phi_{12}\mathbb{I}(c_{i1}=c_{i2}))^n \prod_{i=1}^{n} \gamma_{c_{i1}1}\gamma_{c_{i2}2} && (9)\\
&\propto (1+\phi_{12}\mathbb{I}(c_{i1}=c_{i2}))^n && (10)\\
&= (1+\phi_{12}))^{\sum_{i=1}^{n}\mathbb{I}(c_{i1}=c_{i2})} && (11)\\
&= \sum_{r=0}^{\sum_{i=1}^{n}\mathbb{I}(c_{i1}=c_{i2})} \binom{\sum_{i=1}^{n}\mathbb{I}(c_{i1}=c_{i2})}{r} \phi_{12}^{r} \qquad \text{(from the binomial theorem)} && (12)
\end{aligned}$$

Here $\sum_{i=1}^{n}\mathbb{I}(c_{i1}=c_{i2})$ is the count of observations assigned to the same cluster in both contexts and will be called c.

Now consider b:

$$b = \exp\left(-v \sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \left((1+\phi_{12}\mathbb{I}(j_1=j_2)) \prod_{k=1}^{2} \gamma_{j_k k} \right) \right) \tag{13}$$

We see that for our conditional we can ignore all cases when $j_1 \neq j_2$ as ϕ_{12} is not present in these. This simplifies b to:

$$\begin{aligned}
b &\propto \sum_{j=1}^{N} \gamma_{j1}\gamma_{j2}\,(1+\phi_{12}) && (14)\\
&\propto \sum_{j=1}^{N} \gamma_{j1}\gamma_{j2}\phi_{12} && (15)
\end{aligned}$$

Thus updating (5) accordingly gives us:

$$p(\phi_{12} | \{c_{i1}, c_{i2}\}_{i=1}^{n}, v) \propto \exp\left(-v \sum_{j=1}^{N} \gamma_{j1}\gamma_{j2}\phi_{12} \right) \sum_{r=0}^{c} \binom{c}{r} \phi_{12}^{r} \tag{16}$$

We notice this has the structure similar to a mixture of Gamma distributions. We thus have:

$$\begin{aligned}
p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v | \phi_{12}) &\propto \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} \frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}}{r!} \phi_{12}^{r} \exp\left(-v \sum_{j=1}^{N} \gamma_{j1}\gamma_{j2}\phi_{12} \right) && (17)\\
&= \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} Gamma\left(r+1, v \sum_{j=1}^{N} \gamma_{j1}\gamma_{j2} \right) && (18)
\end{aligned}$$

As we know that $p(\phi_{12}|\{c_{i1}, c_{i2}\}_{i=1}^{n}, v)$ must integrate over ϕ_{12} to 1, we know the normalising constant must be the sum of the integrals of the Gamma distributions, i.e.:

$$Z_{\phi_{12}} = \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} \int \frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}}{r!} \phi_{12}^{r} \exp\left(-v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\phi_{12}\right) d\phi_{12} \tag{19}$$

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} \tag{20}$$

Combining these gives:

$$p(\phi_{12}|\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) = \frac{1}{Z_{\phi_{12}}} \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} Gamma\left(r+1, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right) \tag{21}$$

1.2 Posterior distribution

Now if we consider a prior of $Gamma(a_0, b_0)$ on the ϕ_{12}, we have a prior probability of:

$$p(\phi_{12}) = \frac{b_0^{a_0}}{(a_0-1)!} \phi_{12}^{a_0-1} \exp(-b_0\phi_{12}) \tag{22}$$

Thus our posterior conditional is:

$$p(\phi_{12}|\cdot) \propto p(\phi_{12}) p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v|\phi_{12}) \tag{23}$$

$$\propto \frac{b_0^{a_0}}{(a_0-1)!} \phi_{12}^{a_0-1} \exp(-b_0\phi_{12}) \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} \frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}}{r!} \phi_{12}^{r} \exp\left(-v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\phi_{12}\right) \tag{24}$$

$$\propto \sum_{r=0}^{c} \binom{c}{r} \phi_{12}^{r+a_0-1} \exp\left(\left(-v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2} - b_0\right)\phi_{12}\right) \tag{25}$$

$$\propto \sum_{r=0}^{c} \binom{c}{r} \frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} \frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}{(r+a_0-1)!} \phi_{12}^{r+a_0-1} \exp\left(-\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)\phi_{12}\right) \tag{26}$$

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} Gamma\left(r+a_0, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right) \tag{27}$$

For the normalising constant, we have, similarly to (20):

$$Z'_{\phi_{12}} = \sum_{r=0}^{c} \binom{c}{r} \frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} \int \frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}{(r+a_0-1)!} \phi_{12}^{r+a_0-1} \exp\left(-\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)\phi_{12}\right) d\phi_{12} \tag{28}$$

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} \tag{29}$$

Thus our final posterior on the context similarity parameter ϕ_{12} is:

$$p(\phi_{12}|\,\{c_{i1},c_{i2}\}_{i=1}^{n},v) = \frac{1}{Z'_{\phi_{12}}} \sum_{r=0}^{c} \binom{c}{r} \frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} Gamma\left(r+a_0, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right) \tag{30}$$