DISCUSSION ON

Inferring the number of components in a mixture: dream or reality?

by prof. Christian Robert

Emanuele Aliverti,
University Ca' Foscari Venezia, Department of Economics
emanuelealiverti.github.io

"Statistical methods and models for complex data"

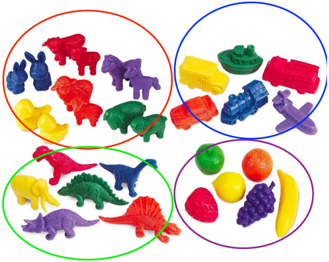
Università Ca'Foscari Venezia

Mixtures for clustering

- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**),

Mixtures for clustering

- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**),



Mixtures for clustering

- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**), by taxonomy (**2 clusters**)...

Mixtures for clustering

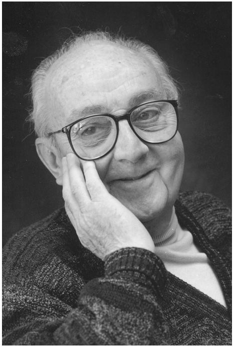
- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**), by taxonomy (**2 clusters**)...
- Each arrangement is potentially correct as there is no ground **“truth”**

Mixtures for clustering

- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**), by taxonomy (**2 clusters**)...
- Each arrangement is potentially correct as there is no ground **“truth”**
- Therefore, estimating the **number of groups** in the sample (or in the population) or the **number of components** is particularly difficult
- Also, recall that these estimates can be different (e.g., McCullagh and Yang, 2008; Miller and Harrison, 2018)

Mixtures for clustering

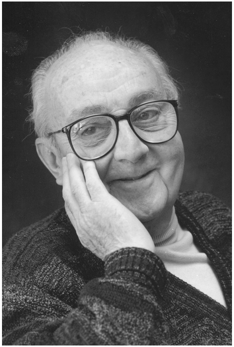
- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**), by taxonomy (**2 clusters**)...
- Each arrangement is potentially correct as there is no ground **“truth”**
- Therefore, estimating the **number of groups** in the sample (or in the population) or the **number of components** is particularly difficult
- Also, recall that these estimates can be different (e.g., McCullagh and Yang, 2008; Miller and Harrison, 2018)



  - Clearly “all models are **wrong**"...

Mixtures for clustering

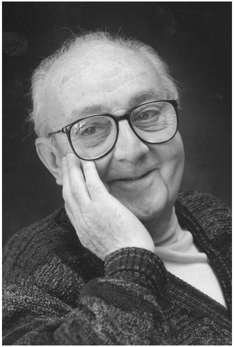
- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**), by taxonomy (**2 clusters**)...
- Each arrangement is potentially correct as there is no ground **"truth"**
- Therefore, estimating the **number of groups** in the sample (or in the population) or the **number of components** is particularly difficult
- Also, recall that these estimates can be different (e.g., McCullagh and Yang, 2008; Miller and Harrison, 2018)



- Clearly "all models are **wrong**"... but mixtures can be "more wrong" than other parametric models, as clusters sometimes are "purely notional" (e.g. Miller and Harrison, 2018; Hennig, 2015)

Mixtures for clustering

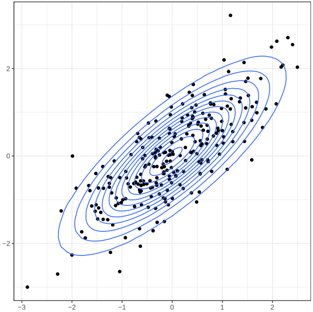
- Clustering is inherently an **ill-posed** problem
- For example: cluster by color (**6 clusters**), by "species" (**4 clusters**), by taxonomy (**2 clusters**)...
- Each arrangement is potentially correct as there is no ground **"truth"**
- Therefore, estimating the **number of groups** in the sample (or in the population) or the **number of components** is particularly difficult
- Also, recall that these estimates can be different (e.g., McCullagh and Yang, 2008; Miller and Harrison, 2018)



- Clearly "all models are **wrong**"... but mixtures can be "more wrong" than other parametric models, as clusters sometimes are "purely notional" (e.g. Miller and Harrison, 2018; Hennig, 2015)
- When we analyze **real data**, this aspect should not be ignored: in particular when we improperly advocate theorems that are developed under **different** conditions (e.g., the "true" distribution is not a mixture, ignore sample size...)

Mixtures for density estimation

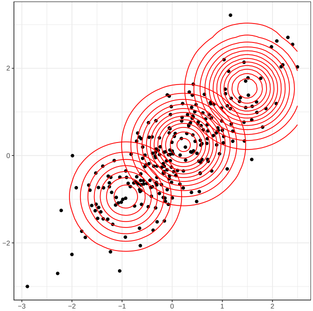
- Mixture models are also widely used for **density estimation** and **density regression**



Mixtures for density estimation

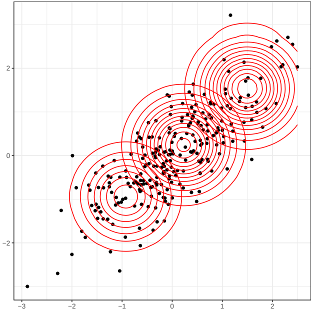
Università Ca'Foscari Venezia

- Mixture models are also widely used for **density estimation** and **density regression**
- A mixture distribution function can be made **arbitrarily close** to **any density**, allowing the number of components to grow (Epanechnikov, 1969)



Mixtures for density estimation

- Mixture models are also widely used for **density estimation** and **density regression**
- A mixture distribution function can be made **arbitrarily close** to **any density**, allowing the number of components to grow (Epanechnikov, 1969)
- Large p: for computational simplicity, little structure is imposed **within** each component (often conditional independence among variables, given cluster membership)



Mixtures for density estimation

- Mixture models are also widely used for **density estimation** and **density regression**
- A mixture distribution function can be made **arbitrarily close** to **any density**, allowing the number of components to grow (Epanechnikov, 1969)
- Large p: for computational simplicity, little structure is imposed **within** each component (often conditional independence among variables, given cluster membership)
- Under a naive approach, this might require **a lot** of components to characterize well enough a complex structure

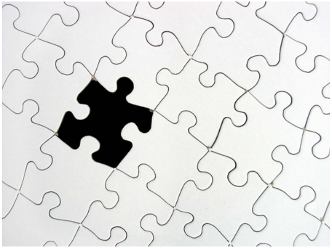
Mixtures for density estimation

- Mixture models are also widely used for **density estimation** and **density regression**
- A mixture distribution function can be made **arbitrarily close** to **any density**, allowing the number of components to grow (Epanechnikov, 1969)
- Large p: for computational simplicity, little structure is imposed **within** each component (often conditional independence among variables, given cluster membership)
- Under a naive approach, this might require **a lot** of components to characterize well enough a complex structure
  - Some **tentatives**: reduce the number of **required components** including more **structure** within each sub-population adaptively, also improving interpretation of the clusters and components
  - non-trivial outside Gaussian world, such as **categorical data**, mixtures of multinomials, latent class models, and many others

Algorithmic issues

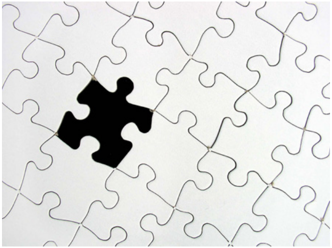
Università Ca'Foscari Venezia

- Data augmentation: a **blessing** or a **curse**?



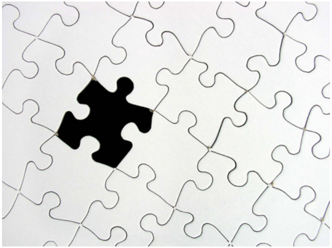
Algorithmic issues

- Data augmentation: a **blessing** or a **curse**?
- We love conditional conjugacy (Gibbs Sampling). But at what **price**? Need to introduce and **update** $\mathcal{O}(n)$ additional latent variables even when the number of parameters is **much smaller** (and sometimes we're not even interested in those augmented data)



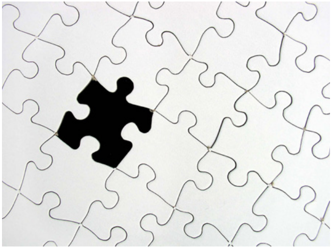
Algorithmic issues

- Data augmentation: a **blessing** or a **curse**?
- We love conditional conjugacy (Gibbs Sampling). But at what **price**? Need to introduce and **update** $\mathcal{O}(n)$ additional latent variables even when the number of parameters is **much smaller** (and sometimes we're not even interested in those augmented data)
- This problem affects mixtures as well as other standard approaches (e.g., binary regression)



Algorithmic issues

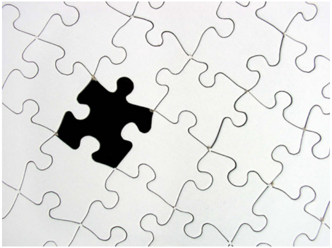
- Data augmentation: a **blessing** or a **curse**?
- We love conditional conjugacy (Gibbs Sampling). But at what **price**? Need to introduce and **update** $\mathcal{O}(n)$ additional latent variables even when the number of parameters is **much smaller** (and sometimes we're not even interested in those augmented data)
- This problem affects mixtures as well as other standard approaches (e.g., binary regression)



- What can we do: find a compromise between **algebraic convenience** and **scalability**, eventually approximating our posterior

Algorithmic issues

- Data augmentation: a **blessing** or a **curse**?
- We love conditional conjugacy (Gibbs Sampling). But at what **price**? Need to introduce and **update** $\mathcal{O}(n)$ additional latent variables even when the number of parameters is **much smaller** (and sometimes we're not even interested in those augmented data)
- This problem affects mixtures as well as other standard approaches (e.g., binary regression)



- What can we do: find a compromise between **algebraic convenience** and **scalability**, eventually approximating our posterior
- For example: some algorithms update only **subsets** of latent variables (e.g., stochastic variational inference; Hoffman et al., 2013), or specify a more **tractable representation** of the latent component to conduct approximate inference (e.g., Daniele's talk)

References

- Epanechnikov, Vassiliy A. (1969). “Non-parametric estimation of a multivariate probability density”. In: *Theory of Probability & Its Applications*.
- Hennig, Christian (2015). “What are the true clusters?” In: *Pattern Recognition Letters*.
- Hoffman, Matthew D et al. (2013). “Stochastic variational inference”. In: JMLR.
- McCullagh, Peter and Jie Yang (2008). “How many clusters?” In: *Bayesian Analysis*.
- Miller, Jeffrey W. and Matthew T. Harrison (2018). “Mixture models with a prior on the number of components”. In: JASA.

References

- Epanechnikov, Vassiliy A. (1969). “Non-parametric estimation of a multivariate probability density”. In: *Theory of Probability & Its Applications*.
- Hennig, Christian (2015). “What are the true clusters?” In: *Pattern Recognition Letters*.
- Hoffman, Matthew D et al. (2013). “Stochastic variational inference”. In: JMLR.
- McCullagh, Peter and Jie Yang (2008). “How many clusters?” In: *Bayesian Analysis*.
- Miller, Jeffrey W. and Matthew T. Harrison (2018). “Mixture models with a prior on the number of components”. In: JASA.

Thanks to prof. Robert for the **wonderful** talk..
and thanks for your attention!