List 3

Kamil Zaborniak

Contents

Introduction

In the following report we will consider statistical models of sparse mixtures. Let's assume sample: $\underline{X} = (X_1, \dots, X_n)$, where $n > 1$, $i \in \{1, \dots, n\}$ and $X_i \sim N(\mu_i, 1)$. In the original model, the hypotheses are:

$$\begin{aligned} H_{0,i} :& X_i \sim N(0,1), \\ H_{1,i} :& X_i \sim N(\mu_i, 1), \text{ where } \mu_i > 0. \end{aligned}$$

We are interested in possibilities within H_1 with a small fraction of non-null hypotheses. Rather than directly assuming that there is some amount of nonzero means under H_1, we assume that our samples follow a mixture of $N(0,1)$ and $N(\mu, 1)$, with $\mu > 0$ fixed and with some mixture parameter ε. This simple model with equals means can be written as:

$$\begin{aligned} H_{0,i} :& X_i \sim N(0,1), \\ H_{1,i} :& X_i \sim (1-\varepsilon)N(0,1) + \varepsilon N(\mu_i, 1). \end{aligned}$$

The expected number of non-nulls under H_1 is $n\varepsilon$.

If ε and μ were known, then the optimal test would be the likelihood ratio test. The likelihood ratio test under the sparse mixture model is:

$$L = \prod_{i=1}^{n} \left[(1-\varepsilon) + \varepsilon \exp\left(\mu X_i - \frac{\mu^2}{2} \right) \right].$$

The dependencies between parameters are:

$$\varepsilon_n = n^{-\beta}, \qquad \text{where } \frac{1}{2} < \beta < 1,$$

$$\mu_n = \sqrt{2r \log(n)}, \qquad \text{where } 0 < r < 1.$$

The parameter β controls the set of non-nulls in the range 1 to $\sqrt{n}$, and thus the sparsity of the alternatives, while r paramterizes the mean shift. If β were large, then our problem would be very sparse, while if β were small, it would be mildly sparse. If $r = 1$, then we get the detection threshold we have seen for Bonferroni. Hence, the ''needle in a haystack" problem corresponds to $\beta = 1$ and $r = 1$.

There is a threshold curve for r of the form:

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta < \frac{3}{4}, \\ \left(1 - \sqrt{1-\beta}\right)^2 & \text{if } \frac{3}{4} \le \beta \le 1. \end{cases}$$

If $r > \rho^*(\beta)$, then we can adjust the Neyman-Pearson test to achieve:

$$P_0(\text{Type I Error}) + P_1(\text{Type II Error}) \to 0$$

and if $r \le \rho^*(\beta)$, then for any test, $\lim_{n\to\infty} \inf_{n\to\infty} P_0(\text{Type I Error}) + P_1(\text{Type II Error}) \ge 1$.

Unfortunately, we generally cannot use the Neyman-Pearson test since we do not know ε or μ.

Higher-Criticism tests

For $r > \rho^*(\beta)$ in the sparse mixture, the higher criticism statistic with a proper threshold has full power asymptotically. Additionally, HC statistic doesn't need knowledge of ε and μ. We have two types of HC statistics:

- Higher Criticism Test (Tukey 1976):

$$HC^* = \max_{\frac{1}{n}<t<\frac{1}{2}} \sqrt{n}\frac{F_n(t)-t}{\sqrt{t(1-t)}},$$

- Modification by Stepanova and Pavlenko (2014):

$$HC^*_{mod} = \max_{0<t<1} \sqrt{n}\frac{F_n(t)-t}{\sqrt{t(1-t)q(t)}}.$$

Where:

$$V_i(t) = \mathbb{I}_{p_i \le t},$$
$$F_n(t) = \frac{\sum V_i}{n},$$
$$q(t) = \log\log\frac{1}{t(1-t)}.$$

Probability of the type I error for HC_{mod}

We know that asymptotic critical value for significance level $\alpha = 0.05$ is equal to 4.14. Let's estimate the probability of the type I error for $n \in \{5000, 50000\}$.

Simulation

In a single iteration we generate n random variables from a uniform distribution (since the null hypothesis is true, the p-value is uniformly distributed on [0,1] and we can replace it with a random variable from a uniform distribution on [0,1]), based on them we determine the test statistic HC_{mod}. We repeat the iteration 1000 times, resulting in a vector of 1000 HC_{mod} statistics. An estimator of the probability of a type I error is the percentage of statistics greater than or equal to 4.14. We repeat such an experiment for each n.

Results

For $n = 5000$, estimated probability of type I error is equal to 0.041. For $n = 50000$, it is equal to 0.049. We see that that modificated HC test controls the probability of the type I error at a significance level α. What is more, this probability decreases as the size of sample increases.

Critical values

Simulation

To find critical values of Higher-Criticism tests for $n = 5000$ and significance level $\alpha = 0.05$, we will run the experiment. In a single iteration, we will generate 5000 random variables from the uniform distribution on [0,1] (the critical value is determined assuming that the null hypothesis is true, therefore we can replace the p-values by random variables from the uniform distribution on [0,1]), based on the obtained vector of variables, we will determine the statistics HC^* and HC_{mod}. We will repeat the iteration 10000 times, resulting in a sample of 10000 HC^* statistics and a sample of 10000 HC_{mod} statistics. The estimators of the critical values will be the 95% quantiles of the generated sample statistics.

Results

The estimated critical value of HC test i equal to 3.2084. The estimated critical value of modified HC test is equal to 4.1339. We see that for modified HC test, critical value is close to 4.14, unlike the standard HC test.

Sparse mixture models testing

Let $n = 5000$, $\underline{X} = (X_1, \dots, X_n)$, $X_i \sim N(\mu_i, 1)$. We will consider 3 statistical models:

- "Needle in a haystack" problem:

$$\begin{aligned} H_0 :& \mu_1 = \dots = \mu_{5000} = 0, \\ H_1 :& \mu_1 = 1.2\sqrt{2\log(n)},\ \mu_2 = \dots = \mu_{5000} = 0. \end{aligned} \tag{A}$$

-

$$\begin{aligned} H_0 :& \mu_1 = \dots = \mu_{5000} = 0, \\ H_1 :& \mu_1 = \dots = \mu_{100} = 1.02\sqrt{2\log\left(\frac{n}{200}\right)},\ \mu_{101} = \dots = \mu_{5000} = 0. \end{aligned} \tag{B}$$

-

$$\begin{aligned} H_0 :& \mu_1 = \dots = \mu_{5000} = 0, \\ H_1 :& \mu_1 = \dots = \mu_{1000} = 1.002\sqrt{2\log\left(\frac{n}{2000}\right)},\ \mu_{1001} = \dots = \mu_{5000} = 0. \end{aligned} \tag{C}$$

Let's assume a significance level of $\alpha = 0.05$ and let's test the above models using Higher-Criticism tests and the tests below:

Bonferroni method , where test statistic has form:

$$B = \min_{1 \le i \le n} p_i,$$

where $p_i = 1 - \Phi(X_i)$. We reject H_0 when it is lower than or equal to $\frac{\alpha}{n}$.

χ^2 **test** with statistic:

$$\chi^2 = \sum_{i=1}^{n} X_i^2,$$

where under H_0: $T_{\chi^2} \sim \chi^2_n$ and we reject H_0 when T_{χ^2} is greater than $1 - \alpha$ quantile of χ^2 distribution with n degrees of freedom.

Fisher test , that has statistic:

$$F = -\sum_{i=1}^{n} 2\log(p_i),$$

where $p_i = 1 - \Phi(X_i)$ and $F \sim \chi^2_{2n}$. We reject global null when F is greater than $1 - \alpha$ quantile of χ^2_{2n} distribution.

Kolmogorov-Smirnov test , where:

$$KS = \sup_{t \in [0,1]} \sqrt{n} \left| \hat{F}_n(t) - t \right| ,$$

where $\hat{F}_n$ is empirical cdf od p-values. We reject H_0 when KS is greater than $1-\alpha$ quantile of Kolmogorov-Smirnov distribution. In this report we will use *ks.test()* function from R, and we will reject H_0 when the obtained p-value is lower than α.

Anderson-Darling test with statistic:

$$A^2 = n \int_0^1 \left(\hat{F}_n(t) - t \right)^2 \frac{1}{t(1-t)}\, dt.$$

We will also use *ad.test()* and we will reject H_0 if the obtained p-value is lower than α.

Let's test models A, B and C with each test and then determine the test powers.

Estimated powers of tests

Simulation

To determine the power of the above texts for each model under consideration, we need to run an experiment. To determine the power of the above texts for each model under consideration, we need to conduct an experiment. Through one iteration, we will generate 5000 observations from a normal distribution assuming the truth of the alternative hypothesis (without loss of generality of the assumptions, we will generate a sample of random variables, where the ith observation will be generated from a normal distribution with mean μ_i). Then we will determine the test statistics according to the above procedures and run the tests. We will repeat this iteration 1000 times, as a result of which we will obtain a zero-one vector of size 1000 for a given test. The power estimator of a given test will be the average of the obtained vector.

Results

	Tests						
models	mod. H-C	H-C	Bonferroni	χ^2	Fisher	Kolmogorov-Smirnov	Anderson-Darling
A	0.093	0.101	0.759	0.094	0.081	0.050	0.053
B	1	1	0.994	1	1	0.805	0.996
C	1	1	0.847	1	1	1	1

Table 1: Estimated powers of tests (Task 3).

We see that in the case of model A, the best test is the Bonferroni procedure, which confirms that the best test for the “needle in a haystack” problem is the Bonferroni method. The worst tests in this case are the Kolmogorov-Smirnov and Anderson-Darling tests. We also see that the Fisher test is slightly worse than the chi-square test, which confirms the thesis from the previous lists.

For model B, the best tests are: modified Higher-Criticism and Higher-Criticism, χ^2, Fisher. The worst test is the Kolmogorov-Smirnov test. However, all tests, even the Bonferroni method, achieve very high powers in this case. The reason for this may be a small percentage of sufficiently strong signals.

In the case of model C, as we expected, the worst test is the Bonferroni method, in this model there are too many too weak signals. All tests except it have a power of 1.

This exercise showed that we should use Higher-Criticism tests for testing sparse mixtures, except for ‘’needle in a haystack” problems.

Sparse mixture model

Let's assume significance level $\alpha = 0.05$ and a sample of random variables: $\underline{X} = (X_1, \ldots, X_n)$, where $n \in \{5000, 50000\}$. We will consider the sparse mixture model:

$$H_0 : \underline{X} \sim N(0,1),$$
$$H_1 : \underline{X} \sim (1-\varepsilon)N(0,1) + \varepsilon N(\mu, 1),$$

with $\varepsilon = n^{-\beta}$, $\mu = \sqrt{2r\log(n)}$. For each of the settings $\beta = \{0.6; 0.8\}$, $r = \{0.1; 0.4\}$ and n, we will test above model using Neyman-Pearson test. We see that $\rho^*(0.6) = 0.1$, $\rho^*(0.8) = 0.3056$. Based on this information, we want to use simulations to check the power of the tests depending on whether r is smaller or larger than $\rho^*(\beta)$.

Critical values for Neyman-Pearson test

Before testing above sparse mixture model, we have to estimate critical values of Neyman-Pearson test. The test statistic of N-P test is:

$$L' = \sum_{i=1}^{n} \log\left((1-\varepsilon) + \varepsilon \exp\left(\mu X_i - \frac{\mu^2}{2}\right)\right).$$

Simulation

As before, the simulation will consist in repeating a single iteration 10000 times for each β, r, n. In a single iteration, we generate n random variables from standard normal distribution (critical values we estimate under H_0). Then, based on them, we determine the L' statistic. After repeating this iteration 10000 times, we obtain a vector of 10000 statistics L'. The estimator of the critical value of the test is the $1-\alpha$ quantile of the obtained statistic vector.

Results

n	5000				50000			
β	0.6		0.8		0.6		0.8	
r	0.1	0.4	0.1	0.4	0.1	0.4	0.1	0.4
Critical value	1.015043	-1,634468	0.2721396	1.088169	1.07162	-7.636816	0.1605569	1.038714

Table 2: Estimated critical values of Neyman-pearson test for $\alpha = 0.05$.

Interestingly, when $r = 0.4$ and $\beta = 0.6$, i.e. r is significantly larger than ρ^*, the critical values are negative. In the case when the values of r and ρ^* are close to each other, the critical value oscillates around 1. If r is significantly smaller than ρ^*, i.e. the cases when $r = 0.1$ and $\beta = 0.8$, then the critical values are close to 0. We also see that with increasing size sample, depending on the β and r values, the critical values decrease.

Powers of test

After determining the critical values of the Neyman-Pearson test, we can compare its power with other tests.

Simulation

To estimate the power of the tests, we will repeat the iteration 1000 times. Using it, we will generate, without loss of generality, $\varepsilon n = n^{1-\beta}$ random variables from the $N(\mu, 1)$ distribution and $(1-\varepsilon)n$ random variables from the standard normal distribution (we estimate the power under the alternative). Then, we will perform each of the following tests on the obtained sample: Neyman-Pearson, both versions of HC, Bonferroni, Fisher and chi-square. The Neyman-Perason test will reject H_0 when L' is greater than the appropriate critical value. Each test will return 1 if the test rejects H_0, otherwise 0. As a result of 1000 repetitions of this iteration, we will obtain a zero-one vector for each test. The mean of the vector will be the estimator of the test power. We will repeat the simulation for each case depends on β and r, n.

Results

β	$\rho(\beta)^*$	r	n	Neyman-Pearson	Higher-Criticism	modified Higher-Criticism	Bonferroni	χ^2	Fisher
0.6	0.1	0.1	5000	0.208	0.175	0.121	0.096	0.146	0.197
			50000	0.233	0.149	0.113	0.098	0.123	0.180
		0.4	5000	0.995	0.964	0.683	0.769	0.626	0.646
			50000	1.000	0.998	0.843	0.930	0.639	0.628
0.8	0.3056	0.1	5000	0.085	0.067	0.060	0.052	0.072	0.063
			50000	0.064	0.048	0.053	0.050	0.079	0.066
		0.4	5000	0.283	0.165	0.058	0.241	0.108	0.097
			50000	0.406	0.096	0.058	0.311	0.067	0.076

Table 3: Powers of tests for testing sparse mixtures (task 4b).

We see that, as in the previous lists, the Neyman-Pearson test has the greatest power of all the others, regardless of chance. We see that when the $r - \rho^*(\beta)$ value increases, the power of the Neyman-Pearson (and other) tests increases and is relatively high. When $r < \rho^*(\beta)$ then no test is appropriate to test the model under consideration. These conclusions support the theory from the lecture in Chapter 1. In cases where $r > \rho^*(\beta)$, we see that the power of the N-P test increases with sample size.

Regarding other tests, we see that when the $r - \rho^*(\beta)$ value is set to 0, the Fisher and chi-square tests perform best. However, if $r - \rho^*(\beta)$ is relatively low (but positive), the best test is the Bonferroni procedure. The power of the H-C test behaves similarly to the power of the N-P test, however, the power of one of the tests: Bonferroni, chi-square, Fisher is always slightly greater than them.

We see that as n increases and $r - \rho^*(\beta)$ increases, the tests powers converge to 1.

Test Functions

```r
p_val<- function(sample){
  return(c(1-pnorm(sample)))
}

HC_mod_stat <- function(p_val_vec){
  n<- length(p_val_vec)
  F_n <- ecdf(p_val_vec)
  t<- seq(0.00001, 0.99999, by=0.00001)
  q<- log(log( 1/(t*(1-t)) ) )

  HC_mod<- max(
    sqrt(n) * (F_n(t)-t) / sqrt(t*(1-t)*q)
  )

  return(HC_mod)
}
HC_mod_test <- function(sample){
  p_vec<- p_val(sample)
  return(as.numeric(HC_mod_stat(p_vec)>C_HC_mod))
}

HC_stat <- function(p_val_vec){
  n<- n<- length(p_val_vec)
  F_n <- ecdf(p_val_vec)
  t<- seq(1/n +0.00001, 1/2-0.00001, by=0.00001)

  HC<- max( sqrt(n) * (F_n(t)-t) / sqrt(t*(1-t)) )

  return(HC)
}
HC_test<- function(sample){
  p_vec<- p_val(sample)
  return(as.numeric(HC_stat(p_vec)>C_HC))
}
Bonferroni_test<- function(sample){
  p_vec<- p_val(sample)
  n<-length(sample)
  B<- min(p_vec)
  return(as.numeric(B < 0.05/n))
}
chi2_test<- function(sample){
  chi<- sum(sample^2)
  n<- length(sample)
  return(as.numeric(chi > qchisq(0.95,n)))
}

Fisher_test<- function(sample){
  p_vec<- p_val(sample)
  n<- length(sample)
  F_stat<- -2*sum(log(p_vec))
  return(as.numeric(F_stat> qchisq(.95,2*n)))
}
KS_test<- function(sample){
  p_vec<- p_val(sample)
  return(as.numeric(ks.test(p_vec, "punif", alternative="greater")$p.value < 0.05))
}
AD_test<- function(sample){
  p_vec<- p_val(sample)
  return(as.numeric(ad.test(p_vec)$p.value < 0.05))
}

L_prim<- function(sample, r, beta){
  n<- length(sample)
  epsilon<- n^(-beta)
  mu<- sqrt(2*r*log(n))
  L_prim<- sum(log(
    (1-epsilon)+epsilon*exp(mu*sample-mu^2/2)
      ))
  return(L_prim)
}

NP_test<- function(crit_v, ...){
  return(as.numeric(L_prim(sample, r, beta)>crit_v))
}
```