The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. 11–35.
doi: 10.1111/ectj.12092

Oracle and adaptive false discovery rate controlling methods for one-sided testing: theory and application in treatment effect evaluation

JIAYING GU[†] AND SHU SHEN[‡]

[†]*Department of Economics, University of Toronto, 150 St George Street, Toronto, M5S3G7, Canada*.
E-mail: jiaying.gu@utoronto.ca

[‡]*Department of Economics, University of California Davis, 1 Shields Ave, Davis, CA 95616, USA*.
E-mail: shushen@ucdavis.edu

First version received: October 2014; final version accepted: March 2017

Summary Economists are often interested in identifying effective policies or treatments together with subpopulations of individuals who respond positively (or with a sign that is expected) to these treatment interventions. In this paper, we propose an optimal false discovery rate controlling method that is especially useful for such one-sided testing problems. The proposed procedure is optimal in the sense of minimizing the false non-discovery rate while controlling the false discovery rate at a pre-specified level; it uses a deconvolution method based on non-parametric maximum likelihood estimation, which allows for a broader class of treatment effect distributions than existing methods do. The proposed test demonstrates good small-sample performance in Monte Carlo simulations and it is applied to study the effect of attending a more selective high school in Romania. The application reveals strong evidence of treatment effect heterogeneity, in that students who marginally gain access to higher-ranked schools are more likely to benefit if the higher-ranked school has a relatively high admission score cut-off – or, in other words, is more selective.

Keywords: *False discovery rate control, Multiple testing, Treatment effect heterogeneity.*

1. INTRODUCTION

It is widely accepted that individuals have heterogeneous responses to policy interventions or exogenous shocks in many economic applications. There are ample examples where making inference on these heterogeneous effects may be of interest. For instance, economists and policymakers are often interested in identifying effective policies or treatments together with subpopulations of individuals who respond positively to these policies or treatment interventions (or with a sign that is expected by these interventions). Such an inference problem often requires consideration of multiple comparisons of many estimators or multiple testing of many hypotheses simultaneously. However, unlike disciplines such as biology and especially genomics, the economics literature has, until recently, focused primarily on multiple testing procedures that

control for family-wise error rate (FWER); see, e.g. Anderson (2008), Lee and Shaikh (2013) and Armstrong and Shen (2015). When the number of hypotheses is large, classic FWER controlling procedures typically have very low power. If the economist or policymaker is interested in determining which subpopulations have a positive effect, then it is not only important to control for multiple testing errors, but also to make sure the multiple testing procedure has sufficient power to correctly find the positively responsive subpopulations. Recent papers such as Hsu et al. (2014) and Fishe and Smith (2012) extend and apply existing false discovery proportion (FDP) and false discovery rate (FDR) controlling procedures to economics and finance settings. We provide formal definitions of various error rates in the Appendix and we refer the readers to Romano et al. (2008) for a comprehensive survey of different multiple testing methods.

The goal of this paper is to propose an FDR controlling testing procedure with optimal power for multiple composite one-sided tests. FDR controlling tests, first proposed by Benjamini and Hochberg (1995), control for the expected value of the FDP, the proportion of Type I errors among all rejections, under some pre-specified level. The counterpart of FDR is the false non-discovery rate (FNR), which is the expected value of the false non-discovery proportion (FNP), the proportion of Type II errors among all non-rejections. The test we propose is optimal in the sense of minimizing the FNR while controlling FDR at a pre-specified level asymptotically as the number of hypotheses goes to infinity. The proposed test is a useful complement to classic FWER control procedures, or other more prudent or conservative testing procedures, when the number of hypotheses becomes large.

The proposed test contributes to the multiple testing literature in several ways. First, most of the recent advances – such as Genovese and Wasserman (2002), Efron (2004a,b) and Sun and Cai (2007), among many others – in the FDR literature focus on testing a simple null hypothesis of zero effect against two-sided or one-sided alternatives. This is because it is natural to assume, in large-scale biology or genomic studies, that the majority of the effects tested are from a simple null of zero effect (the statistics literature often calls this assumption of the null distribution the 'sparsity' condition). However, in many social science applications, one often desires a one-sided test and sparsity is unlikely to hold because it is not desirable to assume a priori that those who do not respond positively to a policy change are unaffected by the policy at all. For example, in the empirical section, we revisit Pop-Eleches and Urquiola (2013) and we study the effect of attending a more selective high school in Romania. As discovered by Pop-Eleches and Urquiola (2013), marginal students who attend a more selective high school can face negative interactions with peers. When testing for a positive effect, it is not desirable to assume away potential negative effects for some specific towns or schools.

For one-sided testing, assuming sparsity is equivalent to utilizing the least favourable configuration when constructing the decision rule of the test. When the null is heterogeneous, it has long been recognized that the use of the least favourable configuration jeopardizes power. Andrews and Soares (2010) propose the use of a moment selection method to avoid the least favourable configuration in the moment (in)equality literature, following the recentring idea first introduced in Hansen (2005). Hsu et al. (2014) adopt the same idea in the context of FDP testing. Our tests do not impose the least favourable configuration either. We take full account of the heterogeneity in the null region. Instead of a step-wise FDP procedure, we propose a single-step thresholding procedure based on a decision theoretic framework that is optimal in the sense of minimizing FNR, while controlling for FDR at a pre-specified level.

Our method is closely related to the FDR controlling procedure proposed in Sun and McLain (2012), which is designed for two-sided multiple testing with composite null hypotheses. Compared to Sun and McLain (2012), we identify an assumption that determines optimality of

the proposed procedure and is at the same time easy to verify. We also find that for one-sided multiple testing, the optimal procedure that depends on the likelihood ratio statistics is equivalent to a procedure that depends on the p-value. This is in contrast to conclusions in the two-sided multiple testing literature where the latter is not optimal, as discussed in Sun and Cai (2007) and Sun and McLain (2012). Additionally, the procedure in Sun and McLain (2012) requires independence between the effect and its associated variances, which is likely to be violated in many economic applications. Our method avoids this independence assumption.

Second, our proposed procedure requires estimation of the effect distribution. For the estimation, we propose to use a deconvolution method based on non-parametric maximum likelihood estimation (NPMLE). Compared to the empirical characteristic function based kernel deconvolution method – see, e.g. Sun and Cai (2007) and Sun and McLain (2012) – that restricts the effect distribution to be continuous with smooth density, our approach admits a larger class of distributions, which includes distributions that are continuous, discrete or a combination of both. When applied to treatment effect evaluation, our test could be used to test for positive treatment effects (or treatment effects exceeding some pre-determined threshold). The NPMLE method that we propose for the testing procedure then provides, as a side product, a consistent estimator for the treatment effect distribution across observationally equivalent subpopulations.

We discuss the use of two NPMLE-based deconvolution methods in testing: the classic, or vanilla, NPMLE method (cf. Kiefer and Wolfowitz, 1956) and a new hybrid method, which first estimates the probability that the effect takes the value of the cut-off of the one-sided null hypothesis and then conducts a restricted NPMLE by plugging in the consistent first-step estimator. We show that when classic NPMLE is used, our testing procedure is asymptotically equivalent to the Oracle procedure that minimizes FNR and controls FDR with known effect distribution, if there is no probability mass at the cut-off value of the one-sided null. When the hybrid NPMLE method is used, our testing procedure is asymptotically equivalent to the Oracle procedure even when there is non-trivial probability mass at the cut-off value of the one-sided null.

We conduct simulation exercises to compare the performance of the proposed tests with other existing tests in the literature. Despite their popularity, FDR controlling procedures are subject to the criticism that they do not account for the variability of the FDP; hence, in practice, such procedures could result in a large realized FDP. However, as discussed in Heller (2010) and further confirmed by our simulation studies, this issue is less of a concern – if at all – when the test statistics are independent compared to when the test statistics are dependent. Besides, our simulation results show that: (a) multiple testing procedures that control FWER or FDP are too conservative and have low power for several data-generating processes (DGPs); (b) existing FDR controlling methods with simple null versus composite alternative, such as the Benjamini and Hochberg (1995) procedure, for example, and its adapted version considered in Benjamini et al. (2006), are conservative when there is heterogeneity under the null; (c) when the effect correlates with its associated variance, the procedure discussed in Sun and McLain (2012) is no longer valid but the proposed procedures control size and have excellent power performance. Details of the results are included in the online Appendix.

In the empirical section, we apply the proposed test to study the effect of attending a more selective high school, using the Romanian administrative data set from Pop-Eleches and Urquiola (2013). Instead of pooling students from different towns and schools together and estimating the pooled effects of attending a more selective high school, we estimate town-specific and school-specific treatment effects and we find strong evidence of treatment effect heterogeneity across towns and schools. Further, the set of schools identified by our test of having positive

treatment effects is also seen to have higher admission score cut-offs, suggesting that students who marginally gain access to higher-ranked schools are more likely to benefit if the school is more selective.

The remainder of this paper is organized as follows. In Section 2, we set up the model and propose new multiple testing methods for one-sided hypotheses with normalization. Both the Oracle and the adaptive procedures are discussed. In Section 3, we revisit Pop-Eleches and Urquiola (2013) and we apply the proposed test to study the effect of going to a better school in Romania. We conclude in Section 4. All proofs are given in Appendix B. Monte Carlo simulations are given in the online Appendix.

2. METHODOLOGY

Following Sun and McLain (2012), consider m independent random variables $(\mu_1, \ldots, \mu_m)$ that follow a normal mixture distribution

$$\mu_i \mid \eta_i, \sigma_i \sim \mathcal{N}(\eta_i, \sigma_i),$$

with a location parameter η_i and a standard deviation σ_i, which are random variables following some joint distribution. In the treatment effect evaluation framework, μ_i can be an unbiased treatment effect estimator for subpopulation i of observationally equivalent individuals whose vector of characteristics X is equal to a fixed value x_i. The location parameter η_i is then the unknown treatment effect for that subpopulation.

We are interested in identifying the set $\mathfrak{I}_+$ associated with a positive location parameter, that is

$$\mathfrak{I}_+ := \{i \mid \eta_i > 0\},$$

or subpopulations with positive treatment effects. Such a task is akin to performing a one-sided hypothesis test simultaneously for all i. The cut-off value of the null hypothesis can be easily extended to a fixed a, or the cut-off rule can be extended to $\eta_i > a$. In such cases, we can shift the random variable by a and define $\tilde{\mu}_i = \mu_i - a$, which reduces the problem back to the above setting. We follow Sun and McLain (2012) in assuming that σ_i is known, but we allow σ_i to be arbitrarily related with η_i. Let $S_i = \mu_i/\sigma_i$. It is easy to see that $S_i \mid \nu_i \sim \mathcal{N}(\nu_i, 1)$ with location parameter $\nu_i = \eta_i/\sigma_i$. The location parameter ν_i is a random variable with distribution G, which is determined by the underlying joint distribution of (η_i, σ_i). The set of interest $\mathfrak{I}_+$ can then be transformed to

$$\mathfrak{I}_+ = \{i \mid \nu_i > 0\}.$$

We aim to construct a decision rule based on S_i, $i = 1, \ldots, m$. Let

$$\delta(Z, z) := \{\delta_i = \mathbf{1}(Z(S_i) < z), i = 1, \ldots, m\},$$

such that $\delta_i = 1$ indicates that we decide that i belongs to the set $\mathfrak{I}_+$. The decision rule consists of two elements: the function $Z(\cdot)$ as some transformation of S_i; and the fixed value z as the universal threshold for all i.

Given the decision rule, let $\hat{\mathfrak{I}}_+ := \{i : \delta_i = 1\}$ be the set that collects all i that are identified to have a positive location parameter. In the treatment effect context, $\hat{\mathfrak{I}}_+$ includes all subpopulations identified by the decision rule to have a positive treatment effect. The optimal decision rule $\delta(Z, z)$ that minimizes the FNR, while controlling FDR under a pre-specified level α for the

one-sided testing problem of interest, can be found via a decision theoretical framework, as discussed in the rest of this section.

2.1. Oracle procedure for one-sided multiple test with composite null hypothesis

We first consider the Oracle case where we assume perfect knowledge of the distribution of ν, $G(\nu)$. We show that the optimal transformation $Z(\cdot)$ can be based on either the likelihood ratio statistics or the p-value. Both methods are optimal in the sense of minimizing the FNR while controlling the FDR at a nominal level. The decision theoretical framework adopted in this section does not rely on S_i following a normal location mixture model. As long as the model for S_i has a hierarchical structure, such that, conditional on ν_i, S_i has a parametric density function $\varphi(\cdot)$, which depends on the individual-specific parameter ν_i, and ν_i are independent across i with a common distribution, then the proposed procedure is optimal.[1]

Let H_i be an independent Bernoulli random variable with success probability p. For testing purposes, we formulate the distribution of S_i, without loss of generality, using a two-group mixture model that distinguishes the null from the alternative,

$$S_i \mid H_i \sim (1 - H_i)F_0 + H_i F_1,$$

where $H_i = 0$ corresponds to the event that $i \in \mathfrak{I}_+^C$ or $\nu_i \in A \equiv (-\infty, 0]$, F_0 corresponds to the distribution under the null with density $f_0(s) = (1/(1-p)) \int_A \varphi(s \mid \nu) dG(\nu)$ and F_1 corresponds to the distribution under the alternative with density $f_1(s) = (1/p) \int_{A^c} \varphi(s \mid \nu) dG(\nu)$ and $1 - p = \int_A dG(\nu)$. Note that this formulation is valid no matter how ν_i is distributed.

The multiple testing problem results from simultaneously testing all m hypotheses. This is a simple generalization of the two-group model for the simple versus composite hypothesis in, for example, Efron et al. (2001) and Sun and Cai (2007) to the composite versus composite hypothesis when A is not a singleton set.

2.1.1. Decision theoretical framework. Let the penalty of making a Type I error (falsely rejecting a null case) be λ and that of making a Type II error (falsely accepting a non-null case) be 1. The loss function for each hypothesis given a decision rule δ_i is $L(H_i, \delta_i, \lambda_i) = \lambda\delta_i(1 - H_i) + (1 - \delta_i)H_i$. The penalty λ is universal as S_i is independent and identically distributed (i.i.d.). The expected loss or the Bayes risk of the compound decision problem is then

$$\begin{aligned} E\Big[\sum_{i=1}^{m} L(H_i, \delta_i, \lambda)\Big] &= m(\lambda \mathbb{P}(\delta_i = 1, H_i = 0) + \mathbb{P}(\delta_i = 0, H_i = 1)) \\ &= m\big((1-p)\lambda \int \delta(s) dF_0(s) + p(1 - \int \delta(s) dF_1(s))\big) \\ &= m\big(p + \int \delta(s)((1-p)\lambda f_0(s) - p f_1(s))\, ds\big). \end{aligned}$$

[1] The density function φ is parametric in the sense that, conditional on ν_i, it is known up to a finite number of structural parameters that are common to all i. We focus on the scalar case for ν_i as this is most common in practice. For identifiability of the mixing distribution G (required later for the adaptive procedure), it suffices for φ to belong to the one-parameter exponential family and the support for S to have a non-empty interior (see Pfanzagl, 1988). The decision theoretical framework can be extended to the multivariate case if the hypotheses in the multiple testing problem involve a vector of parameters. Identifiability of the mixing distribution needs to be discussed case by case.

The optimal Bayes procedure that minimizes the Bayes risk is, for $i = 1, 2, \ldots, m$,

$$\delta_i = \begin{cases} 1 & \text{if } (1-p)f_0(s_i)/f(s_i) \le 1/(\lambda+1) \\ 0 & \text{if } (1-p)f_0(s_i)/f(s_i) > 1/(\lambda+1) \end{cases},$$

where $f(s_i) = (1-p)f_0(s_i) + pf_1(s_i)$ is the marginal density of S_i. The quantity in the optimal Bayes rule implies that the optimal transformation $Z(s)$ in the decision rule $\delta_i = I\{Z(s) \le 1/(\lambda+1)\}$ takes the form $Z(s) = (1-p)f_0(s)/f(s)$, which is a monotonic function of the likelihood ratio. This transformation of s takes the same form as the local false discovery rate (LFDR) statistics, proposed as the optimal test statistics in Efron et al. (2001) and Sun and Cai (2007) for multiple testing with simple versus composite hypotheses. It continues to be the optimal transformation for the composite versus composite case, except that f_0 and f need to be replaced with the corresponding mixture density. It can be interpreted as the posterior probability of the event $\{\nu_i \in A\}$ conditional on the observation of S_i.

Next, we study the choice of λ so that the decision rule δ_i controls the FDR under level α. Denote the optimal cut-off by λ^*. We first introduce a condition on the density of S that is easy to verify, followed by a lemma discussing the implication of the condition. Proposition 2.1 then formalizes the choice of λ^*.

CONDITION 2.1. *Let S_i follow a parametric mixture distribution. Conditional on ν_i, denote the density of S_i as $\varphi(\cdot \mid \nu_i)$. The density of S_i satisfies the following monotonicity condition: $\nabla_s \log \varphi(s \mid \nu)$ is increasing in ν.*

LEMMA 2.1. (MONOTONICITY OF $Z(s)$) *Let $f_0(s)$ be the marginal density of S_i conditional on $\{\nu \le 0\}$ and let $f(s)$ be the marginal density of S_i. Under Condition 2.1, the transformation $Z(s) = (1-p)f_0(s)/f(s)$ is monotonically decreasing in s. Given λ and the decision rule $\delta_i = I\{Z(S_i) \le 1/(\lambda+1)\}$, the rejection region for S_i can be found as $\Gamma_\lambda := [c(\lambda), \infty]$ in which $c(\lambda)$ is the root of $Z(c(\lambda)) = 1/(\lambda+1)$ and $c(\lambda)$ is increasing in λ.*

Condition 2.1 is easy to verify. For example, the density functions of the normal, lognormal and chi-squared distributions all satisfy Condition 2.1. More generally, for distributions belonging to the exponential family whose density functions take the form $\varphi(s \mid v) = h(s)\exp(\eta(v)T(s) - A(v))$, Condition 2.1 is satisfied as long as the derivative $\dot{\eta}(v)\dot{T}(s) > 0$, which is very easy to verify. When the density φ does not have a closed-form derivative, as long as it is possible to evaluate $Z(.)$ on a relevant support of S_i, then the above monotonicity assumption can be verified through a simple numerical analysis.

Lemma 2.1 implies that under Condition 2.1, the thresholding Bayes rule based on $Z(S_i)$ can be equivalently formulated as a thresholding rule on S_i itself. The following proposition proposes an optimal thresholding rule.

PROPOSITION 2.1. *Under Condition 2.1, the decision rule $\delta_i = I\{Z(S_i) \le 1/(\lambda^*+1)\} = I\{S_i \ge c(\lambda^*)\}$ with $\lambda^* = \inf\{\lambda : (1-p)(1-F_0(c(\lambda)))/(1-F(c(\lambda))) = \alpha\}$ yields the Oracle testing procedure that minimizes $mFNR(\lambda) = pF_0(c(\lambda))/F(c(\lambda))$ while controlling $mFDR(\lambda) = (1-p)(1-F_0(c(\lambda)))/(1-F(c(\lambda)))$ at level α. Because $mFDR = FDR + O(1/m)$ and $mFNR = FNR + O(1/m)$, the Oracle procedure controls FDR and minimizes FNR asymptotically as $m \to \infty$.*

The marginal false discovery rate (mFDR) and its counterpart, the marginal false non-discovery rate (mFNR) defined in Proposition 2.1, are frequently used in the multiple testing literature, first introduced by Genovese and Wasserman (2002) and Storey (2002). Under

Condition 2.1, the function $mFDR(\lambda) = (1-p)(1-F_0(c(\lambda)))/(1-F(c(\lambda)))$ is a decreasing function of λ, and hence λ^* can be uniquely found. Additionally, $mFNR(\lambda) = pF_0(c(\lambda))/F(c(\lambda))$ is an increasing function of λ, which explains why λ^* is found by setting the mFDR at an exact level α such that the mFNR is minimized. Lemma B.1 in Appendix B shows that mFDR and mFNR converge to FDR and FNR, respectively, as $m \to \infty$, and characterizes the rate of convergence.

Proposition 2.1 is closely related to the thresholding rule in Sun and Cai (2007) and Sun and McLain (2012) for two-sided multiple testing. Their results rely on the crucial assumption of a monotonic likelihood ratio (MLR) on the density of the LFDR statistics $Z(S_i)$. Violating the MLR condition leads to ill-behaved thresholding procedures, as illustrated by several examples in Cao et al. (2013). However, the MLR condition is very difficult to validate because the density for the LFDR statistics usually does not exhibit an explicit form. Instead of focusing on the distribution of the LFDR statistics, Proposition 2.1 characterizes the rejection region directly for S_i, and imposes an easy-to-verify monotonicity assumption on $Z(S_i)$.[2]

2.1.2. Oracle procedure based on p-values. Any monotonic transformation of the LFDR statistics $Z(s)$ will lead to an equivalent decision rule under Condition 2.1. One particular monotonic transformation of interest links the decision rule to the p-value of S_i under the least favourable condition. The equivalence result is in contrast to the findings in Sun and Cai (2007) and Sun and McLain (2012) for two-sided tests, where procedures based on LFDR statistics are superior to those based on p-values. This is because for multiple testing problems involving two-sided hypotheses, when the distribution $G(\nu)$ or the null set A is not symmetric around zero, the Bayes rule yields a two-tailed rejection region on S_i and the cut-off values at the two tails are not necessarily equal in absolute values. The LFDR statistics adapt to such asymmetry while the p-values, by construction, treat the two tails symmetrically. With one-sided hypotheses and one-tailed rejection regions, however, it is always possible to find a p-value based decision rule that is equivalent to the Oracle procedure using the LFDR statistics.

For a composite null such as $H_0 : \nu_i \leq 0$, the common practice for constructing p-values is first to reduce the composite null to a least favourable simple hypothesis $\bar{H}_0 : \nu_i = 0$. Then, the p-value is defined as $P_i = 1 - \int_{-\infty}^{S_i} \varphi(s \mid 0)ds \equiv 1 - F_{\nu=0}(S_i)$. Let $F_\nu(x) = \int_{-\infty}^{x} \varphi(s \mid \nu)ds$. The distribution for P_i under H_0 is

$$F_P^0(p) = \mathbb{P}(P_i \leq p \mid \nu_i \leq 0) = \int_{-\infty}^{0} 1 - F_\nu(F_{\nu=0}^{-1}(1-p))dG(\nu) / \int_{-\infty}^{0} dG(\nu).$$

Unless $G(\nu)$ has probability mass $1-p$ at $\nu = 0$ (i.e. there is no heterogeneity of ν on the null region $(-\infty, 0]$), $F_P^0(p)$ is stochastically dominated by the uniform distribution. Continuing to use the uniform distribution for the p-value under the null, the procedure still controls for size, but it becomes increasingly conservative as the distribution F_P^0 deviates away from the uniform distribution. However, if we know the distribution $G(\nu)$, we can in fact characterize the exact distribution for P_i under the composite null. This leads to the following Oracle procedure using p-values.

[2] For the two-sided testing problem considered in Sun and Cai (2007), a rejection region based directly on S_i can also be found. We can replace their MLR assumption on the density of $Z(S_i)$ with a condition on the density of S_i for monotonicity of FDR to hold. The latter is an easy-to-verify condition as the distribution of S_i is known.

PROPOSITION 2.2. *Under Condition 2.1, the Oracle procedure places a threshold for P_i at u^* (i.e. $\delta_i = I\{P_i \le u^*\}$). The optimal cut-off u^* is the solution to the equation $(1-p)F_P^0(u)/F_P(u) = \alpha$, where $p = \mathbb{P}(\nu_i > 0)$, $F_P^0(u) = \mathbb{P}(P_i \le u \mid \nu_i \le 0\}$ and $F_P(u) = \mathbb{P}(P_i \le u)$.*

This shows that the optimal cut-off for p-values, u^*, has a one-to-one mapping to λ^* defined in Proposition 2.1, which implies that the two Oracle procedures are equivalent. In practice, the p-value procedure boils down to calculating for each P_i the quantity $(1-p)F_P^0(P_i)/F_P(P_i)$, which is the q-value defined in Storey (2003). Proposition 2.2 generalizes the q-value for the composite null case by replacing the uniform distribution for the p-value assumed in Storey (2003) by its exact distribution F_P^0 under the null.

2.2. *Adaptive procedures*

The Oracle procedures discussed above are not feasible unless we know the distribution $G(\nu)$. Given the equivalence result discussed in the previous section, we focus on introducing adaptive procedures based on p-values. Define the plug-in estimator for the q-value as

$$\hat{q}_i := \frac{\int_{-\infty}^{0} 1 - F_\nu(F_{\nu=0}^{-1}(1-p_i))d\hat{G}(\nu)}{\int_{-\infty}^{\infty} 1 - F_\nu(F_{\nu=0}^{-1}(1-p_i))d\hat{G}(\nu)},$$

with a consistent estimator $\hat{G}$ for G. The adaptive p-value procedure rejects all cases where $\hat{q}_i$ is below level α (i.e. $\delta_i = I\{\hat{q}_i \le \alpha\}$).

2.2.1. NPMLE and the benchmark adaptive procedure. Because ν_i is not directly observed while S_i is, the estimation of the distribution $G(\nu)$ turns into a deconvolution problem. We propose to estimate the distribution G using the NPMLE method. Compared to the empirical characteristic function based kernel method previously adopted in the literature, the NPMLE method does not involve nuisance bandwidth selection and it delivers a consistent estimator $\hat{G}$ for a broader class of the G distributions.

Let $\hat{G}$ be the NPMLE estimator of G, defined as

$$\hat{G} = \underset{G\in\mathcal{G}}{\operatorname{argmax}}\Big\{\sum_{i=1}^{m}\log f_i \mid f_i = \int \varphi(S_i \mid \nu)dG(\nu)\Big\},$$

where $\mathcal{G}$ is the space of distribution functions.

LEMMA 2.2. *Let $\hat{G}$ be the NPMLE for G. Suppose $\varphi(\cdot \mid \nu)$ is a member of the one-parameter exponential family and the support of the dominating measure for S has a non-empty interior. Then, G is identifiable and $\hat{G}$ is strongly consistent, i.e. $\hat{G}(u) \to G(u)$ for all $u \in C_G \equiv \{u :$ G is continuous at $u\}$ with probability one.*

Lemma 2.2 is Theorem 33.10 of DasGupta (2008) adapted to our notation. The consistency property of $\hat{G}$ was first established in Kiefer and Wolfowitz (1956) and further discussed in Pfanzagl (1988) and van de Geer (1993). The normal location mixture model we consider at the beginning of the section is included in the exponential family. Lindsay (1995) provides a comprehensive survey of mixture models in general. The commonly used algorithm for computing the NPMLE is the EM algorithm, proposed in Laird (1978), but the slow convergence of the algorithm makes it inaccessible when the sample size m is large or when the distribution G has a complicated structure. Koenker and Mizera (2014) have recently proposed a more efficient

interior point algorithm for NPMLE that makes it computationally feasible for applications to large-scale problems. In this paper, we adopt the computation method developed by Koenker and Mizera (2014). Next, we describe the procedure in the context of our deconvolution problem (i.e. the estimation of $G(v)$).

Discretizing the domain of v to a fixed grid $\{v_1, \ldots, v_J\}$, the likelihood formulation can be rewritten as a convex programming problem with linear constraints:

$$\max_{g \in \mathcal{S}} \Big\{ \sum_{i=1}^{m} \log(f_i) \mid f \equiv (f_1 \ \ldots \ f_m)^\top = Ag \Big\}.$$

Here, the matrix A denotes an $m \times J$ matrix with elements $A_{ij} = \varphi(S_i \mid v_j)$ and g is a $J \times 1$ vector that belongs to the unit simplex $\mathcal{S} = \{g \in \mathbb{R}^J \mid \sum_{j=1}^{J} g_j \Delta_j = 1, g \geq 0\}$ with Δ_j as the jth grid width. As characterized in Theorem 2 of Koenker and Mizera (2014), the unique solution to the convex programming problem, $\hat{G}$, exists and is an atomic probability measure, with no more than m atoms.

Unlike the kernel density estimation method used in Sun and McLain (2012), the NPMLE method does not involve a bandwidth selection. Although the grid size J is indeed a user input parameter, it is very different from the bandwidth selection problem in kernel density estimation. It is well known that kernel density estimators are sensitive to the bandwidth choice, and there is a trade-off between bias and variance, so the optimal bandwidth often minimizes objectives such as mean squared error or mean integrated squared error. However, the estimator $\hat{G}$ is not sensitive to the choice of J once the grid size is sufficiently large. In other words, the choice of grid size would not be an issue if practitioners were to have unlimited computation resources. Further, because the new algorithm proposed by Koenker and Mizera (2014) is computationally very efficient, practitioners can typically afford a sufficient grid size for problems of moderate m. For example, in our simulation experiments and empirical example, the use of 100 and 500 grid points yields identical results for FDR, FNR and the coverage probability. However, for extremely large-scale hypothesis testing problems (i.e. with millions of hypotheses), the computational burden might be a concern. In such situations, Dicker and Zhao (2016) provide some theoretical justification of choosing grid size $J \approx \sqrt{m}$.

The following theorem shows that the adaptive procedure that replaces $G(v)$ by its estimator $\hat{G}(v)$ is asymptotically equivalent to the Oracle procedure when there is no mass point at the cut-off value of the one-sided null hypothesis.

THEOREM 2.1. *Suppose that $\varphi(\cdot \mid v)$ is a member of the one-parameter exponential family that satisfies Condition 2.1, that the support of the dominating measure for S has a non-empty interior and that the distribution G has no point mass at zero, the cut-off value of the one-sided null hypotheses. The adaptive procedure replacing G by the NPMLE $\hat{G}$ is asymptotically equivalent to the Oracle procedure. That is, the decision rule $\delta_i = I\{\hat{q}_i \leq \alpha\}$ controls mFDR at level α asymptotically, and the associated mFNR converges to the Oracle mFNR as $m \to \infty$.*

The class of distributions for G allowed in Theorem 2.1 is larger than the adaptive procedure that uses a kernel-based method, as in Sun and McLain (2012). The consistency of the NPMLE method applies for distributions that are continuous, discrete, or consist of both a continuous part and discontinuity points. The kernel-based deconvolution method, however, is only suitable for $G(v)$ that are continuous with suitable smoothness conditions.[3]

[3] For instance, for asymptotic validity of the adaptive procedures in Sun and McLain (2012), we need the density for v to be continuous, bounded and twice differentiable, and with a bounded second derivative.

In spite of a large family of distributions for G, Theorem 2.1 requires that G has no discontinuity point at the cut-off value of the one-sided null hypothesis. This is because, as stated in Lemma 2.2, consistency of the NPMLE $\hat{G}$ implies consistency of all continuity points of G. Suppose zero is the cut-off value and $\mathbb{P}_G(\nu = 0) > 0$, then $\hat{G}(0) \not\to G(0)$. Then, this further leads to invalidity of the decision rule based on the classic NPMLE $\hat{G}$. In the next section, we propose a hybrid-NPMLE method that is robust to this special case, so that we can avoid imposing any assumption on the distribution of $G(.)$ at the cut-off value of the one-sided null.

2.2.2. Hybrid-NPMLE method. In this section, we extend the classic NPMLE method to a hybrid version to account for the possibility that, in some applications, there might be a non-trivial probability mass at the cut-off value of the one-sided null hypothesis. The hybrid method imposes the point mass at zero into the linear constraints of the convex optimization problem. Given that the probability mass at $\nu = 0$ is $1 - \omega$ and that for the rest of the support ν follows a distribution H from the class $\mathcal{H}$, which consists of well-defined distributions such that $P_H(\nu \neq 0) = 1$, the marginal density of S_i under these assumptions becomes

$$f(s) = (1 - \omega)\varphi(s \mid 0) + \omega \int \varphi(s \mid \nu) dH(\nu).$$

In the infeasible case where ω is known, the NPMLE estimator for H can be solved by, on a grid for ν of size J,

$$\max_{h \in \mathcal{S}} \Big\{ \sum_{i=1}^{m} \log(f_i) \mid f = \tilde{A}h \Big\},$$

where h is a J-vector in the unit simplex and $\tilde{A}$ denotes an $m \times J$ matrix with elements

$$\tilde{A}_{ij} = (1 - \omega)\varphi(S_i \mid 0) + \omega\varphi(S_i \mid \nu_j).$$

Because $\varphi(S_i \mid 0)$ does not change over the grid for ν, the ith row of the linear constraints $\tilde{A}h$ is the grid approximation of the marginal density $f(.)$, taking the form,

$$f(S_i) = (1 - \omega)\varphi(S_i \mid 0) + \omega \sum_{j=1}^{J} \varphi(S_i \mid \nu_j) h_j,$$

where h_j is the jth element of the vector h.

When ω is unknown but can be consistently estimated by $\hat{\omega}$, we plug $\hat{\omega}$ into the linear constraints in the matrix $\tilde{A}$ and the NPMLE thus solved also leads to a consistent estimator for H, which subsequently provides an asymptotically valid adaptive procedure. Results are formally stated in Theorem 2.2. Consistent estimation of ω can be obtained using methods proposed by Jin (2008) and Cao and Kosorok (2011). One advantage of our hybrid-NPMLE method is that unlike the two-step approach in Sun and McLain (2012), which requires trimming away the test statistics that fall in the range of the $\hat{\omega}/2$ and $1 - \hat{\omega}/2$ quantile of S, denoted as $[Q_{\hat{\omega}/2}(S), Q_{1-\hat{\omega}/2}(S)]$ when applying the kernel-based deconvolution for density H, our hybrid-NPMLE method does not require data trimming and permits a broader class of distribution for H.

THEOREM 2.2. *Suppose $\varphi(\cdot \mid \nu)$ is a member of the exponential family that satisfies Condition 2.1 and the support of the dominating measure for S has a non-empty interior. Suppose the distribution G of ν takes the form $(1 - w)\delta_0 + wH(\nu)$ where $w \in [0, 1]$, δ_0 is a Dirac function*

and H belongs to a family of distributions $\mathcal{H}$. Let $\hat{\omega}$ be a consistent estimator for ω as $m \to \infty$. The adaptive procedure, based on $\hat{G}(v) \equiv (1 - \hat{w})\delta_0 + \hat{w}\hat{H}(v)$ with

$$\hat{H} \equiv \underset{H \in \mathcal{H}}{\operatorname{argmax}} \Big\{ \sum_{i=1}^{m} \log \big((1 - \hat{w})\varphi(s_i \mid 0) + \hat{w} \int \varphi(s_i \mid v) dH(v) \big) \Big\},$$

is asymptotically equivalent to the Oracle procedure.

3. THE HETEROGENEOUS EFFECT OF GOING TO A BETTER HIGH SCHOOL

In recent years, much research – see, e.g. Hoekstra (2009), Duflo et al. (2011), Pop-Eleches and Urquiola (2013), Abdulkadiroglu et al. (2014) and Shen and Zhang (2016) – in labour and education economics has studied the effect of going to a more selective school. Findings in this literature are not conclusive. For example, using an administrative data set from Romania, Pop-Eleches and Urquiola (2013) find strong evidence that attending a more selective high school significantly improves a student's academic outcome, while Abdulkadiroglu et al. (2014) find little evidence that attending a selective exam school in Boston and New York matters. In this section, we revisit the Romanian data set studied by Pop-Eleches and Urquiola (2013). Instead of pooling students from different towns and schools together and estimating the pooled effects of attending a more selective high school, we estimate town-specific and school-specific treatment effects and investigate potential treatment effect heterogeneity.

As is discussed in Pop-Eleches and Urquiola (2013), in Romania, a typical elementary school student takes a nationwide test in the last year of elementary school (eighth grade) and applies to a list of high schools and tracks. The admission decision is entirely dependent on the student's transition score, an average of the student's performance on the nationwide test and grade point average, as well as the student's preference for schools and tracks. A student is admitted to the most selective school and track for which he or she qualifies based on the transition score and admission score cut-offs of different schools and tracks. Following Pop-Eleches and Urquiola (2013), we use the regression discontinuity (RD) approach to identify and estimate the effect of attending a higher-ranked school. The approach, in short, compares the average outcome of students who are marginally admitted to a more selective high school with that of those who marginally missed out and attended a less selective or non-selective school. Figure 1 replicates Panels D and F of Figure 1 in Pop-Eleches and Urquiola (2013) and summarizes the concept of the RD analysis; see Pop-Eleches and Urquiola (2013) for details. The x-axis, or the running variable, in both graphs is the standardized transition score subtracting individual school cut-off. The y-axis, or the outcome variable, in the left graph is the probability of a student taking the Baccalaureate exam and that in the right graph is the Baccalaureate exam grade. We see clear evidence of outcome discontinuity in the right graph but not so much in the left graph. This indicates that, compared with students who marginally miss out, students who marginally gain access to a higher-ranked school have similar exam-taking rate but higher average exam grade. This is called the reduced-form effect. The treatment effect of attending a more selective school is equal to the size of the discontinuity documented in the graphs for reduced-form effect divided by the proportion of compliers at the score cut-off, or the proportion of marginal students who attend a higher-ranked school when eligible to do so. The treatment effect always has the same sign as the reduced-form effect.

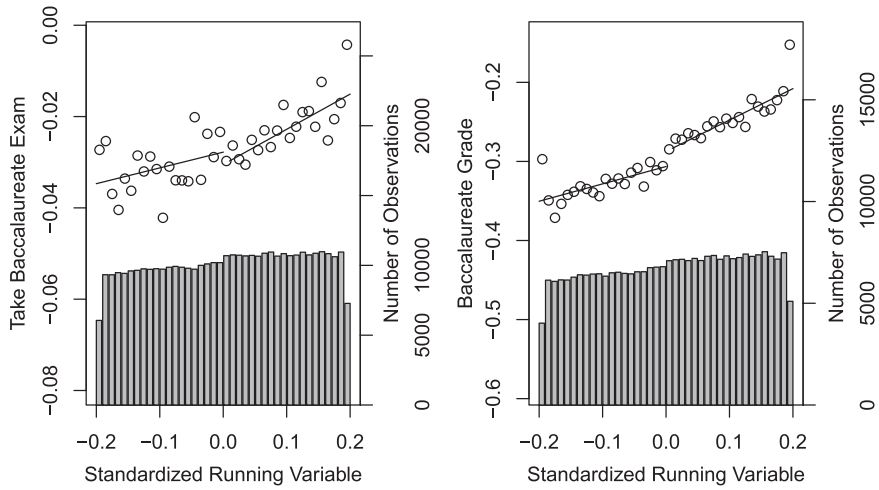

Figure 1. Pooled regression discontinuity analysis.

As discussed earlier, it is natural to expect that the effect of attending a better school varies from town to town and from school to school. To investigate such heterogeneity, we repeat the RD analysis conducted in Pop-Eleches and Urquiola (2013) for each town and then for each school. We use the exact same specification as in Pop-Eleches and Urquiola (2013) with a uniform kernel and a bandwidth equal to 1. We then further restrict our data set to remove schools with an unbalanced RD design (i.e. schools with the left boundary of the running variable falling inside the estimation window $[-1, 1]$). This leaves us with 106 towns and 503 schools. We are interested in testing the null hypothesis that the treatment effect is non-positive for each individual town or school. By doing this, we will be able to select out towns and schools with positive treatment effects while controlling the FDR at various pre-determined levels (1%, 5% and 10%). As studying the reduced-form effect avoids the potential weak first-stage problem in finite sample inference, we carry out our test using the reduced-form estimates that compare students who barely pass the transition score cut-off, and thus have access to a better school, with those who barely miss out. Note that as our problem is a pure inference problem, the conclusion of our multiple testing analysis can always be extended to the treatment effect of attending a better school because the treatment effect and the reduced-form effect always have the same sign. However, if researchers are interested in comparing the magnitude of treatment effects among towns/schools, the reduced-form estimates would be uninformative unless the first stage is homogeneous.

Using the notation in the methodology section, in this application, i is the town (and later school) index; $i = 1, \ldots, 106$ (and later $i = 1, \ldots, 503$ for school heterogeneity). The random variable μ_i is the population reduced-form effect of having access to a better school in town i (and later school i). For reasons specified above, the set of interest $\Im_+$ is the set of all towns (and later schools) that have positive average reduced-form effects, which also implies positive average treatment effect, of attending a better school. The statistic S_i is the studentized estimator for town-specific reduced-form effects. If the standard errors of the estimator are known for each group (towns or schools), then the test statistics follow a standard normal distribution, which

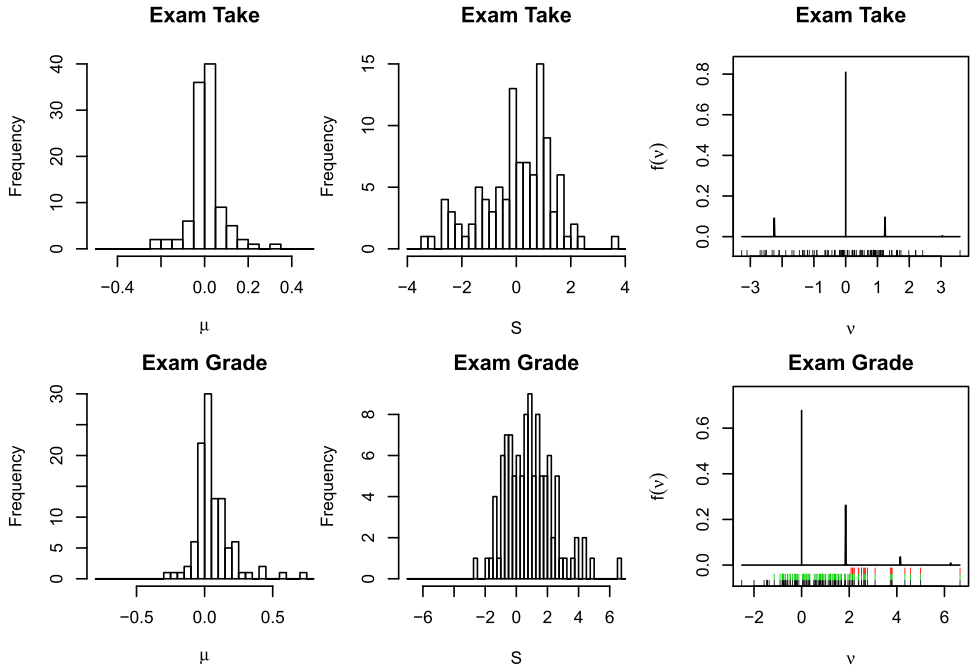

Figure 2. Distribution of treatment effect estimates: town-level heterogeneity. [Colour figure can be viewed at wileyonlinelibrary.com]

belongs to the exponential family and satisfies Condition 2.1. As demonstrated by the simulation included in the online Appendix, we do not expect the estimation of the standard errors to affect the asymptotic properties of the proposed method. We also convert the statistic S_i to a Z-score – following suggestions in Efron et al. (2001) and Sun and Cai (2007) – as a robustness check and we find almost identical results.[4]

Figure 2 plots the histograms of the town-specific μ_i (in the two left panels) and S_i (in the two middle panels) for 106 towns, as well as the hybrid-NPMLE deconvolution estimates for the distribution of $G(.)$ (in the two right panels). In the deconvolution graphs, the rugs at the lower level show values of the studentized statistics in all towns, the rugs at the middle level show values for towns rejected by the method in Sun and McLain (2012) and those at the upper level show values for towns rejected by the proposed hybrid-NPMLE method.

The top three panels report town-specific effects of attending a better school on the probability of a student taking the Baccalaureate exam. The proposed hybrid-NPMLE method does not identify any town as having a positive effect. The bottom three panels capture town-specific effects on average Baccalaureate exam grade. The proposed hybrid method finds 21 towns out of 106 that have significant positive effects.

Figure 2 also reports the multiple testing results using the procedures of Sun and McLain (2012). As discussed earlier and illustrated with the simulation results, the method in Sun and McLain (2012) does not apply and could have substantial inflation in FDR when the effect and

[4] These results are omitted in the interest of space but can be obtained with the companion codes.

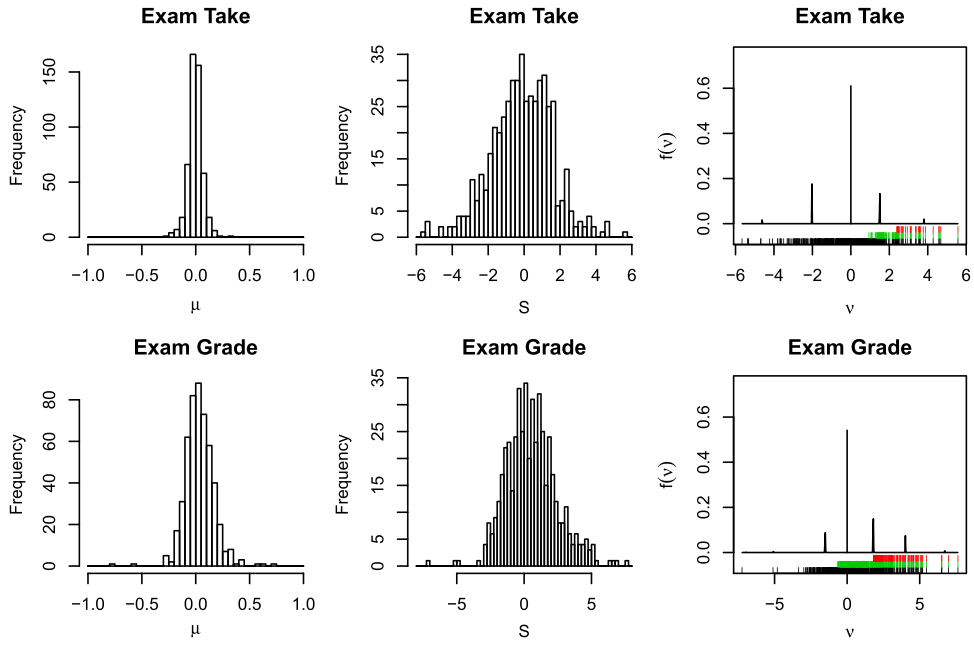

Figure 3. Treatment effect distribution: school-level heterogeneity. [Colour figure can be viewed at wileyonlinelibrary.com]

its associated standard deviation are correlated. This assumption is not likely to hold in this empirical example. For the town-level analysis, the correlation coefficient between the effect estimator and its associated standard error is 0.298 for the exam-taking rate outcome and 0.335 for the exam grade outcome. The procedure of Sun and McLain (2012) agrees with the proposed method regarding towns with significant positive effects on the exam-taking rate, yet it rejects a much larger number of towns (88 towns out of 106) for having a positive effect on exam grade compared to the proposed procedure.

Next, we further disaggregate the data set to study the school-level treatment effect heterogeneity. Comparing the histograms plotted in Figure 3 with those in Figure 2, we find that the school-level estimates and statistics reveal stronger evidence of heterogeneous effects, which is further confirmed by the non-parametric deconvolution results reported in the two graphs in the right panel.

Figure 4 compares the set of schools rejected (represented by the solid circles) by the proposed hybrid-NPMLE method with a 5% FDR control, with those not rejected (represented by the open circles). The x-axis of the figure is the reduced-form effect estimate while the y-axis is the school admission cut-off. Interestingly, for both outcomes (exam-taking rate and average exam grade), the set of schools that are identified to have positive effects also has higher admission score cut-offs. The difference in admission score cut-offs is large and statistically significant. For the exam-taking rate, we find that schools identified by the proposed method to have a positive effect have 0.5 (p-value < 0.0001) standard deviation higher admission score cut-off; for the exam grade outcome, the difference is 0.3 (p-value < 0.0001) standard deviation.

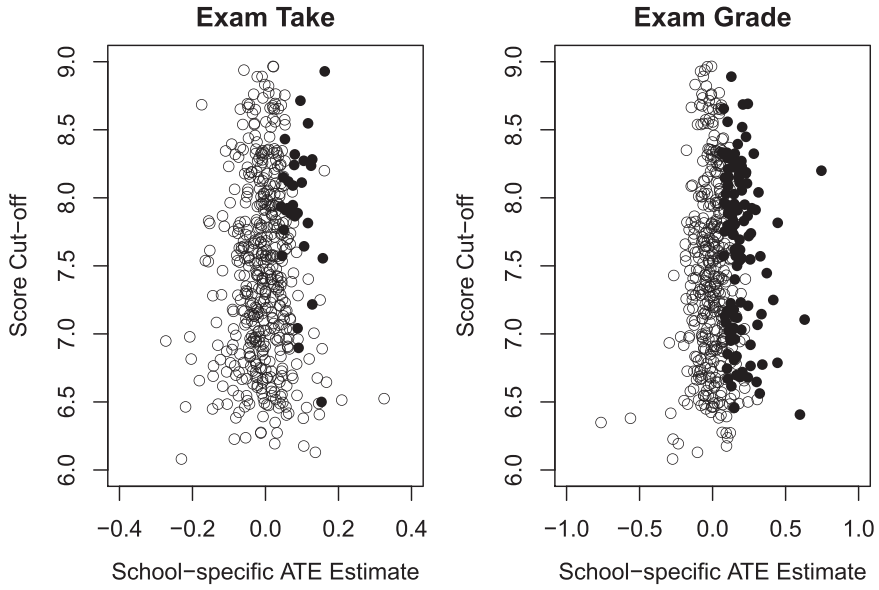

Figure 4. The rejected set of schools.

Table 1. Multiple testing for positive treatment effects.

	Take Baccalaureate exam			Baccalaureate exam score		
	FDR = 1%	FDR = 5%	FDR = 10%	FDR = 1%	FDR = 5%	FDR = 10%
Town level						
Hybrid-NPMLE	0	0	0	7	14	21
Sun and McLain	0	0	0	40	70	88
BH	0	1	1	7	14	21
Holm	0	1	1	7	7	7
No. of hypotheses	106	106	106	106	106	106
School level						
Hybrid-NPMLE	9	19	29	53	87	115
Sun and McLain	23	49	78	109	239	340
BH	11	17	22	45	70	88
Holm	5	7	10	24	31	35
No. of hypotheses	503	503	503	503	503	503

Note: The results for the hybrid-NPMLE, Sun and McLain, and BH procedures are based on tests with FDR controlled at 1%, 5% and 10%, respectively. The results for the Holm procedure are based on tests with FWER controlled at 1%, 5% and 10%, respectively.

All the multiple testing results discussed in this section are summarized in Table 1, where the number of rejected null hypotheses are reported for various testing methods. In summary, the empirical section confirms the simulation evidence that the proposed hybrid-NPMLE procedure has robust small-sample performance. When there is little evidence of negative treatment effects, and the one-sided composite null is homogeneously zero, the proposed method gives rejection

Table 2. Cut-off for p-values: school level.

	Take Baccalaureate exam			Baccalaureate exam score		
	FDR* = 1%	FDR* = 5%	FDR* = 10%	FDR* = 1%	FDR* = 5%	FDR* = 10%
Hybrid-NPMLE	0.0002	0.0032	0.0086	0.0015	0.0161	0.0344
BH	0.0002	0.0017	0.0044	0.0009	0.0069	0.0168
Holm	9.26×10^{-6}	8.35×10^{-5}	0.0002	1.42×10^{-5}	7.55×10^{-5}	0.0002
No. of hypotheses	503	503	503	503	503	503

Note: The results for the hybrid-NPMLE and BH methods are based on tests with FDR controlled at 1%, 5% and 10%, respectively. Results for the Holm procedure are based on tests with FWER controlled at 1%, 5% and 10%, respectively.

sets comparable to the classic BH procedure. When there is evidence for negative treatment effects, the proposed method has better power than the BH procedure. To make the comparison explicit, in Table 2 we report the adaptively estimated optimal cut-off value for p-values for the hybrid-NPMLE, the BH and the Holm procedures, with FDR (or FWER for the Holm procedure) controlled at the 1%, 5% and 10% levels, respectively. We only report the school-level analysis as the cut-offs for the town-level analyses of hybrid-NPMLE and BH are almost identical. Additionally, the procedure in Sun and McLain (2012) is omitted because of the violation of the key independence assumption, as discussed above.

4. CONCLUSION

In this paper, we propose a multiple testing framework for identifying subpopulations with positive responses to the outcome variable. There are many applications where this task is of interest, especially for the treatment effect evaluation. This can be considered as a continuation of the work by Lee and Shaikh (2013) and Armstrong and Shen (2015) who consider multiple testing procedures for treatment effects that control the FWER. As recognized in Lee and Shaikh (2013), their methodology is sufficient when the number of null hypotheses jointly tested is modest. However, in some other applications, including the empirical example of this paper, procedures controlling the FWER are too stringent and have low power once the number of tests becomes large. Our FDR procedure is designed for multiple testing of a one-sided hypothesis with a composite null. The adaptive version of the procedure uses classic NPMLE and a new hybrid-NPMLE estimator for the effect distribution, instead of the conventionally used empirical characteristic function based kernel estimator. NPMLE methods allow for a broader class of treatment effect distributions. Monte Carlo simulations demonstrate that the adaptive procedure has good size and power in comparison to many existing multiple testing procedures. We apply the framework to study the effect of attending a more selective high school in Romania.

ACKNOWLEDGEMENTS

The authors would like to thank Roger Koenker, Stanislav Volgushev and seminar participants at the Midwest Econometrics Group Meeting 2014, the University of Alberta, and the Simon Fraser University for helpful comments. Jiaying Gu acknowledges financial support from the Connaught Fund for the 2016–2018 New Researcher award. Part of the research was carried out

while Jiaying Gu was visiting Ruhr University Bochum. She is very grateful for the hospitality of the Mathematics Department and acknowledges financial support from Project C1 of the SFB 823 of the German Research Foundation.

REFERENCES

Abdulkadiroglu, A., J. Angrist and P. Pathak (2014). The elite illusion: achievement effects at Boston and New York exam schools. *Econometrica 82*, 137–96.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and early training projects. *Journal of the American Statistical Association 103*, 1481–95.

Andrews, D. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*, 119–57.

Armstrong, T. and S. Shen (2015). Inference for optimal treatment assignments. Working Paper, Yale University.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Series B 57*, 289–300.

Benjamini, Y., A. Krieger and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*, 491–507.

Cao, H. and M. Kosorok (2011). Simultaneous critical values for t-tests in very high dimensions. *Bernoulli 17*, 347–94.

Cao, H., W. Sun and M. Kosorok (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika 100*, 495–502.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer Texts in Statistics. Berlin: Springer.

Dicker, L. and S. D. Zhao (2016). High-dimensional classification via nonparametric empirical Bayes and maximum likelihood. *Biometrika 103*, 21–34.

Duflo, E., P. Dupas and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. *American Economic Review 101*(5), 1739–74.

Efron, B. (2004a). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association 99*, 96–104.

Efron, B. (2004b). Local false discovery rate. Working paper, Stanford University.

Efron, B., R. Tibshirani, J. Storey and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association 96*, 1151–60.

Fishe, R. P. and A. D. Smith (2012). Identifying informed traders in futures markets. *Journal of Financial Markets 15*, 329–59.

Genovese, C. and L. Wasserman (2002). Operating characteristic and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society*, Series B *64*, 499–517.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics 23*, 365–80.

Heller, R. (2010). Comment: Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association 105*, 1057–59.

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: a discontinuity-based approach. *Review of Economics and Statistics 91*, 717–24.

Hsu, Y.-C., C.-M. Kuan and M.-F. Yen (2014). A generalized stepwise procedure with improved power for multiple inequalities testing. *Journal of Financial Econometrics 12*, 730–55.

Jin, J. (2008). Proportion of non-zero normal means: universal Oracle equivalences and uniformly consistent estimators. *Journal of Royal Statistical Society*, Series B *70*, 461–93.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics 27*, 887–906.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *Journal of the American Statistical Association 109*, 674–85.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association 73*, 805–11.

Lee, S. and A. M. Shaikh (2013). Multiple testing and heterogeneous treatment effects: re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics 29*, 612–26.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5. Hayward, CA: Institute of Mathematical Statistics.

Pfanzagl, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference 19*, 137–58.

Pop-Eleches, C. and M. Urquiola (2013). Going to a better school: effects and behavioral responses. *American Economic Review 103*(4), 1289–324.

Resnick, S. (1998). *A Probability Path*. Boston, MA: Birkhäuser.

Romano, J., A. Shaikh and M. Wolf (2008). Formalized data snooping based on generalized error rates. *Econometrics Theory 24*, 404–47.

Shen, S. and X. Zhang (2016). Distributional test for regression discontinuity: theory and empirical examples. *Review of Economics and Statistics 98*, 685–700.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, Series B *64*, 479–98.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics 31*, 2013–35.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association 102*, 901–12.

Sun, W. and A. McLain (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association 107*, 673–87.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics 21*, 14–44.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

APPENDIX A: DEFINITION OF VARIOUS ERROR RATES IN MULTIPLE TESTING

We provide the definitions of various multiple testing error rates that are commonly used in the literature for the convenience of the readers. Suppose we are conducting m hypothesis tests simultaneously. We have the following possible outcomes:

	Accept	Reject	Total
Null case ($\mu_i \leq 0$)	U	V	m_0
Non-null case ($\mu_i > 0$)	T	S	$m - m_0$
	$m - R$	R	m

(1) FWER control of level α guarantees $\mathbb{P}(V < 1) \leq \alpha$;
(2) FDP control of level α guarantees $\mathbb{P}(V/R \geq k) \leq \alpha$ for $k \in (0, 1)$;
(3) FDR control of level α guarantees $\mathbb{E}[V/R] \leq \alpha$.

APPENDIX B: PROOFS OF LEMMATA AND THEOREMS

Proof of Lemma 2.1: Note that we can write

$$Z(s) = \frac{\int_{-\infty}^{0} \varphi(s \mid \nu) dG(\nu)}{\int_{-\infty}^{\infty} \varphi(s \mid \nu) dG(\nu)}.$$

It suffices to show $\nabla_s Z(s) \leq 0$:

$$\begin{aligned}
\nabla_s \frac{\int_{-\infty}^{0} \varphi(s \mid \nu) dG(\nu)}{\int_{-\infty}^{\infty} \varphi(s \mid \nu) dG(\nu)} &= \frac{\int_{-\infty}^{0} \nabla_s \log \varphi(s|\nu) \varphi(s|\nu) dG(\nu)}{\int_{-\infty}^{\infty} \varphi(s|\nu) dG(\nu)} \\
&\quad - \frac{\int_{-\infty}^{0} \varphi(s|\nu) dG(\nu)}{\int_{-\infty}^{\infty} \varphi(s|\nu) dG(\nu)} \frac{\int_{-\infty}^{\infty} \nabla_s \log \varphi(s|\nu) \varphi(s|\nu) dG(\nu)}{\int_{-\infty}^{\infty} \varphi(s|\nu) dG(\nu)} \\
&= E[I\{\nu \leq 0\} \nabla_s \log \varphi(s|\nu) \mid s] - E[I\{\nu \leq 0\} \mid s] E[\nabla_s \log \varphi(s|\nu) \mid s] \\
&= \mathrm{Cov}[I\{\nu \leq 0\}, \nabla_s \log \varphi(s|\nu) \mid s] \leq 0
\end{aligned}$$

The last inequality is satisfied under Condition 2.1. Given the decision rule $\delta_i = I\{Z(s) \leq 1/(1+\lambda)\}$ and monotonicity of $Z(s)$, it is obvious that the rejection region takes the form $[c(\lambda), \infty)$. Further, by monotonicity of $Z(s)$, we have $c(\lambda)$ increasing in λ. □

Proof of Proposition 2.1: First, we show that the procedure minimizes mFNR among all procedures that control mFDR at or under α.

Because $c(\lambda)$ is increasing in λ as proved in Lemma 2.1, it is sufficient to show that mFDR is decreasing in c. Taking the derivative with respect to c by applying the fundamental theorem of calculus, we obtain

$$\begin{aligned}
\frac{\partial}{\partial c} mFDR &= \frac{-(1-p) f_0(c)(1 - F(c)) - (1-p)(1 - F_0(c))(-(1-p) f_0(c) - p f_1(c))}{(1 - F(c))^2} \\
&= \frac{(1-p) p \left(f_1(c)(1 - F_0(c)) - f_0(c)(1 - F_1(c)) \right)}{(1 - F(c))^2} < 0.
\end{aligned}$$

The last inequality follows by noticing

$$\frac{1 - F_0(c)}{1 - F_1(c)} = \frac{\int_c^{\infty} f_0(s) ds}{\int_c^{\infty} (f_1(s)/f_0(s)) f_0(s) ds} < \frac{\int_c^{\infty} f_0(s) ds}{\int_c^{\infty} (f_1(c)/f_0(c)) f_0(s) ds} = \frac{f_0(c)}{f_1(c)}.$$

The inequality results from the fact that $f_1(t)/f_0(t)$ is increasing in t, because $1/Z(t) = f(t)/(1-p) f_0(t) = 1 + p f_1(t)/f_0(t)$ is increasing in t, as proved in Lemma 2.1. Using the same argument, we can prove that mFNR is monotonically increasing in c, and hence monotonically decreasing in λ. This suggests

that we should choose mFDR at the maximally allowed level α, and that the chosen λ^* leads to the optimal Oracle procedure that minimizes mFNR while controlling mFDR at the nominal level.

Next we want to show the relationship between mFDR, mFNR and FDR, FNR stated in Lemma B.1.

LEMMA B.1. *For a cut-off value c such that $F_0(c)$, $F_1(c) \in (0, 1)$, under Condition 2.1, we have $E(FDP) = (1-p)(1-F_0(c))/(1-F(c)) + O(m^{-1})$ and $V(FDP) = O(m^{-1})$. Similarly, $E(FNP) = pF_1(c)/F(c) + O(m^{-1})$ and $V(FNP) = O(m^{-1})$.*

Proof: Denote the cardinality of the set $\mathcal{X}_+$ as m_1 and the cardinality of its compliment set as m_0 and let $m = m_0 + m_1$ as the total number of hypotheses and $m_1/m = p$. Given the thresholding rule $\delta_i = I\{S_i \geq c\}$, the FDP based on this decision rule can be found as

$$FDP = \frac{\sum_{i\in\mathcal{X}_+^c} I(S_i \geq c)}{\sum_{i\in\mathcal{X}_+^c} I(S_i \geq c) + \sum_{i\in\mathcal{X}_+} I(S_i \geq c)} I\Big(\sum_{i\in\mathcal{X}} I(S_i \geq c) \neq 0\Big).$$

First, consider the FDP for the case with $m_0 \neq 0$ conditioning on the event that $\sum_{i\in\mathcal{X}} I(S_i \geq c) \neq 0$:

$$\begin{aligned} FDP &= \frac{\sum_{i\in\mathcal{X}_+^c} I(S_i \geq c)}{\sum_{i\in\mathcal{X}_+} I(S_i \geq c) + \sum_{i\in\mathcal{X}_+^c} I(S_i \geq c)} \\ &= \frac{(1/m_0)\sum_{i\in\mathcal{X}_+^c} I(S_i \geq c)}{(1/m_0)\sum_{i\in\mathcal{X}_+} I(S_i \geq c) + (m_1/m_0)(1/m_1)\sum_{i\in\mathcal{X}_+^c} I(S_i \geq c)} \equiv \frac{A}{A + (m_1/m_0)B}. \end{aligned}$$

Given the two-group model for S_i, we have $\sum_{i\in\mathcal{X}_+^c} I(S_i \geq c) \sim \text{binomial}(m_0, 1-F_0(c))$ and $\sum_{i\in\mathcal{X}_+} I(S_i \geq c) \sim \text{binomial}(m_1, 1-F_1(c))$. So, $E[A] = 1 - F_0(c) \equiv A_0$ and $E[B] = 1 - F_1(c) \equiv B_0$, $\text{Var}[A] = (1/m_0)F_0(c)(1-F_0(c))$, $\text{Var}[B] = (1/m_1)F_1(c)(1-F_1(c))$. We know that $A, B \in [0, 1]$, $A + (m_1/m_0)B > 0$ and $A_0, B_0 \in (0, 1)$. Taylor expansion of the random quantity FDP around the ratio of the mean of the binomial random variable leads to

$$\begin{aligned} FDP = \frac{A}{A + (m_1/m_0)B} &= \frac{A_0}{A_0 + (m_1/m_0)B_0} \\ &+ (A - A_0)\Big(\frac{(m_1/m_0)B_0}{(A_0 + (m_1/m_0)B_0)^2}\Big) + (B - B_0)\Big(\frac{-(m_1/m_0)A_0}{(A_0 + (m_1/m_0)B_0)^2}\Big) \\ &+ (A - A_0)^2\Big(\frac{-2(m_1/m_0)\tilde{B}}{(\tilde{A} + (m_1/m_0)\tilde{B})^3}\Big) + (B - B_0)^2\Big(\frac{2(m_1/m_0)\tilde{A}}{(\tilde{A} + (m_1/m_0)\tilde{B})^3}\Big) \\ &+ (A - A_0)(B - B_0)\frac{(m_1/m_0)(\tilde{A} - (m_1/m_0)\tilde{B})}{(\tilde{A} + (m_1/m_0)\tilde{B})^3}, \end{aligned}$$

where $\tilde{A}$ and $\tilde{B}$ are between A and A_0, and B and B_0, respectively. It is clear that $\tilde{A} \in [0, 1]$ and $\tilde{B} \in [0, 1]$ and $\tilde{A} + (m_1/m_0)\tilde{B} > 0$.

Taking expectation on both sides, we obtain

$$\begin{aligned} E\Big[FDP\Big|\sum_{i\in\mathcal{X}} I(S_i \geq c) \neq 0\Big] &= \frac{m_0(1-F_0(c))}{m_0(1-F_0(c)) + m_1(1-F_1(c))} + O(\text{Var}[A]) + O(\text{Var}[B]) \\ &= \frac{m_0(1-F_0(c))}{m_0(1-F_0(c)) + m_1(1-F_1(c))} + O\left(\frac{1}{m}\right). \end{aligned}$$

Because

$$P\Big(\sum_{i\in\mathcal{X}} I(S_i \geq c) = 0\Big) = F_0^{m_0}(c)F_1^{m_1}(c) = o\left(\frac{1}{m}\right).$$

Then

$$E[FDP] = E\Big[FDP\Big|\sum_{i\in\mathcal{X}} I(S_i \geq c) \neq 0\Big]P\Big(\sum_{i\in\mathcal{X}} I(S_i \geq c) \neq 0\Big)$$
$$= \frac{m_0(1-F_0(c))}{m_0(1-F_0(c)) + m_1(1-F_1(c))} + O\left(\frac{1}{m}\right).$$

If $m_0 = 0$, both mFDR and FDR are zero and the lemma is trivial.

Similarly, we can show that $V(FDP) = O(m^{-1})$. The variance behaviour provides some confidence that controlling for FDR is almost as effective as controlling for FDP itself as the number of tests becomes large because FDP will be concentrated around FDR with diminishing variance. Lastly, similar arguments also lead to the results for FNP. □

Given the results in Lemma B.1, it is easy to see that the Oracle method controls FDR asymptotically. Moreover, there does not exist another λ^{**} such that the decision rule $\delta_i = I\{S_i \geq c(\lambda^{**})\}$ controls FDR asymptotically and achieves a lower *FNR* than the *FNR* achieved by λ^*, or $FNR(\lambda^*)$ in the limit. Suppose such an λ^{**} exists. Because $c(\lambda)$ is increasing in λ, $mFDR(c)$ is decreasing in c and $mFNR(c)$ is increasing in c, we only need to consider the case where $\lambda^{**} = \lambda^* - \epsilon$, where a positive sequence of ϵ converges to 0 as $m \to \infty$. Decision rule with such an λ^{**} controls nominal size asymptotically. Further, because both mFDR and mFNR are continuous in λ, $mFDR(\lambda^{**}) = \alpha + \delta_1$ and $mFNR(\lambda) = mFNR(\lambda^*) - \delta_2$ where both δ_1 and δ_2 are positive and converges to 0 as $m \to \infty$. Then we have

$$\begin{aligned}
&\lim_{m\to\infty} |FNR(\lambda) - FNR(\lambda^*)| \\
&\quad = \lim_{m\to\infty} |FNR(\lambda) - mFNR(\lambda) - (FNR(\lambda^*) - mFNR(\lambda^*) - \delta_2)| \\
&\quad \leq \lim_{m\to\infty} |FNR(\lambda) - mFNR(\lambda)| + \lim_{m\to\infty} |FNR(\lambda^*) - mFNR(\lambda^*)| + \lim_{m\to\infty} |\delta_2| \\
&\quad = 0.
\end{aligned}$$

Hence, asymptotically choosing λ^* for the decision rule minimizes FNR while controlling FDR asymptotically. □

Proof of Proposition 2.2: We observe that the optimal procedure based on p-values can be found using the arguments in Proposition 2.1 except that the $Z(s)$ transformation, instead of being the LFDR statistics, now takes the form $Z(s) = 1 - F_{v=0}(s)$. Notice that this new transformation $Z(s)$ is also monotonically decreasing in s. The proposition is then proved by setting $c(\lambda) = F^{-1}_{v=0}(1-u)$ in Proposition 2.1. This provides a direct link between the p-value approach and the LFDR approach. $u^* = 1 - F_{v=0}(c(\lambda^*))$, where λ^* is the optimal penalty for Type I error found in Proposition 2.1. □

Proof of Theorem 2.1: Because the Oracle procedure rejects the ith hypothesis if $q_i \leq \alpha$, it suffices to prove that $\hat{q}_i \to q_i$ uniformly for all i with probability one, where

$$q_i = \frac{\int_{-\infty}^{0} 1 - F_v(F^{-1}_{v=0}(1-p_i))dG(v)}{\int_{-\infty}^{+\infty} 1 - F_v(F^{-1}_{v=0}(1-p_i))dG(v)}.$$

Given the definition of the p-value p_i for each s_i, we can rewrite the above quantity as

$$q_i = \frac{\int_{-\infty}^{0} 1 - F(s_i|v)dG(v)}{\int_{-\infty}^{+\infty} 1 - F(s_i|v)dG(v)},$$

and likewise

$$\hat{q}_i = \frac{\int_{-\infty}^{0} 1 - F(s_i|v)d\hat{G}(v)}{\int_{-\infty}^{+\infty} 1 - F(s_i|v)d\hat{G}(v)},$$

where $F(s_i|v)$ is the distribution function for S_i given v. To prove consistency of $\hat{q}_i$ uniformly in i, it suffices to prove uniform consistency for the numerator and the denominator separately, and then to invoke the continuous mapping theorem.

For the denominator, we have

$$\sup_s \left| \int_{-\infty}^{+\infty} 1 - F(s|v)d\hat{G}(v) - \int_{-\infty}^{+\infty} 1 - F(s|v)dG(v) \right| = \sup_s \left| \hat{F}(s) - F(s) \right| \to 0$$

with probability one, where $\hat{F}(s)$ is the induced marginal distribution for S given $\hat{G}$ and $F(s)$ is the true marginal distribution for S. The convergence of $\hat{F}(s)$ in sup-distance is via consistency of the mixture density estimator in terms of Hellinger distance (cf. Example 4.2.4 of van de Geer, 2000), and the fact that $F(s)$ is continuous (see Theorem 2.1 of DasGupta, 2008).

For the numerator, without loss of generality, assume there is a discontinuity point at k in the interval $(-\infty, 0)$. We need to show

$$\sup_s \left| \int_{-\infty}^{0} 1 - F(s|v)d\hat{G}(v) - \int_{-\infty}^{0} 1 - F(s|v)dG(v) \right| \to 0.$$

Denote $H(s|v) = 1 - F(s|v)$, For some arbitrary $\epsilon > 0$, rewrite the quantity in the absolute sign as

$$\int_{-\infty}^{k-\epsilon} H(s|v)d(\hat{G}(v) - G(v)) + \int_{k+\epsilon}^{0} H(s|v)d(\hat{G}(v) - G(v)) + \int_{k-\epsilon}^{k+\epsilon} H(s|v)d(\hat{G}(v) - G(v)).$$

The first two terms converge to 0 uniformly over s because on the interval $(-\infty, k)$ there is no discontinuity point, and point-wise convergence of $\hat{G}$ implies uniform convergence on all intervals $(-\infty, k - \epsilon]$ for all $\epsilon > 0$. Hence, using integration by parts, we have for the first term

$$\begin{aligned}
&\sup_s \left| \int_{-\infty}^{k-\epsilon} H(s|v)d(\hat{G}(v) - G(v)) \right| \\
&\quad \leq \sup_s H(s|k-\epsilon)|\hat{G}(k-\epsilon) - G(k-\epsilon)| + \lim_{v\to-\infty} \sup_s H(s|v)|\hat{G}(v) - G(v)| \\
&\qquad + \sup_{v\in(-\infty,k-\epsilon]} |\hat{G}(v) - G(v)| \sup_s \int_{-\infty}^{k-\epsilon} dH(s|v) \to 0.
\end{aligned}$$

A similar argument holds for the second term.

For the third term, as ϵ can be arbitrarily small, we have

$$\begin{aligned}
&\sup_s \left| \int_{k-\epsilon}^{k+\epsilon} H(s|v)d(\hat{G}(v) - G(v)) \right| \\
&\quad = \sup_s \left| H(s|k) \left((\hat{G}(k+\epsilon) - G(k+\epsilon)) - (\hat{G}(k-\epsilon) - G(k-\epsilon)) \right) \right| \\
&\qquad + \sup_s \left| \int_{k-\epsilon}^{k+\epsilon} H(s|u) - H(s|k)d(\hat{G}(u) - G(u)) \right| \\
&\quad \leq \sup_s H(s|k) \left(|\hat{G}(k+\epsilon) - G(k+\epsilon)| + |\hat{G}(k-\epsilon) - G(k-\epsilon)| \right)
\end{aligned}$$

$$+\sup_s \int_{k-\epsilon}^{k+\epsilon} |H(s|u) - H(s|k)| d\hat{G}(u) + \sup_s \int_{k-\epsilon}^{k+\epsilon} |H(s|u) - H(s|k)| dG(u)$$

$$\leq |\hat{G}(k+\epsilon) - G(k+\epsilon)| + |\hat{G}(k-\epsilon) - G(k-\epsilon)| + 2\sup_s \sup_{|u-k|\leq\epsilon} |H(s|u) - H(s|k)| \to 0.$$

The last inequality is because $\sup_s H(s|k) \leq 1$ and $\int_{k-\epsilon}^{k+\epsilon} d\hat{G}(u) \leq 1$ and $\int_{k-\epsilon}^{k+\epsilon} dG(u) \leq 1$. The last convergence is because of the uniform convergence of $\hat{G}$ on continuous point $k-\epsilon$ and $k+\epsilon$ and $\sup_s \sup_{|u-k|\leq\epsilon} |H(s|u) - H(s|k)| \to 0$ as long as $F(s|v)$ is uniformly continuous in both arguments.

This establishes that $\hat{q}_i \to q_i$ uniformly for all i with probability one. The mFDR of the decision rule $\hat{\delta}_i = I\{\hat{q}_i \leq \alpha\}$ is $(1-p)\mathbb{P}_{H_0}(\hat{\delta}_i = 1)/\mathbb{P}(\hat{\delta}_i = 1)$, which converges to $(1-p)\mathbb{P}_{H_0}(\delta_i = 1)/\mathbb{P}(\delta_i = 1) = \alpha$. Additionally, the associated mFNR of decision rule $\hat{\delta}_i$ is $p\mathbb{P}_{H_1}(\hat{\delta}_i = 0)/\mathbb{P}(\hat{\delta}_i = 0)$ which converges to $p\mathbb{P}_{H_1}(\delta_i = 0)/\mathbb{P}(\delta_i = 0)$, the mFNR for the Oracle procedure. □

Proof of Theorem 2.2: In the proof, we denote $\hat{w}$ and $\hat{H}$ defined in the theorem by $\hat{w}_m$ and $\hat{H}_m$ to emphasize the feature that they are sequences of m, the number of hypotheses tested. It suffices to show that $\hat{H}_m$ is a consistent estimator of H. First, let us define $\tilde{H} \equiv \underset{H\in\mathcal{H}}{\text{argmax}}(1/m)\sum_{i=1}^m \log\{(1-w)\varphi(s_i \mid 0) + w\int \varphi(s_i \mid v)dH(v)\}$, which corresponds to the NPMLE if we were to know the true w. It is easy to see that $\tilde{H}$ is a consistent estimator of H via the usual argument for consistency of NPMLE; see, e.g. Pfanzagl (1988). The idea of the proof is to show that provided $\hat{w}_m \xrightarrow{p} w$, the criteria function that defines $\tilde{H}$ is close to the criteria function that defines $\hat{H}$ uniformly over $H \in \mathcal{H}$, which further leads to the strong consistency of $\hat{H}$, and hence $\hat{H}$ can be used in replacement of $\tilde{H}$ in the adaptive procedure.

Denote $\ell(s_i, w, H) \equiv (1-w)\varphi(s_i \mid 0) + w\int \varphi(s_i \mid v)dH(v)$. First, we consider the end points of w (i.e. $w = 0$ and $w = 1$). When $w = 0$,

$$\sup_{H\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^m \log \ell(s_i, \hat{w}_m, H) - \frac{1}{m}\sum_{i=1}^m \log \ell(s_i, w, H) \right|$$

$$= \sup_{H\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^m \log \frac{\varphi(s_i \mid 0)}{(1-\hat{w}_m)\varphi(s_i \mid 0) + \hat{w}_m \int \varphi(s_i \mid v)dH(v)} \right|$$

$$\leq \frac{1}{m}\sum_{i=1}^m \log \sup_{H\in\mathcal{H}} \left| \frac{\varphi(s_i \mid 0)}{(1-\hat{w}_m)\varphi(s_i \mid 0) + \hat{w}_m \int \varphi(s_i \mid v)dH(v)} \right|$$

$$\leq \frac{1}{m}\sum_{i=1}^m \log \sup_{H\in\mathcal{H}} \left| \frac{\varphi(s_i \mid 0)}{(1-\hat{w}_m)\varphi(s_i \mid 0)} \right|$$

$$= \log \left| \frac{1}{1-\hat{w}_m} \right| \xrightarrow{p} 0.$$

When $w = 1$,

$$\sup_{H\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^m \log \ell(s_i, \hat{w}_m, H) - \frac{1}{m}\sum_{i=1}^m \log \ell(s_i, w, H) \right|$$

$$= \sup_{H\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^m \log \frac{\int \varphi(s_i \mid v)dH(v)}{(1-\hat{w}_m)\varphi(s_i \mid 0) + \hat{w}_m \int \varphi(s_i \mid v)dH(v)} \right|$$

$$\leq \frac{1}{m}\sum_{i=1}^m \log \sup_{H\in\mathcal{H}} \left| \frac{\int \varphi(s_i \mid v)dH(v)}{(1-\hat{w}_m)\varphi(s_i \mid 0) + \hat{w}_m \int \varphi(s_i \mid v)dH(v)} \right|$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\log\sup_{H\in\mathcal{H}}\left|\frac{\int\varphi(s_i\mid v)dH(v)}{\hat{w}_m\int\varphi(s_i\mid v)dH(v)}\right|$$

$$=\log\left|\frac{1}{\hat{w}_m}\right|\xrightarrow{p}0.$$

For any $w\in(0,1)$, via Taylor expansion of $(1/m)\sum_{i=1}^{m}\log\ell(s_i,\hat{w}_m,H)$ around $(1/m)\sum_{i=1}^{m}\log\ell(s_i,w,H)$, we have, for $\tilde{w}_i=\alpha_i\hat{w}_m+(1-\alpha_i)w$,

$$\frac{1}{m}\sum_{i=1}^{m}\log\ell(s_i,\hat{w}_m,H)$$

$$=\frac{1}{m}\sum_{i=1}^{m}\log\left(\ell(s_i,w,H)+(w-\hat{w}_m)\varphi(s_i\mid 0)+(\hat{w}_m-w)\int\varphi(s_i\mid v)dH(v)\right)$$

$$=\frac{1}{m}\sum_{i=1}^{m}\log(\ell(s_i,w,H))+\frac{(w-\hat{w}_m)\varphi(s_i\mid 0)+(\hat{w}_m-w)\int\varphi(s_i\mid v)dH(v)}{(1-\tilde{w}_i)\varphi(s_i\mid 0)+\tilde{w}_i\int\varphi(s_i\mid v)dH(v)}.$$

Then we have

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\log\ell(s_i,\hat{w}_m,H)-\frac{1}{m}\sum_{i=1}^{m}\log\ell(s_i,w,H)\right|$$

$$=\sup_{H\in\mathcal{H}}\left|(\hat{w}_m-w)\frac{1}{m}\sum_{i=1}^{m}\frac{\varphi(s_i\mid 0)-\int\varphi(s_i\mid v)dH(v)}{(1-\tilde{w}_i)\varphi(s_i\mid 0)+\tilde{w}_i\int\varphi(s_i\mid v)dH(v)}\right|$$

$$\leq\frac{|\hat{w}_m-w|}{m}\sum_{i=1}^{m}\left(\sup_{H\in\mathcal{H}}\left|\frac{\varphi(s_i\mid 0)}{(1-\tilde{w}_i)\varphi(s_i\mid 0)+\tilde{w}_i\int\varphi(s_i\mid v)dH(v)}\right|\right)$$

$$+\frac{|\hat{w}_m-w|}{m}\sum_{i=1}^{m}\left(\sup_{H\in\mathcal{H}}\left|\frac{\int\varphi(s_i\mid v)dH(v)}{(1-\tilde{w}_i)\varphi(s_i\mid 0)+\tilde{w}_i\int\varphi(s_i\mid v)dH(v)}\right|\right)$$

$$\leq|\hat{w}_m-w|\left(\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\varphi(s_i\mid 0)}{(1-\tilde{w}_i)\varphi(s_i\mid 0)}\right|+\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\int\varphi(s_i\mid v)dH(v)}{\tilde{w}_i\int\varphi(s_i\mid v)dH(v)}\right|\right)$$

$$\leq|\hat{w}_m-w|\frac{1}{m}\sum_{i=1}^{m}\left(\left|\frac{1}{1-\tilde{w}_i}\right|+\left|\frac{1}{\tilde{w}_i}\right|\right)$$

$$\leq|\hat{w}_m-w|\left|\frac{1}{(w\vee\hat{w}_m)(1-(w\wedge\hat{w}_m))}\right|\xrightarrow{p}0,$$

with $(a\vee b)\equiv min(a,b)$ and $(a\wedge b)\equiv max(a,b)$. The last inequality is because w_i lies in between $\hat{w}_m$ and w for all $i=1,2,\ldots,m$. The last convergence result is due to the fact that if $\hat{w}_m\xrightarrow{p}w$ as $m\to\infty$ and w is bounded away from zero and one.

When w is known, the following uniform law of large numbers (ULLN) result holds

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\log\ell(s_i,w,H)-E[\log\ell(s_1,w,H)]\right|\xrightarrow{p}0.$$

See detailed discussion in van de Geer (1993) on sufficient conditions for the ULLN to hold (Theorem 2.4) and the satisfaction of these sufficient conditions in the mixture model we consider here (Lemma 5.1). Combining with the results derived above, we have

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\log\ell(s_i,\hat{w}_m,H)-E[\log\ell(s_1,w,H)]\right|\xrightarrow{p}0.$$

The above convergence in probability result is equivalent to (cf. Theorem 6.3.1 of Resnick, 1998) the statement that for every subsequence m_j, there has a further subsequence m_{j_ℓ} for which

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{m_{j_\ell}}\sum_{i=1}^{m_{j_\ell}}\log\ell(s_i,\hat{w}_{m_{j_\ell}},H)-E[\log\ell(s_1,w,H)]\right|\xrightarrow{a.s.}0.$$

Invoking Lemma 1.1 in van de Geer (1993), we have Hellinger convergence of the mixture density estimator $\ell(\cdot,\hat{w}_{m_{j_\ell}},\hat{H}_{m_{j_\ell}})$ to $\ell(\cdot,w,H)$ a.s., where $\hat{H}_{m_{j_\ell}}$ is the subsequence of hybrid-NPMLE estimators that minimizes $(1/m_{j_\ell})\sum_{i=1}^{m_{j_\ell}}\log\ell(s_i,\hat{w}_{m_{j_\ell}},H)$ among $H\in\mathcal{H}$. Lemma 5.2 in van de Geer (1993) then implies $\hat{H}_{m_{j_\ell}}$ as defined in Lemma 2.2 is strongly consistent, provided H is identifiable; see also Example 4.2.4 in van de Geer (2000). Then, for the subsubsequence m_{j_ℓ}, we can use the same argument in the proof of Theorem 2.1 to show that $\hat{q}_i\to q$ uniformly in i with probability one. Because the subsequence m_j is arbitrarily picked, this in turn gives that $\hat{q}_i\xrightarrow{p}q$ uniformly in i. Using a similar argument as at the end of the proof for Theorem 2.1 then implies that mFDR based on decision rule $I\{\hat{q}_i\le\alpha\}$ converges to α while the mFNR based on the adaptive rule converges to the mFNR achieved by the Oracle. □

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication files