Locating Modes in Samples by a Weighted Laplace Convolution

Ulderico Santarelli

Retired Statistician and Mathematician

ulderico.santarelli@gmail.com

Abstract

In the case of continuous random variables, a straight application of the theorem of Fermat is enough to locate modes when they are located inside of the variables' range. The zeros of their Densities' derivatives are in fact the points where they attain their maximum, minimum or inflection points. In order to decide between the three cases, a check on the second derivative is needed. Unfortunately, derivatives are of no help in the search of modes in a sample. Discrete by its very nature, a sample makes derivatives meaningless. However, keeping in mind that by definition modes are local Density's maximums, one can completely overhaul the derivative-based approach and exploit the expectation that sample points would cluster around the sample's modes. This paper shows how this approach can be implemented through a Gravitational model that endows each sample point with a "centrality" score in its surrounding topological neighbourhood. Thus, instead of a direct search of modes, a two-steps method is suggested: (1) searching points that bear a good centrality score and then (2) checking the local density in their neighbourhoods.

Keywords: mode; clustering; mixture decomposition; gravitation; deep learning.

1. How the paper is organized

After an introduction, the theory underpinning the method is presented. Then, the case of the Logistic distribution is addressed. It enjoys a Cumulative in closed form that under the gravitational law $exp(-|w-x|)$ between the couple of points (w, x) allows a closed form representation of the corresponding gravitational field $T(w)$. The zeros of $T(w)$ locate the mode when it is inside of the variable's range. The value of the method lies in the fact that in a sample the gravitational field is easy to compute. Details needed to demonstrate the statements can be found in the Appendix.

2. Introduction

Due to the randomness of data patterns in a sample, even after a preliminary careful smoothing of sample's data, the smartest attempt to locate modes may end in a failure. In Author's opinion the simple approach here sketched could attract some attention from readers. The core of the method stays in the study of the gravitational field generated by a negative exponential gravitational law of points' inter-distance that corresponds to a weighted convolution of the density with the Laplace distribution. Because the convolution acts on the Density of continuous random variables and because Densities are absolutely summable over the whole

real line, any convergence concern disappears. The method can work well when the data body is a sample from a continuous distribution and the mode is an internal point in the variable's range. In this paper, therefore, mode means "a central point in a dense neighbourhood". This definition implies that the method can be viewed as a new variant of Clustering algorithms.

3. **The gravitational model**

The gravitational field $T(w)$ represents the accumulation on each sample point of the information about where the other points are. The value of the method is in its ability to provide each sample point with a "centrality" score in a suitably narrow surrounding neighbourhood. Borrowing from other fields of science, it is possible to endow each sample point with a smooth function of the pairwise point inter-distances from all the other points. We know that, by the notion of mutual attraction between couples of points, the gravitational Newton's model does exactly this. Here the Newton model is taken only as a mental paradigm, without pretending the applicability to Statistics of the gravitational hypothesis as such. Therefore, we won't use the Newton law that, among others, would lead to singular points, the infamous black holes. Narrowly circulated attempts by the Author with a plain application of the Newton's law show that the gravitational force between some couples of sample points may be too big thus remarkably hindering the search. Mechanical paradigms are not new in Statistics (Capra et al. 1970). The simplest case is the interpretation of the mean as the barycenter of data. All sample points, bearing the information of their position in the variable's range, concur to the estimation of the sample mean. In addition, it is known that some multivariate models plunge their roots in mechanics. Principal Components Analysis is the straight replication of the theory of the axis's of permanent rotation of a rigid body. In fact, the variance is the statistical companion of the moment of inertia that measures the resistance of a rigid body to rotation. Coming to samples, the sample variance is the resistance to rotation of a cloud of points keeping their inter-distances invariant during rotation. Viewing a data body as a collection of asteroids is thus certainly not new.

In this paper, keeping the basic characteristics of the Newton approach, we use a suitable law of attraction, both mathematically tractable and free from singularities. Namely:

1. for each couple of points (w, x) in the sample, an attraction force $f(w, x)$ that declines with the distance $|w - x|$, is defined,
2. each point w is endowed with the sum $T(w)$ of all the attraction forces emanating from all the other x points in the sample

Often, we will refer to the $T(w)$'s intensity as $|T(w)|$. A central point in a neighbourhood enjoys a local minimum of the total attraction's intensity $|T(w)|$ because the nearest surrounding points, pivotal in the sum of forces acting on that point, balance their contribution making $|T(w)|$ to decrease to a local minimum. Central points in a neighbourhood lay therefore at the bottom of a potential well. The points where the intensity of the sum of the attraction forces has a local minimum are thus candidates as local modes in the sample. They can be taken as modes only after checking that the local density attains a local maximum there. In fact, the attraction forces can reach a balance because of the mutual disposition of points in the sample without being central in any dense neighbourhood. They wouldn't be candidate as modes, however until directly checked. The centrality effect here described can be theoretically verified when the cumulative Distribution can be written in closed form.

From what has been said, the attraction force $f(w,x)$ between the two points (w,x) is

$$\forall(w,x)\ \ f(w,x) = sgn(w-x)M(w)M(x)e^{-|w-x|} \qquad (1)$$

where

$sgn(w-x)$ is the sign of the difference between the two points w and x. Taking w as the point under consideration, all forces are always directed from x to w. So, they are positive when $w > x$ and negative when $w < x$. This entails that the sum of forces $T(w)$ is the integral in x of (1)

$M(w)$ is the mass of the point w, actually the local Density at w, actually its likelihood

$M(x)$ is the mass of the point x, actually the local density at x, actually its likelihood

$exp(-|w-x|)$ is the law of the force intensity's decline with $|w-x|$

Given the point w, the total force $T(w)$ acting on due to all the other points x is given by

$$T(w) = M(w)\left\{\int_{-\infty}^{w} M(x)e^{-|w-x|}dx - \int_{w}^{+\infty} M(x)e^{-|w-x|}dx\right\} = M(w)\{A(w) - B(w)\} \quad (2)$$

Two integrals are needed because the forces change their direction at the point w. Please note that, being the integral in x, the density $M(w)$ is a constant, so that it can be moved out from the integral symbol.

All symmetric distributions obviously have $T(\mu) = 0$; μ = mean. In fact, taking without loss of generality $\mu = 0$, the force's intensity that links the couple $((w = -x), +x)$ is equal to the force intensity that links the two points $((w = +x), -x)$ because $M(w)$ is fixed, $M(x) = M(-x)$ and also $|w-x|$ is the same. Thus, forces' intensities are also the same with opposite signs, however. When the density is symmetric, the function $T(w)$ is thus antisymmetric with $T(-w) = -T(w)$. Summing over all points symmetrically positioned with respect to the mean 0, one gets the sum of 0. You conclude that, in the case of a unimodal symmetric distribution, the root of $T(w) = 0$ is also the single mode of the distribution. In fact, $T(w) = 0$ happens when w is central to a dense neighbourhood. $T(w)$ starts negative and ends positive so that $T(w)$ admits at least 1 root in the range. The root is actually only one. In fact, both $A(w)$ and $B(w)$ are positive. For a generic $w < 0 \rightarrow A(w) < B(w)$ and similarly $w > 0 \rightarrow A(w) > B(w)$. Therefore, $w < 0 \rightarrow T(w) < 0$ and $w > 0 \rightarrow T(w) > 0$. Buy the Ulysses Dini theorem, $T(w)$ crosses the x-axis. The point of crossing is 0 due to the symmetry between $A(w)$ and $B(w)$.

4. **The Logistic distribution**

Unfortunately, $T(w)$ can't be put in closed form for the normal Distribution. Integrating by parts, the Cumulative $\Phi(x)$ will show up at some point. Though $\Phi(x)$ admits approximations, they are of no help in the search of the roots of $T(w)$. A suitable approach is to leverage Distributions having a Cumulative in closed form. The Logistic distribution can work well with the distances $|w-x|$.

The Logistic Density and the Logistic Cumulative distributions are written as

$$f(x) = \frac{e^x}{(1+e^x)^2}; \quad F(x) = \int_{-\infty}^{x} f(t)dt = \frac{e^x}{1+e^x}$$

For each (w, x) and the $|w - x|$ distance you have

$$f(w,x) = sgn(w-x)\frac{e^w}{(1+e^w)^2}\frac{e^x}{(1+e^x)^2}e^{-|w-x|}$$

For the total force $T(w)$ acting on the point w you have

$$T(w) = \left\{e^{-w}\frac{e^w}{(1+e^w)^2}\int_{-\infty}^{w}\frac{e^x}{(1+e^x)^2}e^x dx - e^w\frac{e^w}{(1+e^w)^2}\int_{w}^{+\infty}\frac{e^x}{(1+e^x)^2}e^{-x}dx\right\}$$

In the Appendix you find the form of $T(w)$. It appears as a Weighted Laplace Convolution that combines the density of the Logistic with that of the Laplace distribution.

5. **An application to a mixture of normal distributions**

The case of a symmetric distribution could be addressed directly without the help of $T(w)$. After computing the skewness of the distribution, one can assume that the distribution is symmetric and take the mean and the median as approximations of the mode. In the case of mixtures, on the opposite, $T(w)$ becomes precious if you act coordinate-wise. You can search for the points where $|T(w)|$ attains some minimums as hints for local modes. You can't expect a plain 0 but just local minimums, because in a mixture the contributing Distributions often overlap, show different ranges and variances with the marginal Distributions asymmetric.
As an example, we take the classical Fisher Iris dataset. The dataset keeps 50 records of 4 variables for 3 species. The variables are

1. Sepal length
2. Sepal width
3. Petal length
4. Petal width

The species are

1. Setosa
2. Versicolor
3. Virginica

The following tables show means and standard deviations of the 4 variables by the 3 species.

Means	Sepal length	Sepal width	Petal length	Petal width
Setosa	50.06	34.28	14.62	2.46
Versicolor	59.36	27.7	42.60	13.26
Virginica	65.88	29.74	55.52	20.26

Table 1. Means of the 4 variables by the three species

Standard dev	Sepal length	Sepal width	Petal length	Petal width
Setosa	3.52	3.79	1.74	1.05
Versicolor	5.16	3.14	4.79	1.98
Virginica	6.36	3.22	5.51	2.75

Table 2. Standard deviations of the 4 variables by the three species

To find the modes, the algorithm first computes $T(w)$ for all sample points w. Then locates the points where $|T(w)|$ attains its minimums. However, in a sample you have to make room form randomness, so that you need an approach that finds a point cloud that includes the points with a minimum of $|T(w)|$. Going coordinate-wise, you start finding the pentiles of the distribution of $|T(w)|$. Here they are:

Sepal length: 65
Sepal width: 64
Petal length: 363
Petal width: 1

Of course, you have to choose between the points you get. You find 30 points that show $|T(w)|$ within a pentile from 0 at least for one of the four variables. This is obviously due to the fact that the points around a mode are in the same potential well so they share the minimality of $|T(w)|$. Therefore, you need to aggregate points that fall within a small radius around each point flagging a point as "no more available as a candidate mode" as soon as it is associated to another point. After aggregation and elimination of too near points you find 3 points surrounded by some other points as shown.

Centers	Surrounding points	Average Sepal length	Average Sepal width	Average Petal length	Average Petal width
1	11	48.09	31.00	14.36	1.91
2	10	56.20	25.50	39.60	11.60
3	7	67.00	31.42	55.28	23.14

Table 3. Candidate Modes found by the algorithm

A direct Euclidean aggregation to the nearest point results in the solution

Means	Sepal length	Sepal width	Petal length	Petal width
Setosa	50.06	34.28	14.62	2.46
Versicolor	58.00	27.02	41.93	13.11
Virginica	66.40	30.10	54.89	19.74

Table 4. Means of the three aggregation clusters

Standard dev	Sepal length	Sepal width	Petal length	Petal width
Setosa	3.52	3.79	1.74	1.05
Versicolor	4.59	3.07	4.74	2.10
Virginica	5.56	2.89	5.63	3.05

Table 5. Standard deviations of the 4 variables by the three species

The following table shows the correspondence between species and aggregation clusters

	Cluster 1	Cluster 2	Cluster 3
Setosa	50		
Versicolor		42	8
Virginica	3		47

Table 6. Cross correspondence between clusters and species

6. **Algorithm details and SAS pseudocode**

The SAS pseudocode uses a libname “grav” where to host algorithm results
Steps and pseudocode follow

a. Compute all points inter-distances

- Seq is the point sequence in the 150 rows of data
- f1_, …, f4_ are the points coordinates. The variables are Sepallength, …. Petalwidth

```
proc sql;
create table grav.dist as
select a.seq as seq1, a.f1_ as w1, a.f2_ as w2, a.f3_ as w3, a.f4_ as w4,
       b.seq as seq2, b.f1_ as x1, b.f2_ as x2, b.f3_ as x3, b.f4_ as x4
from grav.irisdata as a, grav.irisdata as b
order by a.seq, b.seq
 ;
```

b. Compute the force between each couple of pints

```
data grav.laplace;
**manhattan distance;
 set grav.dist;
 dist1 = exp(-abs(w1 - x1));
 dist2 = exp(-abs(w2 - x2));
 dist3 = exp(-abs(w3 - x3));
 dist4 = exp(-abs(w4 - x4));
 F1 = (w1>x1)*dist1-(x1>w1)*dist1; ****change sign when x>w;
 F2 = (w2>x2)*dist2-(x2>w2)*dist2;
 F3 = (w3>x3)*dist3-(x3>w3)*dist3;
 F4 = (w4>x4)*dist4-(x4>w4)*dist4;
 if seq1 eq seq2
    then do;
      dist1 = 0; dist2 = 0; dist3 = 0; dist4 = 0;
      f1 = 0; f2 = 0; f3 = 0; f4 = 0;
      end;
     run;
```

c. Compute the sum of all forces given the point w

```
 proc sort data = grav.laplace;
 by seq1 seq2;
 run;

 proc means data = grav.laplace noprint;
 var f1-f4;   ****f1, …, f4 are the coordinatewise forces;
 by seq1;
 output out = grav.totf2 sum=;
 run;
```

d. Create the basic table where point coordinates are followed by the total forces

```
proc sql;
    create table grav.totfx2 as
      select a.seq, a.f1_, a.f2_, a.f3_, a.f4_,
           b.F1, b.F2, b.F3, b.F4
      from grav.irisdata as a, grav.totf2 as b
```

```
      where a.seq eq b.seq1
      order by a.seq
      ;
```

e. Find the critical points of the absolute value of forces

```
data grav.totfx2a;
 set grav.totfx2;
     af1 = abs(f1); af2 = abs(f2); af3 = abs(f3); af4 = abs(f4);
run;

proc univariate data = grav.totfx2a;
var af1-af4;
run;

**f1 pent = 85  dec = 215;
**f2 pent = 64  dec = 86;
**f3 pent = 363 dec = 513;
**f4 pent = 1   dec = 151;
```

f. Now choose among the 30 points having coordinate-wise a total force within a pentile for at least one coordinate those with the highest number of surrounding points
You find the points 6 with 11 points, 56 with 10 points and 102 with 7 points. A pseudo chisquare test tells you that 6 has a pseudo chisquare of 10.57, 56 of 9.97 and 102 of 6.97. The chisquare is a pseudo one because the expected value in the small cell surrounding the three points admissible.is practically 0, so the Poisson approximation is not applicable.

7. **Appendix**

A detailed computation for the Logistic distribution follows. Three basic integrals underpin the key steps. The first one is a definite integral spanning from $-\infty$ to x and linking the Logistic density to its companion Cumulative. The second and the third ones are indefinite integrals you can retrace in the mathematical literature (ex. Weast et al., 1964). Omitting the arbitrary constant c, you have

$$(a)\ \int_{-\infty}^{x} \frac{e^t}{(1+e^t)^2}\,dt = \frac{e^x}{1+e^x}$$

$$(b)\ \int \frac{1}{1+e^x}\,dx = log\,\frac{e^x}{1+e^x}$$

$$(c)\ \int \log(x) = x\log(x) - x$$

Also, the substitution $z = 1 + e^x$ is used when needed. The total force $T(w)$ is written as

$$T(w) = \frac{e^w}{(1+e^w)^2}\left\{e^{-w}\int_{-\infty}^{w} \frac{e^x}{(1+e^x)^2}e^x dx - e^w \int_{w}^{+\infty} \frac{e^x}{(1+e^x)^2}e^{-x}dx\right\}$$

$$= \frac{e^w}{(1+e^w)^2}\{e^{-w}A - e^w B\}$$

Separately dealing with A and B, you get

$$(A)\ \int_{-\infty}^{w} \frac{e^x}{(1+e^x)^2} e^x dx = e^x \frac{e^x}{1+e^x}\Big|_{-\infty}^{w} - \int_{-\infty}^{w} \frac{e^x}{1+e^x} e^x dx$$
$$= \frac{e^{2w}}{1+e^w} - e^x \log(1+e^x)\Big|_{-\infty}^{w} + \int_{-\infty}^{w} \log(1+e^x) e^x dx = \frac{e^{2w}}{1+e^w}$$
$$- e^w \log(1+e^w) + (1+e^w)\log(1+e^w) - e^w$$
$$= \frac{e^{2w}}{1+e^w} + \log(1+e^w) - e^w;\quad e^{-w}A = \frac{e^w}{1+e^w} + \frac{\log(1+e^w)}{e^w} - 1$$

$$(B)\ \int_{w}^{+\infty} \frac{e^x}{(1+e^x)^2} e^{-x} dx = e^{-x}\frac{e^x}{1+e^x}\Big|_{w}^{+\infty} - \int_{w}^{+\infty} \frac{1}{1+e^x} dx = -\frac{1}{1+e^w} - \log\frac{e^x}{1+e^x}\Big|_{w}^{+\infty} =$$
$$-\frac{1}{1+e^w} + \log\frac{e^w}{1+e^w} = -\frac{1}{1+e^w} + \log(1+e^{-w}) e^w B = -\frac{e^w}{1+e^w} - e^w \log(1+e^{-w})$$

Putting all together

$$T(w) = \frac{e^w}{(1+e^w)^2}\{e^{-w}A - e^w B\}$$
$$= \frac{e^w}{(1+e^w)^2}\left\{\frac{2e^w}{1+e^w} + e^{-w}\log(1+e^w) - 1\ - e^w \log(1+e^{-w})\right\}$$

$T(w)$ is antisymmetric in w. The factor outside the major parenthesis is the density at w, symmetric for a logistic distribution. Easy to prove, anyway. Substituting w with $-w$ you come back to the same form with inverted sign. You easily check that

$$Q(w) = 2\frac{e^w}{1+e^w} - 1 = \frac{2e^w - 1 - e^w}{1+e^w} = \frac{e^w - 1}{e^w + 1} \to Q(-w) = \frac{e^{-w} - 1}{e^{-w} + 1} = \frac{1 - e^w}{1+e^w}$$
$$\to Q(w) = -Q(-w)$$

And also

$$e^{-w}\log(1+e^w)\ - e^w \log(1+e^{-w})$$

is obviously antisymmetric. Therefore, the part within the major parenthesis is antisymmetric. A continuous antisymmetric function admits at least one root $T(w) = 0$. In addition, because 0 is a root of $T(w)$, you should have $A(0) = B(0)$. This is easily checked. Both take the value $-½ + log(2)$ at 0.

10 The cognitive side of Clustering and Deep Learning

Clustering is now a component of the toolset of Machine Learning. This is due to the cognitive side of Clustering that justifies its use in the search of meaningful subsets in a data body. Clustering supports the creation of new terms applicable to the distinct sets in a partition the clustering process is able to find. Machine's knowledge, and our own too, grows when separate

components can be isolated within a given data body so to enrich the language with new connoting terms. As the terms are a key means for communication, the proof of the knowledge's increase stays in the enrichment of our vocabulary with new meaningful terms. Knowledge would be useless if kept buried in somebody's mind. Knowledge can unleash its potential only as a social entity. The real world, however, is both substantive and diversified. While "being substantive" suggests the idea of data homogeneity within the sets in the partition, "diversification" hinders the grouping of data into sets due the ambiguity of the assignment of points to clusters. In fact, with the notable exception of Zadeh's "fuzzy" method, clustering aims at drawing crisp boundaries between sets. The difficulty of the problem is exactly here: how to manage together homogeneity and distinction. Without retracing the history of logic through 25 centuries, the contraposition between belonging to the same category while keeping personal and distinctive characteristics still dominates both logic and human relations. The persistence of the conscience over time, while our cells die at any moment and our body is in continuous change, stays as an item of the actual coexistence of a constant substance with volatile characteristics. In Cantor's view, each set corresponds to a term that connote the elements included in it. For instance, the set of equilateral triangles includes those that span three equal angles. All in all, sets inherit the old Aristotle's problem of unequivocal categorization of objects. The basic assumption is that our mind is able to catch the inner nature of objects, perceived the same notwithstanding they show up in so many variants in our daily experience. We have no doubt in recognizing Mary even if she wears a new dress or has been recently by a hairdresser. Irrespective of small variations in appearance, the nature of objects persists unmodified. Such a bipolar view, same nature shared by different individuals, is almost exactly replicated in the Cantor's naïve definition of a set as a "pool of distinct elements" that "contribute to a whole". Cantor assumes that objects in the same set share something though they keep their distinctiveness. With the Cantor's view in mind, Substantive Clustering appears as a method to provide a dedicated name to groups of points that keep together distinct elements so similar to legitimate the assumption of a shared nature. Substantive Clustering means that a persistent grouping of points, indifferent to sample substitutions, has been found.

Bibliography

[1] Capra, R., Lena, S., Santarelli, U., Vescovi, P. (1976). "Cluster Analysis by Moment of Inertia Method", IBM technical disclosure Bulletin, Vol. 18 No. 8.

[2] Weast R., M. Samuel (1964). "Standard Mathematical Tables", The Chemical Rubber Co.

[3] Guojun G., M. Chaoqun, W. Jianhong (2007). "Data Clustering, Theory, Algorithms and Applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

[4] Hartigan, J. (1975). "Clustering Algorithms", Toronto, John Wiley & Sons.