Mode estimation applied to clustering and scenario reduction via Optimal Transport metric

Diego Fonseca

June 9, 2022

Mode estimation as a stochastic problem

Clustering

Scenario reduction

Mode estimation when the density is known

Let ξ be a random vector with probability distribution $\mathbb{P}$ supported in $\mathbb{R}^d$, we assume that ξ has density function f. The mode estimation problem can be formulated as:

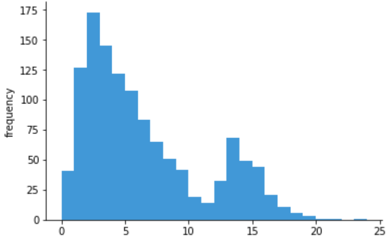
$$\max_{\xi \in \mathbb{R}^d} f\left(\xi\right). \tag{1}$$

Let $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N$ be a sample of ξ. Our focus is on find the point in the sample with the highest density. This problem can be formulated as:

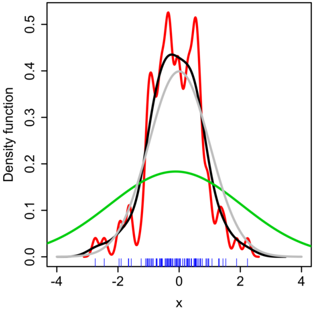
$$J^* := \max_{i=1,\ldots,N} f(\widehat{\xi}_i). \tag{2}$$

Objective: To find the point in the sample with the highest density when $\mathbb{P}$ and the density function f are unknown.

Mode estimation when the density is unknown
What has been done so far?



Based on frequency analysis

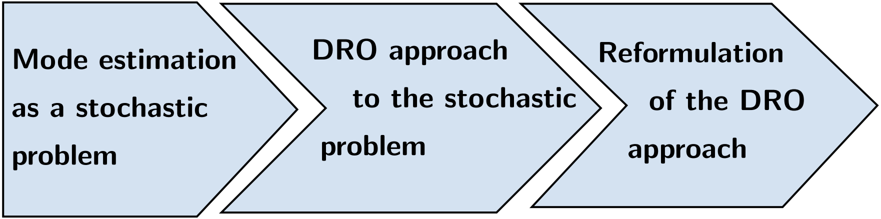


Based on density estimation

What has been done so far?

- Some attempts to replicate frequentist analysis in the continuous case (Burman and Polonik, 2009; Hsu and Wu, 2013; Kirschstein et al., 2016), but these tend to have a high sensitivity to sample contamination and, depending on sample size, are computationally expensive.
- Other attempts consist of estimating the unknown density function f using the sample, and then calculating the mode of this estimated density (Silverman, 1981; Silverman, 1986). However, determining that estimated density and calculating its maximum tends to be computationally complex when the sample size is large, and the random vector has a large dimension (Lee et al., 2019).

Our proposal


Mode estimation as a stochastic problem
DRO approach to the stochastic problem
Reformulation of the DRO approach

* DRO: Distributionally robust optimization.

Mode estimation as a stochastic problem

$$j^* := \underset{i=1,\ldots,N}{\text{máx}} f(\widehat{\xi}_i) \approx J_h := \underset{i=1,\ldots,N}{\text{máx}} \mathbb{E}_{\xi\sim\mathbb{P}}[f_{h,\widehat{\xi}_i}(\xi)] \tag{3}$$

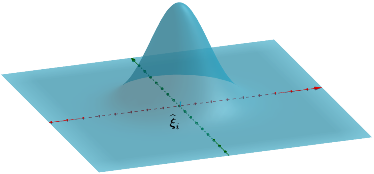
$\Uparrow$

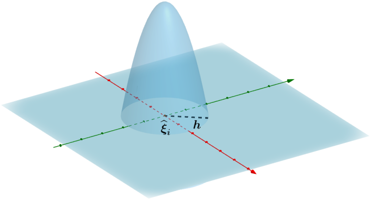
Stochastic problem

for very small values of $h > 0$, where $f_{h,\widehat{\xi}_i}$ is any uni-modal density function with mode in $\widehat{\xi}_i$ such that, if $\mathbb{P}_{h,\widehat{\xi}_i}$ is a probability distribution with density $f_{h,\widehat{\xi}_i}$, then $\mathbb{P}_{h,\widehat{\xi}_i} \to \delta_{\widehat{\xi}_i}$, when $h \to 0$, in the sense of weak convergence of measures, and $\delta_{\widehat{\xi}_i}$ is the Dirac delta distribution supported in $\{\widehat{\xi}_i\}$.

How are the density functions $f_{h,\widehat{\xi}_i}$?

There are several examples of $f_{h,\widehat{\xi}_i}$ densities, but in this work we focus on two types:





The **Gaussian density**, which is defined by

$$f_{h,\widehat{\xi}_i}(\xi) := \frac{1}{h^d(2\pi)^{d/2}} e^{-\frac{\left\|\widehat{\xi}_i-\xi\right\|^2}{2h^2}}.$$

The **Quadratic density**, this is defined by

$$f_{h,\widehat{\xi}_i}(\xi) := \text{máx}\left\{C_{h,d}\left(1-\frac{\left\|\widehat{\xi}_i-\xi\right\|^2}{h^2}\right),0\right\}$$

where $C_{h,d}$ is a constant that allows this function to be a density.

Why does this work?

$$j^* := \underset{i=1,\ldots,N}{\text{máx}} f(\widehat{\xi}_i) \approx \begin{array}{c} J_h := \text{máx}_{i=1,\ldots,N}\, \mathbb{E}_{\xi\sim\mathbb{P}}[f_{h,\widehat{\xi}_i}(\xi)] \\ \text{for a very small } h > 0 \end{array}$$

$$\Updownarrow$$

Note that taking the limit as h approaches zero, (3) becomes (2). Indeed, we have

$$\underset{i=1,\ldots,N}{\text{máx}}\ \underset{h\to 0}{\text{lím}}\ \mathbb{E}_{\xi\sim\mathbb{P}}[f_{h,\widehat{\xi}_i}(\xi)] = \underset{i=1,\ldots,N}{\text{máx}}\ \underset{h\to 0}{\text{lím}}\ \mathbb{E}_{\xi\sim\mathbb{P}_{h,\widehat{\xi}_i}}[f(\xi)] = \underset{i=1,\ldots,N}{\text{máx}}\ \mathbb{E}_{\xi\sim\delta_{\widehat{\xi}_i}}[f(\xi)] = \underset{i=1,\ldots,N}{\text{máx}} f(\widehat{\xi}_i).$$

DRO approach

The drawback is that $\mathbb{P}$ is unknown, so we are looking for a way to overcome this drawback.

$$J_h := \operatorname*{máx}_{i=1,\dots,N} \mathbb{E}_{\xi\sim\mathbb{P}}[f_{h,\widehat{\xi}_i}(\xi)] \approx \underset{\overset{\Uparrow}{\text{Our proposal}}}{\overset{\text{DRO approach}}{\widehat{J}_N(\varepsilon,h) := \operatorname*{máx}_{i=1,\dots,N} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi\sim\mathbb{Q}}[f_{h,\widehat{\xi}_i}(\xi)].}} \tag{4}$$

for very small values of $h > 0$.

In this case, $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ is a ball with respect to the 2-**Wasserstein distance** with radius $\varepsilon > 0$ and center at the empirical distribution $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}$ determined by the sample $\left\{\widehat{\xi}_i\right\}_{i=1}^N$ of ξ.

Note that $J_h \leq \widehat{J}_N(\varepsilon,h)$ if $\mathbb{P}\in\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$.

Wassertein distance

Definition 1 (Wassertein distance)

The *p-Wasserstein distance* $W_p(\mu,\nu)$ between $\mu,\nu \in \mathcal{P}_p(\Xi)$ is defined by

$$W_p^p(\mu,\nu) := \operatorname*{ínf}_{\Pi\in\mathcal{P}(\Xi\times\Xi)} \left\{ \int_{\Xi\times\Xi} \mathbf{d}^p(\xi,\zeta)\Pi(\mathbf{d}\xi, d\zeta) \;\middle|\; \begin{array}{l} \Pi(\cdot\times\Xi) = \mu(\cdot), \\ \Pi(\Xi\times\cdot) = \nu(\cdot) \end{array} \right\}$$

where

$$\mathcal{P}_p(\Xi) := \left\{\mu \in \mathcal{P}(\Xi) \;:\; \int_{\Xi} \mathbf{d}^p(\xi,\zeta_0)\mu(\mathbf{d}\xi) < \infty \text{ for some } \zeta_0 \in \Xi\right\}$$

and $\mathbf{d}$ is a metric in Ξ.

Thus, in the context of $\mathcal{P}_p(\Xi)$ for $p \in [1,\infty)$, the ball with radius $\varepsilon > 0$ and center $\mu \in \mathcal{P}(\Xi)$ is given by

$$\mathcal{B}_\varepsilon(\mu) := \{\nu \in \mathcal{P}(\Xi) \mid W_p(\mu,\nu) \le \varepsilon\}. \qquad (5)$$

Reformulation of the DRO approach

Theorem 2

Assume that $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ *is defined with respect to* W_2 *with* $\mathbf{d} := \|\cdot\|$ *the euclidean distance in* $\mathbb{R}^d$*. If* $f_{h,\widehat{\xi}_i}$ *is* **Quadratic density***, then (4) is equivalent to the optimization problem*

$$\max_{i=1,\ldots,N}\left\{\begin{array}{lll} \inf\limits_{\lambda,s} & \lambda\varepsilon^2+\dfrac{1}{N}\displaystyle\sum_{j=1}^{N}s_j & \\ \textit{subject to} & \lambda C_{h,d}\left(h^2-\left\|\widehat{\xi}_i-\widehat{\xi}_j\right\|^2\right)-\lambda s_j h^2-s_j C_{h,d}\leq -C_{h,d}^2 & \forall j, \\ & s_j\geq 0 & \forall j, \\ & \lambda\geq 0. & \end{array}\right. \tag{6}$$

Moreover, if $i_N^*(\varepsilon)$*,* $\lambda_N^*(\varepsilon)$ *and* $s_N^*(\varepsilon)$ *are optimal solutions in (6), then an optimal measure of (4) has N supports* $\widehat{\varphi}_1(\varepsilon),\widehat{\varphi}_2(\varepsilon),\ldots,\widehat{\varphi}_N(\varepsilon)$ *given by*

$$\widehat{\varphi}_j(\varepsilon)=\left\{\begin{array}{ll} \frac{C_{h,d}\widehat{\xi}_{i_N^*(\varepsilon)}+h^2\lambda_N^*(\varepsilon)\widehat{\xi}_j}{C_{h,d}+h^2\lambda_N^*(\varepsilon)} & \textit{if } s_N^*(\varepsilon)_j>0, \\ \widehat{\xi}_j & \textit{if } s_N^*(\varepsilon)_j=0. \end{array}\right. \quad \textit{for each } j=1,2,\ldots,N. \tag{7}$$

Reformulation of the DRO approach

Theorem 3

Assume that $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ is defined with respect to W_2 with **d** $:= \|\cdot\|$ *the euclidean distance in* $\mathbb{R}^d$*.If* $f_{h,\widehat{\xi}_i}$ *is* **Gaussian density***, then an (4) is equivalent to the optimization problem*

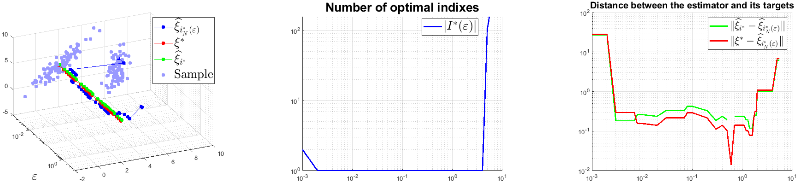
$$\max_{i=1,\dots,N} \left\{ \begin{array}{lll} \inf\limits_{\lambda,s} & \lambda\varepsilon^2 + \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} s_j & \\ \textit{subject to} & \sup\limits_{\xi\in\mathbb{R}^d}\left(\dfrac{1}{h^d(2\pi)^{d/2}}e^{-\frac{\left\|\widehat{\xi}_i-\xi\right\|^2}{2h^2}} - \lambda\left\|\widehat{\xi}_j-\xi\right\|^2\right) \leq s_i & \forall j, \\ & \lambda \geq 0. & \end{array} \right. \tag{8}$$

Moreover, if $i_N^*(\varepsilon)$ and $\lambda_N^*(\varepsilon)$ are optimal solutions in (8), then optimal measure of (4) has N supports $\widehat{\varphi}_1(\varepsilon), \widehat{\varphi}_2(\varepsilon), \dots, \widehat{\varphi}_N(\varepsilon)$ given by

$$\widehat{\varphi}_j(\varepsilon) = \underset{\xi\in\mathbb{R}^d}{\operatorname{argmax}}\left(\frac{1}{h^d(2\pi)^{d/2}}e^{-\frac{\left\|\widehat{\xi}_{i_N^*(\varepsilon)}-\xi\right\|^2}{2h^2}} - \lambda_N^*(\varepsilon)\left\|\widehat{\xi}_j-\xi\right\|^2\right) \quad \textit{for each } j = 1,\dots,N.$$

How does the proposed mode estimator work?

we propose $\widehat{\xi}_{i_N^*(\varepsilon)}$ as the estimator of the point in the sample with the highest density. That is to say, $\widehat{\xi}_{i_N^*(\varepsilon)}$ is an estimator of $\widehat{\xi}_{i^*}$ where $i^* := \underset{i=1,\ldots,N}{\mathrm{argmax}} f\left(\widehat{\xi}_i\right)$.

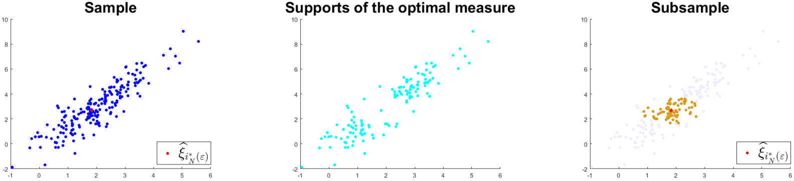


(a) (b) (b)

Figura: (a) Three-dimensional plot of the performance of the proposed estimator with respect to ε. (b) Number of optimal indexes $|I^*(\varepsilon)|$. (c) Distance between the estimator and the true mode, and the estimator and the highest density point. In this example, $h = 0{,}0001$ and Quadratic density were used.

How to choose ε?

Subsample associated to the optimal solution



(a) (b) (b)

Figura: (a) Sample and the sample point $\widehat{\xi}_{i_N^*(\varepsilon)}$ associated with the optimal index $i_N^*(\varepsilon)$ of (4). (b) The supports of the optimal measure in (4). (c) In orange color, subsample associated with the optimal index $i_N^*(\varepsilon)$, and, in opaque blue, the sample. In this example, $\varepsilon = 0{,}4$ and $h = 0{,}0001$ were used.

We denote by $\widehat{\Xi}_{i_N^*(\varepsilon)}$ the set determined by the subsample associated with $i_N^*(\varepsilon)$.

How to choose ε?

Coefficient of variation

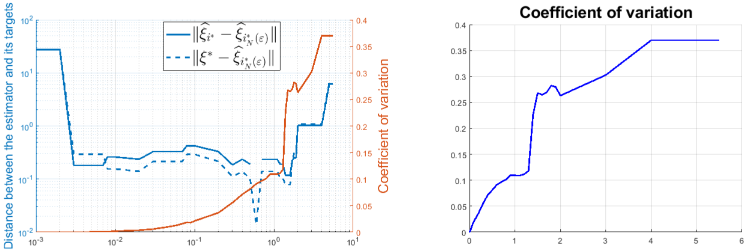
For each ε, let $CH(\widehat{\Xi}_{i_N^*(\varepsilon)})$ be the convex hull of the subsample $\widehat{\Xi}_{i_N^*(\varepsilon)}$ associated to the optimal index $i_N^*(\varepsilon)$ of (4). We define the coefficient of variation of ε as

$$CV(\varepsilon) := \frac{\mathrm{Area}\left(CH(\widehat{\Xi}_{i_N^*(\varepsilon)})\right)}{\left|\widehat{\Xi}_{i_N^*(\varepsilon)}\right|}.$$

This coefficient is interpreted as the area of the convex hull over the number of points in the subsample.

How to choose ε?

Coefficient of variation



(a)

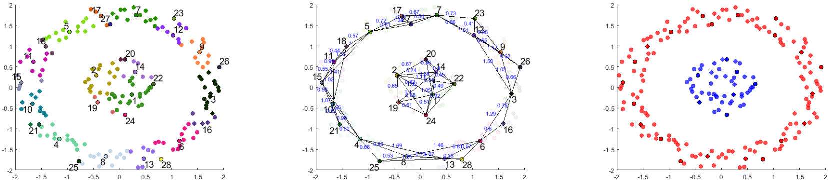
(b)

Figura: (a) Distance between the estimator and the true mode, and the estimator and the highest density point (left axis), and coefficient of variation (right axis). (b) Coefficient of variation. All plots with respect to ε on the horizontal axis.

Clustering

The sample is partitioned

Spectral clustering on the representatives of the partition



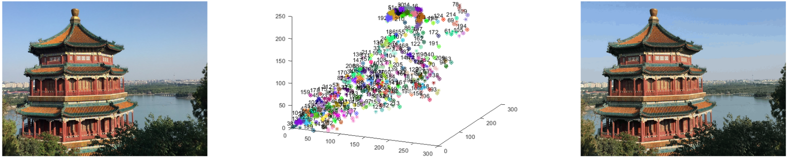
(a)

(b)

(b)

Figura: (a) Partition of the sample. (b) Reduced graph. (c) Resulting clustering. In this example, Quadratic density, $\varepsilon = 0{,}1$ and $h = 0{,}0001$ were used.

Scenario reduction



(a) Original

(b) Partition in the RGB space.

(c) Quantized image

Figura: Outputs of our color quantization algorithm for a image. In this example, Quadratic density, $\varepsilon = 0{,}1$ and $h = 0{,}0001$ were used. The original image is composed of $187,173$ colors while the quantized image is composed of 208 colors.

Thanks for your attention.

Burman, P. and Polonik, P. (2009).
Multivariate mode hunting: Data analytic tools with measures of significance.
Journal of Multivariate Analysis, 100(6):1198–1218.

Hsu, C. and Wu, T. (2013).
Efficient estimation of the mode of continuous multivariate data.
Computational Statistics & Data Analysis, 63:148 – 159.

Kirschstein, T., Liebscher, S., Porzio, G., and Ragozini, G. (2016).
Minimum volume peeling: A robust nonparametric estimator of the multivariate mode.
Computational Statistics & Data Analysis, 93:456–468.

Lee, J., Li, J., Musco, C., Phillips, J., and Tai, W. (2019).
Finding the mode of a kernel density estimate.
arXiv1912.07673.

Silverman, B. (1981).
Using Kernel Density Estimates to Investigate Multimodality.
Journal of the Royal Statistical Society, 43(1):97–99.

Silverman, B. W. (1986).
Density Estimation for Statistics and Data Analysis.
Chapman & Hall.