Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Consensus clustering for Bayesian mixture models

Stephen Coleman [1,*], Paul DW Kirk [1, 2] and Chris Wallace [1,2*]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, CB2 0SR, United Kingdom and
[2]Department of Medicine, University of Cambridge, Cambridge, CB2 0AW, United Kingdom.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Bayesian mixture models have attractive features and been successfully applied in a diverse range of settings. Inference of these models is normally performed using Markov-chain Monte Carlo (MCMC) methods. In high dimensions MCMC methods often explore a limited range of partitions, with a lack of overlap between chains (i.e. a lack of convergence) frequently present.
Results: We extend the ensemble method, Consensus clustering (CC), to Bayesian mixture models. We compare CC to inference performed using MCMC methods and also to `mclust`, a popular mixture model R package. We show that CC can be extended to Bayesian integrative clustering models. CC is then demonstrated on real datasets in both the single dataset and multiple dataset context. CC is shown to capture more modes in the clustering distribution than either the maximum-likelihood estimate (MLE) or any individual Markov chain. CC also reduces the computation time required when a parallel environment is present compared to Bayesian inference.
Availability: None?
Contact: stephen.coleman@mrc-bsu.cam.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

From defining a taxonomy of disease based upon to creating molecular sets, grouping items can help us to understand and make decisions about complex biological problems. For example, clustering patients based upon disease characteristics and personal omics data allows the reaction and progression of individuals within the cluster to inform treatment decisions about other members of the same group. In another setting, defining gene regulatory networks can provide valuable insights into the causal mechanisms driving molecular products and may be used in diagnosis or for drug targets (Emmert-Streib *et al.*, 2014).

In short, clustering data is about maximising the quantity of relevant data for each individual within a specific analysis, reducing from the complexity of the entire set without contracting to the individual level. The clustering approximates complex variation within the dataset, enabling local downstream analysis and decisions upon each group rather than at the global level.

The act of identifying such groups is referred to as "cluster analysis". Traditional methods such as k-means clustering (Lloyd, 1982; Forgy, 1965) or hierarchical clustering condition upon a user inputted choice of K, the number of occupied clusters present. Normally different choices of K are compared under some metric such as silhouette or based upon subjective interpretation of the within-cluster sum of squared errors as a function of K. There exists an alternative school of clustering, model-based clustering, which embeds the cluster analysis within a formal, statistical framework. This means that choice of K is a model selection problem with all the associated literature.

In many analyses or decision making processes, understanding how certain the clustering is van be vital. For example, in clustering patients there might be individuals with almost equal probability of being allocated between a number of clusters. Knowing which individuals have uncertain membership could stronlgy influence decisions about treatment. However in many cluster analyses only a point estimate is provided and thus one is no wiser about which individuals are boundary members of clusters. Bayesian mixture models provide an uncertainty quantification that allows one to include this uncertainty in a formal manner in downstream analyses and decisions.

Furthermore, one might believe that the number of clusters present might itself be uncertain, that it is a random variable that should be inferred from the data. Bayesian mixture models can treat K in this way, thereby incorporating uncertainty about K into the the allocation uncertainty and reducing some of the importance of choice of K. Furthermore, if there is

a prior belief about K, this may be included due to the Bayesian nature of the model.

Inference of these methods are often drive by Markov-chain Monte Carlo (MCMC) methods. However, MCMC methods in general are known to display problematic behaviour in high-dimensions (Robert *et al.*, 2018; Yao *et al.*, 2020; Chandra *et al.*, 2020). One of these problems is the difficulty in exploring a range of modes in the posterior distribution of the clustering. This means that there is a lack of convergence across chains and the distribution in any single chain is highly concentrated upon a very small number (often one) possible clustering(s). Another problem that emerges in large or high-dimensional datasets is that individual iterations of the MCMC sampler become slow and a long chain can become infeasible.

Ensembles of learners are often better at representing modes within parameters than the individual learners Ghaemi *et al.* (2011). Ensembles also offer reductions in computational runtime because most ensemble methods enable use of a parallel environment to improve computation speed Ghaemi *et al.* (2009). Consensus clustering Monti *et al.* (2003) is an ensemble method for cluster analysis, previously implemented using k-means clustering (refer to original paper) in the R package `ConsensusClusterPlus` Wilkerson *et al.* (2010). This flavour of consensus clustering has been applied to cancer subtyping (references, TCGA) and identifying subclones in single cell analysis Kiselev *et al.* (2017). Consensus clustering applies R independent runs of the clustering algorithm to peturbed versions of the dataset (i.e. sampled with replacement) and combines the R final partitions in a *Consensus matrix* (CM) to infer a final clustering. The consensus matrix is a symmetric matrix with the $(i,j)^{th}$ entry being the proportions of model runs for which the i^{th} and j^{th} items are clustered together. For a single partition the *coclustering matrix* represents this information, being a binary matrix with the $(i,j)^{th}$ entry indicating if items i and j are allocated to the same cluster.

Data: $X = (x_1, \ldots, x_N)$
Input: A resampling scheme *Resample*
A clustering algorithm *Cluster*
Number of resampling iterations S
Set of cluster numbers to try $\mathcal{K} = \{K_1, \ldots, K_{max}\}$
Output: A predicted clustering, $\hat{Y}$
The predicted number of clusters present $\hat{K}$
begin
  for $K \in \mathcal{K}$ **do**
    `/* initialise an empty Consensus Matrix */`
    $\mathbf{M}^{(K)} \leftarrow \mathbf{0}_{N \times N}$;
    for $s = 1$ **to** S **do**
      $X^{(s)} \leftarrow Resample(X)$;
      `/* Cluster the peturbed dataset, represented in a coclustering matrix */`
      $\mathbf{B}^{(s)} \leftarrow Cluster(X^{(s)}, K)$;
      $\mathbf{M}^{(K)} \leftarrow \mathbf{M}^{(K)} + \mathbf{B}^{(s)}$;
    end
    $\mathbf{M}^{(K)} \leftarrow \frac{1}{S}\mathbf{M}^{(K)}$;
  end
  $\hat{K} \leftarrow$ best $K \in \mathcal{K}$ based upon all $\mathbf{M}^{(K)}$;
  $\hat{Y} \leftarrow$ partition X based upon $\mathbf{M}^{(\hat{K})}$;
end

Algorithm 1: Consensus Clustering algorithm

We extend consensus clustering to Bayesian mixture models. We show via simulation that ensembles consisting of short chains are sufficient to uncover meaningful structure in a number of scenarios including some within which a Gibbs sampler becomes trapped in individual modes for any reasonable length of runtime. The chains are both short and independent, thus their individual runtime is far shorter than the chains traditionally used for Bayesian inference and may also be run in parallel. This means that consensus clustering of Bayesian mixture models offers significant reductions in runtime without sacrifices in performance. As the ensemble can describe multiple modes, the uncertainty present in the consensus matrix can be more representative of the data than the individual modes captured by any single chain.

We then apply our algorithm to the multiple dataset setting and the extension of Bayesian mixture models, Multiple Dataset Integration (MDI). We show on three datasets from the original MDI paper that Consensus clustering performs similarly to Bayesian inference of this model, and then using more modern, larger data that Consensus clustering enables implementation of such models in previously difficult scenarios.

2 Approach

Equation (**??**) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text .....

3 Methods

3.1 Consensus clustering for Bayesian mixture models

We extend CC to use Bayesian mixture models as the underlying model. This offers the ability to include a prior distribution on parameters and inference of the number of occupied clusters present, maintaining two of the key attractions of Bayesian model-based clustering while losing the asymptotic guarantees of Bayesian inference. In this case the algorithm is simplified as it is not necessary to try a range of possible clusters present. Furthermore, the dataset is not perturbed as described in algorithm 1 for two reasons. First, the overfit mixture can capture individuals in singletons and thus is more robust to outliers than k-means. Secondly, the MCMC method driving each model offers diversity of partitions when combined with different initialisations. The method is described in algorithm 2.

3.2 Predicting a clustering from CMs

We use the `maxpear` function Fritsch *et al.* (2009) from the R package `mcclust` Fritsch (2012) to infer a point clustering from CMs. This function was designed to perform inference upon the posterior similarity matrix (**PSM**) from the samples of a single long chain (this is analogous to a CM, except the partitions are all drawn from a single Markov chain), predicting a clustering that has maximum posterior expected adjusted Rand index Hubert and Arabie (1985) with the true clustering. In the context of the CM, the function does not have this interpretation. However, the method produces a kind of "sample average clustering" based upon all sampled partitions. This appears a sensible method for predicting a point clustering from the CM, averaging over each learner in the ensemble.

3.3 Data generating mechanism

We use the notation:

- N: the number of items generated;
- P: the number of features in the dataset;
- P_n: the number of irrelevant features used;
- K: the true number of clusters present;

Data: $X = (x_1, \ldots, x_N)$
Input: A Bayesian mixture model with membership vector $c = (c_1, \ldots, c_N)$
A clustering algorithm that generates samples *Cluster*
The number of chains to run, S
The number of iterations within each chain, R
Output: A predicted clustering, $\hat{Y}$
The consensus matrix $\mathbf{M}$
begin
  `/* initialise an empty Consensus Matrix */`
  $\mathbf{M} \leftarrow \mathbf{0}_{N \times N}$;
  for $s = 1$ **to** S **do**
    `/* set the random seed controlling initialisation and MCMC moves */`
    $set.seed(s)$;
    `/* initialise a random partition on` X `drawn from the prior distribution */`
    $Y_{(0,s)} \leftarrow Initialise(X)$;
    for $r = 1$ **to** R **do**
      `/* generate a markov chain for the membership vector */`
      $Y_{(r,s)} \leftarrow Cluster(c, r)$;
    end
    `/* create a coclustering matrix from the` R^{th} `sample */`
    $\mathbf{B}^{(s)} \leftarrow Y_{(R,s)}$;
    $\mathbf{M} \leftarrow \mathbf{M} + \mathbf{B}^{(s)}$;
  end
  $\mathbf{M} \leftarrow \frac{1}{S}\mathbf{M}$;
  $\hat{Y} \leftarrow$ partition X based upon $\mathbf{M}$;
end

Algorithm 2: Consensus Clustering for Bayesian mixture models

- $X = (x_1, \ldots, x_N)$: the items generated;
- $\pi = (\pi_1, \ldots, \pi_K)$: the expected proportions of the population belonging to each cluster;
- $c = (c_1, \ldots, c_N)$: the allocation variable for each item;
- $\theta = (\theta_1, \ldots, \theta_K)$: the parameters associated with each component; and
- $\phi = (\phi_1, \ldots, \phi_P)$: the indicator variable of feature relevance.

The data generating model is a finite mixture model with independent features. Within this model there exist "irrelevant features" (Law *et al.*, 2003) that have global parameters rather than component specific parameters:

$$\begin{aligned} p(x, c, \theta, \pi) &= \prod_{i=1}^{N} p(x_i|c_i, \theta_{c_i}) \prod_{i=1}^{N} p(c_i|\pi)p(\pi)p(\theta) \\ &= \prod_{i=1}^{N} \prod_{p=1}^{P} p(x_{ip}|c_i, \theta_{c_i p})^{(1-\phi_p)} p(x_{ip}|\theta_p)^{\phi_p} \times \\ &\quad \prod_{i=1}^{N} p(c_i|\pi)p(\pi)p(\theta) \end{aligned}$$

In the simulation study described here, the model is a mixture of *Gaussian* distribution and thus $\theta_{kp} = (\mu_{kp}, \sigma^2_{kp})$. The prior distributions used on the mixture parameters are

$$\pi \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \qquad \mu_{kp} \sim \mathcal{N}(\xi, \kappa), \qquad \sigma^2 \sim \Gamma^{-1}(a, b).$$

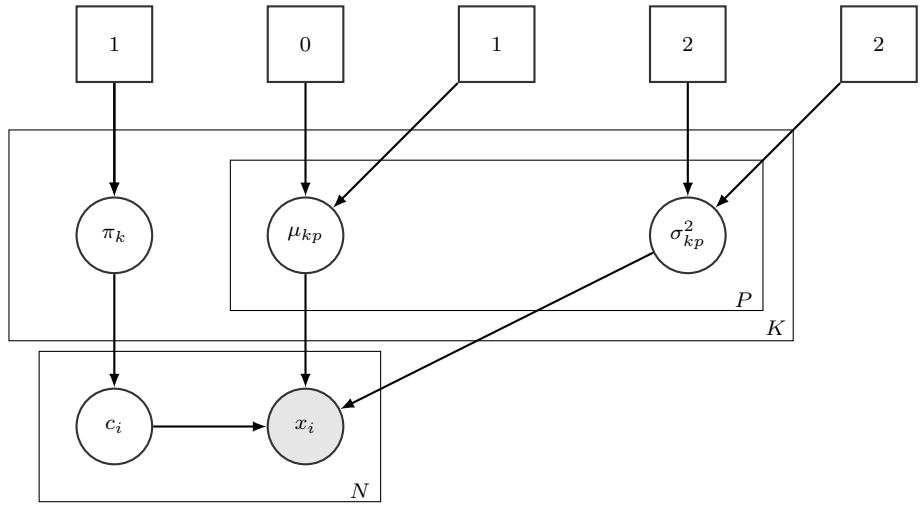

Fig. 1. Directed acyclic graph for the Bayesian mixture of Gaussians used.

The conditional independences and the priors used are shown in the directed acyclic graph (**DAG**) in figure 1. The total number of occupied and empty components is set to $K_{max} = 50$. This and the choice of priors are the defaults in the software provided by Mason *et al.* (2016).

Data is generated using algorithm 3.

We test different scenarios that change various parameters in this model. The scenarios tested and there defining parameters are shown in table 1. In this case the number of relevant features (P_s) is $\sum_p \phi_p$, and $P_n = P - P_s$.

3.4 Performance quantification

We use the Adjusted Rand Index as our metric for the quality of the point clustering inferred by each method, comparing this estimate with the generating labels produced by algorithm 3. This is a measure of "predictive performance", the ability of the methods to infer a single partition that and its similarity to the truth. We also attempt to summarise the uncertainty quantification from each by computing the Frobenius Norm between the true coclustering matrix and the

- consensus matrix for consensus clustering,
- posterior similarity matrix for the Bayesian inference, and
- coclustering matric for `mclust`.

The Frobenius Norm will provide some information if the above matrices correspond at all to the true coclustering matrix, but if no method has performed well then this Norm will reward the *singleton solution* wherein all items are allocated to individual clusters. This means that a visual inspection of the PSMs and CMs is also required. As `mclust` provides only a point estimate the ARI between this and the truth will contain the required information.

The runtime of each MCMC chain is calculated using the terminal command `time`, measured in milliseconds.

3.5 Bayesian model convergence

Within chain Geweke Across chain

Input: Distance between means Δ_μ
A common standard deviation σ^2
A number of clusters K
The number of items to generate in total N
The number of features to generate in total P
An indicator vector of feature relevance $\phi = (\phi_1, \ldots, \phi_P)$
The expected proportion of items in each cluster
$\pi = (\pi_1, \ldots, \pi_K)$
A method for sampling x times from the array y, with weights π: *Sample*(y, x)
A method for permuting a vector x: *Permute*(x)
A method for generating a value from a univariate Gaussian distribution with mean μ and standard deviation σ^2: *Gaussian*(μ, σ^2)
Output: A dataset, X
The generating cluster labels $c = (c_1, \ldots, c_N)$
begin
  `/* initialise the empty data matrix */`
  $X \leftarrow 0_{N \times P}$;
  `/* create a matrix of` K `means */`
  $\mu \leftarrow (\Delta_\mu, \ldots, K\Delta_\mu)$;
  `/* generate the allocation vector */`
  $c \leftarrow$ *Sample*$(1 : K, N, \pi)$;
  $\mathbf{M} \leftarrow \mathbf{0}_{N \times N}$;
  for $p = 1$ **to** P **do**
    `/* Test if the feature is relevant, if relevant generate data from a mixture of univariate Gaussians, otherwise draw all items from the same distribution */`
    if $\phi_p = 1$ **then**
      $\nu \leftarrow$ *Permute*(μ);
      for $n = 1$ **to** N **do**
        $X(n, p) \leftarrow$ *Gaussian*(ν_{c_n}, σ^2)
      end
    end
    if $\phi_p = 0$ **then**
      for $n = 1$ **to** N **do**
        $X(n, p) \leftarrow$ *Gaussian*$(0, \sigma^2)$
      end
    end
  end
  `/* Mean centre and scale the data */`
  $X \leftarrow Normalise(X)$
end

Algorithm 3: Data generation for a mixture of Gaussian with independent features.

4 Examples

We compare consensus clustering of Bayesian mixture models to a traditional inference using several long chains of 1 million iterations (thinning to every thousandth) and an maximum-likelihood estimator as implement in the R package `mclust` Scrucca *et al.*, 2016. These are compared within a range of simulations,

- a low-dimensional dataset,
- a wide dataset representative of the *small N, large P* paradigm prevalent in genetics, and
- a dataset with a large number of irrelevant features.

Table 1. Parameters defining the simulation scenarios as used in generating data and labels. Results for the Simple 2D, the first Small N, large P and final Irrelevant features scenarios are shown in this report, please see the supplementary material for additional results.

Scenario	N	P_s	P_n	K	Δ_μ	σ^2	π
Simple 2D	100	2	0	5	3.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
No structure	100	0	2	1	0.0	1	1
Base Case	200	20	0	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Standard deviation	200	20	0	5	1.0	3	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Standard deviation	200	20	0	5	1.0	5	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	10	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	20	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	100	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Varying proportions	200	20	0	5	1.0	1	$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16})$
Varying proportions	200	20	0	5	0.4	1	$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16})$
Small N, large P	50	500	0	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Small N, large P	50	500	0	5	0.2	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$

The first of these is expected to be the setting where `mclust` and the individual long chains behave well. In the other simulations increasing dimensionality means that mixing problems can emerge and the chains become liable to being trapped in individual modes. Within each simulation 100 datasets are generated following algorithm 3. To each of these the following are run:

- `mclust` for a range of possible K,
- 10 chains of 1 million iterations, thinning to every thousandth sampel for the overfitted Bayesian mixture model, and
- a variety of consensus clustering ensembles defined by inputs of S chains and R iterations within each chain (see algorithm 2) with $S \in \{1, 10, 30, 50, 100\}$ and $R \in \{1, 10, 100, 1000, 10000\}$.

Results

4.1 Simulations

In the 2D dataset `mclust` outperforms the Bayesian inference and consensus clustering in terms of the point estimate of the generating structure.

- Simple example - Mclust and Bayesian win! We see MCMC exploring well.
- Small N large P, problems for Bayesian. Separate modes represented in CM, so consensus is finding sensible modes
- Irrelevant 100, many modes, Mclust collapse, similar to above.
- Real data

Multiple dataset

- MDI yeast - pretty identical
- Full yeast - Bayesian fails, consensus succeeds
- Cancer - hmmmmmm

- for bulleted list, use itemize
- for bulleted list, use itemize
- for bulleted list, use itemize

Figure 2 shows that the above method **?**

Table 2. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

Fig. 2. Caption, caption.

4.2 This is subheading

4.2.1 This is subsubheading

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

4.3 Test1

5 Discussion

6 Conclusion

(Table 2) Text Text Text Text Text Text Text Text Text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Acknowledgements

Text Text Text Text Text Text Text Text. nt to know about text text text text

Funding

This work has been supported by the... Text Text Text Text.

References

Chandra, N. K. *et al.* (2020). Bayesian clustering of high-dimensional data. *arXiv preprint arXiv:2006.02700*.

Emmert-Streib, F. *et al.* (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, **2**, 38.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, **21**, 768–769.

Fritsch, A. (2012). *mcclust: Process an MCMC Sample of Clusterings*. R package version 1.0.

Fritsch, A. *et al.* (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis*, **4**(2), 367–391.

Ghaemi, R. *et al.* (2009). A survey: clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, **50**, 636–645.

Ghaemi, R. *et al.* (2011). A review: accuracy optimization in clustering ensembles using genetic algorithms. *Artificial Intelligence Review*, **35**(4), 287–318.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.

Kiselev, V. Y. *et al.* (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, **14**(5), 483–486.

Law, M. H. *et al.* (2003). Feature selection in mixture-based clustering. In *Advances in neural information processing systems*, pages 641–648.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, **28**(2), 129–137.

Mason, S. A. *et al.* (2016). Mdi-gpu: accelerating integrative modelling for genomic-scale data using gp-gpu computing. *Statistical applications in genetics and molecular biology*, **15**(1), 83–86.

Monti, S. *et al.* (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, **52**(1-2), 91–118.

Robert, C. P. *et al.* (2018). Accelerating mcmc algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, **10**(5), e1435.

Scrucca, L. *et al.* (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 289–317.

Wilkerson *et al.* (2010). Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, **26**(12), 1572–1573.

Yao, Y. *et al.* (2020). Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *arXiv preprint arXiv:2006.12335*.