Rule Based Classification of SPE and IFE report

Li Hanyu

July 17, 2019

1 Labels

Every combination of SPE and IFE will be transformed into a 12-dimension 0-1 vector like [1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1]. Above is the label of lgG λ situation.

component	1	2	3	4	5	6
related columns	lgG,κ	lgG,λ	lgA,κ	lgA,λ	lgM,κ	lgM,λ
wheather 0 or 1	1 if obvious peaks have same position					

component	0	7	8	9	10	11
related columns	overall	lgG	lgA	lgM	κ	λ
wheather 0 or 1	only 0 when 7-11 is 0	1 if obvious peaks are detected				

1. The first component is the overall judgement(1 if abnormal and 0 if normal).

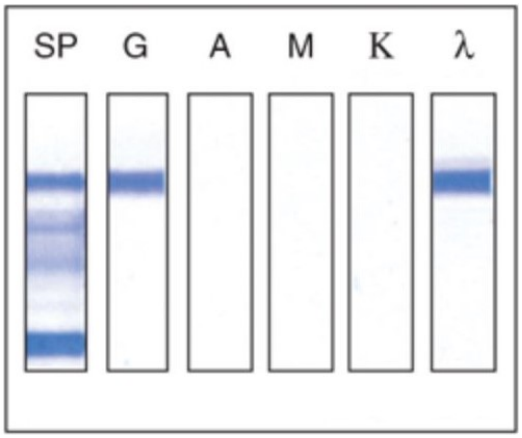
2. The next 6 components are the links between heavy chain(lgG, lgA, lgM) and light chain(κ and λ). For example, if lgG and λ has the same peak position, then the third component will be 1, because the order is:Gκ, Gλ, Aκ, Aλ, Mκ, Mλ.

3. The last 5 components are wheather hard and strong edges(obvious peak) are detected in the corresponding column of IFE.

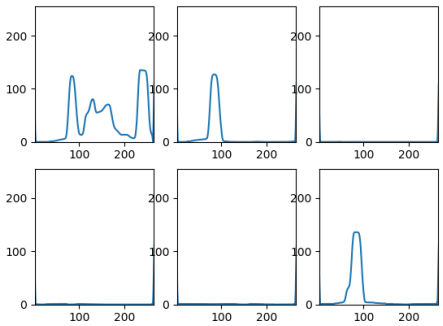
2 Classification Rules

2.1 Peak Detection Method(can only tackle single peak situation)

After reading in a picture of SPE and IFE like:(lgG and λ situation)



cv2.fastNlMeansDenoising are applied to every column of the electrophoresis result. Six density plots are then analysed according to the columns.



Afterwards, we apply cv2.Sobel(take derivative) to the density plots and find the max, max position, min, min position of the derivative curve. The column will be labelled as abnormal in the last five components of the label if the $\frac{max(f')+|min(f')|}{2}$ is higher than a formerly set threshold t_1. The link between certain kinds of heavy and light chain will be noticed if the difference of the peak position (which is determined by the mean of max position and min position) of the two chains are less than another threshold t_2, and both $\frac{max(f')+|min(f')|}{2}$ are higher than threshold t_3.

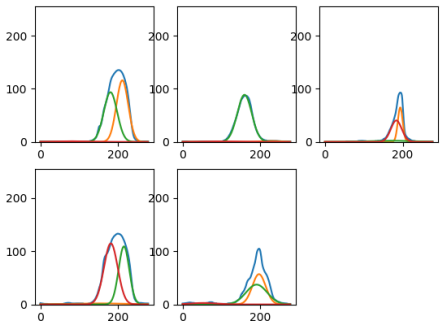
2.2 Gaussian Mixture Model Methods

Using derivative and second order derivative can lead to troublesome results sometime. In addition, coding a edge detection based model that is capable of dealing multiple peaks takes long time and hard work. On contrast, extracting peak information using Gaussian mixture model is both comparatively simple and effective.

Gaussian mixture model is an iteritive algorithm using estimation maximization. It assumes that the data is the weighted conbination of several Gaussian distribution. The number of the components(Gaussian distribution) are usually set manually. Specifically, we can view the density plot of the gel electrophoresis as a sample, where the sample number of certain position is in direct proportion to the density of that position. Thus, ideally the density plot will be divide in to seperate components of Gaussian distribution which approximates the concentration of charged protains(peaks).

After we find the proper decomposition of the density plot(examples in section 3.2), we can use our prior knowledge to match peaks in different columns and determine if a peak indicates monoclonal component.

Firstly we try to preprocess the peak information provided by the GMM(Gaussian Mixture Model). GMM divide the density plot to several distinct Gaussian components, therefore it is possible that it may wrongly treat one polyclonal peak as the sum of two sharp(i.e. monoclonal) peaks. We can name the situation as "far overlap". For instance the 1 and 4 column in the picture below.



However, a sharp peak in the middle of a polyclonal background is also a tricky problem. If the mean of the background peak and the sharp peak is almost the same, the GMM will give us unsatisfying result where the extracted sharp peak is not as sharp as we think. For instance column 5 in the picture above. We can name this case "near overlap".

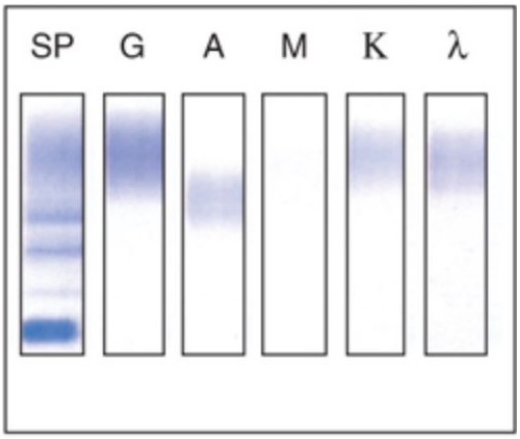
To sum up, the classification rules should be able to identify near and far overlap and of course stay robust when countering these cases. Identifying is easy because all we need to do is considering the means of every pair of components in one column for each column. Hence, the rule has two steps: preprocessing and classifying. In preprocessing step, we supress the far overlap components so that each one of them will be considered normal. Of course when the weight difference or the variance difference is too large this mechanism will not be trigered. At the same time, the weight of each peak of the near overlap case will be replaced by $2\times(sumofweight)$. In the classifying step we can consider every components with the all the components in the rest of the columns, and the rules are similar with the peak detection method in the column matching part, but here we use $t = \frac{Variance}{n\_samples*weight}$ as the Identifying value. If t is lower than a manually set threshold than the classifier will lable it as abnormal.

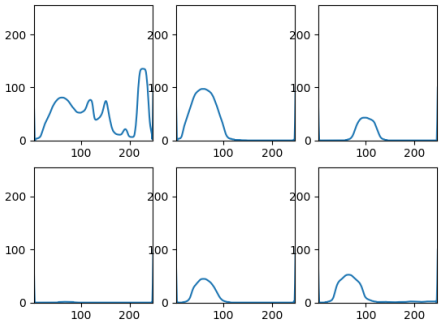
3 Instances

3.1 Peak Detection Methods

For normal and polyclonal increase instances:

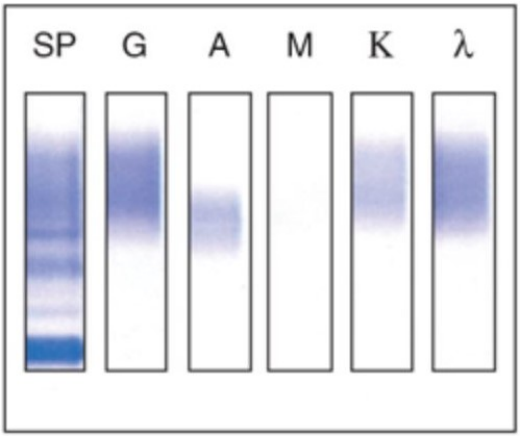
3.1.1 normal

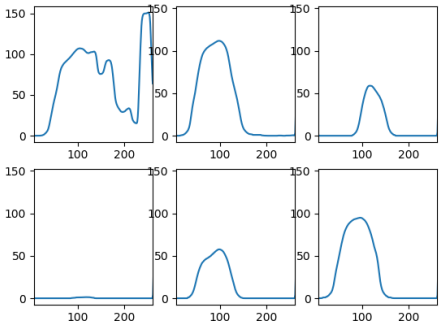




label: [0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

3.1.2 polyclonal increase

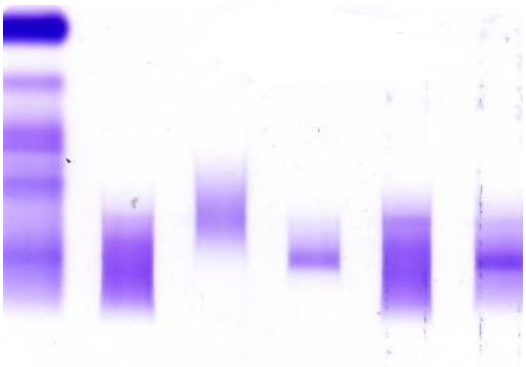


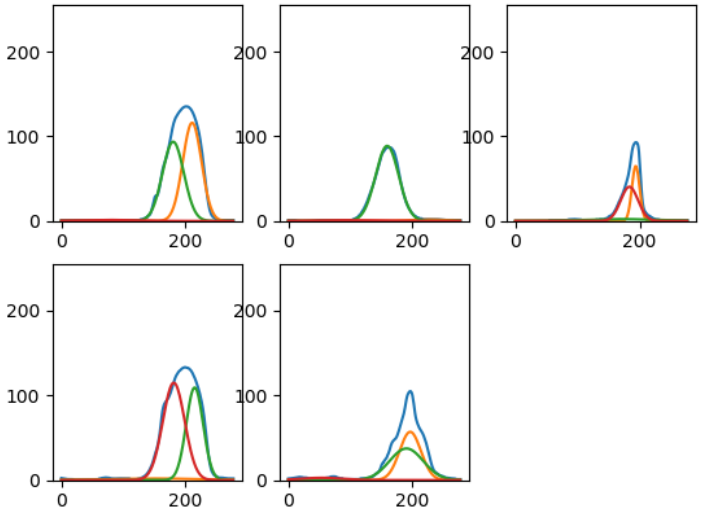


label:[0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

3.2 Gaussian Mixture Model Methods

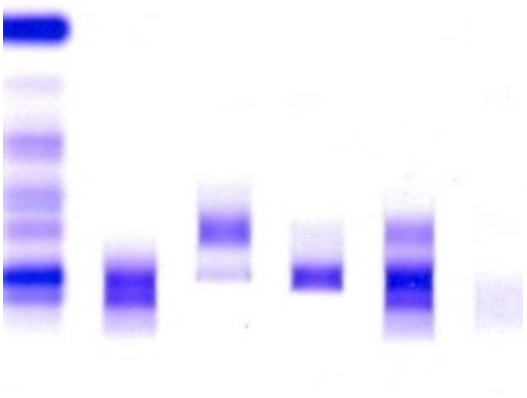
3.2.1 IgM+, Lambda+:

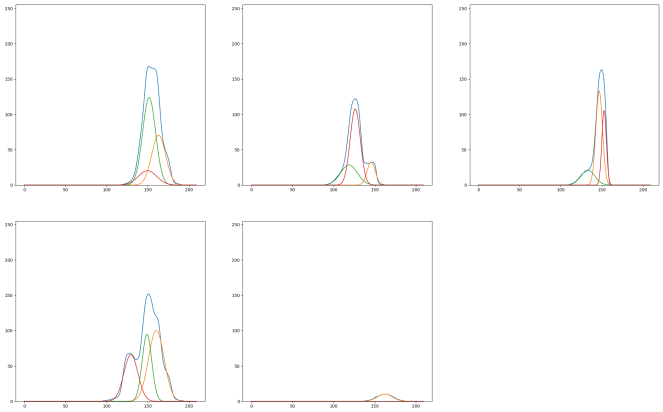




label:[1,0,0,0,0,0,1,0,0,1,0,1]

3.2.2 IgG+, Kappa+; IgA+, Kappa+; IgM+, Kappa+





label:[1,1,0,1,0,1,0,1,1,1,1,0]