Federal State Autonomous Educational Institution for Higher Education
National Research University Higher School of Economics

Faculty of Computer Science
Applied Mathematics and Information Science

BACHELOR'S THESIS

Research project

"Utilizing exact p-values for improved classification and accurate false discovery rate control"

Prepared by the student of group 193, 4th year of study,
Borevsky Andrey Olegovich

Supervisor:
Head of Research Lab, Attila Kertesz-Farkas

Moscow 2023

Contents

Abstract

Artificial Intelligence has been demonstrated as an incredibly useful instrument for a broad range of tasks. One of them — Bioinformatics — proposed fundamentally novel techniques at the intersection of machine learning and statistics. Despite their potential, these methods have not been utilized for more general tasks yet. Accordingly, in our work, we elaborate a drastically new approach called empirical p-values (EPV). Assuming negative training data of classification task to be the null hypothesis distribution, we calculate the corresponding p-values for the test samples. Later, we expand the BH procedure to control FDR, making it possible both to regulate the interrelation of train and test data distributions, as well as to predict new labels based on those already investigated. The major goal is to accurately predict number of accepted discoveries at each level without true labels.

Keywords

ML, EPV, BH, FDR control, LTT

Аннотация

Искусственный интеллект проявил себя как невероятно полезный инструмент для решения широкого спектра задач. Одна из них — биоинформатика — задействовала принципиально новые методы на стыке машинного обучения и статистики. Несмотря на их потенциал, эти подходы пока широко не использовались для решения более общих задач. Соответственно, в нашем исследовании мы предлагаем принципиально новый алгоритм, называемый эмпирическими p-значениями (EPV). Предполагая, что ложные обучающие данные задачи классификации можно представить в виде распределения нулевой гипотезы, мы вычисляем соответствующие p-значения для тестовых выборок. Позже мы расширяем процедуру BH для контроля FDR, что позволяет как регулировать взаимосвязь распределений обучающих и тестовых данных, так и прогнозировать новые метки на основе уже исследованных. Основная цель состоит в том, чтобы точно предсказывать количество принятых открытий на каждом пороге без истинных меток для всех возможных случаев.

Ключевые слова

машинное обучение, EPV, метод Бенджамини-Хохберга, контроль FDR, LTT

1 Introduction

1.1 Background

Classification is a renown machine learning task, being enriched throughout last decades with various metrics. Each of them expects a growth only with the simultaneous enhancement of the model's classification power, revealing a certain level of effectiveness. Meanwhile, a broad family of evaluation approaches has been introduced to present a comprehensive analysis.

It is vital to distinct two basic types of errors, which could be made by a classification method. If implying existence of two classes only, then the model might either propose a negative sample to be positive (false negative, FN) or vice versa (false positive, FP). Usually, the type of mistake is indifferent to the inference. However, some situations require diametrically opposite evaluation policy. For instance, we could dive into the biomedical sphere, such as brain cancer. A neural network processes snapshots of living-patient's brain cells, searching for those infected. Hence, separating two kinds of error is of paramount importance. Incorrectly recognizing healthy cells as rogue ones (FP) is much more hazardous since the surgeon will remove a wrong part of a brain - action, which can not be undone. So, there is a pool of applications that demand extremely strict control of the extent to which the algorithm makes FP mistakes. In such cases, False Discovery Rate (FDR) appears to be preferable [11].

We intend to elaborate a specific approach to produce uniformly distributed p-values for test data that would establish a consequent FDR control. Thus, we strive to take statistical methods of bioinformatics and dive into a broader area of ML classification.

1.2 Problem statement

Acquiring empirical p-values (EPVs) of test samples towards some null distribution, built from the training data, would allow us to calculate the number of

accepted discoveries at each possible error rate without using true labels. However, not only do we want to measure classification results through the FDR metric, but also the latter to be controlled at a certain ground-truth level. Such a task, where inference strategy should be highly compatible both with the alternative solutions and the baseline, must consider all potential ramifications of conditions, including data distribution shifts, alterations of classes' shares, etc. As a preliminary stage to deduce discoveries at each FDR, it is still vital to approach with manually extracted EPVs from test data to uniformly distributed p-values of train data. We will hold our entire investigation with the CNN on the MNIST dataset. Nevertheless, we opt to extend its findings to real-life cases.

1.3 Subject area

The object of the study is the FDR control without labels at a similar to ground- truth level. The subject of the research is the embodiment of empirical p-values and its bundle with Benjamini-Hochberg protocol for solving the presented issue.

1.4 Research structure

The paper is organized classically, where the introduction is followed by an analysis of the scrutinized literature. Then we explore the aforementioned statistical mechanics and finally examine the proposed classification methodology with a brief mention of the current results. The code can be found in our paper's GitHub repository.

2 Definitions

- **Classification**
  - a classic machine learning task implying a supervised approach to elaborate an algorithm that reveals the key features of a specified number of classes to categorize upcoming observations correctly. The neural network is trained on a training dataset, for which true labels are available. Later, its performance is evaluated based on test data, being brand new for the model.

- **P-value**
  - ddenotes the probability of events ' occurrence in the case of the null hypothesis (H_0) truthfulness.

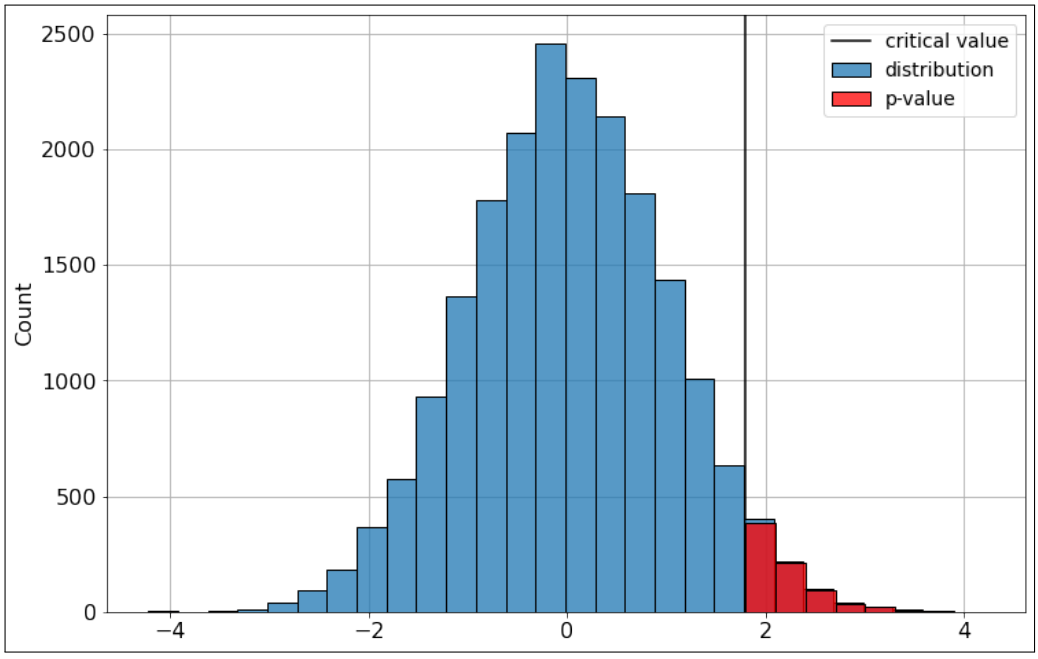


Figure 2.1: Illustration of p-value's correspondence to H_0

  - $$\text{p-value}_t = \mathbb{P}(T \geq t \mid H_0) \tag{2.1}$$

    where t is a particular test sample; T is the explored distribution [10].
  - For example, if we propose an object to belong to the first class and get a score exceeding 95% of those in the first class (the p-value is 5%),

then we can assume our hypothesis is highly likely incorrect. So, the p-value with a certain threshold stands for a chance of getting a false positive result, referring to some null distribution.

- **False discovery rate**
  - a proportion of samples that belong to type I errors. In other words, this is simply a share of observations, falsely rejecting H_0.
  - $$\mathrm{FDR} = \mathbb{E}(V/R \mid R > 0) \times \mathbb{P}(R > 0) \tag{2.2}$$

    where V is the number of false discoveries; R is the total number of discoveries [3].
  - Hence, we are looking for a powerful tool that would produce p-values of test objects in terms of the negative training data distribution to establish a certain FDR control policy. By the way, the use of FDR as an essential metric for research is dictated not only by launching from the biomedical sphere, but also by its widespread success in modern statistics.

- **q-values**
  - are the adjusted p-values in such a manner that they can be interpreted as FDR.
  - the basic technique to acquire such q-values if having access to labels suggests first sorting the values. The FDR is counted iteratively at each possible threshold of this array from greater to lower score. Then, the second run over the sequence is executed, where the respective q-value for i-th sample is produced by the following formula:

$$\mathrm{q}_{i-1} = \min(\mathrm{FDR}_i, \mathrm{FDR}_{i-1}) \tag{2.3}$$

The FDR_i is a share of samples, being negative by labels, among all $j \in \{0, ..., i\}$ samples.

- **FDR control**
  - a family of statistical procedures aimed at formulating a certain α level of FDR for a particular number of accepted discoveries.

$$\text{FDR} = \mathbb{E}(V/R \mid R > 0) \times \mathbb{P}(R > 0) \leq \alpha \tag{2.4}$$

- **Benjamini-Hochberg (BH) protocol**
  - ubiquitously recognized FDR-controlling step-up technique, where the metric is being set to a particular α level without using true labels. After calculating critical values for each observation, the method asserts the boundary of acceptance as the highest rank, where the critical value is greater than the corresponding p-value.
  - for instance, getting a p-value of 5% means that we have only a 5% chance that such a score can be observed for a sample of such a class. However, sometimes we accidentally discard false positives, leading to a higher level of misclassification. Consequently, a special technique — BH procedure — was elaborated to control FDR at particular α level [3].
  - The method assumes that if all hypotheses $H_0, ..., H_m$ come from null distributions, then the corresponding p-values $p_0, ..., p_m$ are uniformly distributed. The algorithm entails first determining p-values by the aforementioned formula. Then, it suggests ordering all p-values in a descending manner. Next stage is to calculate critical values by the following formula:

$$\text{critical value} = \frac{i \cdot \alpha}{m \cdot \pi_0} \tag{2.5}$$

where i is the rank of p-value in ascending order; α is a certain user-

defined FDR control level; m is the total number of test samples; π_0 is the proportion of negative samples in the data;

  - However, no proportions are available for the test data. We solve this difficulty by using such a value according to the predicted labels, expecting their plausibility.
  - This gives us the BH critical values and the boundary: the highest rank, so that the corresponding p-value is below its critical value. All the p-values above are considered significant, with a new threshold asserted [8]. We believe such a method would successfully control FDR at the particular α level in the general case.

- **Family-wise error rate**
  - the probability of making at least one type I error. All the tests are considered independent from each other, which in some cases cause multiple comparisons problem.

- **Bonferroni correction**
  - a statistical method to counteract drawbacks of FWER approach. It compensates the growth of rare event's probability increase. The test is conservative as the type I error rate drops below the nominal level as the number of samples becomes greater.

- **Convolutional neural network**
  - a specific class of neural networks that mostly consist of consecutive convolutional layers, applying each channel's kernel to the input feature map in order to extract key features of data. It has similar characteristics to the human visual cortex and is mostly used for the image analysis task.

3 Literature Review

3.1 Peptide annotation & FDR

Although the key paper's orientation is to infer accurate p-values for the test data without true labels, initially we relied on the doctoral dissertation [6], suggesting computational solutions for peptides' annotation. Directly belonging to the biomedical field, classification of mass spectrometry data requires close attention to the FDR level and its ubiquitous regulation at every threshold. Hence, an in-depth exploration of the possible statistical methods for the FDR control was given. First, the author underscored that the classical FDR control is carried out by managing the metric on the training data at a user defined α level and selecting the corresponding output as a threshold t. Test data, predicted with a score less than t, is rejected. The major issue with this approach is that it does not adapt either to the class distribution or to the data distribution shifts.

Second, the BH procedure was compared to the target-decoy technique practiced in medical classifications. We get an idea of the former's universality, theoretically outweighed by the risks. For instance, if we want to outline correct classes for the test data, we need to evaluate the proportion of the negative predictions there precisely. The worse we estimate, the less successful our forecasting policy will be. So, owing to this study, we explore an extensive review of the narrowly focused BH protocol embodiment, considering all challenges.

3.2 Alternatives to BH

3.2.1 Adversarial information

Since we do make a step from the dissertation lying at the intersection of biology and machine learning, where the FDR control takes a superior role, we need to extend its findings. Sequentially, we investigate similar research by [1], where the error rate is of paramount importance. Fortunately, the authors have presented a wide comparison of methods to preserve a particular FDR in the

context of computational biology. They divide the metrics into classical ones, including the BH protocol, and modern ones, which encapsulate complementary information besides p-values. A novel concept of interaction with a decreased error rate revealed considerable shortcomings of the BH procedure. While it is beneficial in terms of usability and applicability, it lacks the power - ability to detect true positive samples.

3.2.2 Filtering technique

Despite our pursuit to advance the key metric - FDR, [5] proposes persisting in utilizing p-values and supplementing the standard BH procedure with special filtering technique, introducing a new "Focused BH" algorithm. Initially, they identify a serious issue of transferring statistical data from one distribution to another, which may differ significantly. Therefore, the authors propose to filter the negative distribution of training data through various approaches in order to ensure a certain level of FDR control under a given range of assumptions. They not only define a vast number of filtering techniques, but also conduct a profound comparison. At the end of our study, we hope to further explore capacities of the tree-structured filtering and the control, established by "Focused BH", which is deeply described by the microbiome analysis example in the paper. Accordingly, we conduct our research by embodying the classical version of the BH protocol, and then we think of ameliorating its performance by recently emerged extensions.

3.2.3 Learn-then-Test

Obviously, FDR is just a single example of the statistical guarantees we can address to the algorithm. An attempt to establish a unified approach for calibrating models to satisfy such explicit constraints was accomplished by Angelopoulos et al. [2]. Their two-stage Learn-then-Test (LTT) framework proposes acquiring a trained model and modifying its predictions through multiple hypothesis testing. Thus, the acceptable set of hyperparameters is determined to preserve any selected statistical error (including FDR). "*Put plainly, we learn a base model and*

then test which parameter values lead to risk control" [2].

In the FDR case, we are simply searching for an optimal threshold t to be chosen for each desired metric value. Such a strategy, being an extensive way to calibrate the error rate, introduces a brand new technique that will be used as a direct competitor to the introduced method. Speaking more formally, we can take some threshold t. Then, we could say that the corresponding classification function f_t (basically, our neural network) is producing (α, β)-risk-controlling predictions only if:

$$\mathbb{P}(\mathcal{R}(f_t(X)) \leq \alpha) \geq 1 - \beta \tag{3.1}$$

where α and β are particular hyperparameters of risk control; R is the risk function, taking as input the prediction set for the particular t. In our case it is mainly the expectation of loss function.

We find it essential to dive into the underlying mechanism of the embodied algorithm.

1. As the first step, a sequence of the FDR values for each train sample is generated. We then present a set of possible thresholds in a form of linear space with $k = 300$ samples evenly distributed across the entire training array. Consequently, the selected t give us a set of corresponding FDR values from the sequence - risk function $\mathcal{R}$.

2. Then, we get into a cycle, iterating over possible α values (identical to the estimated FDR level).

3. At each step, Hoeffding-Bentkus inequality p-value is estimated. Such a parameter equals minimum of two separate statistical heuristics, both using α and $\mathcal{R}$. The presented by authors formula was proven to offer highly efficient results, formulating the p-value p_j^{HB} for the null hypothesis $\mathcal{H}_j : \mathcal{R}(f_t) > \alpha$. "*A small p-value will indicate disagreement with $\mathcal{H}_j$, implying the risk is controlled.*" [2]

4. The second key stage of the LTT framework assumes multiple hypothesis testing through the Bonferonni correction. It allows to reject all the thresholds out of those selected earlier based on the inferred p-values, limiting the false rejection probability for the given β-level. Out of the preserved hypothesis the last is chosen.

5. After the best threshold-score was selected, we can measure percentage of test scores above it. Consequently, calculating such a share for the test data, we can not only separate distribution into positive and negative subsets, but also identify particular number of accepted discoveries at particular α.

3.2.4 Angular Softmax

However, we additionally introduce a drastically new vision of eliminating the BH procedure's drawbacks. Besides adjusting the statistical method, the output of the model can be impacted. Since we want to separate samples by classes using their comparative scores, we could consider increasing the distance between their distributions intentionally. Such a technique is called ArcFace [4] - a descendant of the angular softmax concept [9]. What we basically do is pushing the model towards creating highly separated embeddings during training by a specific loss function formula:

$$L = -\log \frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)} + \sum_{j,j\neq y_i}^{N} e^{s\cos(\theta_j)}} \tag{3.2}$$

where $x_i \in \mathbb{R}$ is the deep feature of i-th sample, belonging to y_i class; θ_j is the angle between the feature x_j and the corresponding weight; m is an additive angular margin penalty between the weight and the feature.

During the inference, the model works as a standard one, predicting highly distanced scores for different classes. This approach seems promising for the assigned task.

4 Empirical p-values

After a number of various permutations, each being scrutinized for the sake of higher efficiency, a particular approach was developed. It includes following stages:

1. The neural network is trained in a classic manner on training data X^{train}, regardless of the classification type (binary or multiclass).

2. Trained model is switched to inference mode. Particular scores are then accessed for positive training data P^{train}, negative training data N^{train} and the entire test dataset X^{test}.

3. X^{test} could be generalized through the kernel density estimation plot. Hence, a specific function's output is obtained, appearing to be eligible for the task of finding local minima. We assert this to be the boarder between the positive and negative subsets of test data. Such a step from the standard methodology was a response to the potential challenges of data distribution shifts, wreaking havoc among test scores.

4. Mean and standard deviation of both P^{train} and N^{train} are calculated in order to describe the fundamental distributions, on which the model learned initially. We then substitute μ_N^{test} and σ_N^{test} by the training ones. In case of positive subset, only μ_P^{test} is altered.

$$\hat{N}^{\text{test}} = \frac{(N^{\text{test}} - \mu_N^{\text{test}})}{\sigma_N^{\text{test}}} \cdot \sigma_N^{\text{train}} + \mu_N^{\text{train}} \tag{4.1}$$

5. After we acquired the modified test distributions, we stack them together. Since we got our final representation of data, we can run both statistical algorithms, providing us with the ground truth and the EPV. The former is basically calculated via equation2.3.

6. At the same time, the EPV calculation includes several stages. Firstly, we

calculate such p-values that reflect the position of the particular test sample within the N^{train}:

$$\mathrm{p}_i = \frac{L_{\hat{N}^{\text{train}}} - \text{bisect}(\hat{N}^{\text{train}}, X_i^{\text{test}})}{L_{\hat{N}^{\text{train}}}} \tag{4.2}$$

where $L_{\hat{N}^{\text{train}}}$ is the total number of samples in $\hat{N}^{\text{train}}$; bisect is the binary search algorithm, declaring the position of the sample in the sequence.

Next, the sorted sequence of EPV is returned to become an input of the BH procedure. There, we calculate the corresponding q-values by the formula2.5. It requires the ratio of the positive (null hypothesis) and the negative (alternative hypothesis) classifications. Such a value is denoted as π_0. In our case, π_0 is estimated by the share of those samples predicted to be negative among the entire test dataset. Final operation implies same procedure, as shown in equation2.3 (with q-values instead of FDR).

5 Experiments

5.1 Methods & datasets

We are working solely on the ubiquitously known MNIST dataset, a rich collection of handwritten digits, becoming a baseline for any classification task nowadays. It contains 60k samples for the training stage and only 10k for the test, where all classes (from 0 to 9) are evenly distributed. Also to be mentioned is the strict normalization of the digits in the image in terms of size and centering, so the algorithms receive unified examples as input.

We have accomplished all our experiments through a standard convolutional neural network [7]. It comprises two convolutional 2d layers, both with 8 channels, kernels of size 3 and two-step strides. Single rectifier activation function is following each of them. The output layer is represented by a simple linear unit, providing as much classes, as it was initialized by the user.

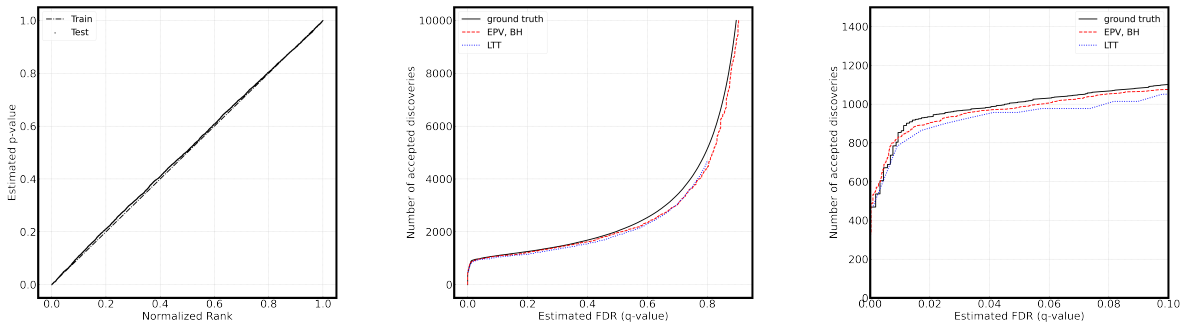
Model is trained with Adam optimizer with a learning rate equal to 0.01 and a BCEWithLogitsLoss cost function unless otherwise specified. The motivation for introducing such a naive model was straightforward: if we find ourselves able to ameliorate a particular quality compared to the ground truth via an elementary model on the resolvable MNIST dataset, then we could extend our approach to more comprehensive real-life tasks and models.

The entire algorithm was written on the PyTorch framework and includes several key stages throughout all separate local experiments. If formalizing our goal, then we could express our dataset as $(X_i, Y_i)_{i\in\{1,\dots,n\}}$, where $X_i \in \mathbb{X}$ is a feature vector and $Y_i \in \mathbb{Y}$ is the labels. For instance, $\mathbb{X}$ can be images and $\mathbb{Y}$ - the corresponding classes. Thus, our machine learning model becomes in some sense a comprehensive function f, such that
$f(\mathbb{X}) :\ \mathbb{X} \to \mathbb{Y}$.

However, besides only generating uniform p-values, we need an understanding of the algorithm's efficiency. Thus, we introduce two types of graphs: the "QQ" plots and the "FDR control" plots. The former depicts p-values from the uniform distribution for visual verification. We plot these p-values obtained with negative training data and their rank along the x-axis for both training (simply a diagonal) and test data. The closer our test samples get to the ideal line, the better our EPV approach operates. The second type of graphs visualizes the number of trusted classifications depending on the FDR when it is controlled with true labels (baseline) and with the BH protocol using the p-values. We also add rival LTT framework, representing the FWER methodology. As we progress towards the ground truth, our ability to function without test labels will improve as well as our capacity to forecast.

5.2 Classical case

Binary classification is a basic task, where $\mathbb{Y} \in 0, 1$. Hence, both training and test sets can be easily divided into negative and positive parts. Speaking of the



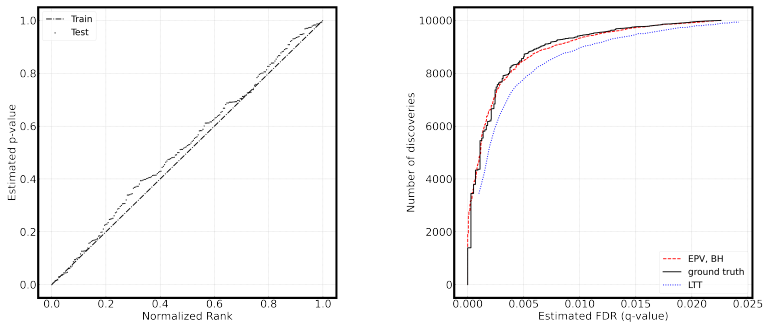
A B C

Figure 5.1: **FDR control with p-values.** (A) a QQ plot of the p-values obtained with using negative training data. (B) The number of trusted classifications as a function of the FDR when it is controlled with true labels (black line), with BHP (red line) and LTT (blue line) using the p-values from panel (A). (C) corresponds to the local range of graph (B).

MNIST dataset, we get 10% for a positive class despite our choice for the target label, not playing any crucial role for our research. We arbitrarily chose the class "2" as a positive one, and all the samples corresponding to other classes were treated as negative data.

The results for this most straightforward situation in context of MNIST dataset is presented on the 5.1. "Classical" stands for the binary classification problem. CNN model easily deals with the task, rapidly achieving almost flawless results on the test set: overall accuracy of 99% with a small drop to 93% for positive class only. The corresponding QQ plot, depicting the extent of EPV's practicality (labeled as "Test") compared to the perfect diagonal line of "Train", presents almost indistinguishable nature of two graphs. Hence, we can assert high quality of the acquired EPV through our approach on the synthetic dataset.

If we go further, diving into two "FDR control" graphs, the ground truth seems to be mostly describable by our method, having a little edge over the rival LTT framework. It is especially visible on a more critical local scale. Meanwhile, the ground truth plot acts as a black two-piece parabola, steadily reaching particular level of made discoveries and then growing quadratically. Each of the presented concepts mostly repeat the required behavior. However, on the major part of the examined range the suggested № of discoveries appears to be slightly lower then



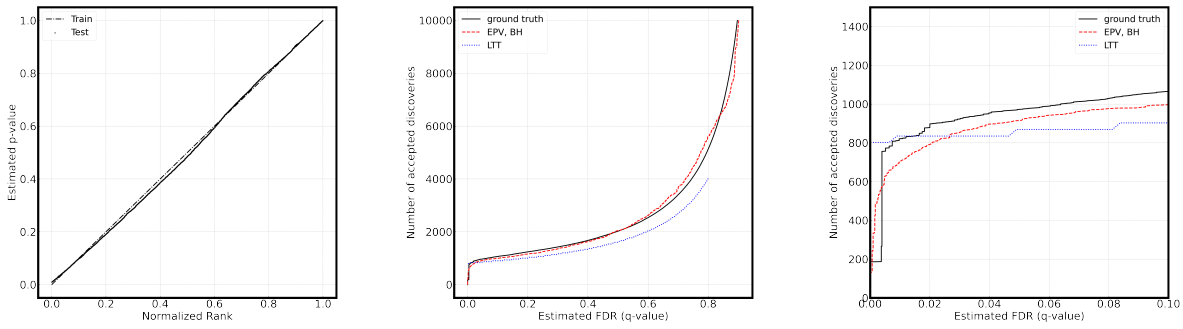
A B

Figure 5.2: **FDR control with p-values for multi-class classification.** (A) a QQ plot of the p-values obtained with using negative training data. (B) The number of trusted classifications as a function of the FDR when it is controlled with true labels (black line), with BHP (red line) and LTT (blue line) using the p-values from panel (A).

it should be. Such a conservative bias could be a consequence of model's worse performance on a test data. However, high quality of the produced EPV makes such an assumption untruthful. Hence, this challenge for both methods needs further exploration.

5.3 Multiclass case

Another frequent problem statement is rather a number of classes with only one true label for each sample. In this case we have ten equiprobable classes. The so-called multi-class classification implies getting simultaneously several scores as the model's output. At the same time, EPV approach necessitates separation of scores into negative and positive only. Thus, we insignificantly reformulate our workflow. Here, the p-value indicates the significance that a sample is correctly classified. Consecutively, we work only with the greatest scores among all presented, and the null distribution is constructed from the maximum values for the entire miss-classified training data. Thereby, we aim to control that whether a test data is correctly classified to its class or not. The π_0, again is estimated by the proportion of the incorrect classification among all samples. It is a tiny number due to the choice of high-performance model in context of resolvable MNIST



A B C

Figure 5.3: **FDR control with p-values for detecting data distribution shift.** (A) a QQ plot of the p-values obtained with using negative training data. (B) The number of trusted classifications as a function of the FDR when it is controlled with true labels (black line), with BHP (red line) and LTT (blue line) using the p-values from panel (A). (C) corresponds to the local range of graph (B).

dataset.

Getting closer to the plots 5.2, we should outline that only two were acquired. Such a situation is still a consequence of CNN superior results, when the error does not exceed several percents. Hence, we could only work with the FDR on a very small range. Nevertheless, a little fall of quality compared to the binary case can be seen on QQ plots. At the same time, EPV still appears to be adjacent to the diagonal at each point with a reasonable maximum distance. If moving to the FDR graph, one can see that this time EPV together with BH protocol noticeably outperform LTT, preserving a high level of interpretability as a smoother version of ground truth.

5.4 Shift case

As it was declared, MNIST has a set of properties, making it restricted for a comprehensive analysis. For instance, it does not allow to check wherever methods are resilient to shifts in data distribution. Therefore, we trained the CNN on described binary task with preserved target label "2". However, for further examination of p-values' nature, we introduce a 1-by-1 southeast pixel move for each test image. The newly born samples with shifted data distribution depict

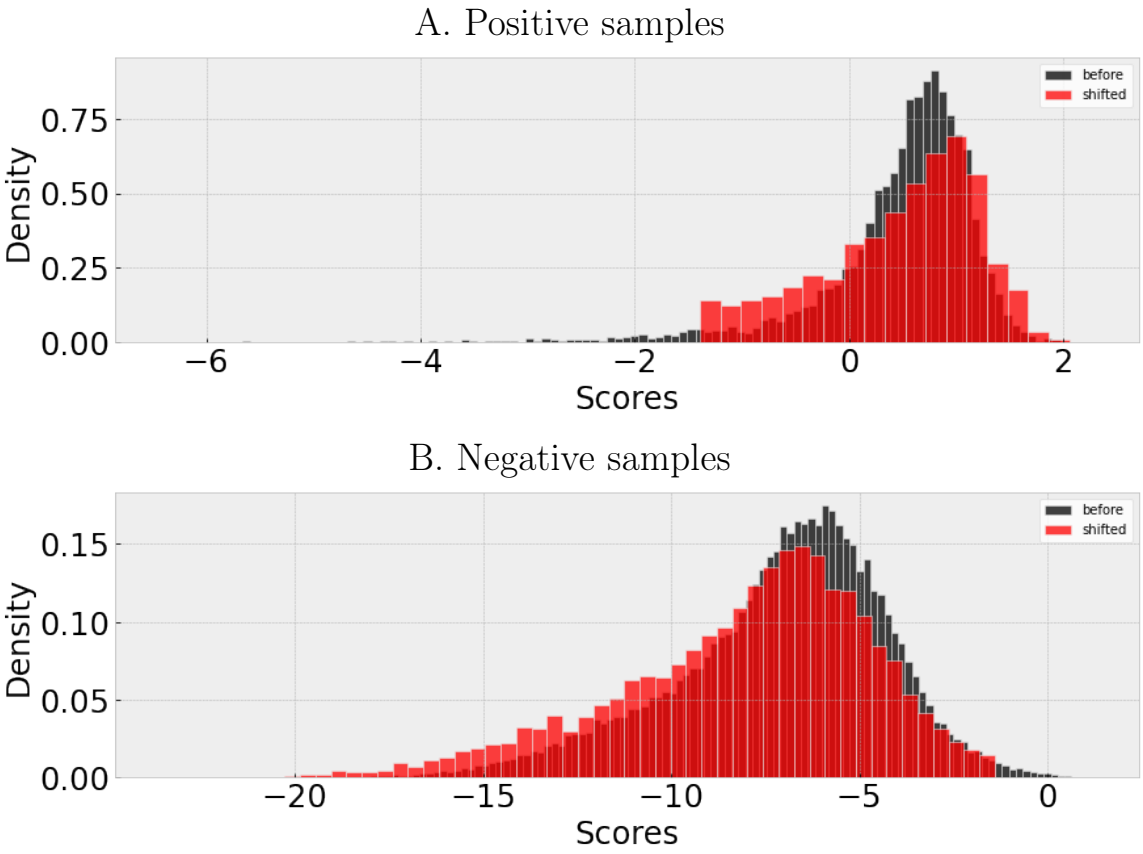
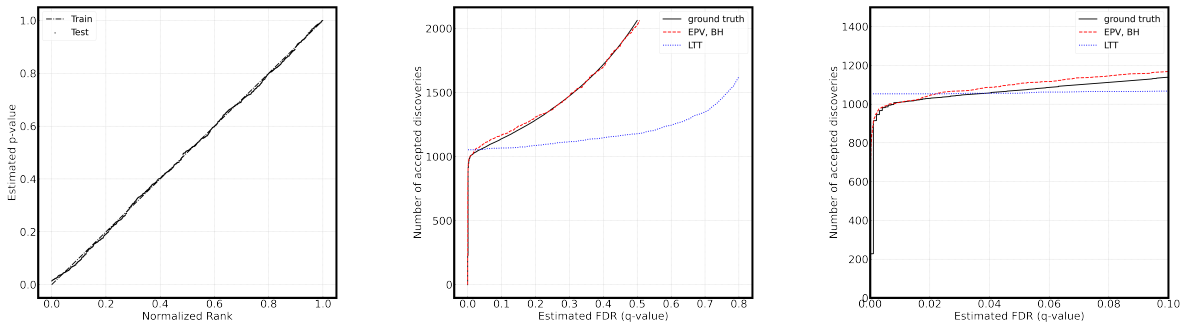


Figure 5.4: **Model's scores before and after shift in test data distribution.** (A) a score distribution of positive test samples according to predictions, where the black color stands for the normal case and the red one - for the shifted case. (B) has identical meaning to panel (A), except for the usage of negative samples.

such a scenario, when upcoming test data has a visible difference from the training phase.

So, the final part of the experiment implies running the altered images through the already trained model. On the resulting graphs 5.3, we would identify the first perceptible limitation of our EPV methodology if eliminating the statistical part with μ and σ. Nevertheless, the stated methodology keeps maintaining a notable proximity to the ground truth plot in terms of the long run. However, the local range depicts a downward shift. At the same time, LTT, also experiencing difficulties, again shows ubiquitously worse results. Another evidence of EPV's consistency in case of data shifts is depicted on the QQ plot, where the test dots do not make any sufficient drift from the diagonal.

The reason for such a behavior is that for our particular experiment's environment a shift can be identified not only in data distribution, but also in scores



A B C

Figure 5.5: **FDR control with p-values for altered class proportions.** (A) a QQ plot of the p-values obtained with using negative training data. (B) The number of trusted classifications as a function of the FDR when it is controlled with true labels (black line), with BHP (red line) and LTT (blue line) using the p-values from panel (A). (C) corresponds to the local range of graph (B).

distribution (shown on 5.4). The left boundary of the initial shifted test results has made a serious leftwards move, meaning far more test cases started obtaining a unit p-value (1.0). This is the reason why our EPV algorithm demands statistical unification of test and train data, separated by positive and negative sets.

5.5 Balanced case

Final conditions to be described imply another form of dataset's alteration. However, this time the proportion of positive and negative classes is changed rather than the data itself. Instead of inferring on the entire available test dataset, we collect all the positive samples (composing only 10 percent) and append the equal amount of negative samples in accordance with true labels for the sake of experiment's requirements. The declared parity differs from the share of classes, which the model has seen during training. Hence, a highly probable real-life scenario is tested here.

Again, straightforward implementation of EPV without any statistical corrections was leading to the low algorithm's ability to describe the ground truth. However, transferring the entire approach into the current form made a dras-

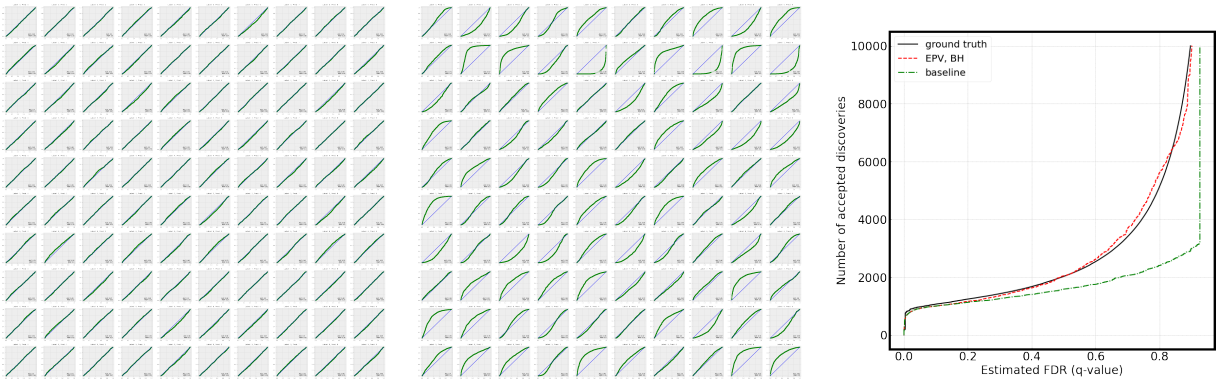


Figure 6.1: **Influence of statistical correction.** (A) shows the QQ plots for each class of predictions, separated by true class labels in the classical case. (B) has an identical meaning to panel (A) except for using the shifted test data. (C) illustrates the number of trusted classifications as a function of the FDR when it is controlled with true labels (black line), with BHP, based on statistically adjusted EPV (red line) and static ones (green line) in the shifted case.

tic amelioration. As it can be seen on figure 5.5, our method plus BH protocol work almost perfectly, with a little bit higher number of discoveries predicted on the local range. On the contrary, the LTT framework, based on FWER-family approach, shows comparatively dreary results, behaving illogically in terms of ground truth. If moving to the 5.5A, we can see ultimately close scatter plot to the "Train" diagonal.

6 Discussion

6.1 Proving statistical adjustment

For the sake of coherence, we find it mandatory to present particular reasons for introducing the aforementioned statistical correction. Actually, the first version of EPV did not involve this step. However, both balanced and shift cases have shown a significant drawback of such an approach. Figure 6.1 represents such a situation. 6.1A shows the QQ plots from the multiclass case, each representing the p-values distribution, with the true label being placed on y-axis and the predicted one - on the x-axis (hence, the p-values for correct predictions are located on the diagonal). We can see positive results for EPV in terms of the

perfect diagonal. However, when the data distribution shift takes place (6.1B), a serious degradation can be identified. If moving further to the FDR control graph, we see a huge difference between poor performance of the static EPV (green line) and the statistically adjusted ones (red line).

The explanation lies in the nature of the model's scores distribution. It was empirically found that when bringing the shift procedure to the experiment, our CNN starts extending the possible range of scores since facing the unexpected data representation. Hence, the static EPV procedure is challenged by lots of negative values being moved leftwards, which is the reason of such a slashing jump at the end of the curve - too many test samples have a unit p-value, appearing below the left border of the null hypothesis.

At the same time, if perceiving our data in terms of two distributions - positive and negative - then we can infer the corresponding statistical features. The mean and standard deviation are the one giving profound description. Hence, we found it rational to adjust the test distributions, separated by a border equal to local minima of kde function, so that they inherit training μ and σ parameters. The results have empirically proven our thesis.

6.2 Searching for accuracy advancement

We also made an attempt to ameliorate the accuracy of the classification itself via EPV. Predicting solely on such p-values seemed to bring no effect. Among many methods, several were the most promising: the k-nearest neighbours algorithm and the angular softmax. However, they all faced significant challenges and did not ameliorate our performance.

7 Conclusion

To sum up, we should outline that the foundation of the work has been laid with a pure understanding of an established goal. We have successfully controlled the FDR level in each of the cases because of the profound implementation of

the entire procedure, taking into account all possible ramifications. While we did not achieve our secondary objective of enhancing the model's classification performance, we still get results very close to the ground truth, surpassing the rival LTT algorithm. The EPV are uniformly distributed, which allows to plot highly efficient curves on the "FDR control" graphs. Hence, we hope to continue our research, bringing the method to more real-life challenges. Such a validation would take our approach to a new level and perhaps make it a common solution for dozens of classification tasks.

References

1. "A practical guide to methods controlling false discoveries in computational biology". In: *Genome Biology* 20.1 (2019), p. 118. DOI: 10.1186/s13059-019-1716-1. URL: https://app.dimensions.ai/details/publication/pub.1116652806.

2. Anastasios N. Angelopoulos et al. *Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control*. Research Papers 4030. Stanford University, Graduate School of Business, Apr. 2022. URL: https://ideas.repec.org/p/ecl/stabus/4030.html.

3. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society series b-methodological* 57.1 (1995), pp. 289–300. DOI: 10.1111/J.2517-6161.1995.TB02031.X.

4. Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *arXiv:1801.07698* (2018).

5. M. Bogomolov E. Katsevich C. Sabatti. "Controlling FDR while highlighting distinct discoveries". In: (2018).

6. A. Kertesz-Farkas. "Computational methods for tandem mass spectrometry data annotation". Higher School of Economics, 2021.

7. Yann LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1 (1989), pp. 541–551.

8. John D. Storey. "The positive false discovery rate: a Bayesian interpretation and the q-value". In: *The Annals of Statistics* 31.6 (2003), pp. 2013–2035. DOI: 10.1214/aos/1074290335. URL: https://doi.org/10.1214/aos/1074290335.

9. Hao Wang et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

10. Ronald L. Wasserstein and Nicole A. Lazar. "The ASA Statement on p-Values: Context, Process, and Purpose". In: *The American Statistician* 70.2 (2016), pp. 129–133. DOI: 10.1080/00031305.2016.1154108. eprint: https://doi.org/10.1080/00031305.2016.1154108. URL: https://doi.org/10.1080/00031305.2016.1154108.

11. J. Zhang et al. "PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification". In: (2012).