

Flow Cytometry Data Assessment
with L2 Discrepancy Learning
Process

Faysal El Khettabi

Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC, Canada.

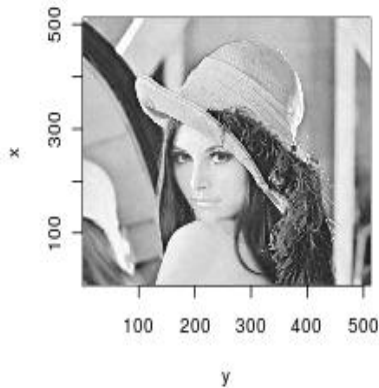
Problem

The raw data has N events where each given event is a s -dimensional vector.

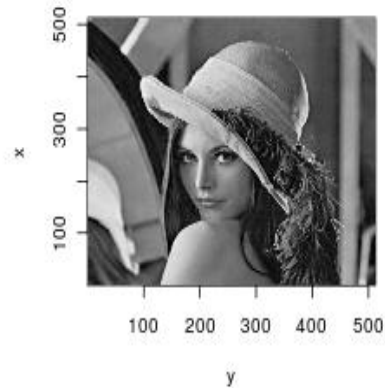
- **Is a given event an insider or an outlier?**
- **How similar or different are two given events?**
- **Without assuming that the events determine a probability measure, can one "emulate" a similar density measure-theoretic concept beginning directly with the event s -variables values ?**
- **Can the events be clustered?**

An Illustrative Introduction

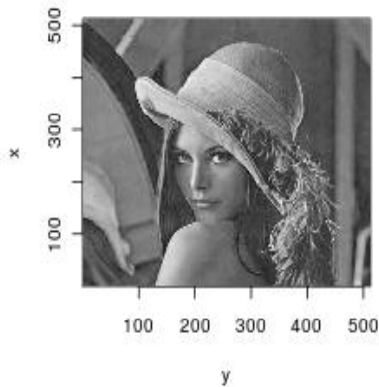
Red Channel



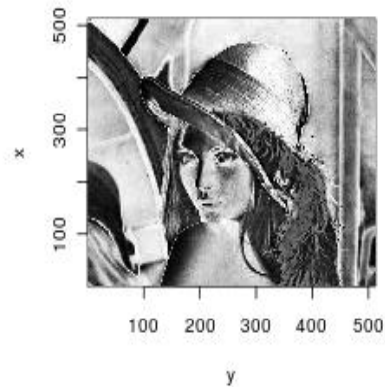
Green Channel



Blue Channel



L2 Discrepancy Channel



L2 Discrepancy

Local Discrepancy

Global Discrepancy

Sensitivity

SENSITIVITY

PROPOSITION 1 *The L_2 discrepancy sensitivity, S_n , is given by the following analytical formulas:*

$$S_n = \frac{(\sigma_n - \sigma) + 2^{-s}(\gamma - \mathbf{g}_n)}{\mathbf{T}_N(X)}, \quad (9)$$

satisfying

$$\sum_{n=1}^N S_n = 0, \quad (10)$$

where $\sigma = \frac{1}{N^2} \sum_{n=1}^N \sum_{m=1}^N \mathbf{f}_{n,m}$, $\sigma_n = \frac{1}{N} \sum_{m=1}^N \mathbf{f}_{n,m}$ and $\gamma = \frac{1}{N} \sum_{m=1}^N \mathbf{g}_m$.

INTERPRETATION

- If $S_n < 0$, the event x_n is in a region quasi-empty of events, i.e. add more events in its immediate neighborhood will minimize L_2 discrepancy.
- If $S_n > 0$, the event x_n is in a region where the events are clumped, i.e. L_2 discrepancy will decrease if the event x_n is removed.
- If $S_n = 0$, the event x_n is in a region where the events are uniformly distributed.

Kernel Density Estimation

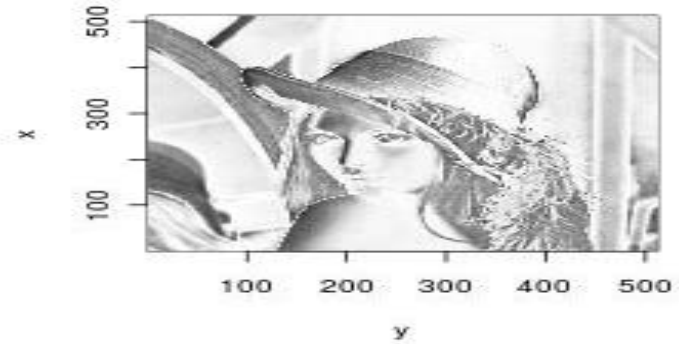
vs.

Discrepancy Sensitivity

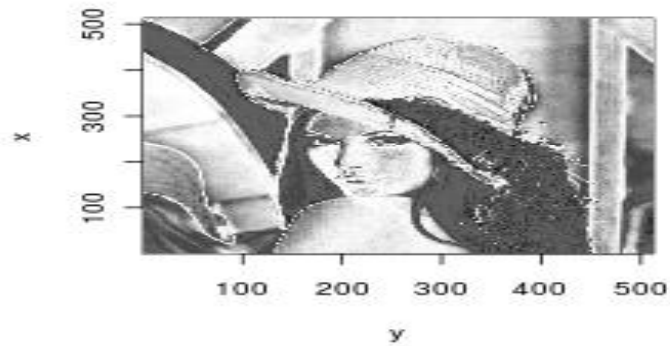
RGB Channel



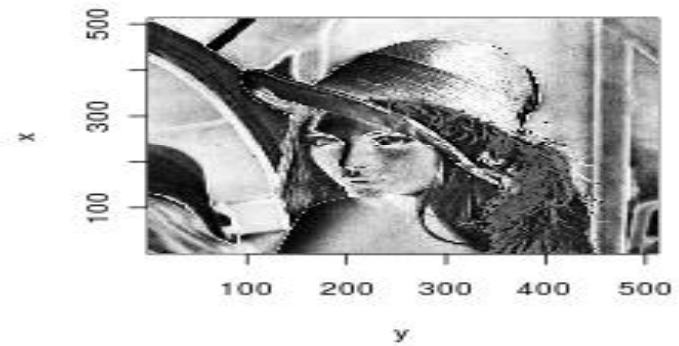
KDE_BW_0.72_0.65_0.07_RE_0.59



KDE_BW_0.94_0.83_0.65_RE_0.19



SENSITIVITY



Discrepancy Cytometry Analytics (DCA)

(Examining raw data with the purpose of drawing conclusions about the events.)

The root mean square expectation of L2 discrepancy(RMSELD) value for all random sequence with N events is equal to: $\text{SQRT}(12^{-s} * (2^s - 1)/N)$.

RF, Randomness Fit = (L2 discrepancy Raw Data) / RMSELD
(if RF is close to one, the raw data is uniformly random.)

PO, Proportion of Outlier events.

PI, Proportion of Insider events.

LA, Logarithmic Average of (1+Sn) (equivalent to Variance of the Sensitivity).

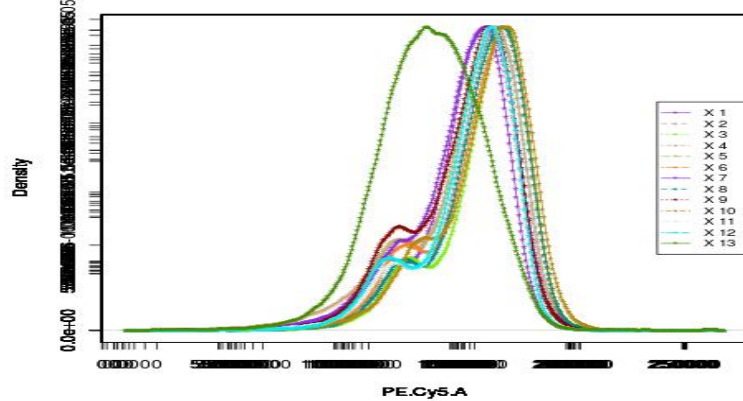
XI, square sensitivity of outlier events / square sensitivity of insider events.

These parameters are intrinsic properties of the events distribution.

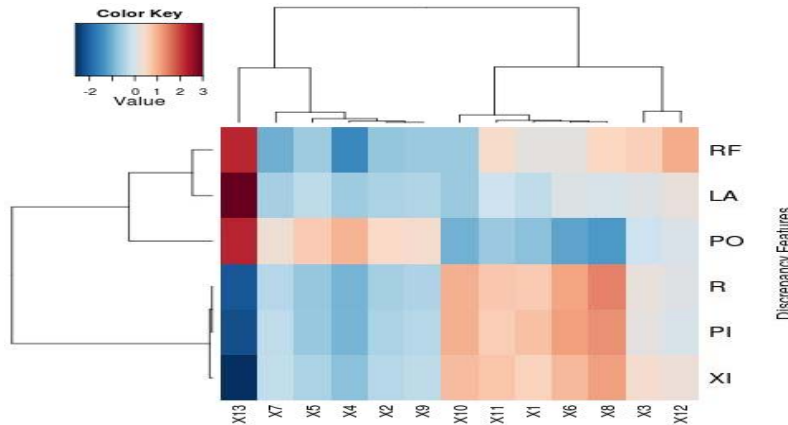
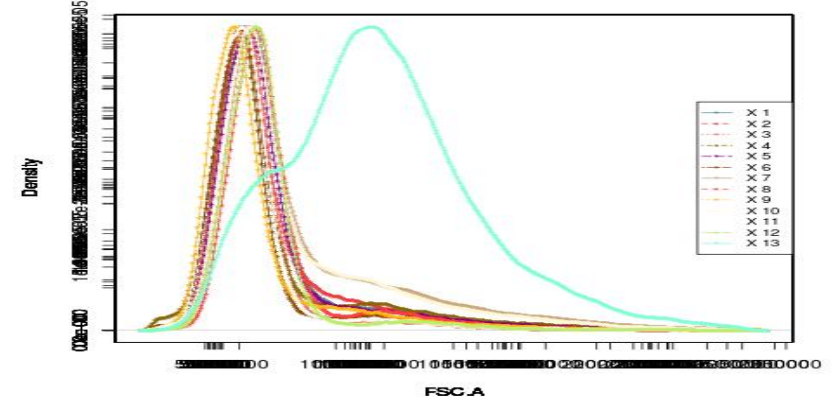
Quality Assesemnt

(CFSE data Set)
(Sample 13 has an issue)

RESULTS_FEK_CAP_/DISCREPANCY_K_MEANS/CFSE_DISCR_/Sampl

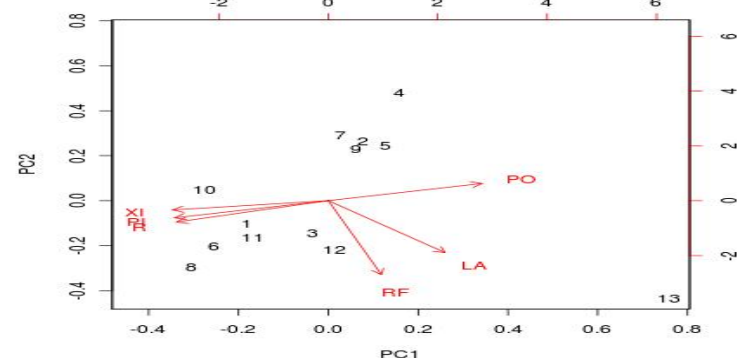


RESULTS_FEK_CAP_/DISCREPANCY_K_MEANS/CFSE_DISCR_/Sampl



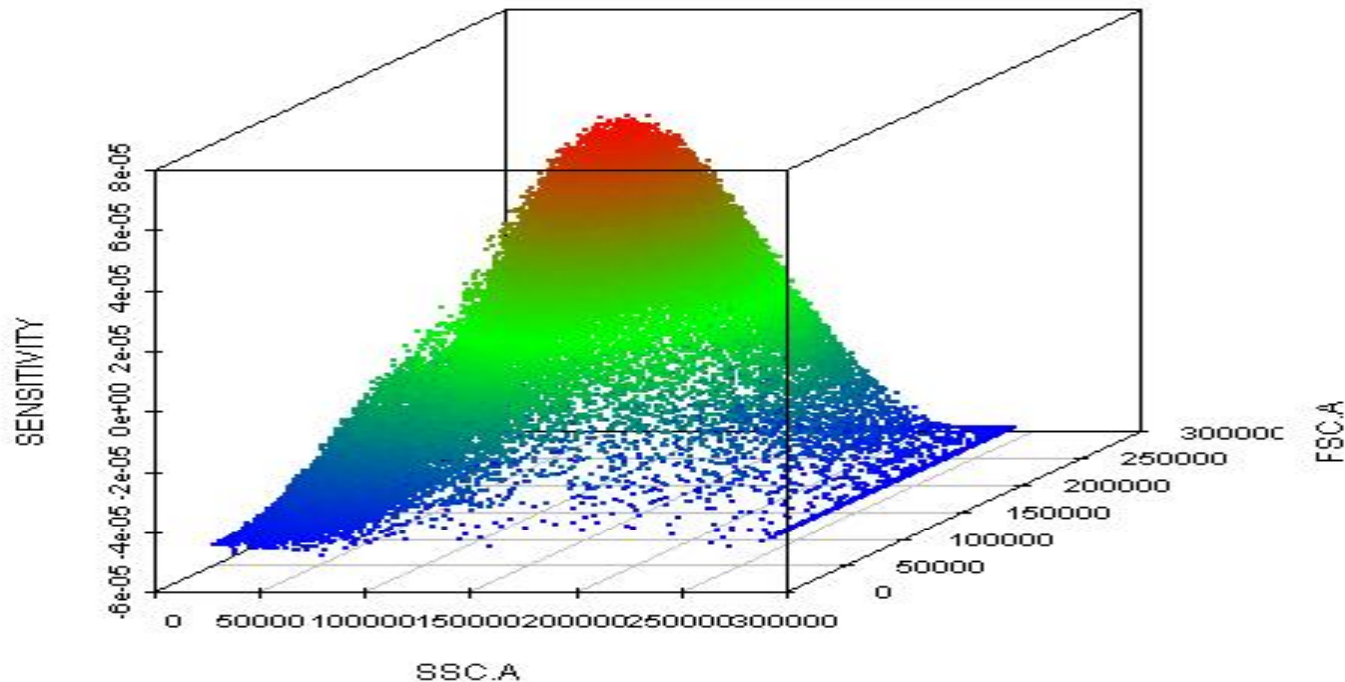
me/faysal/ALL_RESULTS_FEK_CAP_/DISCREPANCY_K_MEANS/CFSE_DISCR_/Sampl

LL_RESULTS_FEK_CAP_/DISCREPANCY_K_MEANS/CFSE_DISCR_/S



Density Emulation NDD with 12 variables

FSC and SSC measurements vs. SENSITIVITY



FEK_L2_DISC_NDDCSV004__P_62310_F_12_RF_601_P_0.49_XI_0.5

Discrepancy K means

K-Means is a least-squares partitioning method. Allowing users to divide a collection of objects into K groups.

The initial centroid guesses and their numbers are very hard to figure out for any given data.

K-Means method based on L2 Discrepancy used the sensitivity to define the most insiders as centroids, the algorithm iterate between the following simple steps:

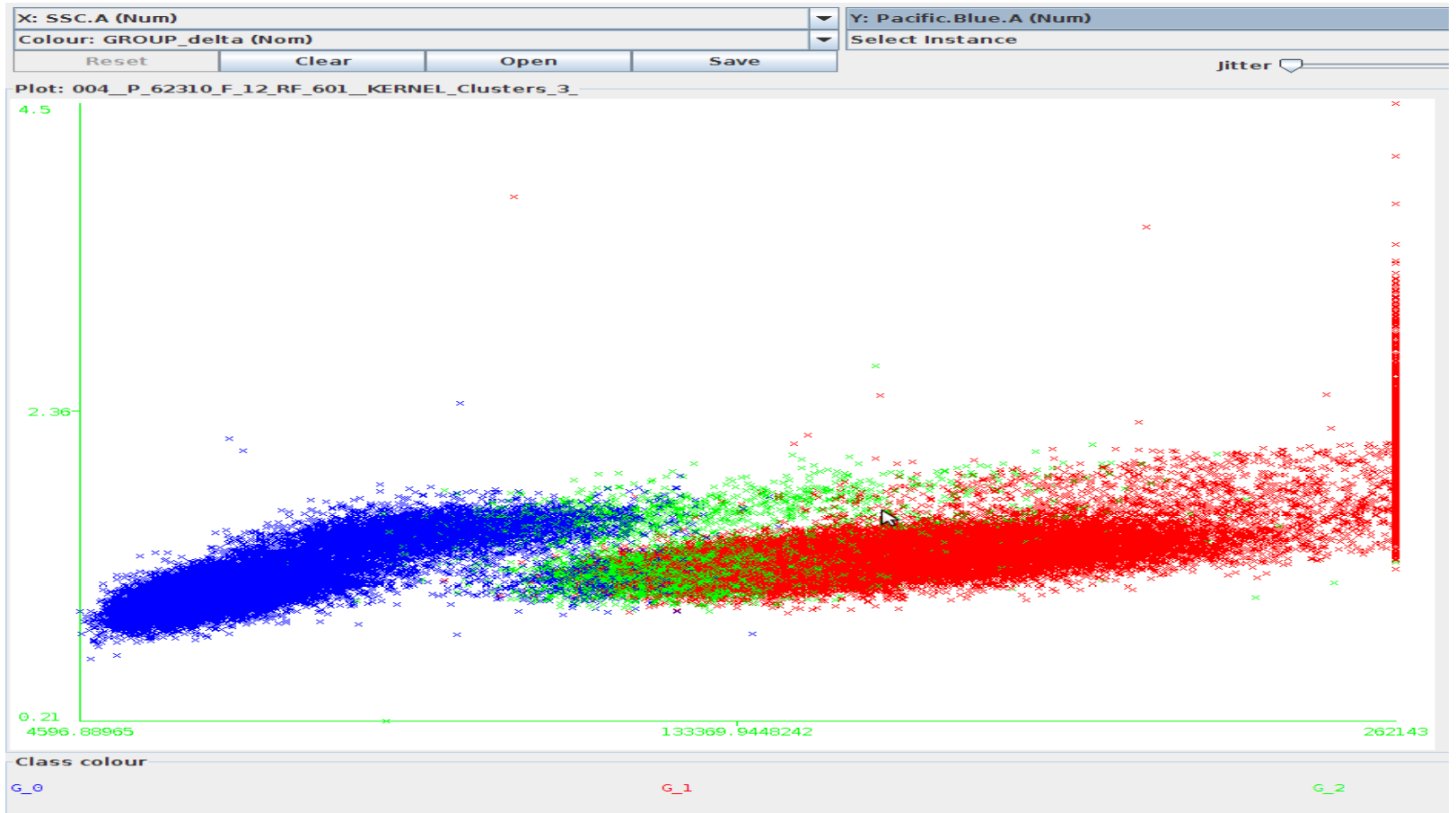
Step 1, assign all events to the set $R=X$

Step 2, pick up the event, $x_{\{n\}}$, with the maximum sensitivity in R.

Step 3, $C=\{\text{all events close to } x_{\{n\}} \text{ , accordingly to a given criteria}\}$

Step 5, set $R=R-C$, go to step 2.

Discrepancy Kmeans



CONCLUSION

- We developed a L2 discrepancy learning process to assess how flow cytometry data are spatially distributed.
- This discrepancy learning process is able to recover the spatial distribution where the individual events are either clumped or scarce.
- It is simple to numerically implement and provides a quantitative level of information to track the most outliers .
- We applied the L2 discrepancy learning process to K-Means clustering method. The discrepancy K-Means does not require the estimation of the number of clusters or other parameters and the L2 discrepancy learning process defines the means/modes as insiders automatically.

SUPPORT

The statistical and bioinformatics approaches to the classification of clinical lymphoma and leukemia data (Canadian Cancer Society grant 700374)

The statistical and computational analysis of flow cytometry data (NIH/NIBIB grant EB008400).