T-augmented Gaussian mixture - multiple dataset integration

Stephen Coleman

December 3, 2018

Abstract

tagmmdi is a Bayesian method for semi-supervised prediction using paired datasets. It can be considered an extension of multiple dataset integration (MDI) [8], an unsupervised clustering method utilising Dirichlet Processes, to allow semi-supervised clustering using the t-augmented Gaussian mixture (TAGM) model. We applied tagmmdi to protein localisation using two datasets, mass spectrometry data and Gene ontology (GO) terms. The MS data had a tagm model applied for prediction, using proteins with experimentally verified organelles as labelled data and a fixed number of clusters. The GO terms were treated as simple categorical data (i.e. we assumed no hierarchy of terms for model parsimony) using an overfitted unsupervised Dirichlet mixture model. The joint model is shown to outperform the state-of-the-art method tagm which itself is compared to other methods in Crook et al. [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 $i \in (1, \dots, n)$ for $n \in \mathbb{N}$ observations.

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 $i \in \{1, \dots, n\}$;
- $L \in \mathbb{N}$: 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;
- $c_{il} \in \{1, \dots, K_l\}$: the latent clustering variable for the ith observation in the lth dataset;
- $Z \in \mathbb{R}$: a relevant normalising constant;
- $\phi \in \mathbb{R} \mid \phi > 0$: the context similarity parameter, a measure of the similarity between datasets 1 and 2;
- $\mathbb{R}_+ = \{x \in \mathbb{R} | x > 0\}$;
- $\pi_j \in [0, 1] \subset \mathbb{R}$: the component proportions within the dataset for some $j \leq K$; and
- $\gamma_{jl} \in \mathbb{R} \mid \gamma_{jl} > 0$: the component weights for the jth component in dataset l.

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 a Markov-Chain Monte Carlo (MCMC) method. Our implementation uses a Gibbs sampler, sampling first the component parameters, θ_k, and associated weights, π_k, from the associated distributions and then sampling component membership using the rejection method on the vector of weighted allocation probabilities.

Basically:

1. Sample θ_k and π_k from the associated distributions based on current memberships, c_i; and
2. Sample c_i based on the new θ_k and π_k.

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

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_\Theta p(X|\theta')p(\theta')d\theta'} \tag{2}$$

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 model we would use if we were to take a frequentist approach to the inference, the best model fit based purely on 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* [9] 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 random variables that enforce conjugacy. This allows us to sample directly from the correct distribution for each posterior distribution.

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. [8] 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:

$$p(c_{i,1}, c_{i,2}|\phi) \propto \pi_{i,1}\pi_{i,2}(1 + \phi\mathbb{I}(c_{i,1} = c_{i,2})) \tag{3}$$

Here $\phi \in \mathbb{R}_+$ controls the strength of association between datasets. $\mathbb{I}(\cdot)$ is the indicator function. (3) 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. Note that Kirk et al. [8] include the generalised case for L datasets for any $L \in \mathbb{N}$.

3 Protein localisation

Protein localisation is a fundamental question as localisation to the correct location is required for interaction with its binding partners and to carry out its function [4]. These cellular components also provide the ideal biochemical environment for the proteins to function. Aberrant localisation is associated with a multitude of diseases including many subtypes of cancer, obesity and cardiovascular disease [10][6][7]. A thorough understanding of the biology behind protein localisation is important to a better understanding of these diseases. Possible translational implications of this understanding are diagnostic tools such as blood or urine tests based on protein localisation or drugs to correct mislocalisation as part of treatment [7][5].

The protein localisation data is produced using synchronous precursor selection (SPS)-based MS^3 technology using the LOPIT and *hyper*LOPIT pipelines [3][2]. In summary:

1. The cells undergo lysis in such a way as to maintain the integrity of their organelles.
2. The cell content is then separated along a density gradient. Thus, organelles and macro-molecular complexes are described by density-specific profiles along the gradient.
3. Discrete fractions along the density gradient are collected.
4. Within these fractions, quantitative protein profiles are measured using high accuracy mass spectrometry.

Thus for each fraction we have a description of the proteins present. The normalised incidence of the protein across fractions is found to follow a specific pattern unique to each organelle's associated proteins. From pre-exiting microscopy experiments we have some proteins with known associations to certain organelles. This allows use of supervised and semi-supervised methods to predict the localisation of the unlabelled data.

4 T-augmented Gaussian mixture models

T-augmented Gaussian mixture (TAGM) models are a semi-supervised prediction method using Gaussian mixture models (i.e. models as described in 1 where the distribution f is restricted to the Normal distribution) with a t-distribution as a "junk" or outlier distribution. The model is defined:

$$p(x_i|\theta, \pi, \kappa, \epsilon, M, V) = \sum_{k=1}^{K} \pi_k((1-\epsilon)f(x_i|\mu_k, \Sigma_k) + \epsilon g(x_i|\kappa, M, V)) \tag{4}$$

Where:

- θ is the component specific parameters for the component Gaussian distribution (here μ_k, Σ_k);
- π_k is the mixture proportion;
- κ is the degrees of freedom for the global outlier distribution;
- ϵ is the outlier component weight;
- M is the global mean used as the mean in the outlier distribution;
- V is half the global covariance, used in the outlier distribution;
- $f(\cdot)$ is the probability density function for the Normal distribution; and
- $g(\cdot)$ is the probability density function for a t-distribution.

Similar to the method outlined in 1, this model iterates over these steps:

1. Sample component parameters based on current allocation;
2. Sample component weights based on current allocation;
3. Sample outlier distribution weight based on current allocation;
4. Sample component allocation for each individual;
5. Given the above allocation, sample membership of the outlier distribution; and
6. Repeat.

The proteins allocated as outliers contribute to the mixing proportions but not the component parameters. One of the advantages associated with tagm models is that they quantify the uncertainty of the allocation. This distribution of allocation probability across the K organelles allows greater interpretation of the allocation and as Crook et al. [1] show can lead to the concept of multiple membership.

5 GO terms

6 TAGMMDI

A MDI model

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

$$p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) \propto \gamma_{c_{i1}1}\gamma_{c_{i2}2}\left(1 + \phi\mathbb{I}(c_{i1} = c_{i2})\right) \tag{5}$$

We assume priors of $\gamma_{1,l}, \ldots, \gamma_{n,l} \overset{i.i.d}{\sim} Gamma(\alpha_l / K_l, 1)\ \forall\, l \in \{1, 2\}$ where K_l is the number of clusters in the lth dataset. Similarly $\phi \sim Gamma(a, b)$.

From /(5) we calculate the normalising constant Z, and find:

$$Z = \sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2}\gamma_{c_{i1}1}\gamma_{c_{i2}2}\left(1 + \phi\mathbb{I}(j_1 = j_2)\right) \tag{6}$$

The joint density is hence:

$$p(\{c_{i1}, c_{i2}\}_{i=1}^{n}|\phi) = \frac{1}{Z}\prod_{i=1}^{n}\gamma_{c_{i1}1}\gamma_{c_{i2}2}\left(1 + \phi\mathbb{I}(c_{i1} = c_{i2})\right) \tag{7}$$

We introduce a strategic latent variable v such that the form is:

$$p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) = \frac{v^{n-1}\exp(-vZ)}{(n-1)!}\prod_{i=1}^{n}\left(\left(1 + \phi\mathbb{I}(c_{i1} = c_{i2})\right)\prod_{l=1}^{2}\gamma_{c_{il}l}\right) \tag{8}$$

Where Z remains as in (6).

A.1 Conditional likelihood

Consider the conditional probability of ϕ, then from (8) and expanding Z:

$$p(\phi|\,\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) \propto \exp\left(-v\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\left(\left(1 + \phi\mathbb{I}(j_1 = j_2)\right)\prod_{k=1}^{2}\gamma_{j_k k}\right)\right)\prod_{i=1}^{n}\left(\left(1 + \phi\mathbb{I}(c_{i1} = c_{i2})\right)\prod_{k=1}^{2}\gamma_{c_{ik}k}\right) \tag{9}$$

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

$$a = \prod_{i=1}^{n} \left((1 + \phi \mathbb{I}(c_{i1} = c_{i2})) \prod_{k=1}^{2} \gamma_{c_{ik}k} \right) \tag{10}$$

$$b = \sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \left((1 + \phi \mathbb{I}(j_1 = j_2)) \prod_{k=1}^{2} \gamma_{j_k k} \right) \tag{11}$$

Beginning with a from above:

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

$$= \left(1 + \phi \mathbb{I}(c_{i1} = c_{i2})\right)^{n} \prod_{i=1}^{n} \gamma_{c_{i1}1} \gamma_{c_{i2}2} \tag{13}$$

$$\propto \left(1 + \phi \mathbb{I}(c_{i1} = c_{i2})\right)^{n} \tag{14}$$

$$= \left(1 + \phi\right)^{\sum_{i=1}^{n} \mathbb{I}(c_{i1} = c_{i2})} \tag{15}$$

$$= \sum_{r=0}^{\sum_{i=1}^{n} \mathbb{I}(c_{i1} = c_{i2})} \binom{\sum_{i=1}^{n} \mathbb{I}(c_{i1} = c_{i2})}{r} \phi^{r} \qquad \text{(from the binomial theorem)} \tag{16}$$

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:

$$b = \exp\left(-v \sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \left((1 + \phi \mathbb{I}(j_1 = j_2)) \prod_{k=1}^{2} \gamma_{j_k k} \right) \right) \tag{17}$$

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:

$$b \propto \sum_{j=1}^{N} \gamma_{j1} \gamma_{j2} \left(1 + \phi\right) \tag{18}$$

$$\propto \sum_{j=1}^{N} \gamma_{j1} \gamma_{j2} \phi \tag{19}$$

Thus updating (9) accordingly gives us:

$$p(\phi | \{c_{i1}, c_{i2}\}_{i=1}^{n}, v) \propto \exp\left(-v \sum_{j=1}^{N} \gamma_{j1} \gamma_{j2} \phi \right) \sum_{r=0}^{c} \binom{c}{r} \phi^{r} \tag{20}$$

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

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

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} Gamma\left(r+1, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right) \tag{22}$$

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.:

$$Z_\phi = \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} \int \frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}}{r!} \phi^r \exp\left(-v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\phi\right) d\phi_{12} \tag{23}$$

$$= \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} \tag{24}$$

Combining these gives:

$$p(\phi|\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) = \frac{1}{Z_\phi} \sum_{r=0}^{c} \binom{c}{r} \frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}} Gamma\left(r+1, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right) \tag{25}$$

A.2 Posterior distribution

Now if we consider a prior of $Gamma(a_0, b_0)$ on the ϕ, we have a prior probability of:

$$p(\phi) = \frac{b_0^{a_0}}{(a_0-1)!}\phi^{a_0-1}\exp\left(-b_0\phi\right) \tag{26}$$

Thus our posterior conditional is:

$$p(\phi|\cdot) \propto p(\phi)p(\{c_{i1}, c_{i2}\}_{i=1}^{n}, v|\phi) \tag{27}$$

$$\propto \frac{b_0^{a_0}}{(a_0-1)!}\phi^{a_0-1}\exp\left(-b_0\phi\right)\sum_{r=0}^{c}\binom{c}{r}\frac{r!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}}\frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\right)^{r+1}}{r!}\phi^{r}\exp\left(-v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}\phi\right) \tag{28}$$

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

$$\propto \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}\frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}{(r+a_0-1)!}\phi^{r+a_0-1}\exp\left(-\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)\phi\right) \tag{30}$$

$$= \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}Gamma\left(r+a_0, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right) \tag{31}$$

For the normalising constant, we have, similarly to (24):

$$Z'_{\phi} = \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}\int\frac{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}{(r+a_0-1)!}\phi^{r+a_0-1}\exp\left(-\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)\phi\right)d\phi_{12} \tag{32}$$

$$= \sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}} \tag{33}$$

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

$$p(\phi|\{c_{i1}, c_{i2}\}_{i=1}^{n}, v) = \frac{1}{Z'_{\phi}}\sum_{r=0}^{c}\binom{c}{r}\frac{(r+a_0-1)!}{\left(v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right)^{r+a_0}}Gamma\left(r+a_0, v\sum_{j=1}^{N}\gamma_{j1}\gamma_{j2}+b_0\right) \tag{34}$$

References

[1] Oliver M Crook, Claire M Mulvey, Paul D. W. Kirk, Kathryn S Lilley, and Laurent Gatto. A Bayesian Mixture Modelling Approach For Spatial Proteomics. *PLOS Computational Biology*, 14(11), November 2018. doi: 10.1101/282269.

[2] T. P. J. Dunkley, R. Watson, J. L. Griffin, P. Dupree, and K. S. Lilley. Localization of Organelle Proteins by Isotope Tagging (LOPIT). *Molecular & Cellular Proteomics*,

3(11):1128–1134, November 2004. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.T400009-MCP200.

[3] Aikaterini Geladaki, Nina Kocevar Britovsek, Lisa M. Breckels, Tom S. Smith, Claire M. Mulvey, Oliver M. Crook, Laurent Gatto, and Kathryn S. Lilley. LOPIT-DC: A simpler approach to high-resolution spatial proteomics. July 2018. doi: 10.1101/378364.

[4] Toby J. Gibson. Cell regulation: Determined to signal discrete cooperation. *Trends in Biochemical Sciences*, 34(10):471–482, October 2009. ISSN 09680004. doi: 10.1016/j.tibs.2009.06.007.

[5] Richard P Horgan and Louise C Kenny. 'Omic' technologies: Genomics, transcriptomics, proteomics and metabolomics: The Obstetrician & Gynaecologist. *The Obstetrician & Gynaecologist*, 13(3):189–195, July 2011. ISSN 14672561. doi: 10.1576/toag.13.3.189.27672.

[6] M.-C. Hung and W. Link. Protein localization in disease and therapy. *Journal of Cell Science*, 124(20):3381–3392, October 2011. ISSN 0021-9533, 1477-9137. doi: 10.1242/jcs.089110.

[7] Tweeny R. Kau, Jeffrey C. Way, and Pamela A. Silver. Nuclear transport and cancer: From mechanism to intervention. *Nature Reviews Cancer*, 4:106, February 2004.

[8] 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.

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

[10] Jacqueline E. Siljee, Yi Wang, Adelaide A. Bernard, Baran A. Ersoy, Sumei Zhang, Aaron Marley, Mark Von Zastrow, Jeremy F. Reiter, and Christian Vaisse. Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nature Genetics*, 50(2):180–185, February 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-017-0020-9.