August 27, 2025

Dear Dr. Alessandro Barbiero:

We thank you and the reviewers for your kind consideration of our manuscript.

In general, the reviewers have cited the difficulties in following the logic flow; one reviewer have asked for additional experimental tests. The reviewers were right in the clarity. We have added a lengthy paragraph to the introduction about our motivation. We have also added several long sentences to the introduction here and there explaining our motivation. We have also added detailed description about the mathematical/statistical assumptions of our method (TNA) to the methods section. We are submitting two versions of our manuscript. One is a clean version of it (`exact-p-value_v3_clean.pdf`), the other tacks the changes with red (`exact-p-value_v3_track_changes.pdf`). We note that Reviewer 3 did not raised any substantial issues, and his critique was not very constructive.

In our manuscript, we are interested in providing statistically valid p-values for a set of discriminative scores (produced by a machine learning classifier) in general classification problems under data shift. Data shift compromises the validity of the p-values, and as a consequence, multiple testing procedures that rely on valid p-values may fail to control the relevant statistical errors. In order to demonstrate the usability of our method (that provides p-values for raw test discriminative scores), we happened to control the FDR with Benjamini-Hochberg procedure, that also involves π_0 estimation. We note that any other statistical applications could have been used for demonstration, which relies on valid p-values. The reviewer 1 has asked for testing additional methods for π_0 estimation besides than the Storey method. We have evaluated additional 6 methods, and we presented our results in our response to the reviewer; but we did not include these results to the main manuscript, because we did not find it very closely relevant to our TNA method. However, if the editor or the reviewer insists, we can append it to our manuscript.

Below, we address each of the points raised by the three reviewers and describe the changes we have made to the manuscript. In what follows, the reviewer's comments are shown in black type, interleaved with our responses (in blue) and, where appropriate, the modified text (in red).

Thank you very much for your consideration.

Best regards,

Prof. Attila Kertesz-Farkas
HSE University

Reviewer 1

This paper incorporates the proportion of the null hypothesis to adjust the distribution shift from training to testing data in order to control the FDR.

Thank you for this accurate summary of our work. We are submitting two versions of our manuscript. One is a clean version of it (`exact-p-value_v3_clean.pdf`), the other tacks the changes with red (`exact-p-value_v3_track_changes.pdf`).

I found the paper not well presented and it is very difficult to follow the logic flow. I have the following questions for the authors to consider:

1. It is not transparent to me regarding the rationale of the proposed method. The goal of multiple testing is to separate the non-null from the null. How can we estimate the density function of null and non-null without knowing which one is from which? Perhaps we can use the empirical distribution of the test statistic, assuming the null distribution, to derive the alternative distribution as that in Gao and Zhao (2024).

We are very grateful for the reviewer for pointing out this issue. We approached this problem entirely from practical point of view during manuscript preparation, and we did not consider readers from the theoretical side. But the reviewer is right, it was our mistake. We also admit that we did not separate the p-value adjustment problem from FDR control. We have revised the whole introduction, and added a few more paragraphs to the introduction that summarizes the problem statement and our method in a more formal way. Additionally, we also created and included a graphical sketch about the aim of the TNA method.

Briefly responding to the reviewer's critiques is the following. The reviewer is right that we do not know which test instance comes from null or non-null distributions In certain practical applications, it is possible to label some of the test data as negative with high confidence. For instance, decoy peptides[1] in tandem mass spectrometry and internal negative control proteins[2] in proteomic profiling serve this purpose. However, this is not the case with general machine learning applications, and we propose to approximate the unknown test null distribution with the empirical training null distribution.

We kindly ask the reviewer to see the introduction of our revised manuscript. Changes are marked with red.

2. Many methods are proposed to estimate the proportion of null hypothesis, for instance, the ones proposed in Meinshausen and Rice (2006); Wang, H.-Q. et al. (2011). In Wang, H.-Q. et al. (2011), the method adjusts for dependence, which should be particularly relevant to the applications considered in this paper. Can the authors try these approaches and compare the results?

Thank you for pointing out this issue. In the revised manuscript we now emphasize that the main aim of the TNA methods is to provide valid p-values for raw test prediction scores under data shift in general classification problems. And then we demonstrate the usability of the TNA adjusted p-values in FDR control with Benjamini-Hochberg (BH) method. At the input of the TNA method we are given a set or raw prediction scores (not p-values) and we use the number of positive and negative predictions, predicted by the given classifier, to estimate the $pi0$. This is done in the step 2 of the TNA method. We also revised step 2 of the TNA method to emphasize this, changes are marked with red. The pi0 estimation methods suggested by the reviewer, and in general statistical methods to estimate pi0, rely on valid p-values (not raw scores). Therefore, we could not use usual statistical methods for pi0 estimation. We note that, once the TNA-adjusted p-values are obtained, the user may use any statistical methods that rely on valid p-values, for instance: to estimate pi0, control various errors, etc. We happened to use Storey method to estimate pi0, and BH to control FDR. In our manuscript we did not aim to develop novel statistical methods to estimate pi0 or control FDR.

However, out of curiosity, we have evaluated several statistical methods for π_0 estimation, and we included these results below, see Table 1. In our opinion, all these methods provide better π_0 estimation with TNA EPVs than with standard EPVs, except in case of standard PCAM dataset. In our opinion, these results do not seem to be closely relevant to our paper, hence we do not include the table to our manuscript. However, if the reviewer thinks otherway, we can append the table to our manuscript.

3. Furthermore, it is assumed that the variance of training and testing are the same, which is a strong assumption. Is there any way to relax this?

Thank you for this comment, this was also raised by another reviewer. We added the following clarification text in the manuscript:

> In the following steps, we describe two versions of TNA methods that both adjusts the prediction scores so that the approximated test null distribution $\hat{T}_0$ aligns to the actual training null distribution X_0. The first variation, called TNA- (TNA minus), relies on that the shapes of the training null (X_0) and test null (T_0) distributions are similar, but there is no assumption on the form of the non-null

Table 1: Several π_0 estimations with different methods.

Dataset Type	True π_0	EPV	TNA Step 2.	Storey[3]	Meinshausen[4]	Jiang[5]	Nettleton[6]	Slim[7]	Pounds[8]	Last hist[9]
PCAM	0.503	Standard	N.A.	0.580	0.422	0.569	0.564	0.626	0.583	0.575
		TNA-	0.608	0.650	0.383	0.622	0.608	0.639	0.624	0.641
		TNA+	0.608	0.653	0.376	0.628	0.615	0.641	0.629	0.640
PCAM w/ shift	0.835	Standard	N.A.	0.967	0.128	0.904	0.863	0.876	0.887	0.953
		TNA-	0.845	0.878	0.153	0.853	0.836	0.866	0.852	0.869
		TNA+	0.845	0.868	0.120	0.884	0.871	0.880	0.884	0.829
CheXpert	0.749	Standard	N.A.	0.760	0.184	0.779	0.785	0.861	0.809	0.763
		TNA-	0.788	0.743	0.187	0.769	0.779	0.866	0.805	0.750
		TNA+	0.788	0.755	0.173	0.791	0.802	0.871	0.820	0.750
CheXpert w/ shift	0.870	Standard	N.A.	0.881	0.087	0.891	0.890	0.925	0.902	0.878
		TNA-	0.854	0.828	0.131	0.825	0.831	0.916	0.859	0.852
		TNA+	0.854	0.841	0.095	0.867	0.873	0.924	0.890	0.837
TissueNet	0.339	Standard	N.A.	0.269	0.716	0.289	0.303	0.355	0.316	0.257
		TNA-	0.313	0.378	0.699	0.352	0.311	0.361	0.341	0.357
		TNA+	0.313	0.353	0.703	0.338	0.308	0.358	0.334	0.352
TissueNet /w shift	0.668	Standard	N.A.	0.950	0.284	0.858	0.714	0.724	0.808	0.744
		TNA-	0.660	0.659	0.348	0.554	0.673	0.693	0.637	0.282
		TNA+	0.660	0.603	0.370	0.584	0.688	0.699	0.662	0.246

We note that Meinshausen seems to estimate $1 - \pi_0$.

distributions. The second variation, called TNA+ (TNA plus), relies on that the shapes of the training non-null (X_1) and test non-null (T_1) distributions are similar, but there is no assumption on the form of the null distributions. The proportion of the nulls may freely vary between the training and test data in the case of both TNA methods, there is no assumption on this.

We kindly ask the reviewer to check this paragraph in the section of TNA protocol. All updates to the text is marked with red.

4. In practice, how do you determine the number of histogram bins and how many bins we should use?

Unfortunately, we do not have good recommendation on the number of bins or on the width of the bins. It depends on the support of the distributions and the number of the test instances. In general, bins should not be empty. We added this comment to the End of the section TNA Protocol in the manuscript, it is marked with red.

Reviewer 2

1. The proposed TNA method aims to address the issue of distribution shifts between the calibration data and target data in conformal classification.

Thank you for this accurate summary of our work. We are submitting two versions of our manuscript. One is a clean version of it (`exact-p-value_v3_clean.pdf`), the other tacks the changes with red (`exact-p-value_v3_track_changes.pdf`).

However, distribution shifts can manifest in various forms, including covariate shift, class-wise covariate shift, label shift, or posterior shift. I suggest the authors clarify which specific types of distribution shifts the proposed method can handle, and provide the rationale for the effectiveness.

Thank you for this comment, TNA does not have any assumptions on the type of the source of the data shift. TNA operates in the space of the raw discriminative scores, that is usually one dimensional per class. TNA does not aim to reverse the data shift in the input data space. However, we note that other reviewers have also raised questions about the assumptions which TNA method relies on. We have explicitly state these assumptions as follows:

> In the following steps, we describe two versions of TNA methods that both adjusts the prediction scores so that the approximated test null distribution $\hat{T}_0$ aligns to the actual training null distribution X_0. The first variation, called TNA- (TNA minus), relies on that the shapes of the training null (X_0) and test null (T_0) distributions are similar, but there is no assumption on the form of the non-null distributions. The second variation, called TNA+ (TNA plus), relies on that the shapes of the training non-null (X_1) and test non-null (T_1) distributions are similar, but there is no assumption on the form of the null distributions. The proportion of the nulls may freely vary between the training and test data in the case of both TNA methods, there is no assumption on this.

This can be found in the section of "TNA protocol", marked with red.

2. The aim of the article is to provide false discovery rate control in the presence of distribution shift. However, with the proposed TNA method, it remains unclear to me whether the FDR is controlled with respect to the distribution of target data or the calibration data, as the target scores are adjusted to align with the calibration scores in TNA. I suggest the authors provide a discussion on this potential issue.

Thank you very much for pointing out this important issue. The first reviewer has also cited issues in logic flow. We have revised the introduction and the methods sections and we hope it is clearer now. The relevant text now reads as follows:

> *Finally, calculate the TNA EPVs for each adjusted $\hat{t_s}$ test score with respect to the training null (X_0) distribution with using Eq. [1]. The TNA EPVs then can be used in any statistical applications that rely on valid p-values. In our experiments in the following sections, we demonstrated the usability of the TNA EPVs in FDR control with BH procedure.*

Therefore, the p-values of the adjusted test prediction scores are calculated with respect to the training (or calibration) null distribution with using eq. [1].

3. The quantity $\hat{\pi}_0$ appears in TNA+ method, but it causes confusion with the Storey null proportion estimator in the final BH procedure. The authors should provide more detailed explanation on how to obtain $\hat{\pi}_0$ in step 3.b.

We thank the reviewer for pointing out this potential confusion. We renamed the $\hat{\pi}_0$ to $\hat{\pi}_{0_T}$ in our manuscript. We also added a clarifying sentence about the $\hat{\pi}_{0_T}$ calculation as the reviewer requested.

4. The introduced LTT in Section 2.1 is misleading, as the primary focus of LTT in Angelopoulos et al. (2021, arxiv) differs from the false discovery rate control that is essentially related with the BH algorithm. Besides, the use of notation λ in LTT causes confusion with the λ in Storey null proportion estimator. I suggest the authors provide a clearer explanation on the LTT if it is necessary for the current problem.

Indeed, The LTT method has been introduced to control risk in general. The risks they can control includes the FDR too. We quote from the abstract of the LTT paper from Angelopoulos, et al. (2021, arxiv):
"The framework [LTT] addresses, among other examples, false discovery rate control in multi-label classification, intersection-over-union control in instance segmentation, and the simultaneous control of the type-1 error of outlier detection and confidence set coverage in classification or regression. "

We thank the reviewer for pointing out the confusions with the Lambdas in Storey and LTT. We introduced λ_s and λ_l, respectively, in order to avoid this confusion.

Reviewer 3

The authors propose a new method for the false discovery rate control in classification problems. They adjust the test prediction scores so that the test null distribution aligns to the training null distribution, and more accuarte p-vlaues of the test samples can be computed.

Thank you for this accurate summary of our work. We are submitting two versions of our manuscript. One is a clean version of it (`exact-p-value_v3_clean.pdf`), the other tacks the changes with red (`exact-p-value_v3_track_changes.pdf`).

Question:
1.TNA- (or TNA+) assumes that $T_0[i]/T_s[i] = X_0[i]/X_S[i]$ (or $T_1[i]/T_s[i] = X_1[i]/X_S[i]$). Why is this assumption made?

About TNA+: TNA+ does NOT assumes $T_1[i]/T_s[i] = X_1[i]/X_S[i]$. In the formula of TNA+ in Step 3b, we use a π_0 estimation of the true nulls in the test set. That is, the $\hat{\pi}_0$ is not derived from the proportion of the null in the training set. It is stated in Step 2 of the TNA method. Therefore,

$$X_S \neq \hat{\pi}_0 \overline{X_0}[i] + (1 - \hat{\pi}_0)\overline{X_1}[i].$$

Once again: X_0 and X_1 corresponds to the training data, π_0 estimation corresponds to the test data. This is already noted in step 3b.

About the TNA-: unfortunately, we made a mistake in the formula, and we missed the correction factor about the change in the proportion of null. We changed the formulation of the TNA method from count histograms to density histograms, but shamefully, we forgot to update the TNA- methods. We even missed the bars above X_0 and X_s We have corrected the TNA- method. The TNA- method correctly: $\hat{T}_0[i] = \overline{T}_s[i] \cdot c[i]$, where $c[i] = X_0[i]/X_s[i] \cdot \pi_0^X/\hat{\pi}_{0_T}$ and π_0^X is the actual proportion of the null samples in the training data calculated with using the true class labels. That is:

$$\pi_0 \hat{T}_0[i]/\overline{T}_s[i] = \pi_0^X X_0[i]/X_s[i]$$

We added a more detailed description about the $\hat{\pi}_0$ estimation in the test set to the Step 2 of the TNA methods. The changes are marked with red, and we kindly ask the reviewer to check this in the revised manuscript. We did double check our code, and the code was correct, so luckily we did not need to redo our experiments.

2. Do the authors assume that the proportion of the true null hypothesis remains constant between the training and test samples?

In summary, TNA method does not assume fixed null proportions in training and test sets, as we argued above.

We are very grateful to the reviewer for noticing this issue.

References

[1] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207, 2007.

[2] Zijun Gao and Qingyuan Zhao. Simultaneous hypothesis testing using internal negative controls with an application to proteomics. *arXiv preprint arXiv:2303.01552*, 2023.

[3] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[4] Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. 2006.

[5] Hongmei Jiang and RW Doerge. Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer informatics*, 6:117693510800600001, 2008.

[6] Dan Nettleton, JT Gene Hwang, Rico A Caldo, and Roger P Wise. Estimating the number of true null hypotheses from a histogram of p values. *Journal of agricultural, biological, and environmental statistics*, 11(3):337–356, 2006.

[7] Hong-Qiang Wang, Lindsey K Tuominen, and Chung-Jui Tsai. Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, 27(2):225–231, 2011.

[8] Stan Pounds and Stephan W Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.

[9] Megan H Murray and Jeffrey D Blume. Fdrestimation: flexible false discovery rate computation in r. *F1000Research*, 10:441, 2021.