Selected Topics in Electrical Engineering: Flow Cytometry Data Analysis

Bilge Karaçalı, PhD
Department of Electrical and Electronics Engineering
Izmir Institute of Technology

Outline

- Normalization
  - Normalization in pattern recognition
  - Normalization of flow cytometry data
    - Log displays
    - Logicle transformation
    - Univariate normalization of individual fluorescence parameters
    - Multivariate normalization

Feature normalization

- In pattern recognition applications, features may vary in magnitude
- Differences in magnitude may bias the learning system to rely more on high-magnitude features
  - Statistical learning applications construct some representation of local neighborhoods around the different feature vectors
  - High magnitude features dominate the others in determining the distances between the feature vectors, and thus, the corresponding neighborhood structure
- Normalization is carried out to equalize the magnitudes of all features
  - Further normalization may also seek to adjust the way features span their equalized dynamic range

Feature normalization

- Linear normalization:
  - Each feature is scaled by the inverse of its standard deviation observed across the data

$$\hat{x}_{i,j} = x_{i,j}/\sigma_j$$

where

- $\{x_i\}$ is the dataset of vectors $x_i \in X$, $\forall\, i = 1, 2, \ldots, \ell$
- σ_j is the standard deviation of the *j*'th feature across the dataset; $\{x_{i,j}\}$ across all I

  - The mean μ_j can also be removed from each feature via

$$\hat{x}_{i,j} = (x_{i,j} - \mu_j)/\sigma_j$$

to achieve feature vectors with zero mean and unit standard deviation

Feature normalization

- Gamma normalization:
  - The features can be normalized to span their dynamic range as uniformly as possible first by letting

$$x'_{i,j} = a_j x_{i,j} + b_j$$

where a_j and b_j are determined so that

$$\min_i \{x'_{i,j}\} = \frac{1}{\ell + 1}$$

and

$$\max_i \{x'_{i,j}\} = \frac{\ell}{\ell + 1}$$

  - And then defining $\hat{x}_{i,j}$ by

$$\hat{x}_{i,j} = \left(x'_{i,j}\right)^{\gamma_j}$$

where γ_j is such that the collection $\{\hat{x}_{i,j}\}$ for each *j* over all *I* is as uniformly distributed as possible

Data normalization

- Feature normalization adjusts the feature magnitudes so that each feature has an equal chance of influencing the learning rule
- Data normalization, in contrast, aims to make different datasets comparable to each other
  - Data collections do not readily submit themselves to comparison
    - Different equipment
    - Different data acquisition settings
    - Variations on sample preparation
    - …
  - Before any comparative analyses, the different datasets must be converted into a standard form that they would have had should the data collection conditions been identical for all
  - Example: Deformable image registration
    - The images that contain the anatomical information of different subjects are not comparable
    - When the images are registered spatially with a template, the corresponding coordinate transformations are comparable

Flow data normalization

- The same considerations are faced when comparing two or more flow cytometry datasets
  - Any flow cytometry experiment is a delicate procedure
  - Many factors can affect the actual collected values into the data
    - Sample preparation
      - Protocols
      - Staining
      - Choice of fluorochromes
    - Equipment setup
      - Lasers
      - Voltage gains
      - Compensation parameters
  - These differences prevent the transfer of subset specifications obtained from one sample to the others
    - Gating to be carried out from scratch separately for each dataset
- Comparing different flow datasets requires annulling the effects of these irrelevant variations from the measurements
  - Irrelevant to the actual biological and/or clinical hypothesis at hand

Flow data normalization

- In practice, the matching of the cell clusters in the different flow datasets is carried out manually via visual inspection
- This requires a suitable display of the flow data that
  - uses the dynamic range between the smallest and the largest measurements efficiently, and
  - allows distinguishing the different cell types from each other as distinct clusters

Flow data displays: The log scale

- Flow data consists of measurements on fluorescence intensities
- In the presence of the intended biomarker, the fluorescence signal can exhibit very large values
- In the absence, the measurements are of the background illumination towards the lower end of the scale
- The goal is not only
  - to separate the cells that possess the biomarker from those that do not, but also
  - to separate the cell clusters that possess the biomarker at different levels
- The general tendency in such cases is to present the data in a logarithmic scale instead of the linear scale in which the measurements are recorded

Flow data displays: The logicle transformation

- The logarithmic scaling overextends the lower ends of the measurement scale
- Based on the hyperbolic sine function

$$\sinh(x) = \frac{1}{2}(e^{x} - e^{-x})$$

- A generalization provides the bi-exponential function

$$S(x) = ae^{bx} - ce^{-dx} + f$$

- A subset of the bi-exponential functions that are linear near zero are called logicle functions
  - with zero second derivative near zero

Flow data displays: The logicle transformation

- Further requirements are incorporated in addition to setting the second derivative of $S(x)$ with respect to x equal to zero when $x \approx 0$:
  - The maximum data value to be displayed (T)
  - The range of the display (m)
  - The range of linearization around 0 (w)
  - The range of negative values to be included in the display
    - Usually set to the linearization range w
- Then, the logicle function becomes

$$S(x) = Te^{-(m-w)}\left(e^{x-w} - p^2 e^{-\frac{x-w}{p}} + p^2 - 1\right)$$

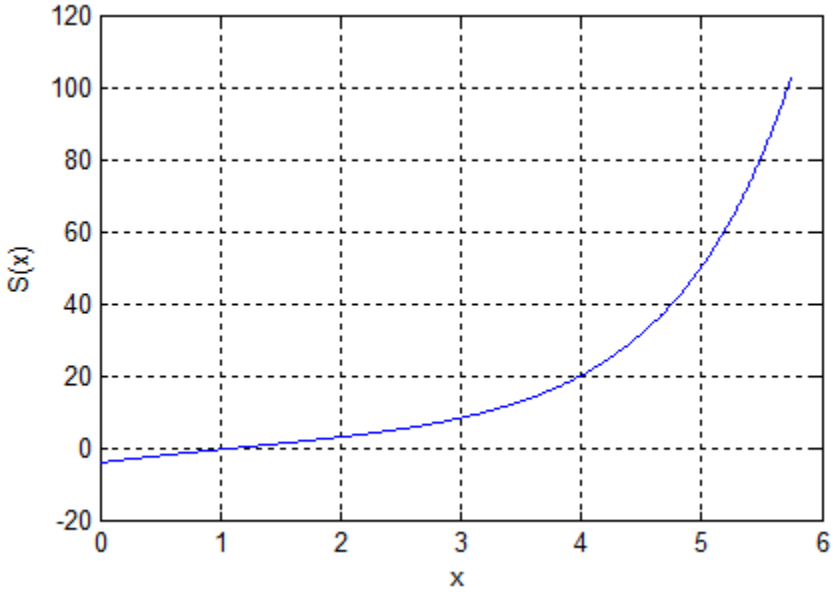
for all $x \geq w$, where p in the expression above is linked to w via

$$w = \frac{2p \ln p}{p+1}$$

- Note: The parameters m and w are in the units of natural logarithm
  - Thus, a range of 10^4 is specified as $m = 4 \ln 10 = 9.23$

Flow data displays: The logicle transformation

- The transformation is then defined by the inverse $S^{-1}(y)$ of the logicle function $S(x)$
  - The function $S(x)$ does not have a closed form inverse
  - The transformation is then carried out by linear interpolation
- The function $S(x)$ is defined for $0 \leq x \leq m$
  - For $x < w$, $-\mathrm{S}(\mathrm{w} - x)$ is computed



The logicle function $S(x)$ with

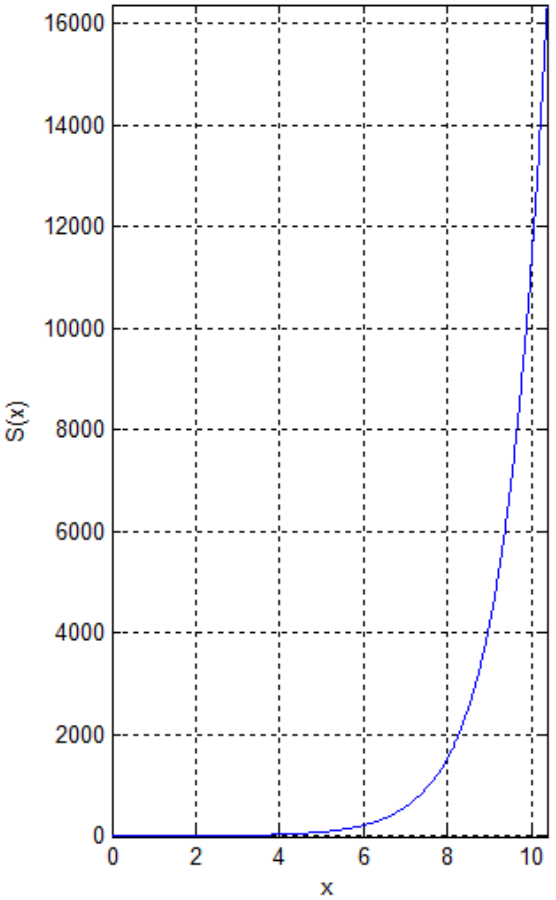
$T = 100$

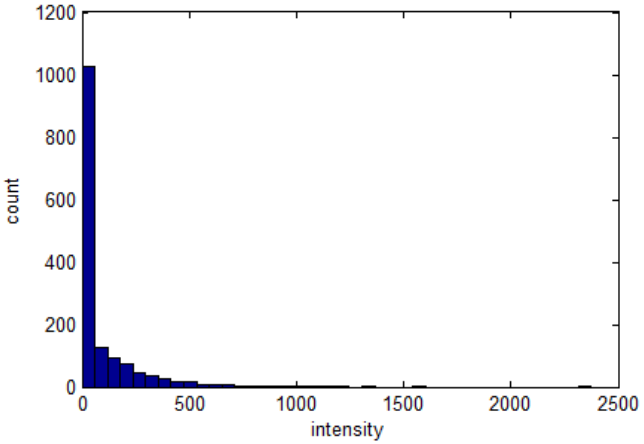
$m = 2.5 \log 10$

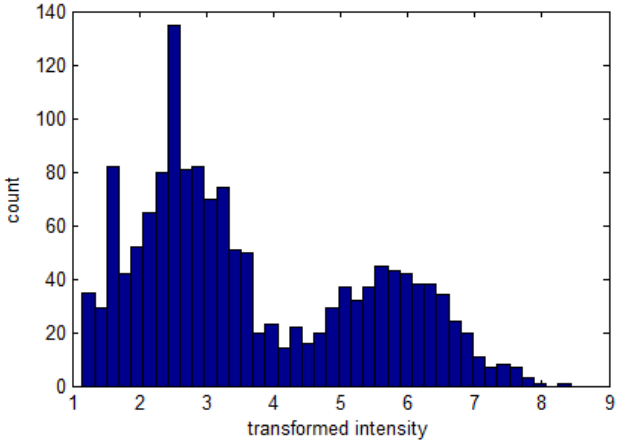
$w = 0.5 \log 10$

obtained using $p = 2.2872$

Flow data displays: The logicle transformation







The logicle transformation

Left: The logicle function using $T = 2^{14}$, $m = 4.5 \log 10$, $w = 0.5 \log 10$ for $p = 2.2872$
Middle: The histogram of the observed intensities
Right: Histogram of the transformed intensities

Univariate normalization of fluorescence intensities

- Once a suitable transformation (“display”) for the flow data is obtained, normalization of the observed intensities in different experiments can be addressed
- The first strategy is to aligned the intensity distributions along each channel individually
  - to ensure that similar cells exhibit similar fluorescence intensities in all experiments

➔ Univariate intensity normalization

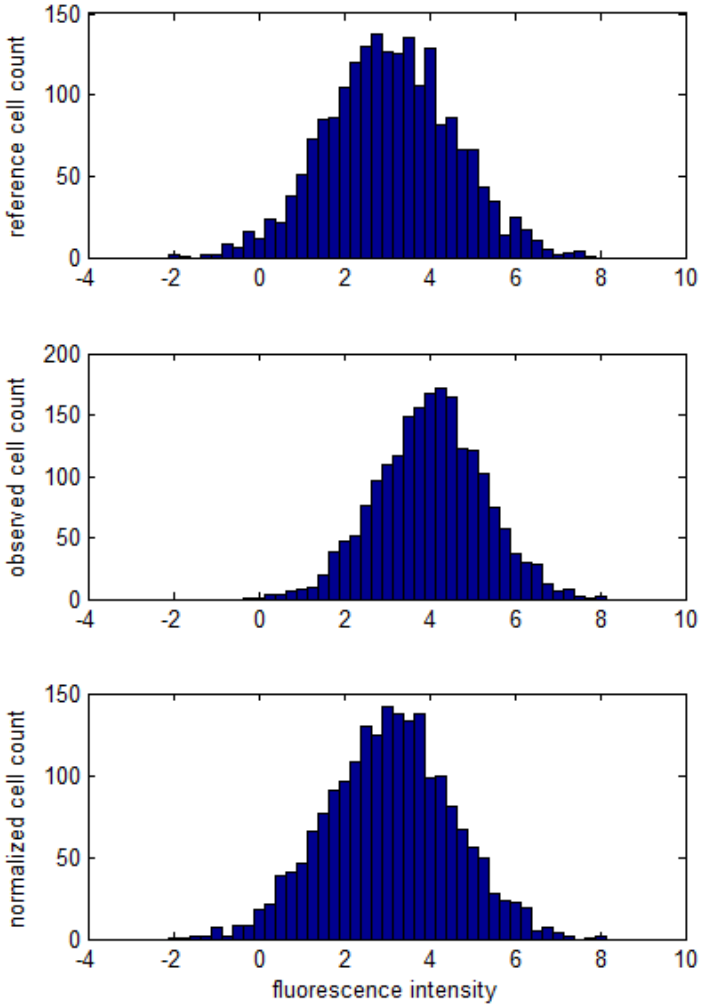
Univariate normalization of fluorescence intensities

- Normalization of fluorescence intensities corresponds to aligning the unknown underlying probability distributions
  - The probability distribution governing the fluorescence intensities in a flow experiment is to be
    - shifted
    - scaled
    - smeared
    - stretched
    - shrunk

  so that it matches a reference distribution
  - The underlying probability distribution is characterized by the observed fluorescence intensities within the experiment
  - The reference distribution is also characterized by another set of fluorescence intensities
- Conceptually, distribution alignment algorithms from the pattern recognition literature can be applied
  - with caution!!

Univariate normalization of fluorescence intensities

- Simplest case: alignment of monomodal intensity distributions
  - The observed fluorescence intensities form one single peak
  - The reference set of fluorescence intensities also form one single peak
  - In this case, intensity normalization can be carried out by a coordinate transformation
    - that aligns the peaks, and
    - adjusts the standard deviations



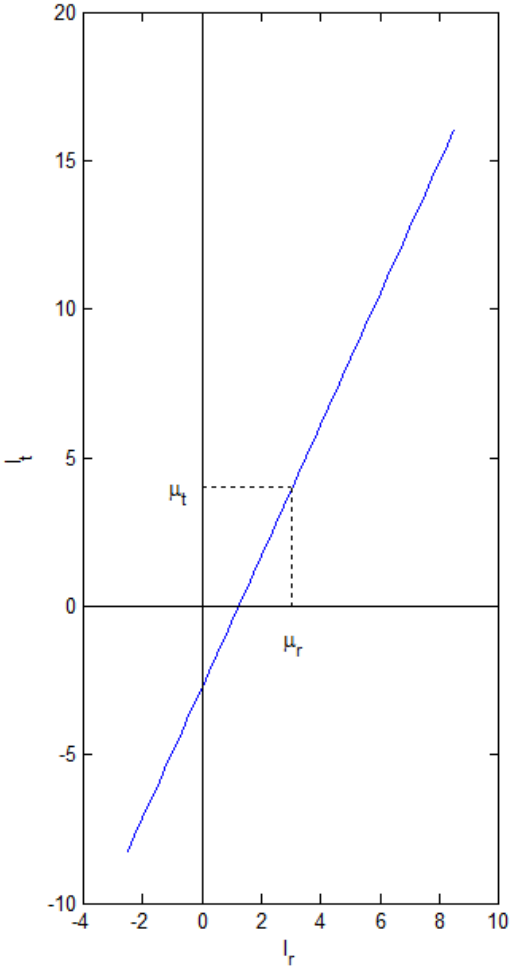
Univariate normalization of fluorescence intensities

- Simplest case (continued):
  - The normalization is carried out by a coordinate transformation $f: I_r \rightarrow I_t$ that links
    - the intensities I_r in the reference cell distribution
    - to the intensities I_t observed in the cell distribution to be normalized
  - For this case, the transformation corresponds to an affine function

$$f(I) = \sigma_t \left(\frac{I - \mu_r}{\sigma_r} \right) + \mu_t$$

where

- μ_r and σ_r are the mean and standard deviation of the reference cell fluorescence intensities, and
- μ_t and σ_t are the mean and standard deviation of the observed intensities to be normalized



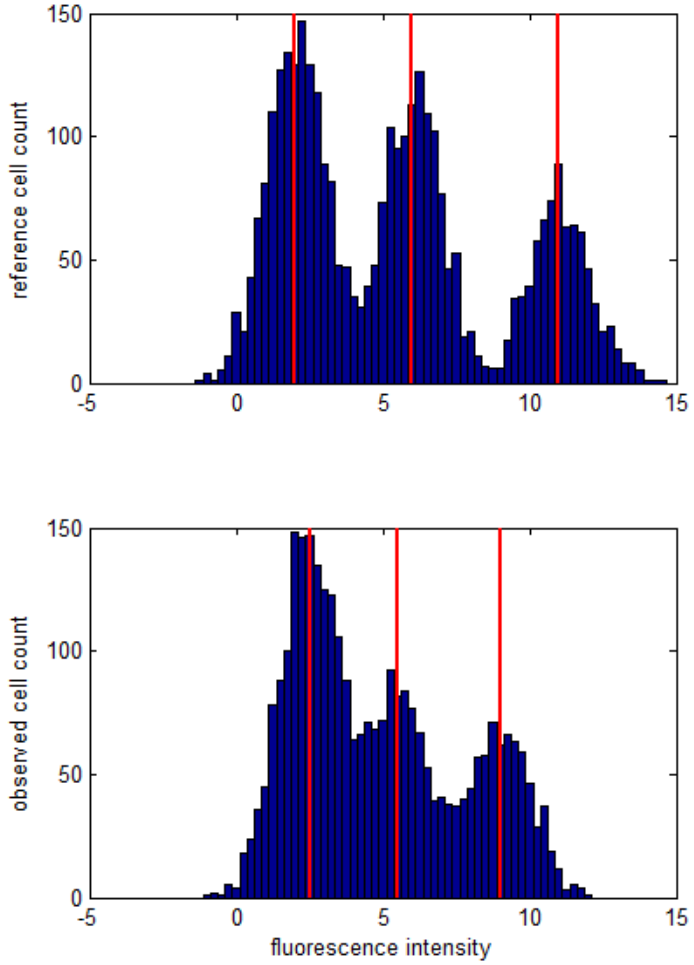
Univariate normalization of fluorescence intensities

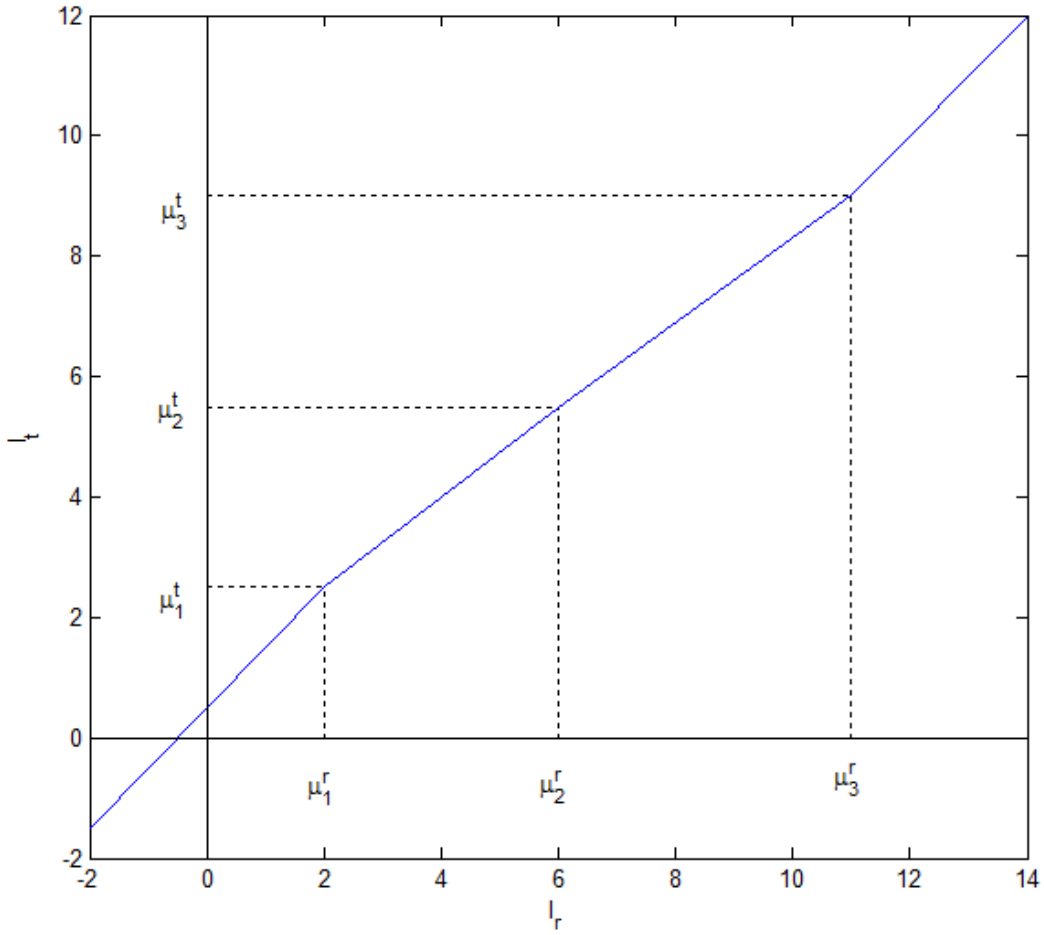
- Aligning multimodal intensities:
  - When the intensities to be aligned present multiple modes, the **mode correspondence problem** must be solved
    - In flow data, each intensity mode represents a distinct cell cluster
    - Normalization must ensure that the same cell clusters in the datasets to be normalized are associated with each other
      - so that they share similar intensities after normalization

Univariate normalization of fluorescence intensities

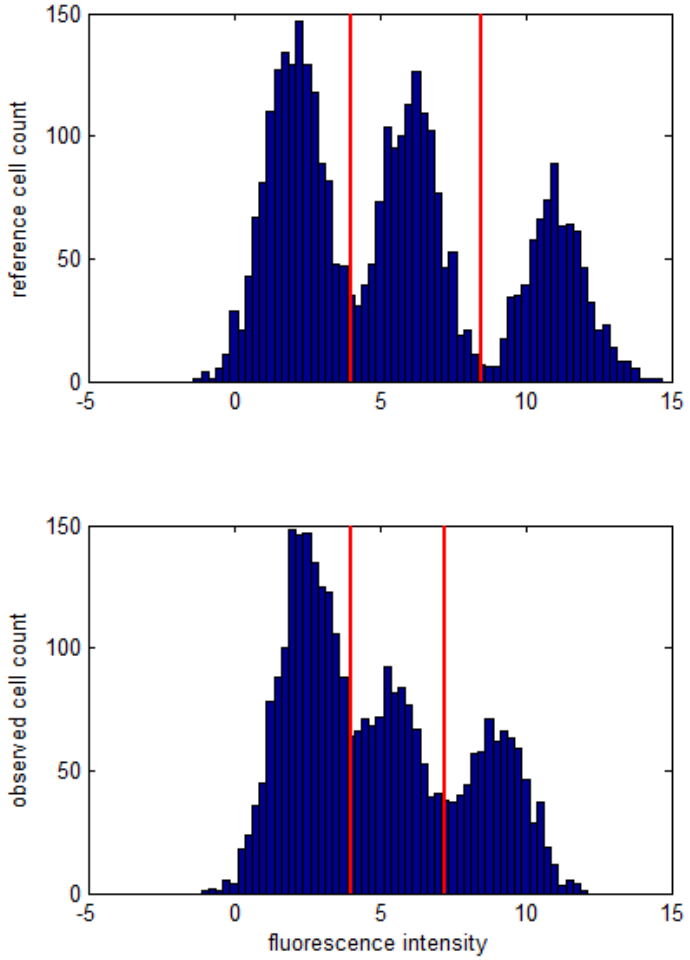
- Aligning multimodal intensities:
  - Assuming that
    - the different modes originated by distinct cell clusters are identified in both samples, and
    - the correspondence between the cell clusters of the two samples is established
  - A piece-wise linear function can be computed that either
    - repositions the mean intensities of the cell clusters in the reference dataset onto the mean intensities of the corresponding clusters in the observed dataset, or
    - matches the boundaries that separate the different cell clusters in their respective datasets

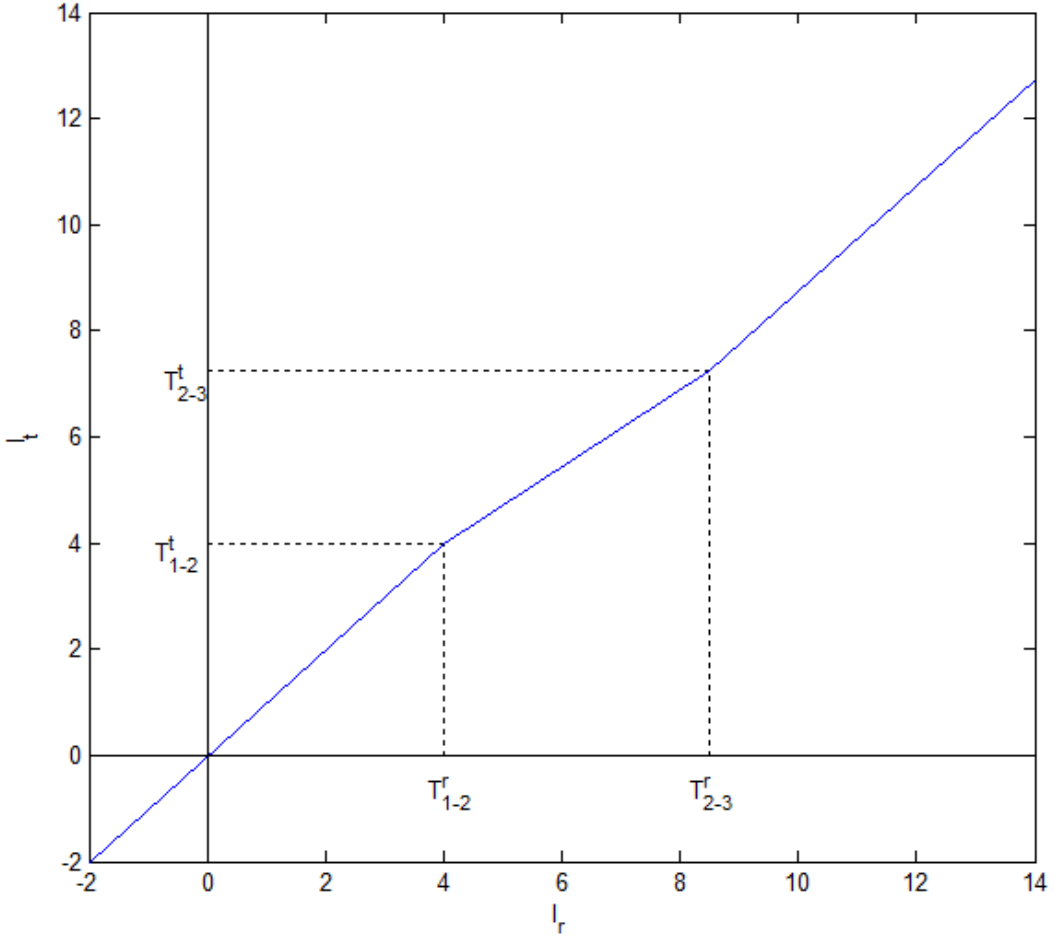
Univariate normalization of fluorescence intensities





Univariate normalization of fluorescence intensities





Univariate normalization of fluorescence intensities

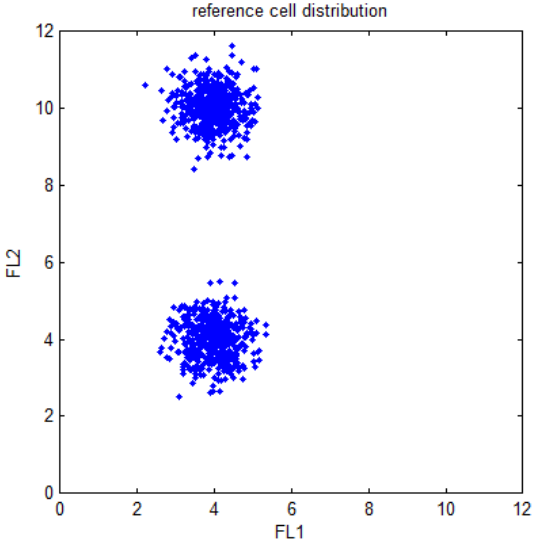
- Remarks:
  - Aligning the multimodal cell fluorescence intensity distributions requires correct identification
    - of the distinct cell clusters, and
    - of the correspondences between the matching cell clusters in the two samples
  - Mismatches inevitably lead to erroneous normalization that can potentially ruin the analysis
  - Missing cell clusters can potentially leave the alignment at an impasse
    - How to match a two-peak distribution onto another distribution that has three peaks?
  - General purpose univariate distribution alignment methods from the pattern recognition literature can be adapted to the task
    - earth movers distance
    - elastic coordinate transformations matching the respective cumulative distributions

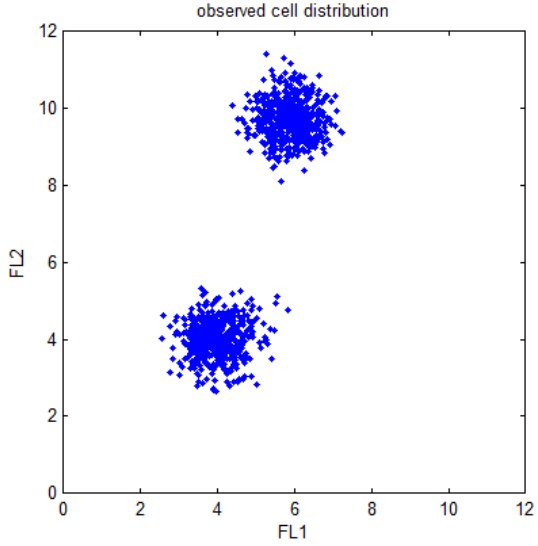
Multivariate normalization

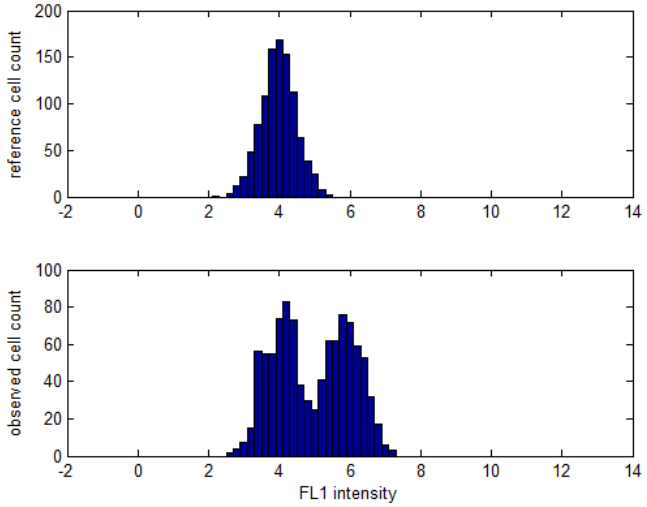
- While univariate normalization can adjust the fluorescence intensity histograms, it is blind to the correlated interactions between the intensities measured from different fluorochromes/biomarkers
  - correlations beyond the spectral overlap to be corrected by compensation
- Cell distributions in multicolor flow data are characterized by a correlated expression of the corresponding biomarkers
  - correlated expression of the biomarkers is observed as correlated fluorescence intensities
- Simple deviations from the experiment settings can throw the univariate intensity normalization procedures off
- A truly adept normalization procedure must take the full scope of the high dimensional flow data into account

Multivariate normalization

- Example: simple rotation
  - Consider two cell distributions
  - Both distributions possess -- and -+ cells
  - But the observed distribution is rotated slightly
    - due to variations in the settings, poor compensation, …
  - The correspondence between the distributions in the scatter plots is clear
  - But the histograms of the first fluorochrome intensities show
    - one peak in the reference cells
    - two peaks in the observed cells to be normalized







Summary

- Normalization of flow data acquired from different sources is still an open question
  - different samples, flow cytometers, setting, …
- Ideally, normalization should cancel out all non-specific and irrelevant variations in the fluorescence intensities
  - so that “had the samples of Experiment A been run at the settings of the Experiment B, the same type of cells would produce statistically identical fluorescence measurements”
- Multivariate normalization is clearly the ultimate goal, but must incorporate some form of recognition of the different cells clusters
  - so that matching cell clusters are aligned via the normalization
- However, the issue may involve combining the compensation and data normalization together
  - as effects of compensation can affect the data normalization