Progress to date: 13/07/2016

Nathan Cunningham

July 13, 2016

1 Read paper: Bayesian cluster analysis: point estimation & credibles balls by Wade & Ghahramani

Summary:

- Paper[1] deals with the problem of summarising the output of a bayesian nonparametric cluster analysis. These analyses provide a posterior over the entire space of clusterings although typically a point estimate along with an associated measure of uncertainty are of interest.
- The authors consider an information and decision theoretic techniques for summarising the posterior clustering allocations via the construction of a loss function:

$$c^* = argmin_{\hat{c}} \Sigma_c \mathbb{L}(c, \hat{c}) p(c|y_{1:N}) \tag{1}$$

  which is a loss associated with approximating the true cluster allocation, c, by an allocation $\hat{c}$ averaged over all possible clusterings.

- Loss functions should satisfy some basic principles such as invariance to permutations of data points indices or cluster labels.
- They first present Binder's loss which penalises for (1) assigning pairs to the same cluster when they should be in different clusters and (2) assigning pairs to different clusters which should be in the same cluster. This is similar to the Rand Index which considers concordance across cluster allocations, whereas this considers discordance across cluster allocations.
- The authors proposal is to use the variation of information (VI) as the loss function. VI compares information within two clusterings with information shared between them. Their proposed loss function is:

$$2H(c, \hat{c}) - H(c) - H(\hat{c}) \tag{2}$$

  where $H(c)$ is the information contained in cluster allocation c, while $H(c, \hat{c}$ is the information contained in each allocation, removing the shared information.

$$\sum_{i=1}^{k_N} \frac{n_{i+}}{N} log\left(\frac{n_{i+}}{N}\right) + \sum_{j=1}^{\hat{k}_N} \frac{n_{+j}}{N} log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\hat{k}_N} \frac{n_{ij}}{N} log\left(\frac{n_{ij}}{N}\right) \tag{3}$$

- A comparison is made of VI and the Binder loss. They are each metrics on the partition space, satisfying non-negativity, symmetry and the triangle inequality. Binder's loss tends to favour splitting clusters rather than merging which can lead to large numbers of clusters.
- Due to impossibility of comparing losses across *all* luster allocations, suggest a greedy search algorithm:
  - Initialise $\hat{c}$
  - For i = 1, ...,
    - Find the l closest partitions that cover $\hat{c}$, and the l closest partitions that $\hat{c}$ covers.
    - Compute $\mathbb{E}\left[L(c, \hat{c}|\mathbb{D}\right]$ for all 2l partitions and select the partition, c^*, which minimises this
    - If the expected loss for c^* is less than for $\hat{c}$, set $\hat{c} = c^*$, otherwise stop

- This also allows for specification of a 'ball' of size ϵ around a clustering:

$$\mathbb{B}_\epsilon(\hat{c}) = c : d(\hat{c}, c) \leq \epsilon \tag{4}$$

Use:

- The original SMC clustering algorithm outputs a separate clustering allocation for each of the particles. Feeding forward into the MDI algorithm would require selection of an optimal allocation. This may be useful.
- May also be useful within the context of deciding the final allocation for the MDI.

2 Integration of MDI and SMC clustering code

- Implementing SMC clustering methods within the MDI.
- Involves the modification of the `DrawNewItemLabel` function in MDI:
  - Originally MDI involves upweighting the allocation probabilities to account for the concordance across datasets. Then considers the allocation of each data point to each of the occupied clusters and one unoccupied cluster and assigns to the cluster best supported by the data. Replace the step with allocating to each cluster with the SMC step.
  - Results at this stage are not promising, (Figure 1). Algorithm appears to assign most observations to the same cluster, differing from the previous output of the MDI algorithm.
  - Still a lot to do with integrating the code.
    - It currently only runs using a single particle, meaning that particle is the one that gets accepted. Need to work on selecting a particular cluster from the particle filter run.
    - Currently begins fresh every run of the algorithm, no parameters carried across.
    - The parameters from the gamma matrix are not being used correctly.
    - Many wasteful calculations being carried out as remnants of the previous algorithm.

References

[1] Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls. *arXiv preprint arXiv:1505.03339*, 2015.

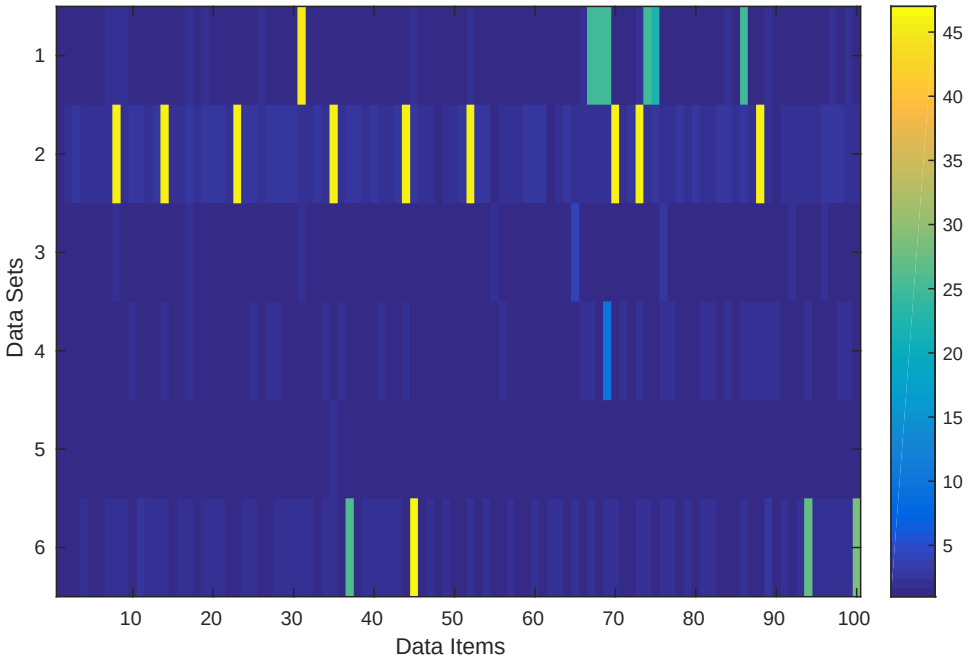

Figure 1: Results of implementation of SMC within the MDI algorithm