Is it possible to construct an asymptotically efficient estimator of the proportion of true null hypotheses in a multiple testing setup?

by

Van Hanh Nguyen and Catherine Matias

Research Report No. 36
December, 2011

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Évry, France
http://www.ssbgroup.fr/

Is it possible to construct an asymptotically efficient estimator of the proportion of true null hypotheses in a multiple testing setup?

Van Hanh Nguyen[*,a,b], Catherine Matias[b]

[a] *Laboratoire de Mathématiques d'Orsay, Université Paris Sud, UMR CNRS 8628, Bâtiment 425, 91 405 Orsay Cedex, FRANCE*

[b] *Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071- USC INRA, 23 bvd de France, 91 037 Évry, FRANCE*

Abstract

One important problem in the multiple testing context is the estimation of the proportion θ of true null hypotheses. This proportion appears in a semiparametric mixture model with two components: a uniform distribution on the interval $[0, 1]$ and a nonparametric density f. A large number of estimators of this proportion exist under different identifiability assumptions but their rate of convergence or asymptotic efficiency has only been partly studied. We shall focus here on two different categories of identifiability assumptions previously introduced in the literature: in the first case, f vanishes on a set with positive Lebesgue measure (and a subcase is obtained when this set is an interval) and in the second case, the set of points where f vanishes has a null Lebesgue measure. We first improve the consistency results of the estimator proposed by Celisse and Robin (2010), by establishing its almost sure convergence as well as $\sqrt{n}$-consistency, under the assumption that f vanishes on an interval. To our knowledge, this is the first result proving that the parametric rate of convergence may be achieved by a consistent estimator of the proportion θ in this semiparametric setup. We also compute a lower bound on the local asymptotic minimax (LAM) quadratic risk of any estimator under the first case. Then, we shall discuss the existence of asymptotically efficient estimators of the proportion θ in the sense of a convolution theorem. In the first case, we conjecture that no $\sqrt{n}$-consistent estimator is efficient. In the second case, we prove that the efficient information matrix for estimating θ is zero. Hence in this case, the LAM quadratic risk is not finite and there is no regular estimator of the proportion θ.

Key words: Asymptotic efficiency; efficient score, false discovery rate; information bound; multiple testing; p-values; semiparametric model.
2008 MSC: 62G20, 62G10

[*] Corresponding author
Email addresses: nvanhanh@genopole.cnrs.fr (Van Hanh Nguyen), catherine.matias@genopole.cnrs.fr (Catherine Matias)

1. Introduction

The problem of estimating the proportion θ of true null hypotheses is of interest in situation where several thousands of (independent) hypotheses can be tested simultaneously. One of the typical applications in which multiple testing problems occur is estimating the proportion of genes that are not differentially expressed in deoxyribonucleic acid (DNA) microarray experiments (see for instance Dudoit and van der Laan, 2008). Among other application domains, we mention astrophysics (Meinshausen and Rice, 2006) or neuroimaging (Turkheimer et al., 2001). A reliable estimate of θ is important when one wants to control multiple error rates, such as the false discovery rate (FDR) introduced by Benjamini and Hochberg (1995). In this work, we shall discuss asymptotic properties of estimators of the true proportion of null hypotheses. We stress that the asymptotic framework is particularly relevant in the above mentioned contexts where the number of tested hypotheses is huge.

In many recent papers (such as Broberg, 2005; Celisse and Robin, 2010; Efron, 2004; Efron et al., 2001; Genovese and Wasserman, 2004, etc), a two-component mixture density is used to model the behavior of p-values $X_1, X_2, \dots, X_n$ associated with n independent tested hypotheses. More precisely, assume the test statistics are independent and identically distributed (iid) with a continuous distribution under the corresponding null hypotheses, then the p-values $X_1, X_2, \dots, X_n$ are iid and follow the uniform distribution $\mathcal{U}([0,1])$ on interval $[0,1]$ under the null hypotheses. The density g of p-values is modeled by a two-component mixture with following expression

$$\forall x \in [0,1], \quad g(x) = \theta + (1-\theta) f(x), \tag{1}$$

where $\theta \in [0,1]$ is the unknown proportion of true null hypotheses and f denotes the density function of p-values generated under the alternative (false null hypotheses).

In order to obtain the parameter's identifiability in model (1), we need to make some assumptions on density f. First, it can be seen that the parameters of the model are identifiable if and only if the infimum of density f on $[0,1]$ is zero. A stronger and sufficient condition for identifiability is the existence of $x_0 \in [0,1]$ with $f(x_0) = 0$. However, from an estimation perspective, these assumptions are too weak and it is hopeless to obtain a reliable estimate of θ without additional assumptions. Note that many works focusing on FDR estimates (instead of θ estimates) do not assume identifiability of the parameters in model (1) and identify thus only the quantity $\bar{\theta} := \theta + (1-\theta)\inf_{x\in[0,1]} f(x)$. As we are going to take an estimation perspective on θ, we need to consider only identifiable models. In the following, we shall classify the setups in two different categories: models assuming that f vanishes on a set of points with positive measure and models where this set of points may have zero measure, and where regularity or monotonicity assumptions are added on f.

One of the most common assumptions made on density f in the literature is that it vanishes on an interval of the form $[\lambda, 1]$. This assumption underlies many estimation procedures further discussed below, such as variants of Efron's estimator. However, the fact that on real data, density f may not be vanishing in the neighborhood of $x = 1$ has previously been noticed in the literature (see Pounds and Cheng, 2006, and the references therein). As a consequence, authors such as Celisse and Robin (2010) propose a milder identifiability assumption that f vanishes on an interval $[\lambda^\star, \mu^\star]$ with $\mu^\star \leq 1$. This assumption is more general than the previous one and contains $\mu^\star = 1$ as a particular case. Nonetheless, we shall remark that when $\mu^\star < 1$, such an assumption makes little sense in the specific context of modeling the distribution of p-values. Indeed, the high values of g (or f) near $x = 1$ that may be observed come from a misspecification of the distribution of the test statistic under the null or the alternative hypotheses. For instance, the test statistics may in fact be discrete, or the performed test is one-sided while some statistics follow the *untested alternative* (Pounds and Cheng, 2006). In this case, rather than generalizing assumptions on the semiparametric mixture model, one should reconsider the use of this mixture model itself (in the context of modeling the distribution of p-values). Indeed, it is then inappropriate to assume that the p-values corresponding to the null hypotheses follow a uniform distribution.

It is important to note that assuming f vanishes on a subset with positive measure is a rather strong assumption. Indeed, it implies that the density of the underlying statistic T under the alternative has bounded support. More precisely, if the rejection region has the form $\{T \geq c\}$, then the density of T under the alternative should have support included in $(-\infty, a]$ for some $a \in \mathbb{R}$.

To avoid such constraint, some authors proposed to assume that the density f is low for x near 1, but not necessarily vanishing on an interval near $x = 1$. In order to obtain an identifiable model, they further assume either that f is decreasing, with $f(1) = 0$ such as in the work of Langaas et al. (2005); or that f is regular near $x = 1$, such as in the work of Neuvial (2010).

Let us now discuss the different estimators proposed in the literature, starting with those assuming (implicitly or not) that f vanishes on a whole interval. First, Schweder and Spjtvoll (1982) suggested a procedure to estimate θ, that has been later used by Storey (2002). This procedure depends on an unspecified parameter $\lambda \in [0, 1)$ and is equal to the proportion of p-values larger than this threshold λ divided by $1 - \lambda$. It is thus consistent only if f vanishes on the interval $[\lambda, 1]$ (an assumption not made in the paper by Schweder and Spjtvoll (1982) nor the one by Storey (2002)). Note that even if such an assumption were made, it would not solve the problem of choosing λ such that f vanishes on $[\lambda, 1]$. Adapting this procedure in order to end up with an estimate of the positive FDR (pFDR), Storey (2002) proposes a bootstrap strategy to pick λ. More precisely, his procedure minimizes the mean-squared error for estimating the pFDR. Note that Genovese and Wasserman (2004) established that, for fixed value λ such that the cumulative distribution function (cdf) G of g satisfies $G(\lambda) < 1$, Storey's estimator converges at parametric rate and is asymptotically normal, but is also asymptotically biased (thus it does not converge to θ at parametric rate). Some other choices of λ are, for instance, based on break point estimation (Turkheimer et al., 2001) or spline smoothing (Storey and Tibshirani, 2003). Recently, Celisse and Robin (2010) proposed an estimator of θ, relying on a histogram estimate of g, under the assumption that f vanishes on an interval $[\lambda^\star, \mu^\star]$. This estimator is shown to converge in probability to θ. We want to stress here an important difference between Storey's and Celisse and Robin's estimators. Indeed, both are constructed using nonparametric estimates $\hat{g}$ of the density g and then estimate θ relying on the value of $\hat{g}$ on a specific interval. However, contrarily to Storey's procedure, the one proposed by Celisse and Robin automatically selects an interval where g is identically equal to θ.

Other estimators of θ are based on regularity or monotonicity assumptions made on f or equivalently on g, combined with the assumption that the infimum of g is attained at $x = 1$ (thus we have $\theta = g(1)$). These estimators rely on nonparametric estimates of g and appear to inherit nonparametric rates of convergence. Langaas et al. (2005) derive estimators based on nonparametric maximum likelihood estimation of the p-value density, in two setups: decreasing and convex decreasing densities f. We mention that no theoretical properties of these estimators are given. Hengartner and Stark (1995) propose a very general finite sample confidence envelope for a monotone density. Relying on this result and assuming moreover that cdf G is concave and that g is Lipschitz in a neighborhood of $x = 1$, Genovese and Wasserman (2004) construct an estimator converging to $g(1) = \theta$ at rate $(\log n)^{1/3}n^{-1/3}$. Under some regularity assumptions on f near $x = 1$, Neuvial (2010) established that by letting $\lambda \to 1$, Storey's estimator may be turned into a consistent estimator of θ, with a nonparametric rate of convergence equal to $n^{-k/(2k+1)}\eta_n$, where $\eta_n \to +\infty$ and k controls the regularity of f near $x = 1$.

We mention that Meinshausen and Bühlmann (2005) discuss probabilistic upper bounds for the proportion of true null hypotheses, which are valid under general and unknown dependence structures between the test statistics. We also mention that, in a very general and nonparametric setup, Swanepoel (1999) proposes a two-step estimator of the minimum of an unknown density h based on the distribution of the spacings between observations. Assuming that in some neighborhood of the value at which the density h achieves its minimum, h has a bounded second derivative h'' satisfying a Lipschitz condition, Swanepoel establishes that for any $\delta > 0$, there exists an estimator of this infimum converging at rate $(\log n)^{\delta}n^{-2/5}$ to the true minimum. However, this procedure assumes that the minimum is achieved in the interior of the support of h. As discussed above, such an assumption is not realistic in the context of estimation of the true proportion of null hypotheses, unless the minimum is achieved in a whole interval of the form $[\lambda, 1]$.

Finally, note that we do not discuss here estimators of the proportion of non null effects in Gaussian mixtures such as in Cai and Jin (2010); Jin (2008); Jin and Cai (2007), a related but although different problem as the one we study.

Despite a large number of different estimation procedures of the proportion of true null hypotheses proposed in the literature, very few results on their convergence properties exist. In particular, none of the above mentioned estimators is shown to achieve a parametric rate of convergence towards θ. Thus, natural open questions are whether it is possible to construct an estimator converging at parametric rate; whether there exists an *asymptotically efficient* estimator and what is the value of an *optimal* variance. Here, asymptotic efficiency stands in the sense of a convolution theorem (see van der Vaart, 1998, Chapter 25, for more details on efficiency theory for semiparametric models). An estimator is said to be *regular* if it converges in distribution at parametric rate. According to a convolution theorem (see Theorem 25.20 in van der Vaart, 1998), regular estimators converge in distribution to the convolution of a Gaussian random variable with some (minimal) variance and another random variable. Thus, an estimator is asymptotically efficient if among regular estimators, it achieves the *best* limit -namely the Gaussian distribution with minimal variance. In fact in our context, at least two different cases occur, whether density f is assumed to vanish on a set with either positive or null Lebesgue measure. Under the minimal identifiability assumption, model (1) can be written in the following form,

$$\mathcal{P} = \big\{p_{\theta,f} : [0,1] \to \mathbb{R}^+;\ p_{\theta,f} := \theta + (1-\theta)f,\ \theta \in (0,1),\\ f : [0,1] \to \mathbb{R}^+ \text{ density with } \inf_{[0,1]} f = 0\big\}.$$

In this semiparametric setup, we aim at estimating the finite-dimensional parameter θ and consider f as a nuisance parameter. The main aim of our work is improving the convergence properties of the estimator proposed by Celisse and Robin (2010) and discussing the existence of an asymptotically efficient estimator of θ under the two categories of identifiability assumptions mentioned above. We recall that Celisse and Robin (2010) established only convergence in probability of their estimator. We shall first see that parametric rate of convergence is achieved by this estimator, when assuming that f vanishes on an interval.

Concerning the study of the semiparametric model, we shall consider two semiparametric submodels of $\mathcal{P}$, respectively denoted by $\mathcal{P}_\delta$ and $\mathcal{P}_0$ and defined in the following sections. Model $\mathcal{P}_\delta$ assumes that the subset of points where f is zero has a positive measure whereas in model $\mathcal{P}_0$, it has a null measure. We compute a lower bound for the local asymptotic minimax (LAM) quadratic risk for estimating θ in model $\mathcal{P}_\delta$ and conjecture that no $\sqrt{n}$-consistent estimator is efficient in this model. Then, we prove that the efficient information matrix for estimating θ in model $\mathcal{P}_0$ is zero. This implies both that the LAM quadratic risk at parametric rate is not finite and that if there exists a $\sqrt{n}$-consistent estimator in this model, it can not have finite asymptotic variance. We stress that this does not necessarily imply that $\sqrt{n}$-consistent estimators do not exist in model $\mathcal{P}_0$. However, the only rates of convergence obtained until now in this case are nonparametric ones.

The paper is organized as follows. In Section 2.1, we establish the consistency properties (almost sure convergence and $\sqrt{n}$-consistency) of the estimator proposed in Celisse and Robin (2010). In Section 2.2, we calculate a lower bound of the LAM quadratic risk for estimating θ in model $\mathcal{P}_\delta$. We also present conditions under which an estimator is asymptotically efficient in the sense of a convolution theorem, in the case where the density f is null on a set with positive measure. Then, we more generally discuss the existence of asymptotically efficient estimators of θ in Section 3. More precisely, we start by recalling in Section 3.1 a general method of construction of efficient estimators relying on first step $\sqrt{n}$-consistent ones. Sections 3.2 and 3.3 further discuss the implications of semiparametric theory on the existence of efficient estimators of θ in models $\mathcal{P}_\delta$ and $\mathcal{P}_0$, respectively. Some technical proofs have been postponed to Appendix A.

2. Estimation of the proportion θ of true null hypotheses

2.1. A $\sqrt{n}$-consistent estimator of θ

In this section, we shall introduce the estimator of the parameter θ proposed by Celisse and Robin (2010). In particular, we specify some assumptions that establish its almost sure convergence and $\sqrt{n}$-consistency. Throughout this section, we shall always assume that the density f belongs to $\mathbb{L}^2([0,1])$.

Assumption 1. *Density f is null on an interval $[\lambda^\star, \mu^\star] \subset (0,1]$ (with unknown values $\lambda^\star$ and $\mu^\star$) and f is monotone outside the interval $[\lambda^\star, \mu^\star]$.*

For example, f is decreasing on $[0, \lambda^\star]$ and increasing on $[\mu^\star, 1]$. This assumption is stronger than (Assumption A' in Celisse and Robin, 2010), the latter not being sufficient to establish their result. The monotonicity part of our assumption is not necessary and we shall explain what is exactly required and how we use the previous assumption in the proof of Lemma 3. Under Assumption 1, the true parameter θ is equal to $g(x)$ for all x in $[\lambda^\star, \mu^\star]$. Note that the case where we impose $\mu^\star = 1$ is included in this setting. The general idea underlying the estimator's construction is the following one. Let us consider a nonparametric density estimator of g. For example, let $\hat{g}_I$ be a histogram estimator corresponding to a partition $I = (I_k)_{1,\dots,D}$ of $[0,1]$, defined by

$$\hat{g}_I(x) = \sum_{k=1}^{D} \frac{n_k}{n|I_k|} \mathbf{1}_{I_k}(x),$$

where $n_k = \text{card}\{i : X_i \in I_k\}$ is the number of observations in I_k, $\mathbf{1}_{I_k}$ is the indicator function of I_k and $|I_k|$ is the width of interval I_k. We estimate θ by the value of $\hat{g}_I$ on an interval $\hat{I}$ which is the closest as possible to the interval $[\lambda^\star, \mu^\star]$.

Let us now recall more precisely the procedure for estimating θ that is presented in Celisse and Robin (2010). For a given integer N, define $\mathcal{I}_N$ as the set of partitions of $[0,1]$ such that for some integers k, l with $2 \le k+2 \le l \le N$, the first k intervals and the last $N-l$ ones are regular of width $1/N$, namely

$$\mathcal{I}_N = \Big\{ I = (I_i)_i : \forall i \neq k+1, |I_i| = \frac{1}{N},\ |I_{k+1}| = \frac{l-k}{N},\ 2 \le k+2 \le l \le N \Big\}.$$

Then for two given integers $m_{min} < m_{max}$, denote by $\mathcal{I}$ the collection of considered partitions, defined by

$$\mathcal{I} = \bigcup_{m_{min} \le m \le m_{max}} \mathcal{I}_{2^m}. \tag{2}$$

Every partition I in $\mathcal{I}$ is characterized by a triplet $(N = 2^m, \lambda = k/N, \mu = l/N)$ and the quality of the histogram estimator $\hat{g}_I$ is measured by its quadratic risk. So in this sense, the *oracle estimator* $\hat{g}_{\hat{I}}$ is obtained through

$$\hat{I} = \underset{I \in \mathcal{I}}{\text{argmin}}\ \mathbb{E}[||g - \hat{g}_I||_2^2] = \underset{I \in \mathcal{I}}{\text{argmin}}\ R(I),\ \text{where } R(I) = \mathbb{E}\Big[||\hat{g}_I||_2^2 - 2\int_0^1 \hat{g}_I(x) g(x) dx\Big].$$

However, for every partition I, the quantity $R(I)$ depends on g which is unknown. Thus $\hat{I}$ is an oracle and not an estimator. It is then natural to replace $R(I)$ by an estimator. In Celisse and Robin (2008, 2010), the authors use leave-p-out (LPO) estimator of $R(I)$ with $p \in \{1, \dots, n-1\}$, whose expression is given by (see Celisse and Robin, 2008, Theorem 2.1)

$$\hat{R}_p(I) = \frac{2n-p}{(n-1)(n-p)} \sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} \Big(\frac{n_k}{n}\Big)^2. \tag{3}$$

The best theoretical value of p is the one that minimizes the mean-square error (MSE) of $\hat{R}_p(I)$, namely

$$p^\star(I) = \underset{p \in \{1,\dots,n-1\}}{\text{argmin}}\ MSE(p, I) = \underset{p \in \{1,\dots,n-1\}}{\text{argmin}}\ \mathbb{E}\Big[\big(\hat{R}_p(I) - R(I)\big)^2\Big].$$

It clearly appears that $MSE(p, I)$ has the form of a function $\Phi(p, I, \alpha)$ (see Celisse and Robin, 2008, Proposition 2.1) depending on the unknown vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_D)$ with $\alpha_k = \mathbb{P}(X_1 \in I_k)$. A natural idea would be to replace the α_ks in $\Phi(p, I, \alpha)$ by their empirical counterparts $\hat{\alpha}_k = n_k/n$ and an estimator of $p^\star(I)$ is therefore given by

$$\hat{p}(I) = \underset{p \in \{1,\dots,n-1\}}{\text{argmin}} \widehat{MSE}(p, I) = \underset{p \in \{1,\dots,n-1\}}{\text{argmin}} \Phi(p, I, \hat{\alpha}).$$

The exact calculation of $\hat{p}(I)$ may be found in (Celisse and Robin, 2008, Theorem 3.1). Hence, the procedure for estimating θ is the following one

1. For each partition $I \in \mathcal{I}$, define $\hat{p}(I) = \underset{p \in \{1,\dots,n-1\}}{\text{argmin}} \widehat{MSE}(p, I)$,
2. Choose $\hat{I} = (\hat{N}, \hat{\lambda}, \hat{\mu}) \in \underset{I \in \mathcal{I}}{\text{argmin}} \hat{R}_{\hat{p}(I)}(I)$ such that the width of the interval $[\hat{\lambda}, \hat{\mu}]$ is maximum,
3. Estimate θ by $\hat{\theta}_n = \text{card}\{i : X_i \in [\hat{\lambda}, \hat{\mu}]\}/[n(\hat{\mu} - \hat{\lambda})]$.

Remark 1. *In our procedure, we consider the set of natural partitions defined by (2), while Celisse and Robin (2010) use the one defined by*

$$\mathcal{I} = \bigcup_{N_{min} \le N \le N_{max}} \mathcal{I}_N.$$

This change is natural for lowering the complexity of the algorithm and has no consequences on the theoretical properties of the estimator. In particular, if we assume the function f vanishes on an interval $[1 - \delta, 1]$, then the complexity of the algorithm is simpler when we consider the following set of partitions

$$\mathcal{I} = \bigcup_{m_{min} \le m \le m_{max}} \mathcal{I}_{2^m},$$

where

$$\mathcal{I}_N = \left\{ I^{(k)} = (I_i)_{i=1,\dots,k+1} : \forall i \le k, |I_i| = \frac{1}{N},\ |I_{k+1}| = \frac{N-k}{N},\ 1 \le k \le N-2 \right\}.$$

Let us now study the almost sure convergence and $\sqrt{n}$-consistency of this estimator. First, we present some lemmas whose proofs have been postponed to Appendix A. The two first ones are proved for all partitions of $[0, 1]$ without relying on Assumption 1, while in Lemma 3 we assume that the function f satisfies Assumption 1. This hypothesis is stronger than (Assumption A' in Celisse and Robin, 2010), the latter not being sufficient to establish their result (see discussion in the proof of Lemma 3). We mention that in their simulations, Celisse and Robin (2010) use a function f satisfying our assumption. In Lemma 2, we improve the results of convergence on LPO risk estimator established in (Celisse and Robin, 2010, Proposition 2.1 and Corollary 2.1). Indeed, we prove its almost sure convergence and asymptotic normality.

For each partition I, let us denote by $\mathcal{F}_I$ the vector space of piecewise constant functions built from the partition I and g_I the orthogonal projection of $g \in L^2([0, 1])$

onto $\mathcal{F}_I$. The mean square error of a histogram estimator $\hat{g}_I$ can be written as the sum of a bias term and a variance term

$$\mathbb{E}[||g - \hat{g}_I||_2^2] = ||g - g_I||_2^2 + \mathbb{E}[||g_I - \hat{g}_I||_2^2].$$

Lemma 1. *Let $I = (I_k)_{k=1}^D$ be an arbitrary partition of $[0,1]$. Then the variance term of the mean square error of a histogram estimator $\hat{g}_I$ is bounded by C/n, where C is a positive constant. In other words,*

$$\mathbb{E}[||g_I - \hat{g}_I||_2^2] = O(\frac{1}{n}).$$

Lemma 2. *Let $I = (I_k)_{1,\dots,D}$ be an arbitrary partition of $[0,1]$. Define $L(I) = ||g_I - g||_2^2$ the bias term of the mean square error of a histogram estimator $\hat{g}_I$ and $\hat{L}_p(I) = \hat{R}_p(I) + ||g||_2^2$. Let $p \in \{1, 2, \dots, n-1\}$ such that $\lim_{n\to\infty} p/n < 1$. Then we have the following results*
i) $\hat{L}_p(I) \xrightarrow[n\to\infty]{as} L(I)$
ii) $\sqrt{n}\big(\hat{L}_p(I) - L(I)\big) = \sqrt{n}\big(\hat{R}_p(I) - R(I)\big) + \frac{1}{\sqrt{n}}(s_{11} - s_{21}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 4\sigma_I^2)$,
where $\xrightarrow{d}$ means convergence in distribution and $\xrightarrow{as}$ is almost sure convergence and

$$\sigma_I^2 = s_{32} - s_{21}^2 \text{ with } s_{ij} = \sum_k \frac{\alpha_k^i}{|I_k|^j}, \forall (i,j) \in \mathbb{N}^2.$$

Lemma 3. *Let I, J be two partitions in $\mathcal{I}$, then I is called a subdivision of J and we denote $I \trianglelefteq J$, if $\mathcal{F}_J \subset \mathcal{F}_I$. Suppose that function f satisfies Assumption 1, let us consider m_{max} large enough such that $\mu^\star - \lambda^\star > 2^{1-m_{max}}$. Define $N = 2^{m_{max}}$ and $I^{(N)} = (N, \lambda_N, \mu_N) \in \mathcal{I}$ with $\lambda_N = \lceil N\lambda^\star \rceil / N$, $\mu_N = \lfloor N\mu^\star \rfloor / N$. Then for every partition $I \in \mathcal{I}$, we have*
i) If I is a subdivision of $I^{(N)}$, then $L(I) = L(I^{(N)})$.
ii) If I is not a subdivision of $I^{(N)}$, then $L(I) > L(I^{(N)})$.

In the two following theorems, we shall improve the properties of estimator $\hat{\theta}_n$ with respect to that obtained in (Celisse and Robin, 2010, Theorem 2.1). Indeed, the latter establish only a convergence in probability, while we shall prove the almost sure convergence and $\sqrt{n}$-consistency of $\hat{\theta}_n$.

Theorem 1. *(Almost sure convergence of $\hat{\theta}_n$). Suppose that density f satisfies Assumption 1 and furthermore,*

$$\forall I \in \mathcal{I}, \quad 8s_{11}s_{21} - 2s_{11}^2 + 8s_{32} - 10s_{21}^2 - 4s_{22} \neq 0, \; s_{21} - s_{22} - s_{32} + 3s_{11} \neq 0. \quad (4)$$

Then for m_{max} large enough, we have $\hat{\theta}_n \xrightarrow[n\to\infty]{as} \theta$.

Proof. First, we remark that under conditions (4), Celisse and Robin prove in their Proposition 2.1 that

$$\frac{\hat{p}(I)}{n} \xrightarrow[n\to\infty]{as} l_\infty(I) \in [0,1).$$

Denoting by $\Lambda^\star = [\lambda^\star, \mu^\star]$ and $\hat{\Lambda} = [\hat{\lambda}, \hat{\mu}]$, we may write

$$\begin{aligned}\hat{\theta}_n &= \theta + (\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\subseteq\Lambda^\star} + (\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\nsubseteq\Lambda^\star} \\ &= \theta + \sum_{I=(2^m,\lambda,\mu):[\lambda,\mu]\subseteq\Lambda^\star} \left[\frac{1}{n(\mu-\lambda)}\sum_{i=1}^n \mathbf{1}\{X_i \in [\lambda,\mu]\} - \theta\right]\mathbf{1}\{\hat{\lambda}=\lambda, \hat{\mu}=\mu\} \\ &\quad + (\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\nsubseteq\Lambda^\star}. \qquad (5)\end{aligned}$$

For each partition $I = (2^m, \lambda, \mu)$ such that $[\lambda, \mu] \subseteq \Lambda^\star$, applying the strong law of large numbers we get that

$$\frac{1}{n(\mu-\lambda)}\sum_{i=1}^n \mathbf{1}\{X_i \in [\lambda,\mu]\} \xrightarrow[n\to\infty]{as} \frac{\mathbb{P}(X_i \in [\lambda,\mu])}{\mu-\lambda} = \theta.$$

Since the cardinality $card(\mathcal{I})$ of $\mathcal{I}$ is finite and does not depend on n, in order to finish the proof, it is sufficient to establish that

$$(\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\nsubseteq\Lambda^\star} \xrightarrow[n\to\infty]{as} 0.$$

In fact, if $\hat{\Lambda} \nsubseteq \Lambda^\star$ then $\hat{I}$ is not a subdivision of $I^{(N)}$. Using Lemma 3, we have $L(\hat{I}) > L(I^{(N)})$. Let

$$\gamma = \min_{I \ntrianglelefteq I^{(N)}} L(I) - L(I^{(N)}) > 0, \qquad (6)$$

where $I \ntrianglelefteq I^{(N)}$ means that I is not a subdivision of $I^{(N)}$. We obtain that

$$\begin{aligned}&|\hat{\theta}_n - \theta|\mathbf{1}_{\hat{\Lambda}\nsubseteq\Lambda^\star} \leq (N+\theta)\mathbf{1}\{L(\hat{I}) - L(I^{(N)}) \geq \gamma\} \leq \\ &(N+\theta)\mathbf{1}\{|\hat{L}_{\hat{p}(\hat{I})}(\hat{I}) - L(\hat{I})| + |\hat{L}_{\hat{p}(I^N)}(I^N) - L(I^N)| + \hat{L}_{\hat{p}(\hat{I})}(\hat{I}) - \hat{L}_{\hat{p}(I^{(N)})}(I^{(N)}) \geq \gamma\} \\ &\qquad \leq (N+\theta)\mathbf{1}\{2\sup_{I\in\mathcal{I}}|\hat{L}_{\hat{p}(I)}(I) - L(I)| + \hat{L}_{\hat{p}(\hat{I})}(\hat{I}) - \hat{L}_{\hat{p}(I^{(N)})}(I^{(N)}) \geq \gamma\}.\end{aligned}$$

By definition of $\hat{I}$, we have $\hat{L}_{\hat{p}(\hat{I})}(\hat{I}) - \hat{L}_{\hat{p}(I^{(N)})}(I^{(N)}) \leq 0$, so that

$$\begin{aligned}|\hat{\theta}_n - \theta|\mathbf{1}_{\hat{\Lambda}\nsubseteq\Lambda^\star} &\leq (N+\theta)\mathbf{1}\{\sup_{I\in\mathcal{I}}|\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}\} \qquad (7)\\ &\leq (N+\theta)\sum_{I\in\mathcal{I}}\mathbf{1}\{|\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}\}.\end{aligned}$$

Since $\forall I \in \mathcal{I}$, we both have $\hat{L}_p(I) \xrightarrow[n\to\infty]{as} L(I)$ and $\hat{p}(I)/n \xrightarrow[n\to\infty]{as} l_\infty(I) \in [0,1)$ as well as the fact that $\hat{R}_p(I)$ (given by (3)) is a continuous function of p/n, we obtain $\hat{L}_{\hat{p}(I)}(I) \xrightarrow[n\to\infty]{as} L(I)$. Therefore,

$$\mathbf{1}\{|\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}\} \xrightarrow[n\to\infty]{as} 0.$$

Indeed, if $X_n \xrightarrow{as} X$ then $\forall\epsilon > 0$, we have $\mathbf{1}\{|X_n - X| \geq \epsilon\} \xrightarrow{as} 0$. It thus follows that $(\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\nsubseteq\Lambda^\star} \xrightarrow{as} 0$. We finally get that $\hat{\theta}_n \xrightarrow{as} \theta$. □

In the following, the expression $O_{\mathbb{P}}(1)$ denotes a sequence that is bounded in probability. A sequence estimator T_n of θ is said to be $\sqrt{n}$-consistent if $\sqrt{n}(T_n - \theta) = O_{\mathbb{P}}(1)$. We remark that if T_n is asymptotically normal then T_n is $\sqrt{n}$-consistent but the converse of this statement is not true. In the following theorem, we shall prove the $\sqrt{n}$-consistency of θ_n but we were not able to prove its asymptotically normality. However, according to Prohorov's theorem, there exists a subsequence of $\{\sqrt{n}(\hat{\theta}_n - \theta)\}_n$ that converges in distribution to some random variable Z.

Theorem 2. *($\sqrt{n}$-consistency of $\hat{\theta}_n$). Suppose that density f satisfies the hypotheses of Theorem 1, then for m_{max} large enough, $\hat{\theta}_n$ is $\sqrt{n}$-consistent.*

Proof. We may write as in the proof of Theorem 1,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sum_{I=(2^m,\lambda,\mu):[\lambda,\mu]\subseteq\Lambda^\star} \sqrt{n}\Big[\frac{1}{n(\mu-\lambda)}\sum_{i=1}^n \mathbf{1}\{X_i \in [\lambda,\mu]\} - \theta\Big]\mathbf{1}_{\{\hat{\lambda}=\lambda,\hat{\mu}=\mu\}} + \sqrt{n}(\hat{\theta}_n - \theta)\mathbf{1}_{\{\hat{\Lambda}\not\subseteq\Lambda^\star\}}.$$

For each partition $I = (2^m, \lambda, \mu)$ such that $[\lambda, \mu] \subset \Lambda^\star$, by applying the central limit theorem, we get that

$$\sqrt{n}\big[\frac{1}{n(\mu-\lambda)}\sum_{i=1}^n \mathbf{1}_{X_i\in[\lambda,\mu]} - \theta\big] \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, \theta\Big(\frac{1}{\mu-\lambda} - \theta\Big)\Big).$$

Hence, using again that $\mathrm{card}(\mathcal{I})$ is finite,

$$\sum_{I=(2^m,\lambda,\mu):[\lambda,\mu]\subseteq\Lambda^\star} \sqrt{n}\big[\frac{1}{n(\mu-\lambda)}\sum_{i=1}^n \mathbf{1}_{X_i\in[\lambda,\mu]} - \theta\big]\mathbf{1}_{\hat{\lambda}=\lambda,\hat{\mu}=\mu} = O_{\mathbb{P}}(1). \tag{8}$$

We shall now prove that $\sqrt{n}(\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\not\subseteq\Lambda^\star} \xrightarrow[n\to\infty]{\mathbb{P}} 0$. In fact, according to (7), for all $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}(\sqrt{n}|\hat{\theta}_n - \theta|\mathbf{1}_{\hat{\Lambda}\not\subseteq\Lambda^\star} > \epsilon) &\le \mathbb{P}(\hat{\Lambda}\not\subseteq\Lambda^\star)\\
&\le \mathbb{P}(\sup_{I\in\mathcal{I}}|\hat{L}_{\hat{p}(I)}(I) - L(I)| \ge \frac{\gamma}{2})\\
&\le \sum_{I\in\mathcal{I}}\mathbb{P}(|\hat{L}_{\hat{p}(I)}(I) - L(I)| \ge \frac{\gamma}{2}) \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

where γ is defined by (6). Therefore, $\sqrt{n}(\hat{\theta}_n - \theta)\mathbf{1}_{\hat{\Lambda}\not\subseteq\Lambda^\star} = o_{\mathbb{P}}(1)$. We finally conclude that $\sqrt{n}(\hat{\theta}_n - \theta) = O_{\mathbb{P}}(1)$. □

Note that the difficulty in establishing asymptotic normality of $\hat{\theta}_n$ comes from the fact that there are many different terms in the left-hand side of equation (8) and we did not succeed in proving that $\hat{I} = I^{(N)}$ a.s for large enough n.

2.2. A lower bound on the LAM quadratic risk for estimating θ when f vanishes on a set with positive Lebesgue measure

For any fixed unknown positive parameter $\delta > 0$, we introduce the set of densities $\mathcal{F}_\delta = \{\text{densities } f \text{ on } [0,1]; \mu(I_f) = \delta\}$, where μ denotes Lebesgue measure and $I_f = \{x \in [0,1]; f(x) = 0\}$. We now consider a semiparametric submodel of $\mathcal{P}$ defined by

$$\mathcal{P}_\delta = \left\{p_{\theta,f} \in \mathcal{P}; p_{\theta,f} = \frac{d\mathbb{P}_{\theta,f}}{d\mu} = \theta + (1-\theta)f, \theta \in (0,1), f \in \mathcal{F}_\delta\right\}.$$

In this model, we assume that the density f vanishes on a set whose Lebesgue measure is exactly $\delta > 0$. We aim at estimating the parameter $\psi(\mathbb{P}_{\theta,f}) = \theta$ and consider f as a nuisance parameter. We shall calculate a lower bound on the LAM quadratic risk of any estimator of θ and we also present conditions under which an estimator is asymptotically efficient in the sense of a convolution theorem.

Let us start by recalling some results from semiparametric theory (we refer to Chapter 25 in van der Vaart, 1998, for more details on this subject). We consider a tangent set, denoted by $\dot{\mathcal{P}}_\delta$, of the model $\mathcal{P}_\delta$ at $\mathbb{P}_{\theta,f}$ (with respect to the two parameters (θ, f)) while $\dot{l}_{\theta,f}$ is the ordinary score function for θ in the model in which f is fixed and $\dot{\mathcal{P}}_{f,\delta}$ is a tangent set for the nuisance parameter f in the model where θ is fixed. We also let $\tilde{l}_{\theta,f}$ be the efficient score function for estimating θ and $\tilde{\psi}_{\theta,f}$ be the efficient influence function relative to the tangent set $\dot{\mathcal{P}}_\delta$. For any function h, let us denote $\mathbb{P}_{\theta,f}h = \int h d\mathbb{P}_{\theta,f}$. We now consider the path

$$f_t(x) = \frac{k(th_0(x))f(x)}{\int k(th_0(u))f(u)du} = c(t)k(th_0(x))f(x), \tag{9}$$

where $k(u) = 2(1+e^{-2u})^{-1}$ and $[c(t)]^{-1} = \int k(th_0(u))f(u)du$. By using (van der Vaart, 2002, Lemma 1.8), we obtain a tangent set for f given by

$$\dot{\mathcal{P}}_{f,\delta} = \left\{h \in \mathbb{L}^2(\mathbb{P}_{\theta,f}); \exists h_0 \in \mathbb{L}^1(\mathbb{P}_{\theta,f}), h = \frac{(1-\theta)fh_0}{\theta + (1-\theta)f} \text{ and } \int fh_0 = 0\right\}.$$

We remark that if $\dot{l}_{\theta,f}$ is the ordinary score function for θ in the model in which f is fixed, then for every $a \in \mathbb{R}$ and for every $h \in \dot{\mathcal{P}}_{f,\delta}$, we have $a\dot{l}_{\theta,f} + h$ is a score function for (θ, f) corresponding to the path $t \mapsto \mathbb{P}_{\theta+ta,f_t}$. Hence, the linear span

$$\dot{\mathcal{P}}_\delta = \text{lin}\big(\dot{l}_{\theta,f} + \dot{\mathcal{P}}_{f,\delta}\big) = \{\alpha\dot{l}_{\theta,f} + \beta h; (\alpha,\beta) \in \mathbb{R}^2, h \in \dot{\mathcal{P}}_{f,\delta}\}$$

is a tangent set for (θ, f) at $\mathbb{P}_{\theta,f}$. This set is a linear subspace of Hilbert space $\mathbb{L}^2(\mathbb{P}_{\theta,f})$ with infinite dimension. For every score function g in the tangent set $\dot{\mathcal{P}}_\delta$, we write $P_{t,g}$ for a submodel with score function g along which the function $\psi : \mathbb{P}_{\theta,f} \to \theta$ is differentiable. An estimator sequence T_n is called regular at $\mathbb{P}_{\theta,f}$ for estimating $\psi(\mathbb{P}_{\theta,f})$ (relative to the tangent set $\dot{\mathcal{P}}_\delta$) if there exists a probability measure L such that

$$\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big) \xrightarrow[P_{1/\sqrt{n},g}]{d} L, \text{ for every } g \in \dot{\mathcal{P}}_\delta.$$

According to a convolution theorem (Theorem 25.20 in van der Vaart, 1998), this limit distribution writes as the convolution between the Gaussian distribution $N(0, P(\tilde{\psi}_P^2))$

and another distribution. Thus we shall say that an estimator sequence is asymptotically efficient at P (relative to the tangent set $\dot{\mathcal{P}}_\delta$) if it is regular at P with limit distribution $L = N(0, P(\tilde{\psi}_P^2))$, in other words it is the best regular estimator. We shall call LAM quadratic risk of an estimator sequence T_n (relative to the tangent set $\dot{\mathcal{P}}_\delta$) the quantity

$$\sup_E \liminf_{n\to\infty} \sup_{g\in E} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2,$$

where the first supremum is taken over all finite subsets E of the tangent set $\dot{\mathcal{P}}_\delta$. According to LAM theorem (see Theorem 25.21 in van der Vaart, 1998), this quantity is lower bounded by the minimal variance $P(\tilde{\psi}_P^2)$.

In the following lemma, we shall calculate the efficient information matrix and the efficient influence function in order to establish conditions under which an estimator is asymptotically efficient.

Lemma 4. *The efficient information matrix $\tilde{I}_{\theta,f}$ for θ and the efficient influence function $\tilde{\psi}_{\theta,f}$ relative to the tangent set $\dot{\mathcal{P}}_\delta$ are respectively given by*

$$\tilde{I}_{\theta,f} = \frac{\delta}{\theta(1-\theta\delta)} \text{ and } \tilde{\psi}_{\theta,f}(x) = \frac{1}{\delta}\mathbf{1}_{\{f(x)=0\}} - \theta.$$

By using (van der Vaart, 1998, Theorem 25.21 and Lemma 25.23), an immediate consequence of this lemma is expressed (without any proof) in the following theorem.

Theorem 3. *i) For any estimator sequence T_n we have,*

$$\sup_E \liminf_{n\to\infty} \sup_{g\in E} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2 \geq \theta\big(\frac{1}{\delta} - \theta\big),$$

where the first supremum is taken over all finite subsets E of the tangent set $\dot{\mathcal{P}}_\delta$.

ii) A sequence of estimators $\hat{\theta}_n$ is asymptotically efficient in the sense of a convolution theorem (best regular estimator) if and only if it satisfies

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \frac{1}{\delta}\mathbf{1}_{\{f(X_i)=0\}} + o_{\mathbb{P}_{\theta,f}}(n^{-1/2}). \tag{10}$$

Remark 2. *i) Note that for fixed parameter value λ such that $G(\lambda) < 1$, Storey's estimator $\hat{\theta}^{Storey}(\lambda)$ satisfies*

$$\sqrt{n}\left(\hat{\theta}^{Storey}(\lambda) - \frac{1-G(\lambda)}{1-\lambda}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{G(\lambda)(1-G(\lambda))}{(1-\lambda)^2}\right)$$

(see for instance Genovese and Wasserman, 2004). In particular, if we assume that f vanishes on $[\lambda, 1]$ then we obtain that $G(\lambda) = 1 - \theta(1-\lambda)$ and $\hat{\theta}^{Storey}(\lambda)$ becomes a $\sqrt{n}$-consistent estimate of θ, which is moreover asymptotically distributed, with asymptotic variance

$$\theta\left(\frac{1}{1-\lambda} - \theta\right).$$

In this sense, the oracle version of Storey's estimator that picks $\lambda = 1-\delta$ (namely choosing λ as the smallest value such that f vanishes on $[\lambda, 1]$) is asymptotically efficient. Note also that $\hat{\theta}^{Storey}(\lambda)$ automatically satisfies (10).

ii) In the same way, an oracle version of Celisse and Robin's estimator that would know the true interval $\Lambda^\star$ and include the partition $I = (N, \lambda^\star, \mu^\star)$ in the collection of possible partitions $\mathcal{I}$ would have asymptotically minimal variance and thus be efficient. However, from an expression such as (5), it is not possible to ensure that $\sqrt{n}\hat{\theta}_n$ will have asymptotic variance equal to $\theta(|\Lambda^\star|^{-1} - \theta)$.

Proof of Lemma 4. First, we can easily compute the ordinary score function $\dot{l}_{\theta,f}$ for θ in the model in which f is fixed:

$$\dot{l}_{\theta,f}(x) = \frac{\partial}{\partial\theta}\log[\theta + (1-\theta)f(x)] = \frac{1-f(x)}{\theta+(1-\theta)f(x)}. \tag{11}$$

Define $\Pi_{\theta,f}$ as the orthogonal projection onto the closure of the linear span of $\dot{\mathcal{P}}_{f,\delta}$ in $\mathbb{L}_2(\mathbb{P}_{\theta,f})$. We write

$$\dot{l}_{\theta,f} = \Big(\frac{1-f}{\theta+(1-\theta)f} + \frac{\delta}{1-\theta\delta}\Big)\mathbf{1}_{\{f>0\}} + \frac{1}{\theta}\mathbf{1}_{\{f=0\}} - \frac{\delta}{1-\theta\delta}\mathbf{1}_{\{f>0\}}.$$

We shall prove that

$$\Pi_{\theta,f}\dot{l}_{\theta,f}(x) = \Big(\frac{1-f(x)}{\theta+(1-\theta)f(x)} + \frac{\delta}{1-\theta\delta}\Big)\mathbf{1}_{\{f(x)>0\}}, \tag{12}$$

and then the efficient score function for θ is

$$\tilde{l}_{\theta,f}(x) = \dot{l}_{\theta,f}(x) - \Pi_{\theta,f}\dot{l}_{\theta,f}(x) = \frac{1}{\theta}\mathbf{1}_{\{f(x)=0\}} - \frac{\delta}{1-\theta\delta}\mathbf{1}_{\{f(x)>0\}}.$$

In fact, we can write

$$\Big(\frac{1-f}{\theta+(1-\theta)f} + \frac{\delta}{1-\theta\delta}\Big)\mathbf{1}_{\{f>0\}} = \frac{(1-\theta)fh_0}{\theta+(1-\theta)f},$$

where

$$h_0(x) = \Big(\frac{1-f(x)}{(1-\theta)f(x)} + \frac{\delta}{1-\theta\delta} \times \frac{\theta+(1-\theta)f(x)}{(1-\theta)f(x)}\Big)\mathbf{1}_{\{f(x)>0\}}.$$

Let us denote by I_f^c the complement of I_f in $[0,1]$. It is not difficult to examine the condition $\int fh_0 = 0$. Indeed,

$$\begin{aligned}
\int_0^1 f(x)h_0(x)dx &= \int_{I_f^c}\Big(\frac{1-f(x)}{(1-\theta)} + \frac{\delta}{1-\theta\delta}\times\frac{\theta+(1-\theta)f(x)}{(1-\theta)}\Big)dx \\
&= \frac{1}{1-\theta}\Big[\int_{I_f^c}dx - \int_{I_f^c}f(x)dx + \frac{\delta}{1-\theta\delta}\int_{I_f^c}(\theta+(1-\theta)f(x))dx\Big] \\
&= \frac{1}{1-\theta}\Big[1 - \int_{I_f}dx - \int_0^1 f(x)dx + \frac{\delta}{1-\theta\delta}\big(1 - \int_{I_f}\theta dx\big)\Big] \\
&= \frac{1}{1-\theta}\Big[-\delta + \frac{\delta}{1-\theta\delta}(1-\theta\delta)\Big] = 0,
\end{aligned}$$

and hence

$$\Big(\frac{1-f}{\theta+(1-\theta)f}+\frac{\delta}{1-\theta\delta}\Big)\mathbf{1}_{\{f>0\}}\in\dot{\mathcal{P}}_{f,\delta}.$$

Therefore, it is necessary to prove that

$$\frac{1}{\theta}\mathbf{1}_{\{f=0\}}-\frac{\delta}{1-\theta\delta}\mathbf{1}_{\{f>0\}}=\frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{\{f=0\}}-\frac{\delta}{1-\theta\delta}\perp\dot{\mathcal{P}}_{f,\delta},$$

where $\perp$ means orthogonality in $\mathbb{L}^2(\mathbb{P}_{\theta,f})$. In fact, for every score function

$$h=\frac{(1-\theta)fh_0}{\theta+(1-\theta)f}\in\dot{\mathcal{P}}_{f,\delta}\text{ with }\int fh_0=0,$$

we have

$$\begin{aligned}
&\int_0^1\Big[\frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{\{f(x)=0\}}-\frac{\delta}{1-\theta\delta}\Big]h(x)d\mathbb{P}_{\theta,f}(x)\\
=&\int_0^1\Big[\frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{\{f(x)=0\}}-\frac{\delta}{1-\theta\delta}\Big]\frac{(1-\theta)f(x)h_0(x)}{\theta+(1-\theta)f(x)}(\theta+(1-\theta)f(x))dx\\
=&\frac{1-\theta}{\theta(1-\theta\delta)}\int_0^1 f(x)h_0(x)\mathbf{1}_{\{f(x)=0\}}dx-\frac{(1-\theta)\delta}{1-\theta\delta}\int_0^1 f(x)h_0(x)dx=0.
\end{aligned}$$

This establishes (12). Let us now calculate the efficient information matrix

$$\begin{aligned}
\tilde{I}_{\theta,f}&=\mathbb{P}_{\theta,f}(\tilde{l}^2_{\theta,f})\\
&=\int_0^1\Big(\frac{1}{\theta^2}\mathbf{1}_{\{f(x)=0\}}+\frac{\delta^2}{(1-\theta\delta)^2}\mathbf{1}_{\{f(x)>0\}}\Big)(\theta+(1-\theta)f(x))dx\\
&=\frac{\delta}{\theta}+\frac{\delta^2}{(1-\theta\delta)^2}(1-\theta\delta)=\frac{\delta}{\theta(1-\theta\delta)}>0.
\end{aligned}$$

Using (van der Vaart, 1998, Lemma 25.25), we remark that the functional $\psi(\mathbb{P}_{\theta,f})=\theta$ is differentiable at $\mathbb{P}_{\theta,f}$ relative to the tangent set $\dot{\mathcal{P}}_\delta$ with efficient influence function given by

$$\begin{aligned}
\tilde{\psi}_{\theta,f}(x)&=\tilde{I}^{-1}_{\theta,f}\tilde{l}_{\theta,f}(x)\\
&=\frac{\theta(1-\theta\delta)}{\delta}\Big(\frac{1}{\theta}\mathbf{1}_{\{f(x)=0\}}-\frac{\delta}{1-\theta\delta}\mathbf{1}_{\{f(x)>0\}}\Big)\\
&=\frac{1-\theta\delta}{\delta}\mathbf{1}_{\{f(x)=0\}}-\theta\mathbf{1}_{\{f(x)>0\}}\\
&=\frac{1}{\delta}\mathbf{1}_{\{f(x)=0\}}-\theta.
\end{aligned}$$

We can thus conclude that a lower bound on the minimax quadratic risk for estimating θ is $\mathbb{P}_{\theta,f}(\tilde{\psi}^2_{\theta,f})=\tilde{I}^{-1}_{\theta,f}=\theta(\delta^{-1}-\theta)$. Moreover a sequence of estimators $\hat{\theta}_n$ is asymptotically efficient if and only if it satisfies

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n-\theta)&=\frac{1}{\sqrt{n}}\sum_{i=1}^n\tilde{\psi}_{\theta,f}(X_i)+o_{\mathbb{P}_{\theta,f}}(1)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\frac{1}{\delta}\mathbf{1}_{\{f(X_i)=0\}}-\theta\Big)+o_{\mathbb{P}_{\theta,f}}(1),
\end{aligned}$$

or equivalently

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\delta}\mathbf{1}_{\{f(X_i)=0\}} + o_{\mathbb{P}_{\theta,f}}(n^{-1/2}).$$

□

3. About asymptotically efficient estimators of θ

3.1. General overview of one-step estimators

In this section, we introduce the one-step method to construct an asymptotically efficient estimator, relying on a $\sqrt{n}$-consistent one (see van der Vaart, 1998, Section 25.8). Let $\hat{\theta}_n$ be a $\sqrt{n}$-consistent estimator of θ, then $\hat{\theta}_n$ can be discretized on grids of mesh width $n^{-1/2}$. Suppose that we are given a sequence of estimators $\hat{l}_{n,\theta}(\cdot) = \hat{l}_{n,\theta}(\cdot; X_1, \dots, X_n)$ of the efficient score function $\tilde{l}_{\theta,f}$. Define with $m = \lfloor n/2 \rfloor$,

$$\hat{l}_{n,\theta,i}(\cdot) = \begin{cases} \hat{l}_{m,\theta}(\cdot; X_1, \dots, X_m) & \text{if} \quad i > m, \\ \hat{l}_{n-m,\theta}(\cdot; X_{m+1}, \dots, X_n) & \text{if} \quad i \leq m. \end{cases}$$

Thus, for X_i ranging through each of the two halves of the sample, we use an estimator $\hat{l}_{n,\theta,i}$ based on the other half of the sample. We assume that, for every deterministic sequence $\theta_n = \theta + O(n^{-1/2})$, we have

$$\sqrt{n}\mathbb{P}_{\theta_n,f}\hat{l}_{n,\theta_n} \xrightarrow[n\to\infty]{\mathbb{P}_{\theta,f}} 0, \tag{13}$$

$$\mathbb{P}_{\theta_n,f}\|\hat{l}_{n,\theta_n} - \tilde{l}_{\theta_n,f}\|^2 \xrightarrow[n\to\infty]{\mathbb{P}_{\theta,f}} 0, \tag{14}$$

$$\int \|\tilde{l}_{\theta_n,f}d\mathbb{P}^{1/2}_{\theta_n,f} - \tilde{l}_{\theta,f}d\mathbb{P}^{1/2}_{\theta,f}\|^2 \xrightarrow[n\to\infty]{} 0. \tag{15}$$

Then, the one-step estimator defined as

$$\tilde{\theta}_n = \hat{\theta}_n - \Big(\sum_{i=1}^{n}\hat{l}^2_{n,\hat{\theta}_n,i}(X_i)\Big)^{-1}\sum_{i=1}^{n}\hat{l}_{n,\hat{\theta}_n,i}(X_i),$$

is asymptotically efficient at (θ, f) (see van der Vaart, 1998, Section 25.8). This estimator $\tilde{\theta}_n$ can be considered a one-step iteration of the Newton-Raphson algorithm for solving an approximation of the equation $\sum_i \tilde{l}_{\theta,f}(X_i) = 0$ with respect to θ, starting at the initial guess $\hat{\theta}_n$.

In the next section, we discuss a converse result on necessary conditions for existence of an asymptotically efficient estimator of θ and its implications in model $\mathcal{P}_\delta$.

3.2. Existence of efficient estimators of θ in model $\mathcal{P}_\delta$

Under condition (15), it is shown in (van der Vaart, 2002, Theorem 7.4) that the existence of an asymptotically efficient sequence of estimators of θ implies the existence of a sequence of estimators $\hat{l}_{n,\theta}$ of $\tilde{l}_{\theta,f}$ satisfying (13) and (14). In our case, it is not difficult to prove that condition (15) holds. As a consequence, we obtain the following proposition.

Proposition 1. *The existence of an asymptotically efficient sequence of estimators of θ in model $\mathcal{P}_\delta$ is equivalent to the existence of a sequence of estimators $\hat{l}_{n,\theta}$ of the efficient score function $\tilde{l}_{\theta,f}$ satisfying (13) and (14).*

Proof. Let us first establish that condition (15) holds. In fact, with the notation $p_{\theta,f} = \theta + (1-\theta)f$, we have

$$\begin{aligned}
&\int \|\tilde{l}_{\theta_n,f} d\mathbb{P}^{1/2}_{\theta_n,f} - \tilde{l}_{\theta,f} d\mathbb{P}^{1/2}_{\theta,f}\|^2 = \int_0^1 \Big(\tilde{l}_{\theta_n,f}(x)\sqrt{p_{\theta_n,f}(x)} - \tilde{l}_{\theta,f}(x)\sqrt{p_{\theta,f}(x)}\Big)^2 dx \\
\leq\ & 2\int_0^1 \big(\tilde{l}_{\theta_n,f}(x) - \tilde{l}_{\theta,f}(x)\big)^2 p_{\theta_n,f}(x)dx + 2\int_0^1 \tilde{l}^2_{\theta,f}(x)\Big(\sqrt{p_{\theta_n,f}(x)} - \sqrt{p_{\theta,f}(x)}\Big)^2 dx \\
\leq\ & 2\int_0^1 \Big[\frac{1}{\theta_n} - \frac{1}{\theta} + \Big(\frac{1}{\theta(1-\theta\delta)} - \frac{1}{\theta_n(1-\theta_n\delta)}\Big)\mathbf{1}_{\{f(x)>0\}}\Big]^2 p_{\theta_n,f}(x)dx \\
& +2\int_0^1 \Big[\frac{1}{\theta} - \frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{\{f(x)>0\}}\Big]^2 \frac{(\theta_n-\theta)^2(1-f(x))^2}{\Big(\sqrt{p_{\theta_n,f}(x)} + \sqrt{p_{\theta,f}(x)}\Big)^2} dx \\
\leq\ & 2\int_0^1 (\theta_n-\theta)^2 \Big[\frac{1}{\theta\theta_n} + \frac{\delta(\theta+\theta_n)+1}{\theta\theta_n(1-\theta\delta)(1-\theta_n\delta)}\mathbf{1}_{\{f(x)>0\}}\Big]^2 p_{\theta_n,f}(x)dx \\
& +2\int_0^1 (\theta_n-\theta)^2 2\Big[\frac{1}{\theta^2} + \frac{1}{\theta^2(1-\theta)^2}\Big]\frac{(1-f(x))^2}{(\sqrt{\theta_n}+\sqrt{\theta})^2} dx \\
\leq\ & (\theta_n-\theta)^2\Big[\frac{C}{\theta^2} + \frac{C(1+2C\theta)}{\theta^2(1-\theta)^2}\Big]^2 + C(\theta_n-\theta)^2\Big[\frac{1}{\theta^3} + \frac{1}{\theta^3(1-\theta)^2}\Big] = O\big(\frac{1}{n}\big),
\end{aligned}$$

where C is some positive constant. Thus, according to (van der Vaart, 2002, Theorem 7.4), the existence of an asymptotically efficient sequence of estimators of θ is equivalent to the existence of a sequence of estimators $\hat{l}_{n,\theta}$ satisfying (13) and (14). □

The estimator $\hat{l}_{n,\theta}$ of the efficient score function $\tilde{l}_{\theta,f}$ must satisfy both a "no-bias" (13) and a consistency (14) condition. The consistency is usually easy to arrange, but the "no-bias" condition requires a convergence to zero of the bias at a rate faster than $1/\sqrt{n}$. In our model, the efficient score function $\tilde{l}_{\theta,f}$ is given by

$$\tilde{l}_{\theta,f}(x) = \frac{1}{\theta} - \frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{\{f(x)>0\}},$$

so that we must estimate the set I_f and its measure δ in order to estimate $\tilde{l}_{\theta,f}$. To our knowledge, the rate of convergence of (the bias of) such level-sets estimators is nonparametric (see Baíllo et al., 2001; Cadre, 2006; Mason and Polonik, 2009). Thus, it is likely that there does not exist an asymptotically efficient sequence of estimators of θ in model $\mathcal{P}_\delta$.

3.3. Existence of efficient estimators of θ in model $\mathcal{P}_0$

We now consider a second submodel of $\mathcal{P}$ given by $\mathcal{P}_0 = \big\{p_{\theta,f} \in \mathcal{P} : \theta \in (0,1),\ f \in \mathcal{F}_0\big\}$, where $\mathcal{F}_0 = \{\text{density } f \text{ on } [0,1] \text{ such that } \mu(I_f) = 0\}$. This model is a limiting case of model $\mathcal{P}_\delta$ when $\delta \to 0$. It corresponds to a difficult case with respect to estimation in model (1), where density f vanishes only on a subset of null Lebesgue measure. We recall that estimators constructed under this setup (and assuming moreover either regularity or monotonicity properties on f) exhibit nonparametric rates of convergence. Besides, the results of Theorem 3 suggest that no estimator of θ can have finite asymptotic variance at rate $\sqrt{n}$ in this case. By using for instance the path given by Equation (9), we obtain a tangent set for f,

$$\dot{\mathcal{P}}_{f,0} = \Big\{ h \in \mathbb{L}^2(\mathbb{P}_{\theta,f}); \exists h_0 \in \mathbb{L}^1(\mathbb{P}_{\theta,f}), h = \frac{(1-\theta) f h_0}{\theta + (1-\theta) f} \text{ with } \int_0^1 f h_0 = 0 \Big\}.$$

In the following lemma, we shall prove that the ordinary score $\dot{l}_{\theta,f}$ belongs to this tangent set $\dot{\mathcal{P}}_{f,0}$. So the efficient information matrix $\tilde{I}_{\theta,f}$ for θ is equal to 0 and we can prove that relative to the tangent set $\dot{\mathcal{P}}_0 = \text{lin}\big(\dot{l}_{\theta,f} + \dot{\mathcal{P}}_{f,0}\big)$, there is no regular estimator sequence for θ or equivalently, the minimax quadratic risk is not finite.

Lemma 5. *The efficient information matrix $\tilde{I}_{\theta,f}$ for θ relative to the tangent set $\dot{\mathcal{P}}_0$ is equal to 0. Moreover, the functional $\psi(\mathbb{P}_{\theta,f}) = \theta$ is not differentiable at $\mathbb{P}_{\theta,f}$ relative to the tangent set $\dot{\mathcal{P}}_0$.*

Proof. First, we prove that the ordinary score function $\dot{l}_{\theta,f}$ for the parameter θ belongs to this tangent set $\dot{\mathcal{P}}_{f,0}$. Indeed, let the function

$$h_0 = \frac{1-f}{(1-\theta) f} \mathbf{1}_{\{f>0\}}.$$

Since $\mu(I_f) = 0$, we have $\int f h_0 = 0$ and according to (11), we obtain

$$\dot{l}_{\theta,f} = \frac{(1-\theta) f h_0}{\theta + (1-\theta) f}, \quad \mu - \text{almost everywhere},$$

so that $\dot{l}_{\theta,f} \in \dot{\mathcal{P}}_{f,0}$. Then, the efficient information matrix $\tilde{I}_{\theta,f}$ is equal to 0. Now, we show that the functional $\psi(\mathbb{P}_{\theta,f}) = \theta$ is not differentiable at $\mathbb{P}_{\theta,f}$ relative to the tangent set $\dot{\mathcal{P}}_0 = \text{lin}\big(\dot{l}_{\theta,f} + \dot{\mathcal{P}}_{f,0}\big) = \dot{\mathcal{P}}_{f,0}$. In fact, if this were true, there would exist a function $\tilde{\psi}_{\theta,f}$ such that

$$a = \frac{\partial}{\partial t} \psi(\mathbb{P}_{\theta+ta,f_t})\Big|_{t=0} = \langle \tilde{\psi}_{\theta,f}, a\dot{l}_{\theta,f} + h \rangle, \ \forall a \in \mathbb{R}, h \in \dot{\mathcal{P}}_{f,0},$$

where $\langle \cdot, \cdot \rangle$ denotes scalar product in $\mathbb{L}^2(\mathbb{P}_{\theta,f})$. Choosing $h = \dot{l}_{\theta,f}$, we obtain $a = (a+1)\langle \tilde{\psi}_{\theta,f}, \dot{l}_{\theta,f} \rangle$ for every value $a \in \mathbb{R}$, which is impossible. □

Theorem 4. *For any estimator sequence T_n we have,*

$$\sup_E \liminf_{n\to\infty} \sup_{g\in E} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2 = +\infty,$$

where the first supremum is taken over all finite subsets E of the tangent set $\dot{\mathcal{P}}_0$.

Proof. We first remark that the tangent set $\dot{\mathcal{P}}_0$ is a linear subspace of $\mathbb{L}^2(\mathbb{P}_{\theta,f})$ with infinite dimension. So we can choose an orthonormal basis $\{h_i\}_{i=1}^{\infty}$ of $\dot{\mathcal{P}}_0$ such that for every m, $\dot{l}_{\theta,f} \notin \dot{\mathcal{P}}_{0,m} := \mathrm{lin}(h_1, h_2, \dots, h_m)$. We thus have

$$\sup_E \liminf_{n\to\infty} \sup_{g\in E} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2 \geq \sup_F \liminf_{n\to\infty} \sup_{g\in F} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2,$$

where E and F range through all finite subsets of the tangent sets $\dot{\mathcal{P}}_0$ and $\mathrm{lin}\big(\dot{l}_{\theta,f} + \dot{\mathcal{P}}_{0,m}\big)$, respectively. The efficient score function for θ corresponding to the tangent set $\mathrm{lin}\big(\dot{l}_{\theta,f} + \dot{\mathcal{P}}_{0,m}\big)$ is

$$\tilde{l}_{\theta,f,m} = \dot{l}_{\theta,f} - \sum_{i=1}^{m} < \dot{l}_{\theta,f}, h_i > h_i \neq 0,$$

where $< \cdot,\cdot >$ stands for the scalar product in $\mathbb{L}^2(\mathbb{P}_{\theta,f})$. Moreover, the efficient information matrix $\tilde{I}_{\theta,f,m} = \mathbb{P}_{\theta,f}(\tilde{l}^2_{\theta,f,m})$ is nonsingular. Using (van der Vaart, 1998, Lemma 25.25), we remark that the functional $\psi(\mathbb{P}_{\theta,f}) = \theta$ is differentiable at $\mathbb{P}_{\theta,f}$ relative to the tangent set $\mathrm{lin}\big(\dot{l}_{\theta,f} + \dot{\mathcal{P}}_{0,m}\big)$ with efficient influence function $\tilde{\psi}_{\theta,f,m} = \tilde{I}^{-1}_{\theta,f,m}\tilde{l}_{\theta,f,m}$. So we can apply (Theorem 25.21 in van der Vaart, 1998) to obtain that

$$\sup_F \liminf_{n\to\infty} \sup_{g\in F} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2 \geq \tilde{I}^{-1}_{\theta,f,m}.$$

Since $\tilde{I}_{\theta,f,m} \xrightarrow[m\to\infty]{} 0$, we have

$$\sup_E \liminf_{n\to\infty} \sup_{g\in E} \mathbb{E}_{P_{1/\sqrt{n},g}} \big[\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\big]^2 \geq \tilde{I}^{-1}_{\theta,f,m} \xrightarrow[m\to\infty]{} +\infty.$$

□

Remark 3. *Using (Chamberlain, 1986, Theorem 2), we can conclude that there is no regular estimator sequence for θ relative to the tangent set $\dot{\mathcal{P}}_0$. This result implies that if there exists a $\sqrt{n}$-consistent estimator in model $\mathcal{P}_0$, it can not have finite asymptotic variance. It does not imply that $\sqrt{n}$-consistent estimators do not exist in model $\mathcal{P}_0$, namely, we could have $\sqrt{n}(\hat{\theta} - \theta) = O_{\mathbb{P}}(1)$ for some estimator $\hat{\theta}$ but then $\mathbb{V}ar(\sqrt{n}\hat{\theta}) \to +\infty$. However, we note that the only rates of convergence obtained until now in this case are nonparametric ones.*

Acknowledgments The authors are grateful to Cyril Dalmasso, Elisabeth Gassiat and Pierre Neuvial for fruitful discussions concerning this work.

A. Appendix. Proofs of technical lemmas

A.1. Proof of Lemma 1

Note that Celisse and Robin (2010) prove that $\mathbb{E}[||g - \hat{g}_I||_2^2] \xrightarrow[n\to\infty]{} 0$, while we further establish that it is $O(1/n)$. By a simple bias-variance decomposition, we may write

$$\mathbb{E}[||g_I - \hat{g}_I||_2^2] = \mathbb{E}[||g - \hat{g}_I||_2^2] - ||g_I - g||_2^2.$$

As for the bias term, it is easy to show that

$$
\begin{aligned}
||g - g_I||_2^2 &= \inf_{h\in\mathcal{F}_I} ||g-h||_2^2 \\
&= \inf_{(a_k)_k\in\mathbb{R}} \Big[||g||_2^2 - 2\int_0^1 \big(\sum_k a_k \mathbf{1}_{I_k}(x)\big) g(x)dx + \int_0^1 \big(\sum_k a_k \mathbf{1}_{I_k}(x)\big)^2 dx\Big] \\
&= \inf_{(a_k)_k\in\mathbb{R}} \Big[||g||_2^2 - 2\sum_k a_k\alpha_k + \sum_k a_k^2 |I_k|\Big] \\
&= ||g||_2^2 - \sum_k \frac{\alpha_k^2}{|I_k|} = ||g||_2^2 - s_{21}. \qquad (16)
\end{aligned}
$$

Let us now calculate the mean square error of $\hat{g}_I$

$$
\begin{aligned}
\mathbb{E}[||g-\hat{g}_I||_2^2] &= ||g||_2^2 + \mathbb{E}\Big[||\hat{g}_I||_2^2 - 2\int_0^1 \hat{g}_I(x)g(x)dx\Big] \\
&= ||g||_2^2 + \mathbb{E}\Big[\int_0^1 \big(\sum_k \frac{n_k}{n|I_k|}\mathbf{1}_{I_k}(x)\big)^2 dx - 2\int_0^1 \sum_k \frac{n_k}{n|I_k|}\mathbf{1}_{I_k}(x)g(x)dx\Big] \\
&= ||g||_2^2 + \mathbb{E}\Big[\sum_k \frac{n_k^2}{n^2|I_k|} - 2\sum_k \frac{n_k\alpha_k}{n|I_k|}\Big].
\end{aligned}
$$

Since n_k follows a Binomial distribution $\mathcal{B}(n,\alpha_k)$, we have

$$\mathbb{E}[n_k] = n\alpha_k \text{ and } \mathbb{E}[n_k^2] = n^2\alpha_k^2 + n\alpha_k(1-\alpha_k).$$

Therefore,

$$
\begin{aligned}
\mathbb{E}[||g-\hat{g}_I||_2^2] &= ||g||_2^2 + \sum_k \frac{n^2\alpha_k^2 + n\alpha_k(1-\alpha_k)}{n^2|I_k|} - 2\sum_k \frac{n\alpha_k^2}{n|I_k|} \\
&= ||g||_2^2 - s_{21} + \frac{1}{n}(s_{11} - s_{21}). \qquad (17)
\end{aligned}
$$

Using (16) and (17), we obtain the desired result, namely

$$\mathbb{E}[||g_I - \hat{g}_I||_2^2] = \mathbb{E}[||g-\hat{g}_I||_2^2] - ||g_I - g||_2^2 = \frac{1}{n}(s_{11}-s_{21}) = O(\frac{1}{n}).$$

A.2. Proof of Lemma 2

i) Since

$$\lim_{n\to\infty} \frac{p}{n} < 1 \text{ and } \frac{n_k}{n} \xrightarrow[n\to\infty]{as} \alpha_k, \text{ for all } k,$$

we obtain that

$$
\begin{aligned}
\hat{L}_p(I) &= ||g||_2^2 + \frac{2n-p}{(n-1)(n-p)}\sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)}\sum_k \frac{1}{|I_k|}\big(\frac{n_k}{n}\big)^2 \\
&\xrightarrow[n\to\infty]{as} ||g||_2^2 - \sum_k \frac{\alpha_k^2}{|I_k|} = ||g||_2^2 - s_{21} = ||g_I - g||_2^2 = L(I).
\end{aligned}
$$

ii) By definition of $R(I)$ and using (17), we have

$$R(I) = \mathbb{E}[||g - \hat{g}_I||_2^2] - ||g||_2^2 = -s_{21} + \frac{1}{n}(s_{11} - s_{21}).$$

This gives that

$$\begin{aligned}
\sqrt{n}[\hat{R}_p(I) - R(I)] &= \sqrt{n}\Big[\frac{2n-p}{(n-1)(n-p)}\sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)}\sum_k \frac{1}{|I_k|}\big(\frac{n_k}{n}\big)^2 \\
&\quad + s_{21} - \frac{1}{n}(s_{11} - s_{21})\Big] \\
&= \frac{2n-p}{(n-1)(n-p)}\sum_k \frac{1}{|I_k|}\big[\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big)\big] + \frac{(2n-p)\sqrt{n}}{(n-1)(n-p)} s_{11} \\
&\quad - \frac{n(n-p+1)}{\sqrt{n}(n-1)(n-p)}\sum_k \frac{1}{|I_k|}\big[\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big)\big]^2 - \frac{(2n-p)\sqrt{n}}{(n-1)(n-p)} s_{21} \\
&\quad - \frac{2n(n-p+1)}{(n-1)(n-p)}\sum_k \frac{\alpha_k}{|I_k|}\big[\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big)\big] - \frac{1}{\sqrt{n}}(s_{11} - s_{21}) \\
&= T_1 - \frac{2n(n-p+1)}{(n-1)(n-p)}\sum_k \frac{\alpha_k}{|I_k|}\big[\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big)\big]. \qquad (18)
\end{aligned}$$

Then, using the central limit theorem and the continuity of the function $x \mapsto x^2$, we have

$$\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \alpha_k(1-\alpha_k)),$$

$$\big[\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big)\big]^2 \xrightarrow[n\to\infty]{d} Z_k^2 \text{ with } Z_k \sim \mathcal{N}(0, \alpha_k(1-\alpha_k)).$$

It thus follows that $T_1 = o_{\mathbb{P}}(1)$. We now consider the remaining term in (18). We have

$$\begin{aligned}
\sum_k \frac{\alpha_k}{|I_k|}\big[\sqrt{n}\big(\frac{n_k}{n} - \alpha_k\big)\big] &= \frac{1}{\sqrt{n}}\sum_k \frac{\alpha_k}{|I_k|} n_k - \sqrt{n}\sum_k \frac{\alpha_k^2}{|I_k|} \\
&= \frac{1}{\sqrt{n}}\sum_k \frac{\alpha_k}{|I_k|}\big(\sum_{i=1}^n \mathbf{1}_{X_i \in I_k}\big) - \sqrt{n}\; s_{21} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\sum_k \frac{\alpha_k}{|I_k|}\mathbf{1}_{X_i \in I_k} - s_{21}\big).
\end{aligned}$$

Let us denote

$$Y_i = \sum_k \frac{\alpha_k}{|I_k|}\mathbf{1}_{X_i \in I_k} - s_{21}.$$

Then the random variables $Y_1, Y_2, \ldots, Y_n$ are iid centered with variance

$$\sigma_I^2 = \mathbb{E}(Y_1^2) = \mathbb{E}\Big(\sum_k \frac{\alpha_k^2}{|I_k|^2}\mathbf{1}_{X_1 \in I_k} - 2s_{21}\sum_k \frac{\alpha_k}{|I_k|}\mathbf{1}_{X_1 \in I_k} + s_{21}^2\Big) = s_{32} - s_{21}^2.$$

By the central limit theorem, we obtain

$$\sum_k \frac{\alpha_k}{|I_k|}\left[\sqrt{n}\left(\frac{n_k}{n}-\alpha_k\right)\right] \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma_I^2).$$

Combining this with (18) implies that

$$\sqrt{n}[\hat{R}_p(I)-R(I)] \xrightarrow[n\to\infty]{d} \mathcal{N}(0,4\sigma_I^2).$$

It is easy to calculate that

$$\sqrt{n}\big(\hat{L}_p(I)-L(I)\big) = \sqrt{n}\big(\hat{R}_p(I)-R(I)\big) + \frac{1}{\sqrt{n}}(s_{11}-s_{21}).$$

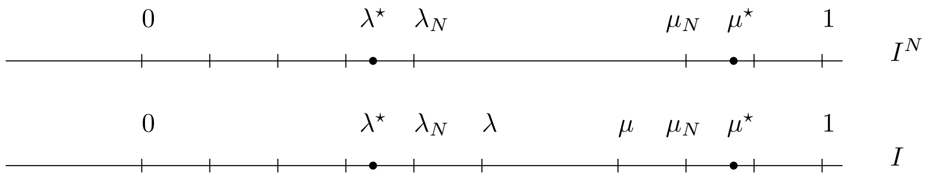
Hence, we have

$$\sqrt{n}[\hat{L}_p(I)-L(I)] \xrightarrow[n\to\infty]{d} \mathcal{N}(0,4\sigma_I^2),$$

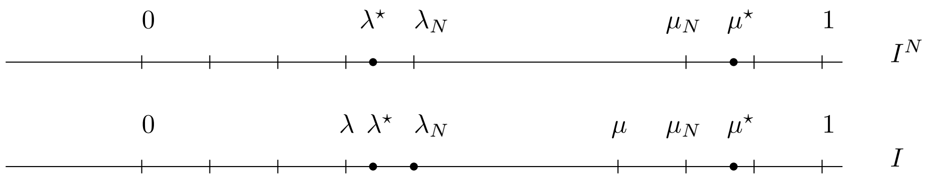
which completes the proof.

A.3. Proof of Lemma 3

i) If I is a subdivision of $I^{(N)}$, then $I=(N,\lambda,\mu)$ with $[\lambda,\mu]\subset[\lambda^\star,\mu^\star]$. For example, we may have the following situation



Since g is constant on the interval $[\lambda^\star,\mu^\star]\supset[\lambda_N,\mu_N]\supset[\lambda,\mu]$, we have $g_I=g_{I^{(N)}}=g$ on the interval $[\lambda_N,\mu_N]$. This implies that $||g_I-g||_2^2=||g_{I^{(N)}}-g||_2^2$.
ii) If $I=(2^m,\lambda,\mu)$ is not a subdivision of $I^{(N)}$, then there are two cases to consider: If $m=m_{max}$ then $[\lambda,\mu]\nsubseteq[\lambda_N,\mu_N]$. For example, we may have



Since $g_I=g_{I^{(N)}}=g$ on the interval $[\lambda_N,\mu_N]$ and the two partitions I and $I^{(N)}$ restricted to the interval $[\lambda,\mu]^c\cap[\lambda_N,\mu_N]^c$ are the same, we thus have

$$||g_I-g||^2_{2,[\lambda,\mu]^c} = ||g_{I^{(N)}}-g||^2_{2,[\lambda,\mu]^c},$$

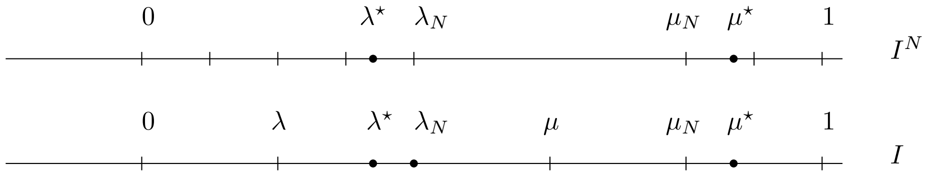
so that

$$||g_I-g||_2^2 - ||g_{I^{(N)}}-g||_2^2 = ||g_I-g||^2_{2,[\lambda,\mu]} - ||g_{I^{(N)}}-g||^2_{2,[\lambda,\mu]}.$$

Using the monotonicity of f on the intervals $[0, \lambda^\star]$ and $[\mu^\star, 1]$, we get that

$$||g_I - g||^2_{2,[\lambda,\mu]} > ||g_{I^{(N)}} - g||^2_{2,[\lambda,\mu]}, \text{ which implies that } L(I) > L(I^{(N)}).$$

If $m < m_{max}$, we may have for example

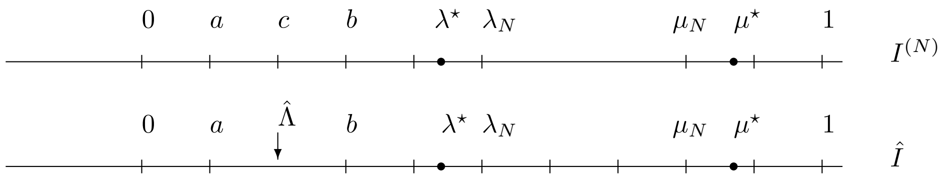


As before, we may show that

$$||g_I - g||^2_2 - ||g_{I^{(N)}} - g||^2_2 \geq ||g_I - g||^2_{2,[\lambda,\mu]^c} - ||g_{I^{(N)}} - g||^2_{2,[\lambda,\mu]^c} > 0,$$

which completes the proof.

We remark that the assumptions in Lemma 2.1 or Theorem 2.1 in Celisse and Robin (2010) are not sufficient to show these results. In fact, the assumption "g is non-constant outside $\Lambda^\star$" is not sufficient to imply that $\|g - g_{I^{(N)}}\|^2_2 < \|g - g_{\hat{I}}\|^2_2$ in the case where $\hat{I}$ is not a subdivision of $I^{(N)}$. For example, let us consider the following situation



We may then calculate that

$$\|g - g_{\hat{I}}\|^2_2 - \|g - g_{I^{(N)}}\|^2_2 = (c - a)(\alpha_1 - \alpha)^2 + (b - c)(\alpha_2 - \alpha)^2,$$

where

$$\alpha = \frac{1}{b-a}\int_a^b g(x)dx, \quad \alpha_1 = \frac{1}{c-a}\int_a^c g(x)dx, \quad \alpha_2 = \frac{1}{b-c}\int_c^b g(x)dx.$$

So that if the function g satisfies $\alpha = \alpha_1 = \alpha_2$ (and g is non-constant outside $\Lambda^\star$) then $\|g - g_{I^{(N)}}\|^2_2 = \|g - g_{\hat{I}}\|^2_2$.

References

Baíllo, A., Cuesta-Albertos, J. A., Cuevas, A., 2001. Convergence rates in nonparametric estimation of level sets. Statist. Probab. Lett. 53 (1), 27–35.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57 (1), 289–300.

Broberg, P., 2005. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. BMC Bioinformatics 6 (1), 199.

Cadre, B., 2006. Kernel estimation of density level sets. J. Multivariate Anal. 97 (4), 999–1023.
Cai, T., Jin, J., 2010. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. Ann. Statist. 38 (1), 100145.
Celisse, A., Robin, S., 2008. Nonparametric density estimation by exact leave-p-out cross-validation. Comput. Statist. Data Anal. 52 (5), 2350–2368.
Celisse, A., Robin, S., 2010. A cross-validation based estimation of the proportion of true null hypotheses. J. Statist. Plann. Inference 140 (11), 3132–3147.
Chamberlain, G., 1986. Asymptotic efficiency in semiparametric models with censoring. J. Econometrics 32 (2), 189–218.
Dudoit, S., van der Laan, M. J., 2008. Multiple testing procedures with applications to genomics. Springer Series in Statistics. Springer, New York.
Efron, B., 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J. Amer. Statist. Assoc. 99 (465), 96–104.
Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96 (456), 1151–1160.
Genovese, C., Wasserman, L., 2004. A stochastic process approach to false discovery control. Ann. Statist. 32 (3), 1035–1061.
Hengartner, N. W., Stark, P. B., 1995. Finite-sample confidence envelopes for shape-restricted densities. Ann. Statist. 23 (2), 525–550.
Jin, J., 2008. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (3), 461493.
Jin, J., Cai, T., 2007. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. J. Amer. Statist. Assoc. 102 (478).
Langaas, M., Lindqvist, B. H., Ferkingstad, E., 2005. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (4), 555–572.
Mason, D. M., Polonik, W., 2009. Asymptotic normality of plug-in level set estimates. Ann. Appl. Probab. 19 (3), 1108–1142.
Meinshausen, N., Bühlmann, P., 2005. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. Biometrika 92 (4), 893–907.
Meinshausen, N., Rice, J., 2006. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. Ann. Statist. 34 (1), 373–393.
Neuvial, P., 2010. Intrinsic bounds and false discovery rate control in multiple testing problems. Tech. rep., arXiv:1003.0747.
Pounds, S., Cheng, C., 2006. Robust estimation of the false discovery rate. Bioinformatics 22 (16), 1979–1987.
Schweder, T., Spjtvoll, E., 1982. Plots of p-values to evaluate many tests simultaneously. Biometrika 69 (3), 493–502.
Storey, J. D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3), 479–498.
Storey, J. D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100 (16), 9440–9445 (electronic).
Swanepoel, J. W. H., 1999. The limiting behavior of a modified maximal symmetric $2s$-spacing with applications. Ann. Statist. 27 (1), 24–35.
Turkheimer, F., Smith, C., Schmidt, K., 2001. Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. NeuroImage 13 (5), 920 – 930.
van der Vaart, A., 2002. Semiparametric statistics. Bolthausen, Erwin et al., Lectures on probability theory and statistics. Ecole d'été de probabilités de Saint-Flour XXIX - 1999, Saint-Flour, France, July 8-24, 1999. Berlin: Springer. Lect. Notes Math. 1781, 331-457 (2002).
van der Vaart, A. W., 1998. Asymptotic statistics. Vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.