Clustering by IF-PCA for sequencing data

Jiashun Jin and Wanrong Zhang

Jiashun Jin(CMU) Wanrong Zhang(PKU)

November 29, 2015

Higher Criticism(HC)

In his Class Notes for Statistics 411 at Princeton University in 1976, Tukey introduced the notion of the higher criticism by means of a story. A young psychologist administers many hypothesis tests as part of a research project, and finds that, of 250 tests 11 were significant at the 0.05 level. The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance. In that sense, finding only 11 significant results is actually somewhat disappointing!

He then proposed a sort of second-level significance testing, based on the statistic
$HC_{0.05,n} = \sqrt{FractionSignificantat 0.05) - 0.05}/\sqrt{0.05 * 0.95}$, and suggested that values of (say) 2 or greater indicate a kind of significance of the overall body of tests.

Higher Criticism(HC)

- Proposed by Donoho and Jin(2004) for sparse signal detection
- Higher Criticism is effective at resolving a very subtle testing problem: testing whether n normal means are all zero versus the alternative that a small fraction is nonzero:

$$H_0 : X_i \sim N(0,1), i = 1, 2, \ldots, n$$

$$H_1 : X_i \sim (1-\lambda)N(0,1) + \lambda N(\mu, 1), i = 1, 2, \ldots, n$$

Higher Criticism(HC)

Let $p_i = P(N(0,1) > X_i)$ be the p-value for the ith component null hypothesis, and let the p(i) denote the p-values sorted in increasing order, so that under the intersection null hypothesis the p(i) behave like order statistics from a uniform distribution.

$$HC_n = max_{1 \leq i \leq n\alpha_0} \sqrt{n}(i/n - p_{(i)})/\sqrt{p_{(i)}(1 - p_{(i)})}$$

To use HC_n to conduct a level- test, we must find a critical value $h(n, \alpha)$:

$$P_{H0}(HC_n \geq h(n, \alpha)) \geq \alpha$$

Higher Criticism(HC)

THEOREM 1 Under the null hypothesis H_0,

$$\frac{HC_n}{\sqrt{2loglog(n)}} \xrightarrow{p} 1$$

THEOREM 2 Consider the higher criticism test that rejects H0 when

$$HC_n > h(n, \alpha_n)$$

where $h(n, \alpha_n) = \sqrt{2loglog(n)}(1 + o(1))$

Higher Criticism(HC)

The asymptotic detection boundary

$$\lambda_n = n^{-\beta}, \mu_n = \sqrt{2rlog(n)}$$

The detection boundary is

$$f(\beta) = \beta - \frac{1}{2}, \frac{1}{2} \leq \beta \leq \frac{3}{4}$$

$$f(\beta) = (1 - \sqrt{1-\beta})^2, \frac{3}{4} \leq \beta \leq 1$$

if $r > f(\beta)$, H_0 and H_1 separate asymptotically, if $r < f(\beta)$, H_0 and H_1 merge asymptotically.

Clustering by IF-PCA

Sequencing data:

- High dimensional data $p >> n$
- Discrete(read counts data)

Dimension reduction: Idea: PCA applied to a small fraction of selected features

Negative binomial model

Let Y be an negative binomial random variable with mean μ and dispersion ϕ, denoted $Y \sim NB(\mu, \phi)$. The probability function is

$$f(y; \mu, \phi) = P(Y = y) = \frac{\Gamma(y + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y + 1)}(\frac{1}{1 + \mu\phi})^{\phi^{-1}}(\frac{\mu}{\phi^{-1} + \mu})$$

giving $E(Y) = \mu$ and $var(Y) = \mu + \phi\mu^2$. Consider $Y_1, \dots, Y_n$ follow $NB(\mu_i = m_i\lambda_j, \phi)$, where m_i is the library size (the total number of RNA-seq reads in sample i) and λ_j represents the proportion of the library that can map to the particular gene j. For K different groups, we can model

$$Y_i \sim NB(m_i\lambda_k, \phi), i \in groupK$$

Calibration between samples

Denote the library size of sample i by m_i , scaling each row of X and denote the resultant matrix by W:

$$W(i,j) = round[X(i,j)/m_i * \bar{m}]$$

We round W to integer in order to fit negative binomial model we presented before.
Denote the counts matrix by:

$$X = [X_i, X_2, \ldots, X_n]^T$$

which each column represents the feature and each row represents the sample.

Important features

Classical PCA is ineffective in such setting that $p >> n$, because noise can be strong. But we can think of a method to reduce the dimension, only left some useful features. If j is a useless feature for clustering, we have

$$m\lambda_1(j) = m\lambda_2(j) = \ldots = m\lambda_K(j)$$

Important features

$$H_{0,j} : X_{ij} \sim NB(m\lambda(j), \phi), i = 1, 2, \ldots, n$$

$$H_{1,j} : X_{ij} \sim \sum_{k=1}^{K} NB(m(\lambda(j) + \lambda_k(j)), \phi), i = 1, 2, \ldots, n$$

The null sampling distribution may be approximated by the Monte Carlo estimate. A p-value can be obtained as the proportion of simulated samples that produce a Pearson statistic as extreme or more extreme than the observed one.

Implement of Monte Carlo goodness of fit test

- Fit an NB model from the data $Y^0 = (Y_1, \ldots, Y_n)^T = W(:, j)$,estimate all the parameters by MLE and calculate Pearson residuals $r^0 = (r_1^0, \ldots, r_n^0)$ $(r_i = \frac{Y_i - \mu}{s}$,where μ is estimated mean and s is estimated standard error.
- Simulate R random vectors: For $h = 1, \ldots, R$, simulate a random vector Y^h from $NB(\mu, \theta)$ and compute Pearson residual r^h as step 1.
- Compute the sum of squared deviation of each residual vector from the median of its sampling distribution$d^h = \sum_{i=1}^n (r_i^h - \bar{r}^h)^2$ for $h = 0, 1, \ldots, R$
- Compute a Monte Carlo GOF test p-value by$\pi_j = \frac{\sum_{h=1}^R I(d^h \geq d^0)}{R}$, where $I()$ is the indicator function.

Choosing threshold by HC

- Sort P-values in the ascending order $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(p)}$.
- Define $HC_j = \sqrt{p}(j/p - \pi_{(j)})/\sqrt{max\sqrt{n}(j/p - \pi_{(j)}), 0 + j/p}$ and let $\hat(j) = argmax_{j:\pi_{(j)}>log(p)/p, j<p/2} HC_j$. HC threthold t_p^{HC} is the j-smallest p-value.

Clustering by PCA and K-means

Let $W_{(HC)}$ be the matrix formed by restricting the columns of W: $W_{(HC)} = W(:, \pi_j < t_p^{HC})$, and U be the matrix of the first $K-1$ left singular vectors of $W_{(HC)}$. Clustering U by classical K-means.

Clustering results

Data set	p	n	selected features	K
Mouse	10250	21	2483	2
Human	8247	41	1287	2
Mont	8361	129	217	2
Kidney	20531	144	1607	2

Table: basic information of data sets

Clustering results

Data set	K	classical	IF-PCA
Mont	2	.201	.116
Kidney	2	.035	.021

Table: comparison of clustering error rate

Clustering results

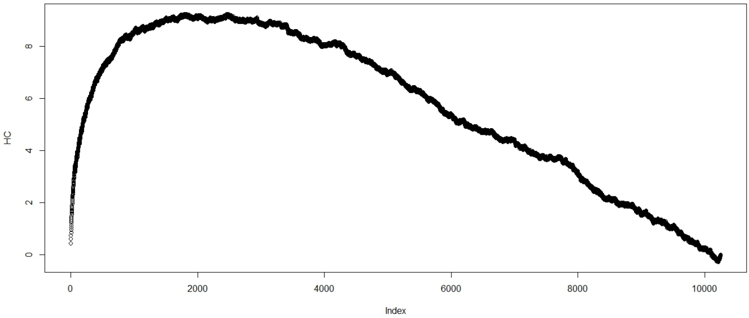

Figure: HC plot(mouse data set)

clustering results

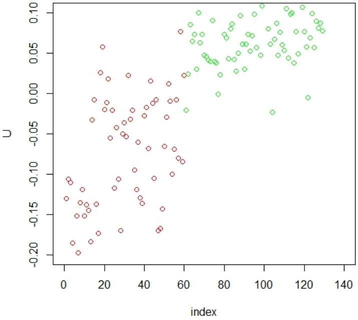

Figure: result of IF-PCA

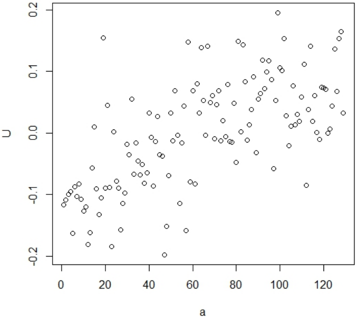

Figure: result of classical PCA

clustering results

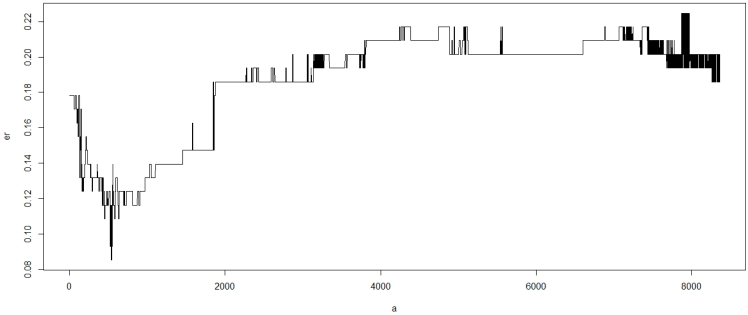

Figure: error rate by IF-PCA with different number of selected features(mont)

Acknowledgements

Main reference:

- Jin J, Wang W. Important Feature PCA for high dimensional clustering[J]. arXiv preprint arXiv:1407.5241, 2014.
- D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist., pages 962C994, 2004
- Mi G, Di Y, Schafer DW. Goodness-of-Fit Tests and Model Diagnostics for Negative Binomial Regression of RNA Sequencing Data. Rapallo F, ed. PLoS ONE. 2015;10(3):e0119254. doi:10.1371/journal.pone.0119254.