Discussion for "Bayesian Cluster Analysis: Point Estimation and Credible Balls" by Sara Wade

Clara Grazian

Universitá degli Studi "Gabriele D'Annunzio"

clara.grazian@unich.it

OBayes 2019

Warwick

29 June 2019

Literature

Based on the paper:

"Bayesian Cluster Analysis: Point Estimation and Credible Balls"

by **Sara Wade and Zoubin Ghahramani**

appeared on Bayesian Analysis in 2018....

Literature

Based on the paper:

"Bayesian Cluster Analysis: Point Estimation and Credible Balls"

by **Sara Wade and Zoubin Ghahramani**

appeared on Bayesian Analysis in 2018....

...with discussion!

The approach

The method proposed in Wade & Gharamani (2018) has the goal of proposing summary estimates (both point and interval estimates) about the **partitions** in an inifinite mixture by using (for example) the standard Dirichlet Process prior.

- Estimation is proposed within a **decision-theoretic** context, by using the Binder's loss and the variation of information (VI) loss.
- The first part of the paper is a review of **Meilă [2007]**, about the properties of these two loss functions in the setting of clustering
- **Credible balls** and point estimates are proposed
- A **greedy search algorithm** is applied in order to investigate the huge partition space

Some points already discussed

1. the bounds may consist of more than one partition (but in practice, this rarely happens) (*Monni*)
2. other definitions of the balls are possible: HPD (*Friel and Rastelli*), "frequentist", entropy (*Nipoti and Shen*)
3. alternatives to the greedy search algorithm: Rastelli & Friel [2017] propose an algorithm which does not need any likelihood approximation
4. other applications: finite mixture models (*Früwirth-Schnatter, Grün, and Malsiner–Walli*) or DAGs (*Castelletti and Peluso*)

"It's the question that drives us"

1) What is the relationship between

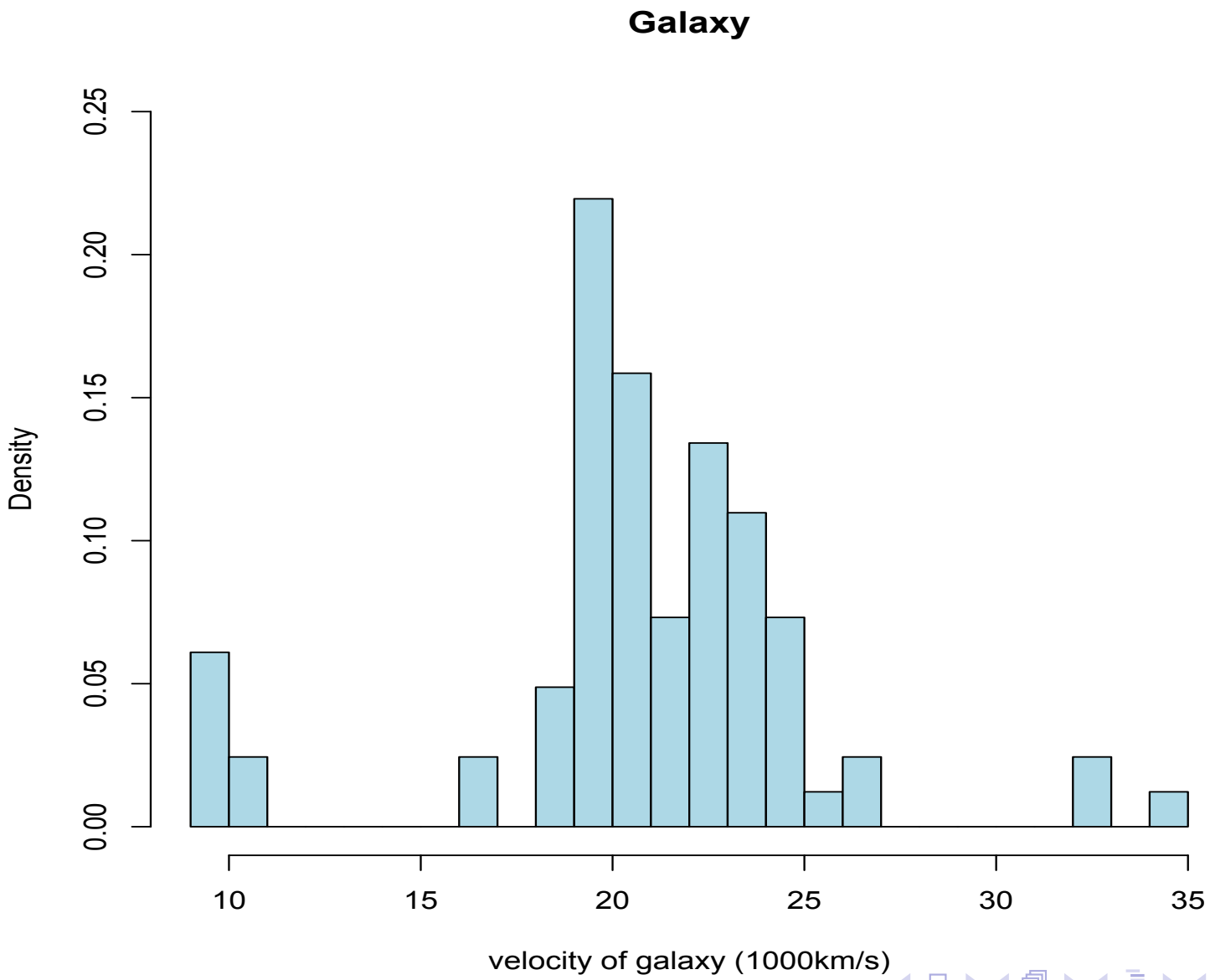
estimation of the number of clusters

and

estimation of the partition

in the decision-theoretic approach?

The galaxy dataset



The question in mind

Until the 90s, it was believed that the galaxy consisted of two stellar populations (the *disk* and the *halo*).

At a certain point, someone hypothesized that there are three stellar populations, (the *thin disk*, the *thick disk* and the *halo*) distinguished by their

- spatial distributions
- velocities
- metallicities

These theories are important because it has different implications in the formation of the Galaxy.

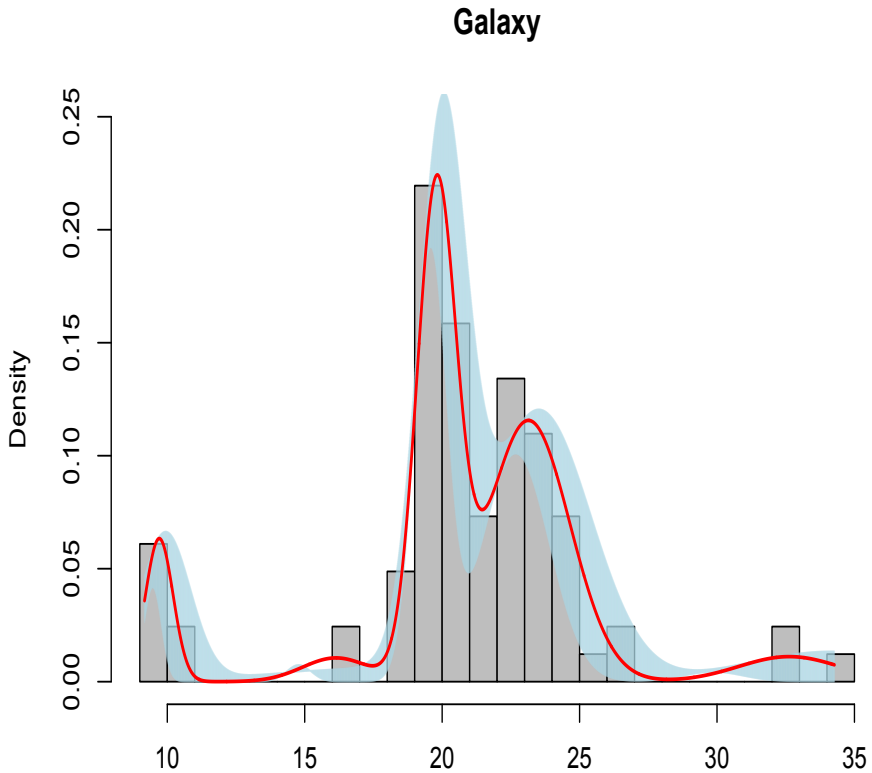
The question is always central clustering: examples in genomics [Armstrong et al., 2001] or networks [Tangari et al., 2019]

When K is known

In overfitted mixtures, it is necessary to define a prior distribution on the mixture weights with certain characteristics [Rousseau & Mengersen, 2011]

Jeffreys prior and $K = 10$

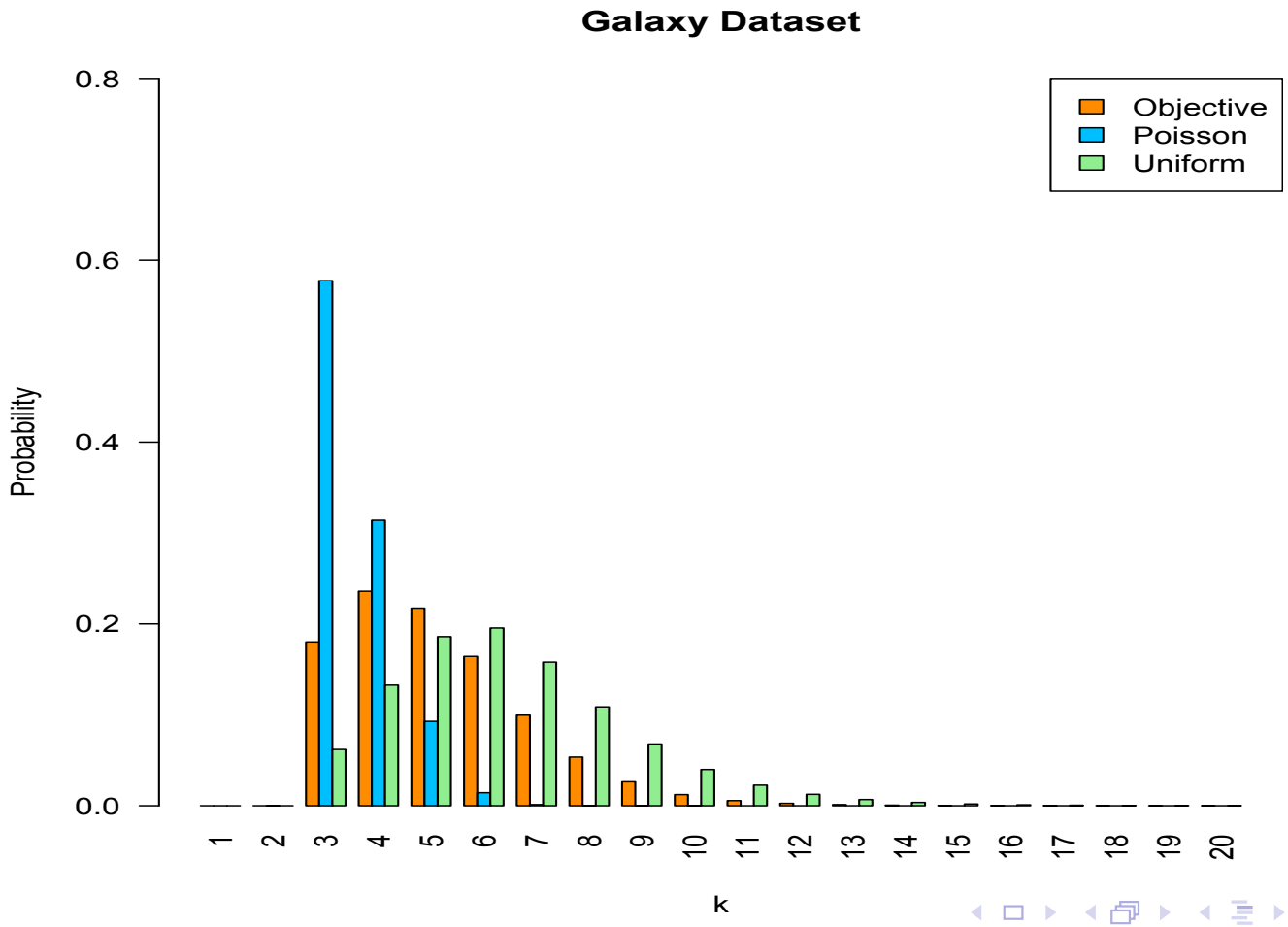
[Grazian & Robert, 2018]



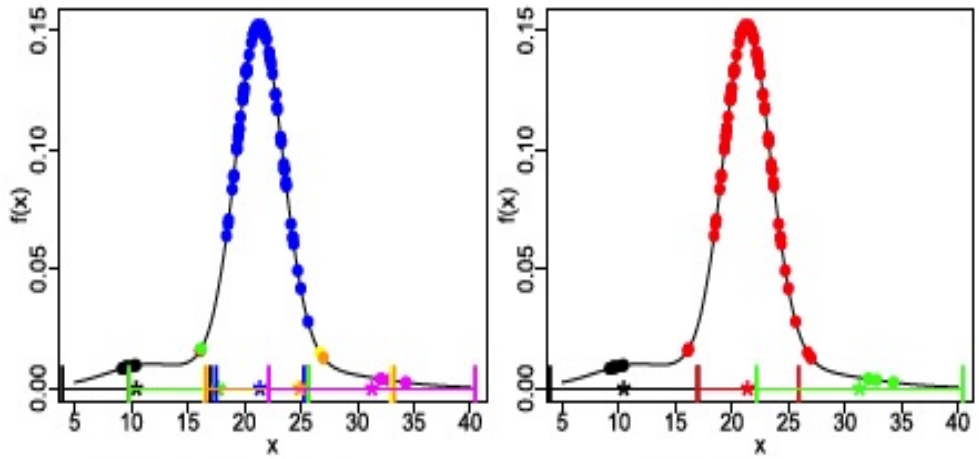
Dataset	Galaxy
p_1	0.437
	(23.139, 1.507)
p_2	0.390
	(19.790, 0.715)
p_3	0.080
	(9.709, 0.503)
p_4	0.056
	(32.630, 1.842)
p_5	0.037
	(16.138, 1.226)
$\sum_{\ell=6}^{10} p_\ell$	0.000

When K is unknown

When the number of components is considered unknown and has its own prior distribution, some conservativeness is welcome
[Grazian, Villa & Liseo, 2018], [Nobile, 2004]



Nonparametric estimation based on VI loss



(a) Binder's loss: 7 clusters (b) VI: 3 clusters

Figure 10: Galaxy example: optimal clustering estimate with color representing cluster membership for Binder's loss and VI, with correspondingly colored stars and bars along the x-axis representing the posterior mean and variance within cluster.

	Upper		Lower		Horizontal	
	k_N^u	$d(\mathbf{c}^*, \mathbf{c}_u)$	k_N^l	$d(\mathbf{c}^*, \mathbf{c}_l)$	k_N^h	$d(\mathbf{c}^*, \mathbf{c}_h)$
Galaxy	2	1.364	15	1.669	8	1.832

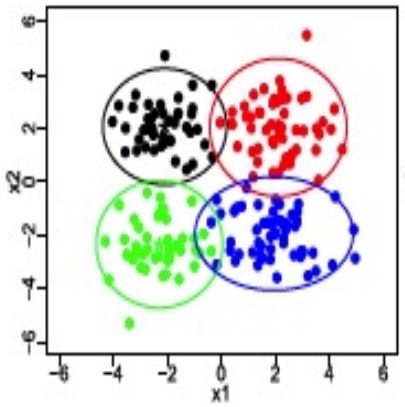
What happens with other priors?
Gnedin [2010], Bassetti, Casarin & Rossini [2018]

Applications

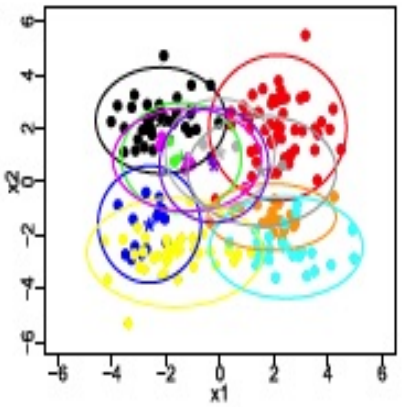
A standard question practitioners ask is:

Are these two observations in the same cluster?

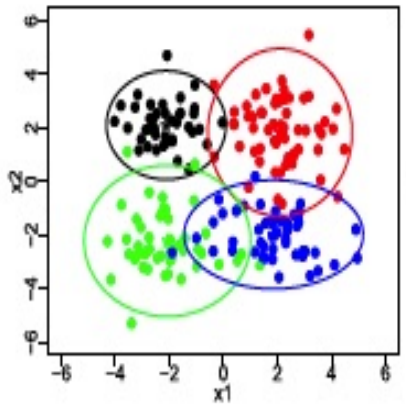
What is the questions answered by using the VI loss?



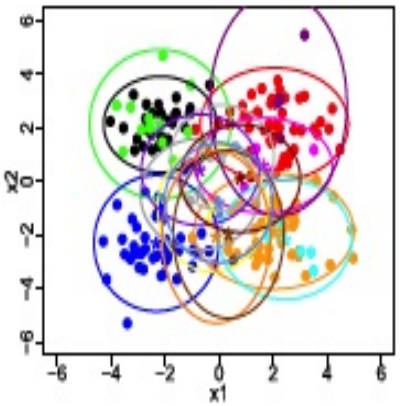
(a) VI estimate: 4 clusters



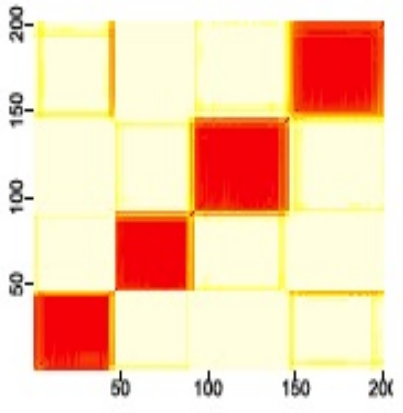
(b) VI horizontal bound: 11 clusters



(c) VI upper vertical bound: 4 clusters



(d) VI lower vertical bound: 16 clusters



(e) Posterior similartiy matrix

Thanks!

Bibliography I

Armstrong, S. et al. (2001) MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. *Nature Genetics*, 30, 41–47.

Bassetti, F., Casarin, R. and Rossini, L. (2018) Hierarchical Species Sampling Models *arXiv:1803.05793v1*.

Gnedin, A. (2010) A species sampling model with infitely many types *Electron. Commun. Probab.*, 15, 79-88.

Grazian, C., Villa, C., and Liseo, B. (2018) On a Loss-based prior for the number of components in mixture models. *arXiv:1807.07874*.

Grazian, C., and Robert, C.P. (2018) Jeffreys priors for mixture estimation: properties and alternatives. *Computational Statistics and Data Analysis*, 121: 149–163.

Meilă, M. (2007) Comparing clusterings – an information based distance. *JMA*, 98: 873–895.

Nobile, A. (2004) On the posterior distribution of the number of components in a finite mixture. *Ann. of Statist.* **32**, 2044–2073

Rastelli, R. and Friel, N. (2017) Modeling with normalized random measure mixture models. *Statistics and Computing*.

Rousseau, J., Mengersen, K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models *JRSS - B*, 73 (5), 689–710.

Tangari, G., Charalambides, M., Tuncer, D., Grazian, C., Pavlou, G. (2019) Accuracy-Aware Adaptive Traffic Monitoring for Software Dataplanes. *submitted*.