An enriched mixture model for functional clustering

Tommaso Rigon

PwC - 2023-06-07

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

A bit about myself

- (2010 - 2013) **B.Sc. in Statistics, Economics & Finance**
- (2013 - 2015) **M.Sc. in Statistical Sciences**
- (2015 - 2019) **Ph.D. in Statistics**
- (2019 - 2020) **Research/Postdoctoral Associate**
- (2020 - Present) **Assistant Professor of Statistical Sciences**

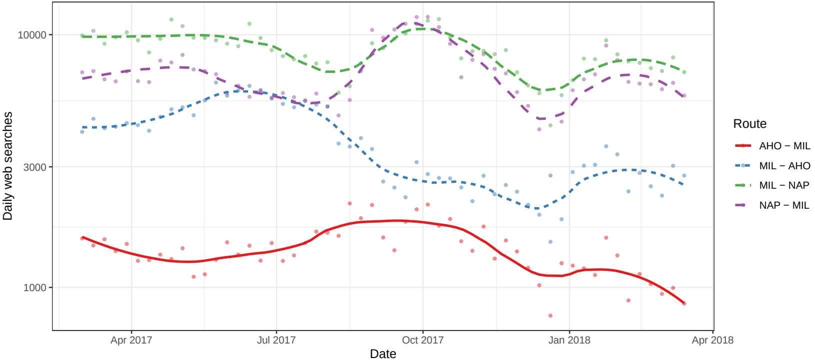
Customer segmentation: a case study

- A private company selling flight tickets is instered in understading its customers' preferences and need.
- In this case study, each statistical unit is a **flight route**, i.e. the number of times that a specific route has been searched on the website of an **e-commerce** company.
- We aim at **clustering** functional observations to perform **market segmentation**.

Statistical challenges

- **Functional data**. Data points are functions (time series in this case), so traditional algorithms (e.g. k-means) cannot/should not be directly applied.
- **Bounding the complexity**. We do not want too many clusters, but at the same time we would like to automatically identify some optimal number.
- **Constrained estimation**. Prior knowledge about the shapes of the functions is available, but it is not easy to incorporate.

The e-commerce dataset



- The total number of flight routes is $n = 214$.
- Each trajectory is observed over a **weekly time grid** $\boldsymbol{t}_i = (1, \ldots, 55)$. Hence, the dataset can be represented as a 214×55 matrix with 11770 entries.

Preliminary considerations

Why do focus on web-searches?

- Different and potentially more interesting **metrics** could be considered.

- However, private companies are (rightly!) worried about disclosing their data.

- In principle, other metrics might include:
  - Route prices;
  - Route marginal earnings;
  - Route-specific customer satisfaction;
  - Conversion rates;
  - ...

Clustering average levels vs clustering shapes

- A very crude but operative summary of each time series is its **average**. Market segmentation according to the average could be useful, but it is not the focus here.

- Missing part of the story (this talk): **clustering shapes** and not average levels.

Mixture models I

- Functional observations are **standardized**, i.e. they have zero mean and unit variance.

- The **clustering** method is **model-based**: not just an algorithm!

- The **model** we assume is:

$$y_i(t) = f_i(t) + \epsilon_i(t), \qquad i = 1, \dots, n,$$

  where $\epsilon_i(t)$ is a Gaussian error with variance σ^2 and $t \in \mathbb{R}^+$.

- Clustering is induced through a **discrete distribution** $\tilde{p}$ for the latent trajectories $f_i(t)$, namely

$$(f_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} = \sum_{h=1}^{H} \xi_h \delta_{\phi_h}, \qquad i = 1, \dots, n.$$

- Two functional observations $y_i(t)$ and $y_j(t)$ both belong to the ***h*th group** whenever they share the same latent trajectory, that is $f_i(t) = f_j(t) = \phi_h(t)$.

Mixture models II

- The **mixture model** of the previous slide can be expressed in an equivalent manner.
- The random variable $S_i \in \{1, \dots, H\}$ is an unknown **cluster indicator**, so that $f_i(t)$ and $f_j(t)$ belong to the same group if $S_i = S_j$.
- **Generative step 1**. Sample the cluster indicators from

$$\mathbb{P}(S_i = h) = \xi_h, \qquad i = 1, \dots, n.$$

- **Generative step 2**. Suppose that $S_i = h$ and assign to the ith observation the latent function $\phi_h(t)$. Then, sample the data points $y_i(t)$ from a $\mathcal{N}(\phi_h(t), \sigma^2)$.
- **Clustering step**. Using Bayes theorem, we obtain the distribution of

$$\mathbb{P}(S_i = h \mid y_1(t), \dots, y_n(t)) = \tilde{\xi}_h = \frac{\xi_h \prod_{s=1}^{T_i} \mathcal{N}(y_i(t_{is}); \phi_h(t_{is}), \sigma^2)}{\sum_{h=1}^{H} \xi_h \prod_{s=1}^{T_i} \mathcal{N}(y_i(t_{is}); \phi_h(t_{is}), \sigma^2)},$$

from which we obtain out clustering solution.

Mixture models III

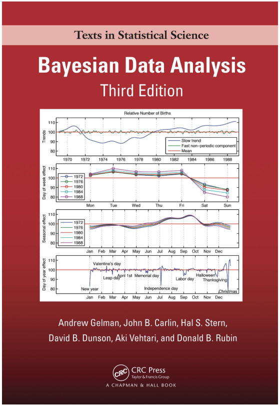
Why model-based clustering?

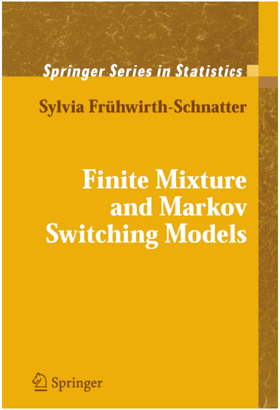
- The underlying **assumptions** are often much more **transparent**.
- Functional observations are noisy and this requires **smoothing**; however, we want to avoid two-step procedures.
- Probabilistic method. For example, you could compute the probability that two observations belong to the same group and/or **estimate** the **number of clusters**.

Why Bayesian?

- It's a natural choice for mixture model, being based on a **data augmentation**.
- You can easily **incorporate prior information**, which is often available, and/or **control the complexity** of the estimates in a natural fashion (prior penalty).
- The estimation of H can be performed together with the estimation of the other parameters: i.e. you need to fit only a single model.

Source material





- **CRAN task view**: `https://cran.r-project.org/web/views/Cluster.html`

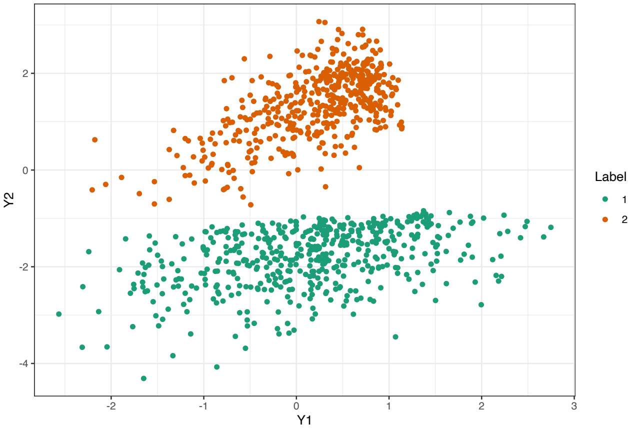
Learning the number of clusters

Normal deviate: Larry Wasserman's blog

"I have decided that mixtures, like tequila, are inherently evil and should be avoided at all costs."

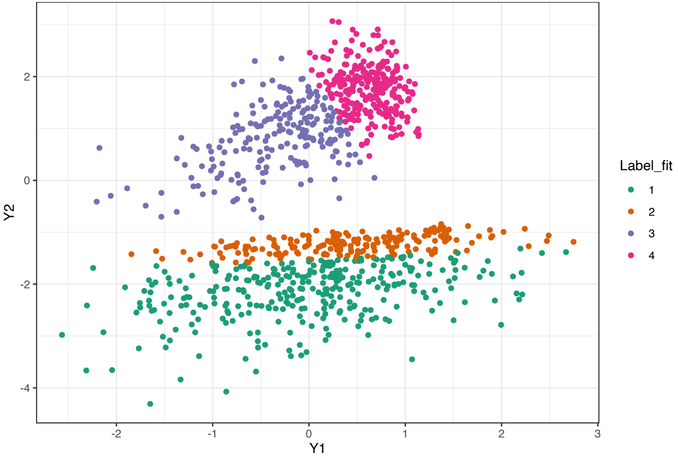
- Mixture models are **powerful** but **delicate** tools.
- Reliably learning the number of clusters has entertained a generation of statisticians!
- **Caveat**. The **number of clusters** K_n does not coincide with the number of components H. The quantity $K_n \le H$ is the number of **non-empty** groups among the cluster indicators.
- This is quite evident in Bayesian nonparametrics, where could have $H = \infty$.
- Can we learn the **true number of clusters** H_0 from the data? **Yes**, but under many **assumptions** and being very careful to prior choices, identifiability issues, etc.

Overclustering and misspecification I



- Data displayed above are the "true labels".
- If the **kernel** is wrong, the estimation of K_n using a mixture model is unreliable.

Overclustering and misspecification II



- In practice, one often get **too many clusters**, compared to H_0. This is exacerbated in high-dimensional settings when misspecifications are more likely to occur.

Better kernels?

- If the multivariate Gaussian kernel is inappropriate, can't we use something else? Yes, but that's not easy!

- Parametric choices (e.g., skew-normals, etc.) may **mitigate** the problem and/or protect against outliers, often at the price of increasing the computational burden.

- What about **nonparametric kernels**? **Mixture of mixtures** are fully nonparametric models, but some serious **identifiability** difficulties must be addressed.

References

- Mukhopadhyay, M., Li, D., & Dunson, D. B. (2020). Estimating densities with non-linear support by using Fisher–Gaussian kernels. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(5), 1249–1271.
- Scarpa, B., & Dunson, D. B. (2014). Enriched stick-breaking processes for functional data. *Journal of the American Statistical Association*, 109(506), 647–660.
- Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2), 285–295.

An enriched discrete prior

- Let's get back to the original clustering problem for **functional data**.
- The proposed process is a **mixture of mixtures**:

$$\tilde{p} = \sum_{\ell=1}^{L} \Pi_\ell \tilde{p}_\ell = \sum_{\ell=1}^{L} \Pi_\ell \sum_{h=1}^{H_\ell} \pi_{\ell h} \delta_{\theta_{\ell h}(t)}, \quad \theta_{\ell h}(t) \overset{\text{ind}}{\sim} P_\ell,$$

  for $h = 1, \ldots, H_\ell$ and $\ell = 1, \ldots, L$.
- Each P_ℓ is a diffuse probability measure taking values on a given **functional class** (monotone, cyclical, linear, S-shaped functions, etc).
- This is closely related to the **enriched processes** of Wade et al. (2011) and Scarpa and Dunson (2014), but the number of clusters is **bounded**.

A nested clustering process

- The random variable $G_i \in (\ell, h)$ is a latent **cluster indicator**, so that $f_i(t)$ and $f_j(t)$ belong to the same group if $G_i = G_j$.

- The random variable $F_i \in \{1, \dots, L\}$ is a latent **functional class indicator**.

- **Generative step 1.a**. Functional class allocation:

$$\mathbb{P}(F_i = \ell) = \Pi_\ell,$$

- **Generative step 1.b**. Within-class allocation:

$$\mathbb{P}(G_i = (\ell, h) \mid F_i = \ell) = \pi_{\ell h},$$

  meaning that $\mathbb{P}(G_i = (\ell, h)) = \Pi_\ell \pi_{\ell h}$.

- **Generative step 2**. Suppose that $G_i = (\ell, h)$ and assign to the ith observation the latent function $\theta_{\ell h}(t)$. Then, sample the data points $y_i(t)$ from a $\mathcal{N}(\theta_{\ell h}(t), \sigma^2)$.

Theoretical corner: enriched urn scheme

- The **prior specification** is as in Rousseau and Mengersen (2011), so that

$$(\Pi_1, \dots, \Pi_{L-1}) \sim \text{DIRICHLET}(\alpha_1, \dots, \alpha_L),$$

and

$$(\pi_{\ell 1}, \dots, \pi_{\ell H_\ell - 1}) \overset{\text{ind}}{\sim} \text{DIRICHLET}\left(c_\ell / H_\ell, \dots, c_\ell / H_\ell\right).$$

- Observations can be sampled **sequentially**:

$$\mathbb{P}(F_{n+1} = \ell \mid F^{(n)}) = \frac{\alpha_\ell + n_\ell}{\alpha + n},$$

$$\mathbb{P}(f_{n+1} \in \cdot \mid f^{(n)}, F^{(n)}, F_{n+1} = \ell) = \left(1 - \frac{k_\ell}{H_\ell}\right) \frac{c_\ell}{c_\ell + n_\ell} P_\ell(\cdot) + \sum_{j=1}^{k_\ell} \frac{n_{j\ell} + c_\ell / H_\ell}{c_\ell + n_\ell} \delta_{f^*_{j\ell}}(\cdot),$$

where the notation is as follows:

  - $n_\ell = \sum_{i=1}^n I(F_i = \ell)$ is the number of elements belonging to the ℓth functional class;
  - $k_\ell \leq n_\ell$ is the number of distinct values observed in the ℓth class;
  - $f^*_{11}, \dots, f^*_{1n_1}, \dots, f^*_{L1}, \dots, f^*_{Ln_L}$ are the **distinct functions** in the sample;
  - $n_{j\ell}$ is the frequency of each distinct function.

Baseline measure specification I

- We need to choose a specification for $\theta_{\ell h}(t) \sim P_\ell$.

- Note that Each P_ℓ can be interpreted as a functional prior guess, since

$$\mathbb{E}\{\tilde{p}(\cdot)\} = \sum_{\ell=1}^{L} \mathbb{E}(\Pi_\ell) P_\ell(\cdot) = \frac{1}{\alpha} \sum_{\ell=1}^{L} \alpha_\ell P_\ell(\cdot), \qquad \alpha = \sum_{\ell=1}^{L} \alpha_\ell.$$

- We assume that $\theta_{\ell h}(t)$ is **linear in the parameters**:

$$\theta_{\ell h}(t) = \sum_{m=1}^{M_\ell} \mathcal{B}_{m\ell}(t) \beta_{m\ell h},$$

where each $\mathcal{B}_{1\ell}(t), \dots, \mathcal{B}_{M_\ell \ell}(t)$ for $\ell = 1, \dots, L$ is a set of **pre-specified basis functions** and where $(\beta_{1\ell h}, \dots, \beta_{M_\ell \ell h})^\intercal$ have Gaussian prior.

- For example, you could use **B-splines**, I-splines and related ideas.

Baseline measure specification II

- The first functional class ($\ell = 1$) captures yearly **cyclical patterns** and characterizes the routes having **one peak** of web-searches during either the summer or the winter.

$$\theta_{1h}(t) = \sum_{m=1}^{8} \beta_{m1h}\mathcal{S}_m(t) + \beta_{91h}\cos\left(2\pi\frac{7}{365}t\right) + \beta_{10,1h}\sin\left(2\pi\frac{7}{365}t\right),$$

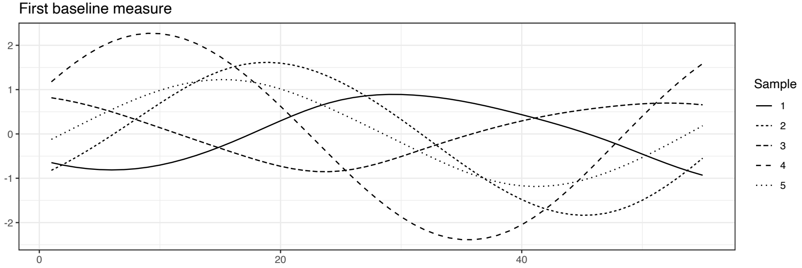
where $\mathcal{S}_1(t), \dots, \mathcal{S}_8(t)$ are deterministic cubic spline basis functions.

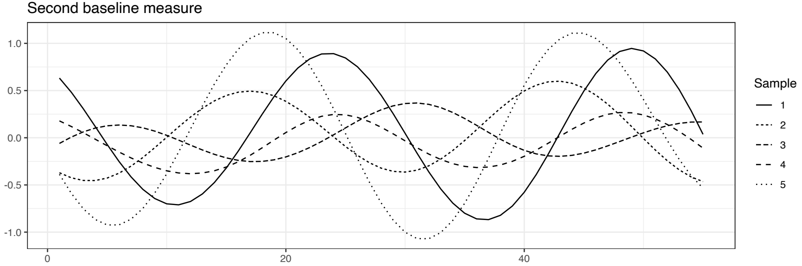
- The second functional class ($\ell = 2$) characterizes functions having **two peaks per year**, which amounts to let

$$\theta_{2h}(t) = \sum_{m=1}^{8} \beta_{m2h}\mathcal{S}_m(t) + \beta_{92h}\cos\left(2\pi\frac{14}{365}t\right) + \beta_{10,2h}\sin\left(2\pi\frac{14}{365}t\right).$$

- We select a **Gaussian prior** with diagonal covariance for $\boldsymbol{\beta}_\ell$ (a.k.a. a **ridge penalty**).

Baseline measure specification III





Variational inference

- Bayesian mixture models are routinely estimated using **Markov Chain Monte Carlo**.
- However, this might be **computationally very expensive** and complicated by the label-switching
- We employ a **mean-field variational approximation** of the posterior distribution, which is easy to get because of conjugacy.
- The variational posterior generally leads to accurate point estimates but also it typically **underestimates the variability**.
- However, in our motivating application we are only interested in a single cluster solution.
- An efficient algorithm (CAVI) is available.

The CAVI algorithm

[1] Update $q(G_i)$ for each $i = 1, \dots, n$;

$$\rho_{i\ell h} \propto \exp\left[\mathbb{E}_q\{\log(\Pi_\ell \pi_{\ell h})\} + \sum_{s=1}^{T_i} \mathbb{E}_q\{\log \mathcal{N}(y_i(t_{is}); \theta_{\ell h}(t_{is}), \sigma^2)\}\right].$$

[2] Update the variational distribution $q(\boldsymbol{\Pi})$ according to

$$q(\boldsymbol{\Pi}) = \textsc{Dirichlet}\left(\boldsymbol{\Pi}; \alpha_1 + \sum_{i=1}^{n}\sum_{h=1}^{H_1} \rho_{i1h}, \dots, \alpha_L + \sum_{i=1}^{n}\sum_{h=1}^{H_L} \rho_{iLh}\right).$$

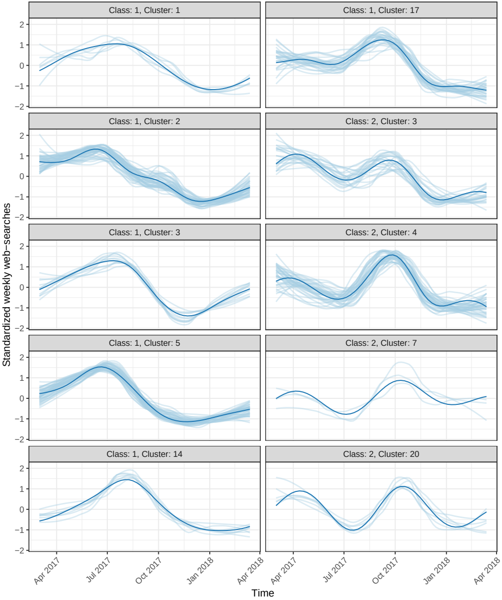
[3] Update $q(\boldsymbol{\pi}_\ell)$ for each $\ell = 1, \dots, L$;

$$q(\boldsymbol{\pi}_\ell) = \textsc{Dirichlet}\left(\boldsymbol{\pi}_\ell; \frac{c_\ell}{H_\ell} + \sum_{i=1}^{n} \rho_{i\ell 1}, \dots, \frac{c_\ell}{H_\ell} + \sum_{i=1}^{n} \rho_{i\ell H_\ell}\right).$$

[4] Update $q(\beta_{\ell h})$ for each $h = 1, \dots, H_\ell$ and $\ell = 1, \dots, L$;

$$q(\beta_{\ell h}) = \mathcal{N}_{M_\ell}\left(\beta_{\ell h}; \tilde{\boldsymbol{\mu}}_{\ell h}, \tilde{\boldsymbol{\Sigma}}_{\ell h}\right).$$

Clustering solution



Macro clusters A and B

Macro-cluster A: labels 2,3,5 ($\ell = 1$)

		Arrival		
		North	Center	South & Islands
Departure	North	0	0	59
	Center	0	0	26
	South & Islands	0	0	13

Macro-cluster B: label 17 ($\ell = 1$) and labels 3,4,7 ($\ell = 2$)

		Arrival		
		North	Center	South & Islands
Departure	North	0	4	2
	Center	9	0	0
	South & Islands	46	22	6

Grazie!

- The proposed model is allows nested clustering of the observations
- The modeling choices reflect a balance between **flexibility** and **pragmatism** in developing an efficient algorithm that can easily handle thousands of data points.
- Crucially, this is because closed-form expressions for the CAVI algorithm and "smart" choices in the model specification.

Main reference

Rigon, T. (2022). An enriched mixture model for functional clustering. *Applied Stochastic Models in Business and Industry*, forthcoming.