Multiple dataset integration

Stephen Coleman

April 10, 2019

1 Theory

We explain some of the concepts upon which the mutliple dataset integration (MDI) model is built.

1.1 Mixture models

Given some data $X = (x_1, \ldots, x_n)$, we assume a number of unobserved processes generate the data, and membership to a process for individual i is represented using the latent variable c_i. It is assumed that each of the K processes can be modelled by a parametric distribution, $f(\cdot)$ with associated parameters θ and that the full model density is then the weighted sum of these probability density functions where the weights are the component proportions, π_k:

$$p(x_i) = \sum_{k=1}^{K} \pi_k f(x_i|\theta_k) \tag{1}$$

We carry out Bayesian inference of this model using MCMC methods. Specifically we use a Gibbs sampler. We sample first the component parameters, θ_k, and associated weights, π_k, from the associated distributions and then sample component membership.

Basically:

1. For each of K clusters sample θ_k and π_k from the associated distributions based on current memberships, c_i; and
2. For each of n individuals sample c_i based on the new θ_k and π_k.

Each individual's membership probabilities are conditionally independent of the other memberships given the cluster parameters:

$$p(c_i|c_{-i}, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) = p(c_i|\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \tag{2}$$

Where $c_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$. Thus our problem is *embarrassingly parallel*. This is part of the reason we use this method rather than a *collapsed Gibbs sampler*. Instead of sampling the parameters each iteration a collapsed Gibbs sampler marginalises them (i.e.

integrates over them) and updates them as each individual's allocation is updated. This method tends to reduce the number of iterations required before stationarity is reached [2], but each iteration is slower and the method is more difficult to implement.

The distribution we sample from for each parameter, θ, is updated after observing data X using Bayes' theorem:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_\Theta p(X|\theta')p(\theta')d\theta'} \tag{3}$$

Here Θ is the entire sample space for θ.

- We refer to $p(\theta|X)$ as the *posterior* distribution of θ as it is the distribution associated with θ *after* observing X.
- $p(\theta)$ is the *prior* distribution of θ and captures our beliefs about θ before we observe X.
- $p(X|\theta)$ is the *likelihood* of X given θ, the probability of data X being generated given our model is true. It is the criterion we focus on in our model if we would use a frequentist approach to the inference; maximising this quantity in our model generates the curve that best describes the observed data.
- $\int_\Theta p(X|\theta')p(\theta')d\theta'$ is the *normalising constant*. This quantity is also referred to as the *evidence* [3] or *marginal likelihood* and is normally represented by Z. It is referred to as the marginal likelihood as we marginalise the parameter θ by integrating over its entire sample space.

In terms of sampling the prior is very useful as it allows us to ensure that the posterior is always solvable, that we do not encounter singularities in our distribution.

Our implementation uses distributions on the priors that enforce conjugacy. This allows us to sample directly from the correct distribution for each posterior distribution.

1.2 Multiple dataset integration

If we have observed paired datasets $X_1 = (x_{1,1}, \ldots, x_{n,1}), X_2 = (x_{1,2}, \ldots, x_{n,2})$, where observations in the ith row of each dataset represent information about the same individual. We would like to cluster using information common to both datasets. One could concatenate the datasets, adding additional covariates for each individual. However, if the two datasets have different clustering structures this would reduce the signal of both clusterings and probably have one dominate. If the two datasets have the same structure but different signal-to-noise ratios this would reduce the signal in the final clustering. In both these cases independent models on each dataset would be preferable. Kirk et al. [1] suggest a method to carry out clustering on both datasets where common information is used but two individual clusterings are outputted. This method is driven by the allocation prior:

$$p(c_{i,1}, c_{i,2}|\phi) \propto \pi_{i,1}\pi_{i,2}(1 + \phi\mathbb{I}(c_{i,1} = c_{i,2})) \tag{4}$$

Here $\phi \in \mathbb{R}_+$ controls the strength of association between datasets. (4) states that the probability of allocating individual i to component $c_{i,1}$ in dataset 1 and to component $c_{i,2}$ in

dataset 2 is proportional to the proportion of these components within each dataset and up-weighted by ϕ if the individual has the same labelling in each dataset. Thus as ϕ grows the correlation between the clusterings grow and we are more likely to see the same clustering emerge from each dataset. Conversely if $\phi = 0$ we have independent mixture models.

The generalised case for L datasets, $X_1 = (x_{1,1}, \ldots, x_{n,1}), \ldots, X_L = (x_{1,l}, \ldots, x_{n,l})$ for any $L \in \mathbb{N}$ is simply a matter of combinatorics. In this case, (4) extends to:

$$p(c_{i,1}, \ldots, c_{i,L}|\boldsymbol{\phi}) \propto \prod_{l_1=1}^{L} \pi_{c_{il_1} l_1} \prod_{l_2=1}^{L-1} \prod_{l_3=l_2+1}^{L} (1 + \phi_{l_2 l_3} \mathbb{1}(c_{il_2} = c_{il_3})) \tag{5}$$

Here $\boldsymbol{\phi}$ is the $\binom{L}{2}$-vector of all ϕ_{ij} where ϕ_{12} is the variable ϕ in (4).

MDI has been applied to precision medicine, specifically glioblastoma sub-typing [4], in the past showing its potential as a tool.

References

[1] Paul Kirk, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24): 3290–3297, December 2012. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/bts595. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts595.

[2] Jun S Liu. The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *Journal of the American Statistical Association*, 89(427):958–966, September 1994.

[3] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. September 2003.

[4] Richard S. Savage, Zoubin Ghahramani, Jim E. Griffin, Paul Kirk, and David L. Wild. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv:1304.3577 [q-bio, stat]*, April 2013. URL http://arxiv.org/abs/1304.3577. arXiv: 1304.3577.