MDI - what is it?

Stephen Coleman[1,*]

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

ABSTRACT

Multiple dataset integration (**MDI**) is a Bayesian approach to sharing information from multiple datasets of observations on the same individuals. The method avoids creating a new, larger dataset that is the naive combination of the original datasets and instead carries out clustering on each dataset *in parallel*. If any of the clusters emerging in the datasets contain common individuals, we can consider these clusters similar and can use this information to more upweigh the allocation probability of an individual who belongs to the cluster in one dataset to the corresponding cluster in the second dataset. The magnitude of this upweighting is determined by the "similarity" of the clusters (i.e. local similarity) and the datasets (a global similarity). We define similarity around this concept of corresponding clusters across datasets sharing the same individuals. This means that if there is no common information between the datasets that the final clustering will be approximately the same as if clustering was done using traditional methods. Throughout this implementation we limit the number of datasets present to 2 but this is not required as is shown in [1].

To implement MDI we require that each dataset share common members of the population in the same order, i.e. observation i in dataset 1 corresponds to observation i in dataset 2 for all $i \in (1, \ldots, n)$ for $n \in \mathbb{N}$ observations. Thus each observation has an associated pair of measurement vectors, $(\vec{x}_{i1}, \vec{x}_{i2})$. We do not require that these measurement vectors share any common nature, but we do require that the type of measurements in each dataset is the same, i.e. either continuous or categorical.

A comment on notation

We use the following symbols throughout this piece.

- $n \in \mathbb{N}$: the number of observations in each dataset;
- x_i: the ith observation for some $i \in \{1, \ldots, N\}$;
- $L \in \mathbb{N}$: the number of datasets;
- $K_l \in \mathbb{N}$: the number of components in dataset l for some $l \leq L$. If $L = 1$ we do not include the subscript;
- $c_{il} \in \{1, \ldots, K_l\}$: the latent clustering variable for the ith observation in the lth dataset;
- $Z \in \mathbb{R}$: a relevant normalising constant;
- $\phi \in \mathbb{R} \,|\, \phi > 0$: the context similarity parameter, a measure of the similarity between datasets 1 and 2;
- $\pi_j \in [0, 1] \subset \mathbb{R}$: the component proportions within the dataset for some $j \leq K$; and
- $\gamma_{jl} \in \mathbb{R} \,|\, \gamma_{jl} > 0$: the component weights for the jth component in dataset l.

1 Multiple dataset integration

1.1 Dataset-level modelling

For the context-specific clustering assume there exists some model of the form:

$$p(x) = \sum_{i=1}^{K} \gamma_i f(x|\theta_i) \tag{1}$$

where $p(x)$ denotes the probability denisty model, here a K-component density model where for the i-th component we have component weight γ_i and some parametric density f (currently limited to a Categorical or Gaussian density) defined by the vector of parameters, θ_i.

As is traditional [2], we introduce the latent component allocation variables $\{c_i\}_{i=1}^n$ where x_i is associated with component c_i and define the model:

$$x_i|c_i \sim F(\theta_i), \tag{2}$$

$$c_i|\pi \sim Multinomial(\pi_1, \ldots, \pi_K), \tag{3}$$

$$\pi_1, \ldots, \pi_K \sim Dirichlet(\alpha/K, \ldots, \alpha/K), \tag{4}$$

$$\theta_c \sim G^{(0)} \tag{5}$$

where F is the distribution corresponding to density f, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the vector of K componenet weights, α is the mass parameter which itself may be given a hyperparameter, and $G^{(0)}$ is the prior on the component parameters.

We consider the c_i to define a *clustering* on the x_i where the number of clusters present is at most K.

1.2 Dependent data sets

Consider now the case defined initially of L different datasets for the same n observations. Ideally we can use information from each context to more confidently allocated observations in all other contexts. We thus directly model the dependence between datasets using some parameter, ϕ, also referred to as the context similarity parameter.

To do this we must consider L context-specific mixture models as referred to above, but augmented by this new parameter ϕ. If we consider now the case $L = 2$ where we have K_1 and K_2 components present in the datasets respectively, we can consider two components likely to be aligned if they contain a relatively large number of the same points. The more clusters we have strongly aligned (i.e. with very similar membership), the greater we expect ϕ to be and the more likely we expect a point present in component l in dataset 1 to be present in component l in dataset 2. Thus we define a model:

$$p(c_{i1}, c_{i2}|\phi) \propto \pi_{c_{i1}1}\pi_{c_{i2}2}(1 + \phi\mathbb{I}(c_{i1} = c_{i2})) \tag{6}$$

where c_{il} is the component x_i is associated with in dataset l and π_{kl} are the weights for component k in dataset l. $\mathbb{I}$ is the indicator function defined for any $x, y \in \mathbb{R}$:

$$\mathbb{I}(x = y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{else} \end{cases} \tag{7}$$

Notice how the original mixture model emerges when $L = 1$. Furthermore, if we construct ϕ such that if the datasets have no similar clustering there is approximately no contribution to the clustering then the contribution of the other context on the current clustering is small. Another point to notice here is that at most $\min(K_1, K_2)$ components receive an upweighting as we enforce at most a pairwise association. The decision to link the datasets at the level of the latent variable c_{il} avoids the difficulties if the contexts have different types of data present with differing levels of noise.

1.3 General model

If we shift from to a weight variable $\boldsymbol{\gamma} = (\gamma_{1l}, \ldots, \gamma_{K_l l}) \overset{i.i.d}{\sim} \text{Ga}(\alpha_l/K_l, 1)$ where Ga denotes the gamma distribution, and $\boldsymbol{\gamma}$ is related to $\boldsymbol{\pi}$ by $\pi_{il} = \frac{\gamma_{il}}{\sum_{j=1}^{K_l}\gamma_{jl}}$ we then have:

$$p(\{c_{i1}, c_{i2}\}_{i=1}^n) \propto \gamma_{c_{i1}1}\gamma_{c_{i2}2}\,(1 + \phi\mathbb{I}(c_{i1} = c_{i2})) \tag{8}$$

From (8) we calculate the normalising constant Z, and find:

$$Z = \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} \gamma_{j_1 1}\gamma_{j_2 2}\left((1+\phi\mathbb{I}(j_1 = j_2))\right) \tag{9}$$

The joint density is hence:

$$p(\{c_{i1}, c_{i2}\}_{i=1}^{n}) = \frac{1}{Z}\prod_{i=1}^{n}\left(\gamma_{c_{i1}1}\gamma_{c_{i2}2}\left(1+\phi\mathbb{I}(c_{i1} = c_{i2})\right)\right) \tag{10}$$

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

$$p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) = \frac{v^{n-1}\exp(-vZ)}{(n-1)!}\prod_{i=1}^{n}\left((1+\phi\mathbb{I}(c_{i1} = c_{i2}))\prod_{l=1}^{2}\gamma_{c_{il}l}\right) \tag{11}$$

Where Z remains as in (9).

1.4 Conditionals

Conditional for v: The strategic latent variable has the following associated distribution:

$$v \sim \mathrm{Ga}(n, Z) \tag{12}$$

Conditional for $\gamma_{j_m m}$: The weight for component j_m in context m is $\mathrm{Ga}(a_\gamma, b_\gamma)$ where, for $l \in (1,2) \subset \mathbb{N}$ such that $l \neq m$:

$$a_\gamma = 1 + \sum_{i=1}^{n}\mathbb{I}(c_{im} = j_m) \tag{13}$$

$$b_\gamma = v\sum_{j=1}^{K_l}\Big(\gamma_{c_{j_l}l}(1+\mathbb{I}(j_1 = j_2)\Big) \tag{14}$$

1.5 Conditional for ϕ

Consider the conditional probability of ϕ, then from (11) and expanding Z:

$$p(\phi \,|\, \{c_{i1}, c_{i2}\}_{i=1}^{n}, v) \propto \exp\left(-v\sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2}\left((1+\phi\mathbb{I}(j_1 = j_2))\prod_{l=1}^{2}\gamma_{j_l l}\right)\right)\prod_{i=1}^{n}\left((1+\phi\mathbb{I}(c_{i1} = c_{i2}))\prod_{l=1}^{2}\gamma_{c_{il}l}\right) \tag{15}$$

Now, consider the coefficients of the two occurences of ϕ:

$$a = \prod_{i=1}^{n}\left((1+\phi\mathbb{I}(c_{i1} = c_{i2}))\prod_{l=1}^{2}\gamma_{c_{il}l}\right) \tag{16}$$

$$b = \sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2}\left((1+\phi\mathbb{I}(j_1 = j_2))\prod_{l=1}^{2}\gamma_{j_l l}\right) \tag{17}$$

Beginning with a from above:

$$a = \prod_{i=1}^{n} \gamma_{c_{i1}1} \gamma_{c_{i2}2} \left(1 + \phi \mathbb{I}(c_{i1} = c_{i2})\right) \tag{18}$$

$$= \left(1 + \phi \mathbb{I}(c_{i1} = c_{i2})\right)^{n} \prod_{i=1}^{n} \gamma_{c_{i1}1} \gamma_{c_{i2}2} \tag{19}$$

$$\propto \left(1 + \phi \mathbb{I}(c_{i1} = c_{i2})\right)^{n} \tag{20}$$

$$= (1 + \phi))^{\sum_{i=1}^{n} \mathbb{I}(c_{i1} = c_{i2})} \tag{21}$$

$$= \sum_{r=0}^{\sum_{i=1}^{n} \mathbb{I}(c_{i1} = c_{i2})} \binom{\sum_{i=1}^{n} \mathbb{I}(c_{i1} = c_{i2})}{r} \phi^{r} \qquad \text{(from the binomial theorem)} \tag{22}$$

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:

$$b = \exp\left(-v \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} \left((1 + \phi \mathbb{I}(j_1 = j_2)) \prod_{l=1}^{2} \gamma_{j_l l} \right) \right) \tag{23}$$

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

$$b \propto \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} (1 + \phi) \tag{24}$$

$$\propto \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \phi \tag{25}$$

where $K_{min} = \min(K_1, K_2)$. Thus updating (15) accordingly gives us:

$$p(\phi \,|\, \{c_{i1}, c_{i2}\}_{i=1}^{n}, v) \propto \exp\left(-v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \phi \right) \sum_{r=0}^{c} \binom{c}{r} \phi^{r} \tag{26}$$

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

$$p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v | \phi) \propto \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \right)^{r+1}} \frac{\left(v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \right)^{r+1}}{r!} \phi^{r} \exp\left(-v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \phi \right) \tag{27}$$

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v \sum_{j=1}^{N} \gamma_{j1} \gamma_{j2} \right)^{r+1}} \mathrm{Ga}\left(r+1, v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \right) \tag{28}$$

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

$$Z_{\phi} = \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \right)^{r+1}} \int \frac{\left(v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \right)^{r+1}}{r!} \phi^{r} \exp\left(-v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \phi \right) d\phi_{12} \tag{29}$$

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v \sum_{j=1}^{K_{min}} \gamma_{j1} \gamma_{j2} \right)^{r+1}} \tag{30}$$

Combining these gives:

$$p(\phi|\{c_{i1},c_{i2}\}_{i=1}^{n},v) = \frac{1}{Z_\phi}\sum_{r=0}^{c}\binom{c}{r}\frac{r!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}\right)^{r+1}}\mathrm{Ga}\left(r+1, v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}\right) \tag{31}$$

1.5.1 Posterior distribution

Now if we consider a prior of $\mathrm{Ga}(a_0,b_0)$ on the ϕ, we have a prior probability of:

$$p(\phi) = \frac{b_0^{a_0}}{(a_0-1)!}\phi^{a_0-1}\exp(-b_0\phi) \tag{32}$$

Thus our posterior conditional is:

$$p(\phi|\cdot) \propto p(\phi)p(\{c_{i1},c_{i2}\}_{i=1}^{n},v|\phi) \tag{33}$$

$$\propto \frac{b_0^{a_0}}{(a_0-1)!}\phi^{a_0-1}\exp(-b_0\phi)\sum_{r=0}^{c}\binom{c}{r}\frac{r!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}\right)^{r+1}}\frac{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}\right)^{r+1}}{r!}\phi^{r}\exp\left(-v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}\phi\right) \tag{34}$$

$$\propto \sum_{r=0}^{c}\binom{c}{r}\phi^{r+a_0-1}\exp\left(\left(-v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}-b_0\right)\phi\right) \tag{35}$$

$$\propto \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}\frac{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}{(r+a_0-1)!}\phi^{r+a_0-1}\exp\left(-\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)\phi\right) \tag{36}$$

$$= \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}\mathrm{Ga}\left(r+a_0, v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right) \tag{37}$$

For the normalising constant, we have, similarly to (30):

$$Z'_\phi = \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}\int\frac{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}{(r+a_0-1)!}\phi^{r+a_0-1}\exp\left(-\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)\phi\right)d\phi_{12} \tag{38}$$

$$= \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} \tag{39}$$

Thus our final posterior on the context similarity parameter ϕ is:

$$p(\phi|\{c_{i1},c_{i2}\}_{i=1}^{n},v) = \frac{1}{Z'_\phi}\sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}\mathrm{Ga}\left(r+a_0, v\sum_{j=1}^{K_{min}}\gamma_{j1}\gamma_{j2}+b_0\right) \tag{40}$$

References

1. P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, Dec. 2012.

2. Tevye and Company. Prologue: Tradition. *Fiddler on the Roof*, 1905.