Augmented quantization : a general approach to mixture models

Charlie Sire[†,1,2,3], Didier Rullière[§,3], Rodolphe Le Riche[§,3], Jérémy Rohmer[§,2], Yann Richet[§,1], Lucie Pheulpin[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2020 – Sep. 2023

[1] IRSN
{charlie.sire, yann.richet, lucie.pheulpin}@irsn.fr

[2] BRGM
J.Rohmer@brgm.fr

[3] LIMOS: CNRS, Mines Saint-Etienne and Univ. Clermont Auvergne
{drulliere,leriche}@emse.fr

Abstract

Quantization methods classically provide a discrete representation of a continuous set. This type of representation is relevant when the objective is the visualisation of weighted prototype elements representative of a continuous phenomenon. Nevertheless, more complex descriptions may be investigated. In this sense, mixture models identify subpopulations in a sample, arising from different distributions. The Gaussian mixture model is particularly popular and relies on the Expectation-Maximisation (EM) algorithm [1] for maximum likelihood estimation. The computation of the likelihood limits the type of distributions in the mixture; more specifically, for the Dirac distributions and even uniform components despite their high interest in practice for processing computer experiments. Their visualization is convenient and can lead to a sensitivity analysis where variables with largest marginals are least sensitive and vice versa, as shown by our application to a flooding real case in [2].

The objective of our study is to build a very general method to provide a mixture model that approximates a sample $(x_i)_{i=1}^n \in \mathcal{X}^n \subset \mathbb{R}^n$ from a random variable X. The representatives of the sample are the calculated components of the mixture, chosen in a parameterized family of laws denoted $\mathcal{R}$. We investigate, for a given number of representatives $\ell \in \mathbb{N}$, the mixture $\tilde{X}_\ell = R^{(J)}$ approximating X. The representatives $(R^{(j)})_{j=1}^\ell$ and the discrete random variable $J \in \{1, \dots, \ell\}$ need to be optimised.

To overcome the restrictions related to likelihood-based methods, the approach is based on a classical quantization formulation similar to K-means [3]. Its

objective is to find $\Gamma = (\gamma_1, \ldots, \gamma_\ell) \in \mathcal{X}^\ell$ minimizing the quantization error

$$\epsilon_p(\Gamma) = \left(\frac{1}{N} \sum_{i=1}^{N} \|x^{(i)} - \underset{\gamma \in \Gamma}{\arg\min} \|x^{(i)} - \gamma\|\|^p \right)^{\frac{1}{p}} = \left(\sum_{j=1}^{\ell} \frac{N_j}{N} \hat{\mathcal{W}}_p(X \mid X \in C_j^\Gamma, \delta_{\gamma_j})^p \right)^{\frac{1}{p}},$$

with $C_j^\Gamma = \{x, j = \underset{i \in \{1,\ldots,\ell\}}{\arg\min} \|x - \gamma_i\|\}$, $N_j = \text{card}(C_j^\Gamma)$ and $\hat{\mathcal{W}}_p(X \mid X \in C_j^\Gamma, \delta_{\gamma_j})$ the empirical p-Wasserstein distance between $X \mid X \in C_j^\Gamma$ and the Dirac distribution at γ_j. In the K-means approach, we have $p = 2$.

This formulation paves the way to an augmented quantization, as the p-Wasserstein distance can be computed between various types of distributions, which allows to investigate $\{R^{(1)}, \ldots, R^{(\ell)}\} \in \mathcal{R}^\ell$ minimizing the quantization error

$$\epsilon_p(\Gamma) = \left(\sum_{j=1}^{\ell} \frac{\text{card}(C_j)}{N} \hat{\mathcal{W}}_p(X \mid X \in C_j, R^{(j)})^p \right)^{\frac{1}{p}}.$$

The link between the clusters $(C_j)_{j=1}^\ell$ and the representatives $(R^{(j)})_{j=1}^\ell$ is a key question. Starting from an initial set of representatives, our method is broken down into the following steps, where each step is justified by reductions in the above, Wasserstein-based, quantization error. The representatives are associated to distinct subgroups (clusters). These clusters are perturbated by identifying elements to move from clusters to others. New representatives are associated to these perturbated clusters. The steps are repeated until convergence. The algorithm is first tested on synthetic points in 2D, illustrating the proof of concept with uniform components. Then an application to the characterization of the sea and weather conditions leading to severe floodings is performed. The marginals of the uniform distributions are interpreted as sensitivities.

Short biography (PhD student)

My three years at Ecole Centrale de Lyon including a Master Degree in Mathematics and Risk Engineering, have given me a strong motivation to keep studying mathematics and statistics applied to industrial problems. My PhD deals with statistical modelling applied to flooding risk. It is part of the chair CIROQUO that gathers academical and technological partners to work on problems involving costly-to-evaluate numerical simulators for uncertainty quantification.

References

[1] Frank Dellaert. “The Expectation Maximization Algorithm”. In: (July 2003).

[2] H. Muller et al. “Assessing storm impact on a French coastal dune system using morphodynamic modeling”. In: *J.Coast Res.* 33.2 (2016), pp. 254–272.

[3] Gilles Pagès and Jun Yu. “Pointwise convergence of the Lloyd algorithm in higher dimension”. Dec. 2013.