Final Project: False Discovery Rate

Zhong C.F. Li

December 2024

1 Introduction

This report aims to summarize and reproduce the key results of the paper *False Discovery Rate: A New Deal* (ND). The codes for my own implementation can be found at the following link: `https://github.com/FlyingBeyondUp/CASIA_FDR`.

For clarity, the key contributions of the ND paper are summarized, and the arrangements of codes are explained in the second section. After that, the two-group model introduced in the lecture note is reviewed, which can be used as an example to compare with the algorithm proposed by the ND paper. Lastly, the reproduced results of the ND paper are shown in detail.

2 Summary

Due to the unavoidable noise in measurement, the main challenge of large-scale hypothesis testing is the trade-off between declaring more discoveries while avoiding false discoveries. The (local) false discovery rate provides a measure of the trade-off. More detailed reviews and discussions are presented in next section.

The main contributions of the ND paper are introducing a new **Bayesian** approach to large-scale hypothesis testing with the following features

- Unimodal assumption: The prior alternative (non-null) distribution is explicitly modeled as a linear combination of a series of unimodal distributions.
- Two input numbers intead of a single p-value: The observed effect strength $\hat{\beta}$ and the estimated standard error $\hat{s}$ are used.
- New concept local false sign rate: The probability of getting the sign of an effect wrong.

The advantages related with these features are

- Unimodal assumption: predicts nonzero alternative distribution near zero instead of assuming that all z scores near zero are null, meaning that

"small" effects can also be captured; the estimation of π_0, which is the weight of the null component, remains to be conservative, meaning that the algorithm still tries to avoid declaring false discovery; provides a simple way to estimate the prior distribution of the observed data set, which is generally difficult due to its de-convolution nature.

- Two numbers to summarize the measurement: allowing the measurement precision to be different for different observations;
- Local false sign rate: more robust to modeling assumptions than local false discovery rate, less sensitive to the estimation of π_0, which is hard for the data set generated from specific prior distribution (shown latter); down weights the observations with poor measurement precision, meaning that it is more robust to the noise.

The arrangements of codes are

- Two Group Model: includes the function that obtain marginal, null, alternative distributions and lfdr by the two-group model.
- Distributions: defines the different prior alternative distributions used in the numerical simulations.
- NewDeal: includes a data generation function and the class NewDeal, which accept two lists (numpy.array) $\hat{\beta}$ and $\hat{s}$ as inputs and provides methods to estimate π and other related quantities. Two different prior alternative distributions are realized, which are normal and uniform distributions, respectively.
- PlotFigures: includes all the codes to plot the figures shown in this report.

3 False Discovery Rate and Two-Group Model

In this section, the concept and numerical simulations of the two-group model are presented as a review and baseline of the local false discovery rate.

Given M noisy observations $\{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_M\}$ of effects $\{\beta_1, \beta_2, ..., \beta_M\}$, a natural question is which of the effects are indeed zero, which can be depicted by the null hypothesis $\mathrm{H}_j : \beta_j = 0$, $j = 1, 2, ..., M$. A nonzero effect is called a discovery, thus a falsely rejected null hypothesis corresponds to a falsely declared discovery. The target is to find more discoveries while avoiding false discoveries.

The p-value was created to test whether each null hypothesis was valid or not. Under the null hypothesis, if the noise added to the j-th observation is assumed to follow a normal distribution with zero mean and $\hat{s}_j$ standard deviation, then the z score is defined as $z_j = \hat{\beta}_j / \hat{s}_j \sim \mathcal{N}(0, 1)$ and the corresponding p_j value is deined as $p_j = 1 - \Phi(z_j)$, where $\Phi(z)$ represents the cumulative distribution function of the standard normal distribution. A smaller p value indicates that it is less probable to observe a test statistic as extreme as $\hat{\beta}_j$ under the assumption

that the null hypothesis is true. Usually, the null hypothesis H_j is rejected for $p_j < 0.05$.

However, the idea of p-value only focuses on a single hypothesis. If the number of hypothesis M is large, then the multiple comparison problem will present. To be specific, the probability that at least one of H_j be rejected becomes significant for a large M. Therefore, family-wise error rate (FWER) was proposed to control the number of falsely rejected null hypothesis. The definition of FWER is $P(V \geq 1)$, where V is the number of falsely rejected hypothesis. Meanwhile, we also face the ***trade-off*** that we want to find more discoveries while controlling the rate of false positive test, meaning that a statistic that is less conservative than the FWER is also worthwhile. This trade-off highlights the importance of false discovery rate (FDR) that is defined as $E(V/R)$ if $R > 0$ and 0 if $R = 0$, where R is the number of total rejected null hypothesis. It can be proved that FWER is an upper bound of the FDR, thus FDR is less conservative than FWER while it is also effective to reduce the rate of falsely rejected null hypothesis.

In this report, according to the content of the ND paper, it is the local false discovery rate (lfdr) that is the key concept, which measures the significance of observation $\hat{\beta}_j$. The definition of lfdr is

$$\text{lfdr}(\hat{\beta}_j) = \Pr(\beta_j = 0 | \hat{\beta}_j, \hat{s}_j, \pi), \tag{1}$$

where π consists of the weights of different components of distributions that generate observations $\hat{\beta}_j$, which is exemplified latter. lfdr is the probability that, under the probability distribution defined by π and noise strength s_j, the observed $\hat{\beta}_j$ corresponds to a zero effect $\beta_j = 0$.

In two-group model, the observed data are divided into two classes, null and non-null, according to whether or not it corresponds to a false discovery. Thus, the marginal distribution $f(z)$ of $z = \hat{\beta}/s$ is modeled as a linear combination of two different parts called null and alternative distributions

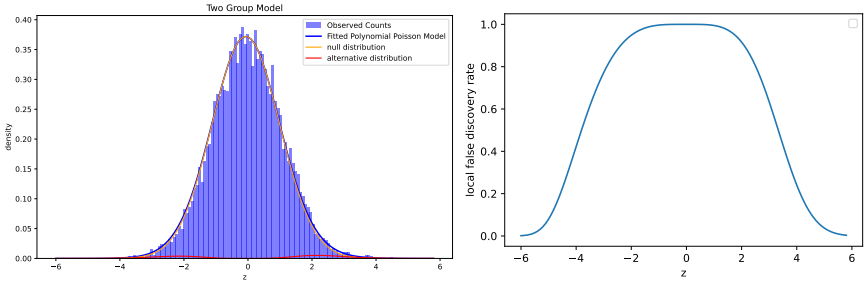
$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z), \tag{2}$$

where π_0 and π_1 are respectively the prior probability that the observed z belons to null and non-null class, while $f_0(z)$ and $f_1(z)$ are the corresponding likelihood. The local false discovery rate can be calculated by

$$\text{lfdr}(z) = \frac{\pi_0 f_0(z)}{f(z)}. \tag{3}$$

The value of π_0 and the mathematical expressions of $f_0(z)$ and $f(z)$ are needed in order to calculate the lfdr. The marginal distribution $f(z)$ can be estimated through standard Poisson regression. Besides, the following two assumpitions are adopted in the numerical simulations to estimated π_0 and $f_0(z)$

- $f_0(z)$ is a normal distribution.
- $f_0(z)$ is the dominant part of $f(z)$ for z near the peak of $f(z)$.



(a) (b)

Figure 1: (a) The estimated marginal distribution, null distribution and alternative distribution as well as the generated data set, where $\pi_0 = 0.8$ and the number of data is 12000. (b) The local false discovery rate estimated by the two-group model.

If $f(z)$ has a single peak at $z = 0$, then the mean and variance of $f_0(z)$ can be obtained by matching $\ln(\pi_0 f_0(z))$ with the second order expansion of $\ln f(z)$.

To test the two-group model, a data set consisting of $\hat{\beta} = \beta + e$ is generated, where $e/\hat{s} \sim \mathcal{N}(0, 1)$ is the noise, $\hat{s}$ is the estimated standard deviation of the noise that is assumed to be known, and β is sampled from the prior distribution

$$p(\beta|\pi_0) = \pi_0\delta_0 + (1 - \pi_0)\mathcal{N}(0, 1), \tag{4}$$

where δ_0 is a point mass at the origin. The numerical simulations are shown in Fig. 1. **The alternative distribution is always zero at $z = 0$ due to the method by which $\pi_0 f_0(z)$ is fitted.**

4 The New Deal

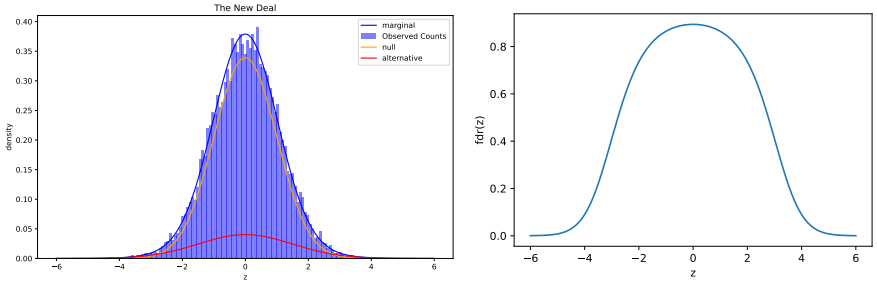
In this section, the reproduction of the theoretical descriptions of the algorithm and the numerical results shown in the ND paper are presented.

The posterior distribution of β is

$$p(\beta) = \frac{p(\hat{\beta}|\beta, \hat{s})p(\beta|\pi)}{p(\hat{\beta})}, \tag{5}$$

where $p(\hat{\beta}) = \int_{-\infty}^{\infty} p(\hat{\beta}|\beta, s)p(\beta|\pi)d\beta$ is the marginal distribution of $\hat{\beta}$. The unimodal assumption can be expressed as

$$\begin{aligned} p(\beta|\pi) &= \prod_{j=1}^{M} g(\beta_j|\pi) \\ g(\beta_j|\pi) &= \pi_0\delta_0 + \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma_k^2) \end{aligned}, \tag{6}$$



(a) (b)

Figure 2: (a) The marginal distribution, null distribution and alternative distribution estimated by the ND algorithm, as well as the generated data set, where $\pi_0 = 0.8$ and the number of data is 12000. (b) The local false discovery rate estimated by the algorithm proposed in the ND paper.

where σ_k are determined by the generated data at the beginning of the algorithm as initial input. The likelihood $p(\hat{\beta}|\beta, \hat{s})$ is assumed to be normal distribution

$$p(\hat{\beta}|\beta, \hat{s}) = \prod_{j=1}^{M} \mathcal{N}(\hat{\beta}_j; \beta_j, \hat{s}_j^2). \tag{7}$$

Then, the weights π of different normal distributions in the prior distribution Eq. (6) can be obtained by maximize the log likelihood $L(\pi)$. Besides, to ensure that the method gives conservative estimation of lfdr, a penalty term $h(\pi; \lambda) = \prod_{k=0}^{K} \pi_k^{\lambda_k - 1}$, where $\lambda_0 = 10$ and $\lambda_{k\neq 0} = 1$ by default, is added to the log likelihood to construct the objective function

$$\tilde{L}(\pi) = \sum_{j=1}^{M} \ln(\sum_{k=0}^{K} \pi_k l_{kj}) + \sum_{k=0}^{K} (\lambda_k - 1) \ln \pi_k, \tag{8}$$

where $l_{kj} = \int_{-\infty}^{\infty} \mathcal{N}(\hat{\beta}_j; \beta, \hat{s}_j^2) \mathcal{N}(0, \sigma_k^2) d\beta$ is the marginal distribution, corresponding to the k-th component of the prior distribution, evaluated at β_j. The objective function defined by Eq. (8) can be optimized by standard EM algorithm.

After determining the optimal π, the null component can be obtained by $\pi_0 \mathcal{N}(\hat{\beta}_j; 0, \hat{s}_j^2)$, and the marginal component is $\sum_{k=0}^{K} \pi_k l_{kj}$. The ND algorithm is tested on the data set generated in the same way as that stated in Section 3. **Thus, $\hat{s}_j = 1$ at this stage**. The numerical results are shown in Fig. 2. Compared with the results given by the two-group model, ***the ND predicts nonzero alternative distribution near*** $z = 0$ because it explicitly models the alternative distribution by a set of normal distributions with different standard deviations and estimates the weights of these distributions from the data set instead of directly assuming that all observed data near $z = 0$ are from the

null distribution. In other words, the ND with unimodal assumption (UA) can produce smaller estimates of π_0 than existing methods. As a consequence, the estimated lfdrs are also smaller. This feature helps to declare more discoveries.

On the other hand, **the estimations given by UA are still conservative**. In order to prove it, a set of data sets are generated according to Eq. (4) but with different alternative prior distributions as shown in Fig. 4(a). The independent data sets generated with different true π_0 are used to test that the estimated π_0s are still conservative, and the max, mean and min estimated values for different true π_0 are shown in Fig. 4. The comparisons of the estimated lfdr and lfsr with the true lfdr and lfsr are shown in Fig. 4(c) and Fig. 4(d), respectively. ***It should be noted that***

- The estimated π_0 and lfdr are anti-conservative for the bimodal prior distributions, which is caused by the mismatch between the unimodal assumption of the ND algorithm and the bimodal nature of the data set.
- In other cases, the estimations remain to be conservative.
- The algorithm substantially overestimates π_0 in the case of spiky prior alternative distribution because the generated non-null β_j are also close to zero, makint it difficult to distinguish them with those generated by the null distribution.

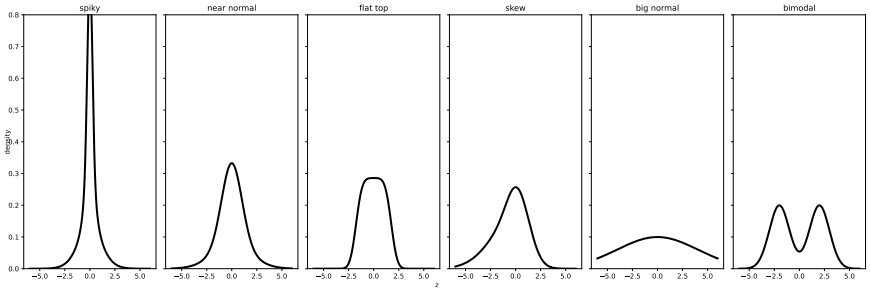
However, in most cases, except for the data sets generated with normal alternative distributions, the algorithm still overestimates the π_0, **especially when there are appreciable fraction of "small non-null effects", which are essentially indistinguishable from 0**, making accurate estimation of π_0 impossible. From the deinition, it can be known that the estimated lfdr is sensitive to the estimated π_0, thus the lfdrs are also overestimated when π_0s are overestimated. Therefore, local false sign rate (lfsr) was proposed in the ND paper since it is less sensitive to the estimated value of π_0. The definition of lfsr is

$$\text{lfsr}(\beta_j) = \min(\Pr(\beta_j \geq 0|\hat{\beta}_j, \hat{s}_j, \pi), \Pr(\beta_j \leq 0|\hat{\beta}_j, \hat{s}_j, \pi)) \tag{9}$$

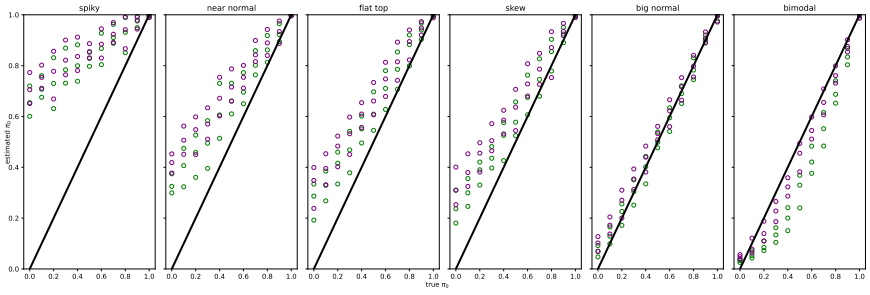
Smaller values of $\text{lfdr}(\beta_j)$ indicate that β_j is more probable to be nonzero, while smaller values of $\text{lfsr}(\beta_j)$ indicate that the estimated sign of β_j is more probable to be right. As shown in Fig. 4(d),

- The overestimation of lfsr is obviously less than that of lfdr.
- For the bimodal case, the anti-convervative estimation of lfsr is less serious than that of lfdr.

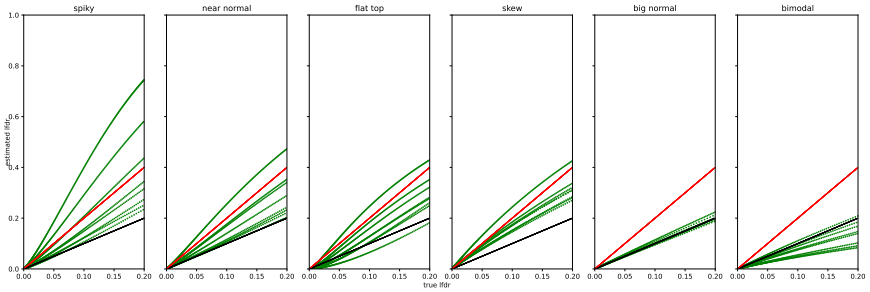
After obtaining the estimated π, the prior distribution $g(\beta)$ can also be estimated. In general, the estimation of g is a kind of "de-convolution" problem that is hard to tackle, but the unimodal assumption, in other words, the idea to approximate g by a set of "basis " functions, as shown in Eq. (6), greatly simplies this problem. The cumulative distribution functions of the estimated prior distributions for different true prior distributions are shown in Fig. 4(a).



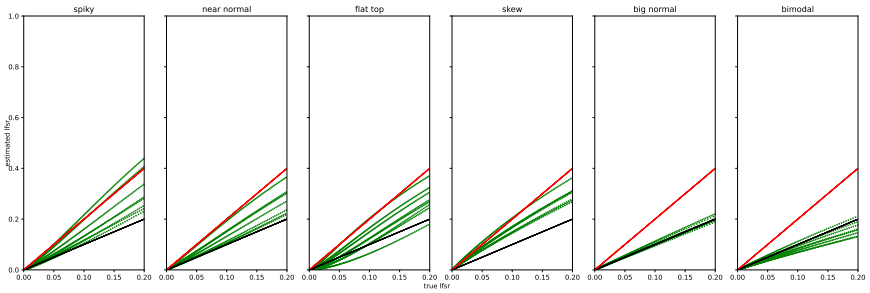
(a)



(b)



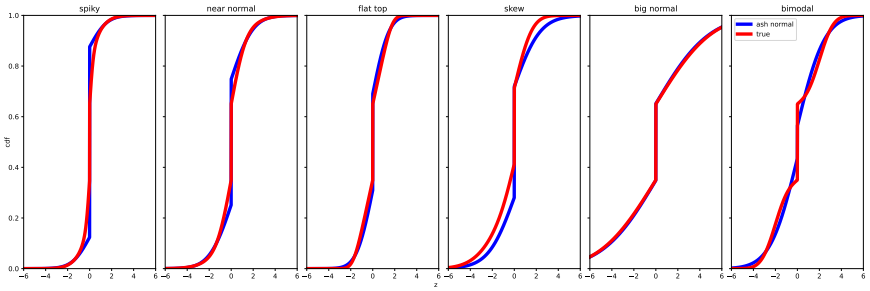
(c)



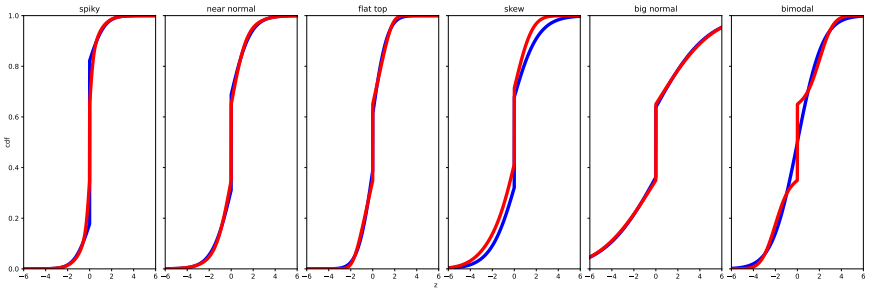
(d)

Figure 3: (a) Densities of different non-null prior distributions. (b) True π_0 versus estimated π_0, the green circles and purple circles are corresponding to using normal and uniform distributions to model $g(\beta_j)$ in Eq. (6). (c) True lfdr versus estimated lfdr. (d) True lfsr versus estimated lfsr. In (c) and (d), g is modeled by normal distributions.

- Due to the presence of the penalty term $h(\pi;\lambda)$, the algorithm tends to overestimate the mass of g at zero.
- By removing the penalty term, the overestimation of the mass of g at zero can be solved for the cases where the unimodal assumption is indeed held for the data set.



(a)



(b)

Figure 4: (a) Comparison between the estimated (blue line) and true (red line) cumulative distribution functions in the presence of penalty term. (b) Comparison between the estimated (blue line) and true (red line) cumulative distribution functions in the absence of penalty term.

The Bayesian approach also allows different measurement precisions for different observations. At this stage, it is assumed that half of the data is generated with good measurement precision $\hat{s}_j = 1$ and the other half is generated with poor measurement precision $\hat{s}_j = 10$. Meanwhile, the true $\pi_0 = 0.5$ for both of these two cases. The p-values of the good, poor and combined data set are shown in Fig. 5(a). The poor measurements contaminate the data set and weaken the signals. However, as shown in Fig. 5(b), the algorithm down weights the poor measured data by assigning a lfsr higher than the data with good measurement precision. The Fig. 5(c) is obtained by the data set which is generated by adding a normal distributed noise to all data with the same noise standard deviation. The result indicates that the algorithm can give precise estimation of the true π_0

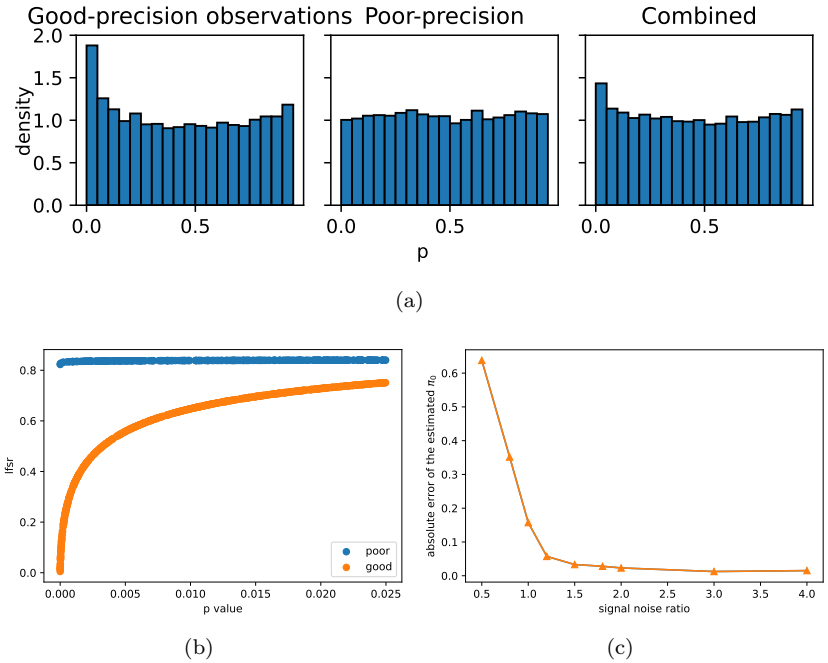


Figure 5: (a) p-values for data sets with different measurement precision. (b) The lfdr versus p-values for data set with good and poor measurement precisions respectively. (c) $|\pi_0^{\text{true}} - \pi_0^{\text{estimated}}|$ for data sets with different noise strengths.

if the noise is small enough. ***The details can be found through the Github link.***