Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 41: 1167–1194, 2014
doi: 10.1111/sjos.12091

On Efficient Estimators of the Proportion of True Null Hypotheses in a Multiple Testing Setup

VAN HANH NGUYEN

Laboratoire de Mathématiques d'Orsay, Université Paris Sud
Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne

CATHERINE MATIAS

Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne

ABSTRACT. We consider the problem of estimating the proportion θ of true null hypotheses in a multiple testing context. The setup is classically modelled through a semiparametric mixture with two components: a uniform distribution on interval $[0, 1]$ with prior probability θ and a nonparametric density f. We discuss asymptotic efficiency results and establish that two different cases occur whether f vanishes on a non-empty interval or not. In the first case, we exhibit estimators converging at a parametric rate, compute the optimal asymptotic variance and conjecture that no estimator is asymptotically efficient (i.e. attains the optimal asymptotic variance). In the second case, we prove that the quadratic risk of any estimator does not converge at a parametric rate. We illustrate those results on simulated data.

Key words: asymptotic efficiency, efficient score, false discovery rate, information bound, multiple testing, p-values, semiparametric model

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 & van der Laan, 2008). Among other application domains, we mention astrophysics (Meinshausen & 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 & Hochberg (1995). In this work, we discuss asymptotic efficiency of estimators of the true proportion of null hypotheses. We stress that the asymptotic framework is particularly relevant in the aforementioned contexts where the number of tested hypotheses is huge.

In many recent articles (such as Broberg, 2005; Celisse & Robin, 2010; Genovese & Wasserman, 2004; and Langaas *et al.*, 2005), a two-component mixture density is used to model the behaviour of p-values $X_1, X_2, \dots, X_n$ associated with n independent tested hypotheses. More precisely, assume that the test statistics are independent and identically distributed (i.i.d.) with a continuous distribution under the corresponding null hypotheses, then the p-values $X_1, X_2, \dots, X_n$ are i.i.d. and follow the uniform distribution $\mathcal{U}([0, 1])$ on interval $[0, 1]$ under the null hypotheses. The density g of p-values is modelled by a two-component mixture with the 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 of p-values generated under the alternative (false null hypotheses).

1.1. Identifiability

Many different identifiability conditions on the parameter (θ, f) in model (1) have been discussed in the literature. For example, Genovese & Wasserman (2004) introduced the concept of purity that corresponds to the case where the essential infimum of f on $[0, 1]$ is zero. They proved that purity implies identifiability but not *vice versa*. Langaas *et al.* (2005) supposed that f is decreasing with $f(1) = 0$, while Neuvial (2013) assumed that f is regular near $x = 1$ with $f(1) = 0$, and Celisse & Robin (2010) considered that f vanishes on a whole interval included in $[0, 1]$. These are sufficient but not necessary conditions on f that ensure identifiability. Now, if we assume more generally that f belongs to some set $\mathcal{F}$ of densities on $[0, 1]$, then a necessary and sufficient condition for parameters identifiability is stated in the next result, whose proof is given in Section 5.1.

Proposition 1. *The parameter* (θ, f) *is identifiable on a set* $(0, 1) \times \mathcal{F}$ *if and only if for all* $f \in \mathcal{F}$ *and for all* $c \in (0, 1)$*, we have* $c + (1 - c)f \notin \mathcal{F}$*.*

This very general result is the starting point in considering explicit sets $\mathcal{F}$ of densities that ensure the parameter's identifiability on $(0, 1) \times \mathcal{F}$. In particular, if $\mathcal{F}$ is a set of densities constrained to have essential infimum equal to zero, one recovers the purity result of Genovese & Wasserman (2004). However, from an estimation perspective, the purity assumption is very weak, and it is hopeless to obtain a reliable estimate of θ based on the value of f at a unique value (or at a finite number of values). Because the p-values that are associated with the false null hypotheses are likely to be small and a large majority of the p-values in the interval $[1-\delta, 1]$, for δ not too large, should correspond to the true null hypotheses, the assumption that f is non-increasing with $f(1) = 0$ is reasonable. Recall that this assumption is used in Langaas *et al.* (2005) and partially in Celisse & Robin (2010). In the following, we explore asymptotic efficiency results for the estimation of θ by assuming that the function f belongs to a set of densities (with respect to the Lebesgue measure μ) defined as

$$\begin{aligned}\mathcal{F}_\delta = \Big\{ f : [0, 1] \mapsto \mathbb{R}^+,\ &\text{continuously non-increasing density, positive on } [0, 1 - \delta) \\ &\text{and such that} f_{|[1-\delta,1]} = 0 \Big\}.\end{aligned} \tag{2}$$

1.2. Existing estimators of θ

Let us now discuss the different estimators of θ proposed in the literature, starting with those assuming (implicitly or not) that f attains its minimum value on a whole interval. First, Schweder & Spjøtvoll (1982) suggested a procedure to estimate θ, which has been later used by Storey (2002). This estimator 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 $1 - \lambda$. Storey established that it is a conservative estimator, and one can note that it is consistent only if f attains its minimum value on the interval $[\lambda, 1]$ (an assumption not made in the article by Schweder & Spjøtvoll (1982) nor the one by Storey (2002)). Note that even if such an assumption was made, it would not solve the problem of choosing λ such that f attains its infimum on $[\lambda, 1]$. Adapting this procedure in order to end up with an estimate of the positive FDR, Storey (2002) proposed a bootstrap strategy to pick λ. More precisely, his procedure minimizes the mean squared error (MSE) for estimating the positive FDR. Note that Genovese & Wasserman (2004) established that, for fixed value λ such that the cumulative distribution function F of f

satisfies $F(\lambda) < 1$, Storey's estimator converges at a parametric rate and is asymptotically normal, but is also asymptotically biased: thus, it does not converge to θ at a parametric rate. Some other choices of λ are, for instance, based on break point estimation (Turkheimer *et al.*, 2001) or spline smoothing (Storey & Tibshirani, 2003). Another natural class of procedures in this context is obtained by relying on a histogram estimator of g (Mosig *et al.*, 2001; Nettleton *et al.*, 2006). Among this kind of procedures, we mention the one proposed recently by Celisse & Robin (2010) who proved convergence in probability of their estimator (to the true parameter value) under the assumption that f vanishes on an interval. Note that both Storey's and histogram-based estimators of θ are constructed using non-parametric estimates $\hat{g}$ of the density g and then estimate θ relying on the value of $\hat{g}$ on a specific interval. The main issue with those procedures is to automatically select an interval where the true density g is identically equal to θ. As a conclusion on the existing results for this setup (f vanishing on a non-empty interval), we stress the fact that none of these estimators was proven to be convergent to θ at a parametric rate.

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 $x = 1$. These estimators rely on non-parametric estimates of g and appear to inherit non-parametric rates of convergence. Langaas *et al.* (2005) derived estimators based on non-parametric 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 & Stark (1995) proposed a very general finite sample confidence envelope for a monotone density. Relying on this result and assuming moreover that the cumulative distribution function G of g is concave and that g is Lipschitz in a neighbourhood of $x = 1$, Genovese & Wasserman (2004) constructed 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 (2013) established that by letting $\lambda \to 1$, Storey's estimator may be turned into a consistent estimator of θ, with a non-parametric rate of convergence equal to $n^{-k/(2k+1)}\eta_n$, where $\eta_n \to +\infty$ and k controls the regularity of f near $x = 1$. Our results are in accordance to the literature: no $\sqrt{n}$-consistent estimator has been constructed yet (that is to say, estimators $\hat{\theta}_n$ such that $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ is bounded in probability, denoted by $\sqrt{n}\left(\hat{\theta}_n - \theta\right) = O_{\mathbb{P}}(1)$), as is expected from the fact that the quadratic risk of any estimator of θ cannot converge at a parametric rate in this case (theorem 1).

To finish this tour on the literature about the estimation of θ, we mention that Meinshausen & Bühlmann (2005) discussed probabilistic lower bounds for the proportion of true null hypotheses, which are valid under general and unknown dependence structures between the test statistics.

1.3. *Our results*

We consider the model (2) and distinguish two different cases: δ is positive, and δ is equal to zero. In the first case, we exhibit $\sqrt{n}$-consistent estimators and also compute the asymptotic optimal variance for this problem. In proposition 2, we prove that a very simple histogram-based estimator is $\sqrt{n}$-consistent, while in proposition 3, we establish that this is also true for the more elaborate procedure proposed by Celisse & Robin (2010), which has the advantage of automatically selecting the 'best' partition among a fixed collection. However, we are not aware of a procedure for estimating θ that asymptotically attains the optimal variance in this context. Besides, one might conjecture that such a procedure does not exist for regular models (Section 3.3). In the second case, while the existence of an estimator $\hat{\theta}_n$ of θ converging at a parametric rate has not been established yet, we prove that if such a $\sqrt{n}$-consistent estimator of θ exists, then the variance $\mathbb{V}\text{ar}\left(\sqrt{n}\hat{\theta}_n\right)$ cannot have a finite limit. In other words, the quadratic risk of $\hat{\theta}_n$ cannot converge to zero at a parametric rate. Note that these results are also true when we consider the more general case where the function f either vanishes on a non-empty interval included in $[0, 1]$ (thus not necessarily of the form $[1-\delta, 1]$) or not.

The article is organized as follows. Section 2 establishes lower bounds on the quadratic risk for the estimation of θ, while Section 3 explores corresponding upper bounds, that is, the existence of $\sqrt{n}$-consistent estimators of θ and the existence of asymptotically efficient estimators. Section 4 illustrates our results relying on simulations. The proofs of the main results are postponed to Section 5, while some technical lemmas are proved in Appendix A.

2. Lower bounds for the quadratic risk and efficiency

In this section, we give lower bounds for the quadratic risk of any estimator of θ. For any fixed unknown parameter $\delta \in [0, 1)$, we introduce an induced set of semiparametric distributions $\mathcal{P}_\delta$ defined as

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

where $\mathcal{F}_\delta$ has been defined in (2). Note that for any fixed value $\delta \in [0, 1)$, the condition stated in proposition 1 is satisfied on the set $\mathcal{F}_\delta$; namely, for all $f \in \mathcal{F}_\delta$ and for all $c \in (0, 1)$, we have $c + (1-c)f \notin \mathcal{F}_\delta$. Thus, the parameter (θ, f) is identifiable on $(0, 1) \times \mathcal{F}_\delta$.

We follow notation form Chapter 25 and more particularly Section 25.4 in van der Vaart (1998) and refer to this book. More precise definitions of the objects involved will also be given in Section 5.2 together with the proof of the main result. We let $\dot{\mathcal{P}}_\delta$ denote a tangent set of the model $\mathcal{P}_\delta$ at $\mathbb{P}_{\theta,f}$ with respect to the parameter (θ, f). For every score function g in the tangent set $\dot{\mathcal{P}}_\delta$, we write $P_{t,g}$ for a path with score function g. Namely, $P_{t,g}$ equals $\mathbb{P}_{\theta+ta,f_t}$ for some path $t \mapsto f_t$ and some $a \in \mathbb{R}$.

Now, an estimator sequence $\hat{\theta}_n$ is called regular at $\mathbb{P}_{\theta,f}$ for estimating θ (relative to the tangent set $\dot{\mathcal{P}}_\delta$) if there exists a probability measure L such that for any score function $g \in \dot{\mathcal{P}}_\delta$ corresponding to a path of the form $t \mapsto (\theta + ta, f_t)$, we have

$$\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right) = \sqrt{n}\left[\hat{\theta}_n - \left(\theta + \frac{a}{\sqrt{n}}\right)\right] \xrightarrow{d} L, \text{ under } P_{1/\sqrt{n},g},$$

where $\xrightarrow{d}$ denotes convergence in distribution. According to a convolution theorem (see theorem 25.20 in van der Vaart, 1998), this limit distribution can be written as the convolution between some unknown distribution and the centred Gaussian distribution $N\left(0, \mathbb{P}_{\theta,f}\left(\tilde{\psi}^2_{\theta,f}\right)\right)$ with variance

$$\mathbb{P}_{\theta,f}\left(\tilde{\psi}^2_{\theta,f}\right) = \int \tilde{\psi}^2_{\theta,f} d\mathbb{P}_{\theta,f},$$

where $\tilde{\psi}_{\theta,f}$ is the efficient influence function. Thus, we say that an estimator sequence is asymptotically efficient at $\mathbb{P}_{\theta,f}$ (relative to the tangent set $\dot{\mathcal{P}}_\delta$) if it is regular at $\mathbb{P}_{\theta,f}$ with limit distribution $L = N\left(0, \mathbb{P}_{\theta,f}\left(\tilde{\psi}^2_{\theta,f}\right)\right)$; in other words, it is the best regular estimator.

We define the quadratic risk of an estimator sequence $\hat{\theta}_n$ (relative to the tangent set $\dot{\mathcal{P}}_\delta$) as

$$\sup_{E_\delta} \liminf_{n\to\infty} \sup_{g \in E_\delta} P_{1/\sqrt{n},g}\left[\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right)\right]^2,$$

where the first supremum is taken over all finite subsets E_δ of the tangent set $\dot{\mathcal{P}}_\delta$. According to the local asymptotic minimax theorem (see theorem 25.21 in van der Vaart, 1998), this quantity is lower bounded by the minimal variance $\mathbb{P}_{\theta,f}\left(\tilde{\psi}^2_{\theta,f}\right)$.

Moreover, according to lemma 25.23 in van der Vaart (1998), an estimator $\hat{\theta}_n$ of θ is asymptotically efficient if and only if

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_{\theta,f}(X_i) + o_{\mathbb{P}_{\theta,f}}(1).$$

Hence, an asymptotically efficient estimator is asymptotically normal with asymptotic variance equal to the optimal variance.

Theorem 1.

(1) *When $\delta = 0$, there is no regular estimator for θ relative to the tangent set $\dot{\mathcal{P}}_0$, and any estimator sequence $\hat{\theta}_n$ has an infinite quadratic risk, namely*

$$\sup_{E_0} \liminf_{n\to\infty} \sup_{g\in E_0} \mathbb{E}_{P_{1/\sqrt{n},g}} \left[\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right)\right]^2 = +\infty,$$

where the first supremum is taken over all finite subsets E_0 of the tangent set $\dot{\mathcal{P}}_0$.

(2) *When $\delta > 0$, we obtain that*

(i) *For any estimator sequence $\hat{\theta}_n$,*

$$\sup_{E_\delta} \liminf_{n\to\infty} \sup_{g\in E_\delta} \mathbb{E}_{P_{1/\sqrt{n},g}} \left[\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right)\right]^2 \geq \theta\left(\frac{1}{\delta} - \theta\right),$$

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

(ii) *A sequence of estimators $\hat{\theta}_n$ is asymptotically efficient if and only if it satisfies*

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\delta} \mathbf{1}_{X_i \in [1-\delta,1]} + o_{\mathbb{P}_{\theta,f}}\left(n^{-1/2}\right). \tag{3}$$

Let us now comment on this theorem. The case where f vanishes on a non-empty interval ($\delta > 0$) appears to be easier from an estimation perspective. Otherwise (f vanishing at most on isolated points), it is usual to add assumptions on f. Here, we choose to consider the case where f is assumed to be non-increasing (see definition (2) of $\mathcal{F}_\delta$). Similar results may be obtained by replacing this assumption with a regularity constraint on f. Note also that when $\delta > 0$, the assumption that f is non-increasing could be removed without any change in our results.

When $\delta = 0$, we obtain that if there exists a $\sqrt{n}$-consistent estimator in model $\mathcal{P}_0$, it can not have finite asymptotic variance. In other words, we could have $\sqrt{n}\left(\hat{\theta}_n - \theta\right) = O_{\mathbb{P}}(1)$ for some estimator $\hat{\theta}_n$ but then $\mathbb{V}\mathrm{ar}\left(\sqrt{n}\hat{\theta}_n\right) \to +\infty$. However, we note that the only rates of convergence obtained until now in this case are non-parametric ones.

When $\delta > 0$, for fixed parameter value λ such that $G(\lambda) < 1$, Storey's estimator $\hat{\theta}^{\text{Storey}}(\lambda)$ satisfies

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

(see, for instance, Genovese & Wasserman, 2004). In particular, if we assume that f vanishes on $[\lambda, 1]$, then we obtain that $G(\lambda) = 1-\theta(1-\lambda)$ and $\hat{\theta}^{\text{Storey}}(\lambda)$ becomes a $\sqrt{n}$-consistent estimator of θ, which is moreover asymptotically normal, with asymptotic variance $\theta\left((1-\lambda)^{-1} - \theta\right)$. 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 $\hat{\theta}^{\text{Storey}}(\lambda)$ automatically satisfies (3).

3. Upper bounds for the quadratic risk and efficiency (when $\delta > 0$)

In this section, we investigate the existence of asymptotically efficient estimators for θ, in the case where $\delta > 0$. We consider histogram-based estimators of θ where a non-parametric histogram estimator $\hat{g}$ of g is combined with an interval selection that aims at picking an interval where g is equal to θ. We start by establishing the existence of $\sqrt{n}$-consistent estimators: a simple histogram-based procedure is studied in Section 3.1, while a more elaborate one is the object of Section 3.2. Finally, in Section 3.3, we explain the general one-step method to construct an asymptotically efficient estimator relying on a $\sqrt{n}$-consistent procedure and discuss conditions under which an asymptotically efficient estimator could be obtained in model $\mathcal{P}_\delta$.

Note that we will assume that the density f belongs to $\mathcal{F}_\delta$ with $\delta > 0$ throughout the current section. However, the results are easily generalized to the case where f vanishes on a non-empty interval included in [0, 1] and is monotone outside this interval.

3.1. A histogram-based estimator

Let $\hat{g}_I$ be a histogram estimator corresponding to a partition $I = (I_k)_{1,\ldots,D}$ 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 $n_k = \operatorname{card}\{i : X_i \in I_k\}$ is the number of observations in I_k and $|I_k|$ is the width of interval I_k. We estimate θ by the minimal value of $\hat{g}_I$; that is,

$$\hat{\theta}_{I,n} = \min_{1\le k\le D} \frac{n_k}{n|I_k|} = \frac{n_{\hat{k}_n}}{n\left|I_{\hat{k}_n}\right|}, \tag{4}$$

where we let

$$\hat{k}_n \in \operatorname*{Argmin}_{1\le k\le D} \left\{ \frac{n_k}{n|I_k|} = \frac{1}{n|I_k|} \sum_{i=1}^{n} \mathbf{1}_{X_i \in I_k} \right\}.$$

Note that histogram estimators are natural non-parametric estimators for g when assuming that $f \in \mathcal{F}_\delta$ with $\delta > 0$; that is, g is constant on an interval. It is easy to see that $\hat{\theta}_{I,n}$ is almost surely consistent as soon as the partition I is fine enough. We moreover establish that this estimator has the MSE of the order $1/n$. The proof of this result appears in Section 5.3.

Theorem 2. *Fix $\delta > 0$ and suppose that $f \in \mathcal{F}_\delta$. Assume moreover that the partition I is such that $\max_k |I_k|$ is small enough, then the estimator $\hat{\theta}_{I,n}$ has the following properties*

(i) *$\hat{\theta}_{I,n}$ converges almost surely to θ,*

(ii) $\limsup_{n\to\infty} n\mathbb{E}\left[\left(\hat{\theta}_{I,n} - \theta\right)^2\right] < +\infty.$

Note that because $\hat{\theta}_{I,n}$ has the MSE of the order $1/n$, we can deduce that $\hat{\theta}_{I,n}$ is $\sqrt{n}$-consistent and has a variance of the order $1/n$. However, asymptotic normality of $\hat{\theta}_{I,n}$ or the value of its asymptotic variance are difficult to obtain. Indeed, for any deterministic interval I_k, the central limit theorem (CLT) applies on the estimator $n_k/(n|I_k|)$. But, a histogram-based estimator such as $\hat{\theta}_{I,n}$ is based on the selection of a random interval $\hat{I}$, and the CLT fails to apply directly on $n_{\hat{I}}/\left(n|\hat{I}|\right)$. Note also that the choice of the partition I is not solved here.

From a practical point of view, decreasing the parameter $\max_k |I_k|$ will in fact increase the variance of the estimator. In the next section, we study a procedure that automatically selects the best partition among a given collection.

3.2. Celisse and Robin's procedure

We recall here the procedure for estimating θ that is presented in Celisse & Robin (2010). It relies on an elaborate histogram approach that selects the best partition among a given collection. As it will be seen from the simulations experiments (Section 4), its asymptotic variance is likely to be smaller than for the previous estimator, justifying our interest into this procedure. Unfortunately, from a theoretical point of view, we only establish that this estimator should be as good as the previous one. Note that because not many estimators of θ have been proved to be $\sqrt{n}$-convergent, this is already a non-trivial result.

For a given integer M, define $\mathcal{I}_M$ as the set of partitions of [0, 1] such that for some integer k with $1 \leq k \leq M-2$, the first k intervals are regular of width $1/M$, and the last one is of width $(M-k)/M$, namely,

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

These partitions are motivated by the assumption that f vanishes on a set $[1-\delta, 1] \subset [0,1]$. Then for two given integers $m_{\min} < m_{\max}$, denote by $\mathcal{I}$ the following collection of partitions

$$\mathcal{I} = \bigcup_{m_{\min} \leq m \leq m_{\max}} \mathcal{I}_{2^m}. \tag{5}$$

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

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

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

$$\hat{R}_p(I) = \frac{2n-p}{(n-1)(n-p)} \sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} \right)^2. \tag{6}$$

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

$$p^\star(I) = \operatorname*{argmin}_{p \in \{1,\dots,n-1\}} MSE(p, I) = \operatorname*{argmin}_{p \in \{1,\dots,n-1\}} \mathbb{E}\left[\left(\hat{R}_p(I) - R(I) \right)^2 \right].$$

It clearly appears that $MSE(p, I)$ has the form of a function $\Phi(p, I, \alpha)$ (see Celisse & 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 is then 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

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

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

(1) For each partition $I \in \mathcal{I}$, define $\hat{p}(I) = \underset{p \in \{1,\dots,n-1\}}{\operatorname{argmin}} \widehat{MSE}(p, I)$,

(2) Choose $\hat{I} = \left(\hat{M}, \hat{\lambda}\right) \in \underset{I \in \mathcal{I}}{\operatorname{argmin}} \hat{R}_{\hat{p}(I)}(I)$ such that the width of the interval $[\hat{\lambda}, 1]$ is maximum,

(3) Estimate θ by $\hat{\theta}_n^{CR} = \operatorname{card}\left\{i : X_i \in \left[\hat{\lambda}, 1\right]\right\} / \left[n\left(1 - \hat{\lambda}\right)\right]$.

Remark 3.1. In our procedure, we consider the set of natural partitions defined by (5), while Celisse & Robin (2010) used the one defined by

$$\mathcal{I} = \bigcup_{M_{\min} \le M \le M_{\max}} \mathcal{I}_M,$$

where $\mathcal{I}_M$ is the set of partitions of [0, 1] such that the first k intervals and the last $M - l$ ones are regular of width $1/M$, for some integers k, l with $2 \le k + 2 \le l \le M$,

$$\mathcal{I}_M = \left\{ I = (I_i)_i : \forall i \neq k+1, |I_i| = \frac{1}{M}, |I_{k+1}| = \frac{l-k}{M}, 2 \le k+2 \le l \le M \right\}.$$

This change is natural for lowering the complexity of the algorithm and has no consequences on the theoretical properties of the estimator.

In Celisse & Robin (2010), the authors only established convergence in probability of this estimator. Here, we prove its almost sure convergence, $\sqrt{n}$-consistency and establish that its variance is of the order $1/n$. We now introduce a technical condition that comes from Celisse & Robin (2010). We let

$$\forall (i, j) \in \mathbb{N}^2, \quad s_{ij} = \sum_{k=1}^{D} \frac{\alpha_k^i}{|I_k|^j},$$

and further assume that the collection of partitions $\mathcal{I}$ and density f are such that

$$\forall I \in \mathcal{I}, \quad 8s_{11}s_{21} - 2s_{11}^2 + 8s_{32} - 10s_{21}^2 - 4s_{22} \neq 0,\ s_{21} - s_{22} - s_{32} + 3s_{11} \neq 0. \tag{7}$$

This technical condition is used in Celisse & Robin (2010) to control the behaviour of the minimizer $\hat{p}(I)$. We are now ready to state our result, whose proof can be found in Section 5.4.

Theorem 3. *Suppose that f satisfies the technical condition (7) and f belongs to $\mathcal{F}_\delta$. Assume moreover that $m_{\max}$ is large enough, then the estimator $\hat{\theta}_n^{CR}$ has the following properties:*

(i) $\hat{\theta}_n^{CR}$ converges almost surely to θ;

(ii) $\hat{\theta}_n^{CR}$ is $\sqrt{n}$-consistent, that is, $\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right) = O_{\mathbb{P}}(1)$;

(iii) If p is fixed, then $\limsup_{n \to \infty} n\mathbb{E}\left[\left(\hat{\theta}_n^{CR} - \theta\right)^2\right] < +\infty$.

Here again, asymptotic normality of $\hat{\theta}_n^{CR}$ or the exact value of its asymptotic variance is difficult to obtain. Heuristically, one can explain that this procedure outperforms the simpler histogram based with fixed partition approach described in the previous section. Indeed, when considering a fixed partition, the latter should be fine enough to obtain convergence but refining the partition increases the variance of $\hat{\theta}_{I,n}$. Here, Celisse and Robin's approach realizes a compromise on the size of the partition that is used.

3.3. Existence of asymptotically efficient estimators

In this section, we introduce the one-step method, a general procedure that aims at constructing an asymptotically efficient estimator relying on a $\sqrt{n}$-consistent one (see van der Vaart, 1998, Section 25.8). Note that if an asymptotically efficient estimator exists, then it can always be constructed by the one-step method, but the method works under conditions that are not always satisfied. Here again, we use terminology from semiparametric theory. 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 $\hat{l}_{n,\theta}(\cdot) = \hat{l}_{n,\theta}(\cdot; X_1, \ldots, X_n)$ of the efficient score function $\tilde{l}_{\theta,f}$ (an expression of the efficient score function in our context is given in Section 5.2). Define with $m = \lfloor n/2 \rfloor$,

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

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

$$\sqrt{n}\mathbb{P}_{\theta_n,f}\hat{l}_{n,\theta_n} \xrightarrow[n\to\infty]{\mathbb{P}_{\theta,f}} 0, \tag{8}$$

$$\mathbb{P}_{\theta_n,f}\|\hat{l}_{n,\theta_n} - \tilde{l}_{\theta_n,f}\|^2 \xrightarrow[n\to\infty]{\mathbb{P}_{\theta,f}} 0, \tag{9}$$

$$\int \|\tilde{l}_{\theta_n,f}d\mathbb{P}^{1/2}_{\theta_n,f} - \tilde{l}_{\theta,f}d\mathbb{P}^{1/2}_{\theta,f}\|^2 \xrightarrow[n\to\infty]{0} . \tag{10}$$

Note that in the aforementioned notation, the term $\mathbb{P}_{\theta_n,f}\hat{l}$ for some random function $\hat{l}$ is an abbreviation for the integral $\int \hat{l}(x)d\mathbb{P}_{\theta_n,f}(x)$. Thus, the expectation is taken with respect to x only and not the random variables in $\hat{l}$. Now under the aforementioned assumptions, the one-step estimator defined as

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

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 $\hat{\theta}_n$.

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

Under condition (10), it is shown in theorem 7.4 from van der Vaart (2002) 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 (8) and (9). Thus, in this case, if an asymptotically efficient estimator sequence exists, then it can always be constructed by the one-step method. In our case, it is not difficult to prove that condition (10) holds. Then, the estimator $\hat{l}_{n,\theta}$ of the efficient score function $\tilde{l}_{\theta,f}$ must satisfy both a 'no-bias' (8) and a consistency (9) 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 $1/\sqrt{n}$. We thus obtain the following proposition, whose proof can be found in Section 5.3.

Proposition 2. *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 (8) and (9). Moreover, if the efficient score function $\tilde{l}_{\theta,f}$ is estimated through a plug-in method that relies on an estimate $\hat{\delta}_n$ of the parameter δ, then this condition is equivalent to $\sqrt{n}\left(\hat{\delta}_n - \delta\right) = o_{\mathbb{P}}(1)$.*

Let us now explain the consequences of this result. The proposition states that efficient estimators of θ exist if and only if estimators of $\tilde{l}_{\theta,f}$ that satisfy (8) and (9) can be constructed. As there is no general method to estimate an efficient score function, such an estimator should rely on the specific expression (15). Although we cannot claim that all estimators of $\tilde{l}_{\theta,f}$ are plug-in estimates based on an estimator of the parameter $\hat{\delta}$ plugged into expression (15), it is likely to be the case. Then, existence of efficient estimators of θ is equivalent to existence of estimators of δ that converge faster than the parametric rate. Note that this is possible for irregular models (see Chapter 6 in Ibragimov & Has'minskiĭ, 1981, for more details). However, for regular models, such estimators cannot be constructed, and one might conjecture that efficient estimators of θ do not exist in regular models.

4. Simulations

In this section, we give some illustrations of the previous results on some simulated experiments and explore the non-asymptotic performances of the estimators of θ previously discussed. We choose to compare three different estimators: the histogram-based estimator $\hat{\theta}_{I,n}$ defined in Section 3.1 through (4), the more elaborate histogram-based estimator $\hat{\theta}_n^{CR}$ proposed in Celisse & Robin (2010) and finally Langaas *et al.* (2005)'s estimator, denoted by $\hat{\theta}_n^{L}$ and defined as the value $\hat{g}(X_{(n)})$ where $X_{(n)}$ is the largest p-value and $\hat{g}$ is Grenander's estimator of a decreasing density. We investigate the behaviour of these three different estimators of θ under two different setups: $\delta = 0$ and $\delta \in (0,1)$. More precisely, we consider the alternative density f given by

$$f(x) = \frac{s}{1-\delta}\left(1 - \frac{x}{1-\delta}\right)^{s-1} \mathbf{1}_{[0,1-\delta]}(x),$$

where $\delta \in [0,1)$ and $s > 1$. This form of density is introduced in Celisse & Robin (2010) and covers various situations when varying its parameters. Note that f is always decreasing, convex when $s \geq 2$ and concave when $s \in (1,2]$. In the experiments, we consider a total of 8 different models corresponding to different parameter values. These models are labelled as described in Table 1, distinguishing the cases $\delta = 0$ and $\delta > 0$. As an illustration, we represent some of the densities obtained for the p-values corresponding to 4 out of the 8 models in Figure 1. For each estimator $\hat{\theta}_n$ of θ, we compare the quantity $n\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$ with the optimal variance $\theta\left(\delta^{-1} - \theta\right)$ when this bound exists. Equivalently, we compare the logarithm of MSE, $\log(\text{MSE}) = \log \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$ for each estimator $\hat{\theta}_n$ with $-\log(n) + \log\left[\theta\left(\delta^{-1} - \theta\right)\right]$.

Table 1. *Labels of the 8 models with different parameter values*

(s,θ)	$\delta = 0.3$	$\delta = 0$
(3,0.6)	(a_1)	(a_2)
(3,0.8)	(b_1)	(b_2)
(1.4,0.7)	(c_1)	(c_2)
(1.4,0.9)	(d_1)	(d_2)

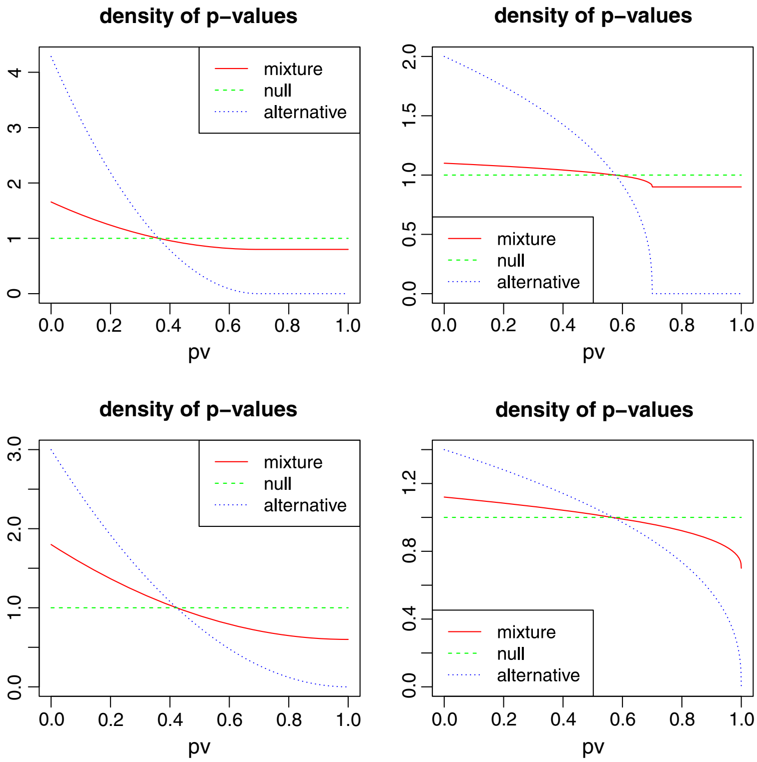


Fig. 1. Density function of the p-values. Top left: model (b_1); top right: model (d_1); bottom left: model (a_2); bottom right: model (c_2).

When $\delta = 0$, we only compare the slope of the line induced by log(MSE) with the parametric rate corresponding to a slope -1. In each case, we simulated data with sample size $n \in \{5000, 7000, 9000, 10000, 12000, 14000, 15000\}$ and perform $R = 100$ repetitions.

When computing the estimator $\hat{\theta}_{I,n}$, the choice of the partition I surely affects the results. Here, we have chosen a regular partition I such that it is fine enough (we fixed $|I_k| < \delta$) but not too fine (choosing a too small value of $|I_k|$ increases the variance). The choice of the partition in the simple procedure $\hat{\theta}_{I,n}$ is an issue for real data problems. Our goal here is to show that on simulated experiments, the 'best' of these estimators still has a larger variance than $\hat{\theta}_n^{CR}$. Note that the partition I is always included in the collection $\mathcal{I}$ of partitions from which $\hat{\theta}_n^{CR}$ is computed.

The results are presented in Figure 2 for the case $\delta > 0$ and Figure 3 for the case $\delta = 0$. First, we note that in both cases ($\delta > 0$ and $\delta = 0$), the estimator of Langaas *et al.* $\hat{\theta}_n^L$ has non-parametric rate of convergence (null slope) and performs badly compared with $\hat{\theta}_{I,n}$ and $\hat{\theta}_n^{CR}$. In particular, when $\delta = 0$, the two histogram-based procedures $\hat{\theta}_{I,n}$ and $\hat{\theta}_n^{CR}$ have better performances than the estimator $\hat{\theta}_n^L$ despite the fact that the latter is dedicated to the convex decreasing setup. Now, when $\delta > 0$, both estimators $\hat{\theta}_{I,n}$ and $\hat{\theta}_n^{CR}$ exhibit a parametric rate of convergence (slope equal to -1). Moreover, $\hat{\theta}_n^{CR}$ has a smaller variance than $\hat{\theta}_{I,n}$ (smaller intercept), and this variance is very close to the optimal one $\theta(\delta^{-1} - \theta)$. Now, when $\delta = 0$, we observe two different behaviours depending on whether f is convex or not. Indeed, for models (a_2) and (b_2) corresponding to the convex case, we observe that both estimators $\hat{\theta}_{I,n}$ and $\hat{\theta}_n^{CR}$ still exhibit a parametric rate of convergence, with a smaller variance for $\hat{\theta}_n^{CR}$. These estimators

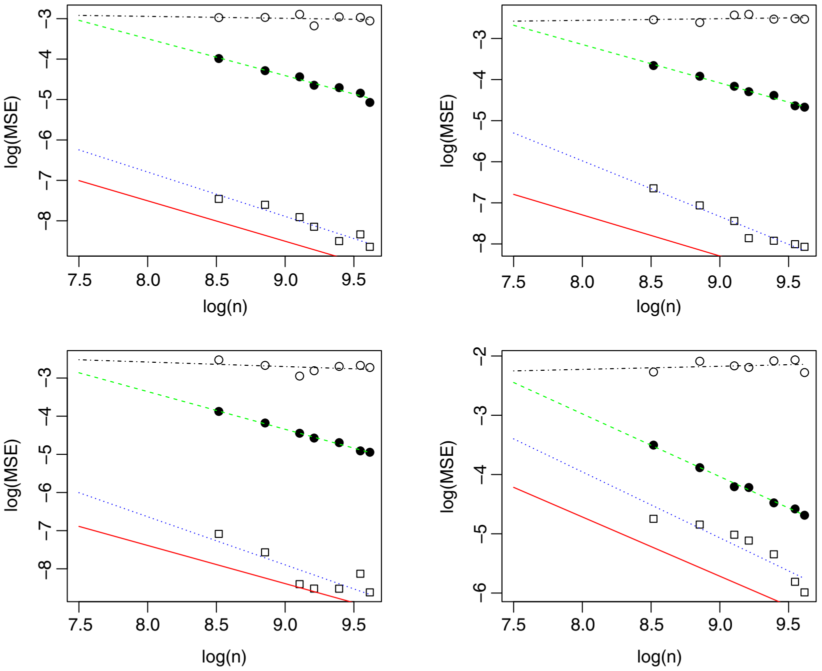


Fig. 2. Logarithm of the mean squared error as a function of $\log(n)$ and corresponding linear regression for $\hat{\theta}_n^L$ (○ and black line, respectively), $\hat{\theta}_n^{CR}$ (□ and blue line, respectively) and $\hat{\theta}_{I,n}$ (● and green line, respectively) in the case $\delta = 0.3$, for different parameter values: (a_1) top left; (b_1) top right; (c_1) bottom left; (d_1) bottom right. Red line represents the line $y = -\log(n) + \log\left[\theta(\delta^{-1} - \theta)\right]$.

are thus robust to the assumption that f vanishes on an interval in the convex setup. The results are slightly different when considering models (c_2) and (d_2) where f is now concave. These estimators have a more erratic behaviour, exhibiting either parametric rate of convergence ($\hat{\theta}_n^{CR}$ in model (c_2) and $\hat{\theta}_{I,n}$ in model (d_2)) or non-parametric rates. Their respective performances in terms of variance are also less clear. Nonetheless, we conclude that $\hat{\theta}_n^{CR}$ seems to exhibit the overall best performances, with the parametric rate of convergence and almost optimal asymptotic variance.

5. Proofs

5.1. Proof of proposition 1

Sufficiency: Let us suppose that for all $f \in \mathcal{F}$ and for all $c \in (0, 1)$, we have $c + (1 - c)f \notin \mathcal{F}$. We prove that the parameters θ and f are identifiable on the set $(0, 1) \times \mathcal{F}$ by contradiction. Suppose that there exist (θ_1, f_1) and $(\theta_2, f_2) \in \mathcal{F}$, $(\theta_1, f_1) \neq (\theta_2, f_2)$ such that

$$\theta_1 + (1 - \theta_1) f_1(x) = \theta_2 + (1 - \theta_2) f_2(x), \text{ for all } x \in [0, 1]. \tag{11}$$

We can always consider $\theta_1 > \theta_2$. Let us denote by $c = (\theta_1 - \theta_2)/(1 - \theta_2)$, then $c \in (0, 1)$. We obtain that

$$\theta_1 + (1 - \theta_1) f_1(x) = \theta_2 + (1 - \theta_2)(c + (1 - c) f_1(x)), \text{ for all } x \in [0, 1]. \tag{12}$$

From (11) and (12), we have $f_2 = c + (1 - c) f_1$, it means that there exist $f_1 \in \mathcal{F}$ and $c \in (0, 1)$ such that $c + (1 - c) f_1 \in \mathcal{F}$. So we have a contradiction.

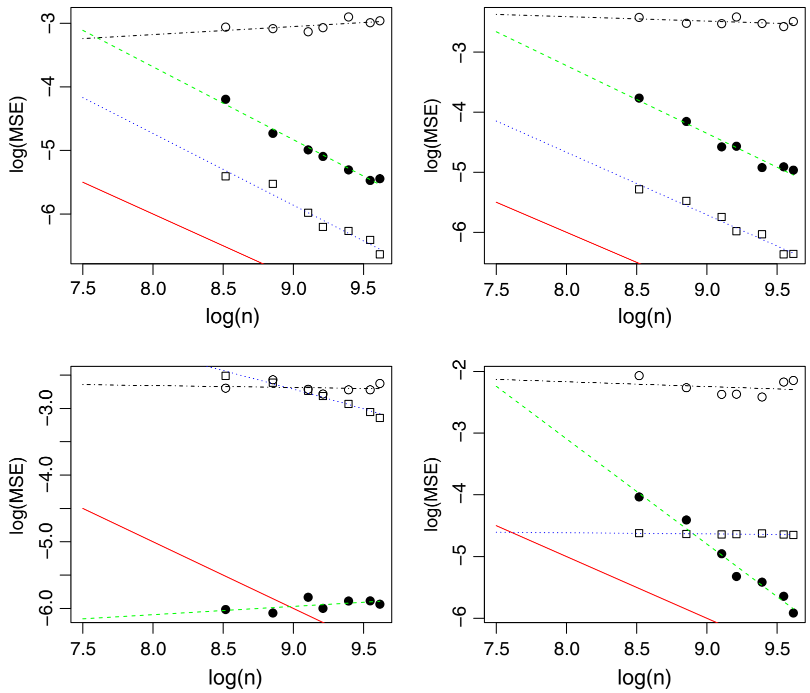


Fig. 3. Logarithm of the mean squared error as a function of $\log(n)$ and corresponding linear regression for $\hat{\theta}_n^L$ (○ and black line, respectively), $\hat{\theta}_n^{CR}$ (□ and blue line, respectively) and $\hat{\theta}_{I,n}$ (● and green line, respectively) in the case $\delta = 0$, for different parameter values: (a_2) top left; (b_2) top right; (c_2) bottom left; (d_2) bottom right. Red line represents the line $y = -\log(n) + c$ for some well-chosen constant c.

Necessity: Suppose that the parameters θ and f are identifiable on the set $(0,1) \times \mathcal{F}$. We prove by contradiction that for all $f \in \mathcal{F}$ and for all $c \in (0,1)$, we have $c + (1-c)f \notin \mathcal{F}$. Indeed, suppose that there exist $f \in \mathcal{F}$ and $c \in (0,1)$ such that $c + (1-c)f \in \mathcal{F}$. For all $\theta_1 \in (0,1)$, we denote $\theta_2 = c + (1-c)\theta_1$, then we obtain

$$\theta_1 + (1-\theta_1)(c + (1-c)f(x)) = \theta_2 + (1-\theta_2)f(x), \text{ for all } x \in [0,1].$$

This implies that θ and f are not identifiable on the set $(0,1) \times \mathcal{F}$.

5.2. *Proof of theorem 1*

Let us first describe more precisely the objects arising from semiparametric theory in our setting. Fix a parameter value (θ, f) and consider first a parametric submodel of $\mathcal{F}_\delta$ induced by the following path:

$$t \mapsto f_t(x) = c(t)k(th_0(x))f(x), \tag{13}$$

where h_0 is a continuous and non-increasing function on $[0,1]$, the function k is defined by $k(u) = 2(1+e^{-2u})^{-1}$ and the normalizing constant $c(t)$ satisfies $c(t)^{-1} = \int k(th_0(u))f(u)du$. A tangent set ${}_f\dot{\mathcal{P}}_\delta$ for the parameter f is composed of the score functions associated to such parametric submodels (as h_0 varies). It is easy to see that the path (13) is differentiable and that its corresponding score function is obtained by differentiating $t \mapsto \log[\theta + (1-\theta)f_t(x)]$ at $t = 0$. We thus obtain a tangent set for f given by

$$_f\dot{\mathcal{P}}_\delta = \left\{ h = \frac{(1-\theta) f h_0}{\theta + (1-\theta) f}; h_0 \text{ is continuous and non-increasing on } [0, 1-\delta) \text{ with}, \int f h_0 = 0 \right\}.$$

We consider parametric submodels of $\mathcal{P}_\delta$ induced by paths of the form $t \mapsto \mathbb{P}_{\theta+ta, f_t}$ where the paths $t \mapsto f_t$ in $\mathcal{F}_\delta$ are given by (13). 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 $h \in {}_f\dot{\mathcal{P}}_\delta$, we have that $a\dot{l}_{\theta,f} + h$ is a score function for (θ, f) corresponding to the path $t \mapsto \mathbb{P}_{\theta+ta, f_t}$. Hence, a tangent set $\dot{\mathcal{P}}_\delta$ of the model $\mathcal{P}_\delta$ at $\mathbb{P}_{\theta,f}$ with respect to the parameter (θ, f) is given by the linear span

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

Moreover, the ordinary score function $\dot{l}_{\theta,f}$ for θ in the model in which f is fixed is given by

$$\dot{l}_{\theta,f}(x) = \frac{\partial}{\partial\theta} \log[\theta + (1-\theta) f(x)] = \frac{1 - f(x)}{\theta + (1-\theta) f(x)}. \tag{14}$$

Now we let $\tilde{l}_{\theta,f}$ be the efficient score function and $\tilde{I}_{\theta,f}$ be the efficient information for estimating $\psi(\mathbb{P}_{\theta,f}) = \theta$. These quantities are defined respectively as

$$\tilde{l}_{\theta,f} = \dot{l}_{\theta,f} - \Pi_{\theta,f} \dot{l}_{\theta,f} \text{ and } \tilde{I}_{\theta,f} = \mathbb{P}_{\theta,f}\left(\tilde{l}^2_{\theta,f}\right),$$

where $\Pi_{\theta,f}$ is the orthogonal projection onto the closure of the linear span of ${}_f\dot{\mathcal{P}}_\delta$ in $\mathbb{L}_2(\mathbb{P}_{\theta,f})$. The functional $\psi : \mathbb{P}_{\theta,f} \mapsto \theta$ is said to be differentiable at $\mathbb{P}_{\theta,f}$ relative to the tangent set $\dot{\mathcal{P}}_\delta$ if there exists a continuous linear map $\tilde{\psi}_{\theta,f} : \mathbb{L}_2(\mathbb{P}_{\theta,f}) \mapsto \mathbb{R}$, called the efficient influence function, such that for every path $t \mapsto f_t$ with score function $h \in {}_f\dot{\mathcal{P}}_\delta$, we have

$$\forall a \in \mathbb{R}, \quad a = \int \tilde{\psi}_{\theta,f}(x) \left[a^\intercal \dot{l}_{\theta,f}(x) + h(x)\right] d\mathbb{P}_{\theta,f}(x).$$

Setting $a = 0$, we see that this efficient influence function must be orthogonal to the tangent set ${}_f\dot{\mathcal{P}}_\delta$. Finally, note that under some assumptions, the efficient influence function $\tilde{\psi}_{\theta,f}$ equals $\tilde{I}^{-1}_{\theta,f} \tilde{l}_{\theta,f}$ (see lemma 25.25 in van der Vaart, 1998). The following proposition provides expressions for these quantities in our setup.

Proposition 3. *The efficient score function $\tilde{l}_{\theta,f}$ and the efficient information $\tilde{I}_{\theta,f}$ for estimating θ in model $\mathcal{P}_\delta$ are given by*

$$\tilde{l}_{\theta,f}(x) = \frac{1}{\theta} - \frac{1}{\theta(1-\theta\delta)} \mathbf{1}_{[0,1-\delta)}(x) \quad \textit{and} \quad \tilde{I}_{\theta,f} = \frac{\delta}{\theta(1-\theta\delta)}, \tag{15}$$

where $\mathbf{1}_A(\cdot)$ is the indicator function of set A. When $\delta > 0$, the efficient influence function $\tilde{\psi}_{\theta,f}$ relative to the tangent set $\dot{\mathcal{P}}_\delta$ is given by

$$\tilde{\psi}_{\theta,f}(x) = \frac{1}{\delta} \mathbf{1}_{[1-\delta,1]}(x) - \theta.$$

Proof of proposition 3. The ordinary score function $\dot{l}_{\theta,f}$ can be written as

$$\begin{aligned}\dot{l}_{\theta,f}(x) &= \frac{\partial}{\partial\theta}\log[\theta+(1-\theta)f(x)]\\ &= \left(\frac{1-f(x)}{\theta+(1-\theta)f(x)}+\frac{\delta}{1-\theta\delta}\right)\mathbf{1}_{[0,1-\delta)}(x)+\frac{1}{\theta}\mathbf{1}_{[1-\delta,1]}(x)-\frac{\delta}{1-\theta\delta}\mathbf{1}_{[0,1-\delta)}(x).\end{aligned} \tag{16}$$

Let us recall that $\Pi_{\theta,f}$ is the orthogonal projection onto the closure of the linear span of ${}_f\dot{\mathcal{P}}_\delta$ in $\mathbb{L}_2(\mathbb{P}_{\theta,f})$. We prove that the orthogonal projection of $\dot{l}_{\theta,f}$ onto this space is equal to the first term appearing in the right-hand side of (16), namely,

$$\Pi_{\theta,f}\dot{l}_{\theta,f}(x)=\left(\frac{1-f(x)}{\theta+(1-\theta)f(x)}+\frac{\delta}{1-\theta\delta}\right)\mathbf{1}_{[0,1-\delta)}(x), \tag{17}$$

and then the efficient score function for θ is

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

In fact, we can write

$$-\left(\frac{1-f}{\theta+(1-\theta)f}+\frac{\delta}{1-\theta\delta}\right)\mathbf{1}_{[0,1-\delta)}=\frac{(1-\theta)fh_0}{\theta+(1-\theta)f},$$

where

$$h_0(x)=\frac{1}{(1-\theta)(1-\theta\delta)}\left(1-\delta-\frac{1}{f(x)}\right)\mathbf{1}_{[0,1-\delta)}(x).$$

The function h_0 is continuous and decreasing on $[0,1-\delta)$. It is not difficult to examine the condition $\int fh_0=0$. Hence,

$$\left(\frac{1-f}{\theta+(1-\theta)f}+\frac{\delta}{1-\theta\delta}\right)\mathbf{1}_{[0,1-\delta)} \text{ belongs to } \overline{\operatorname{lin}\left({}_f\dot{\mathcal{P}}_\delta\right)}.$$

Now, to conclude the proof of (17), it is necessary to establish that the second term in the right-hand side of (16) is orthogonal to the closure of the linear span of ${}_f\dot{\mathcal{P}}_\delta$, namely

$$\frac{1}{\theta}\mathbf{1}_{[1-\delta,1]}-\frac{\delta}{1-\theta\delta}\mathbf{1}_{[0,1-\delta)}=\frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{[0,1-\delta)}-\frac{\delta}{1-\theta\delta}\perp\ \overline{\operatorname{lin}\left({}_f\dot{\mathcal{P}}_\delta\right)},$$

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

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

the scalar product between h and the remaining term in (16) is given by

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

This establishes (17). Let us now calculate the efficient information

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

We now turn to the particular case where $\delta = 0$. In this case, the previous computations show that $\dot{l}_{\theta,f}$ belongs to the closure of the linear span of ${}_f\dot{\mathcal{P}}_\delta$ and that the Fisher information is zero. When $\delta > 0$, the Fisher information is positive and the efficient influence function is given by

$$\tilde{\psi}_{\theta,f}(x) = \tilde{I}_{\theta,f}^{-1}\tilde{l}_{\theta,f}(x) = \frac{1}{\delta}\mathbf{1}_{[1-\delta,1]}(x) - \theta,$$

which concludes the proof. □

We are now ready to conclude the proof of theorem 1.

Proof of theorem 1. We start by dealing with the case $\delta = 0$. Let us recall that in this case, the ordinary score $\dot{l}_{\theta,f}$ belongs to ${}_f\dot{\mathcal{P}}_0$ and the Fisher information is zero. Then, using theorem 2 in Chamberlain (1986), we conclude that there is no regular estimator for θ relative to the tangent set $\dot{\mathcal{P}}_0$. We remark that the tangent set ${}_f\dot{\mathcal{P}}_0$ 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 ${}_f\dot{\mathcal{P}}_0$ such that for every m, we have $\dot{l}_{\theta,f} \notin {}_f\dot{\mathcal{P}}_{0,m} := \text{lin}(h_1, h_2, \ldots, h_m)$. We thus have

$$\begin{aligned}&\sup_{E_0}\liminf_{n\to\infty}\sup_{g\in E_0}\mathbb{E}_{P_{1/\sqrt{n},g}}\left[\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right)\right]^2\\ &\geq \sup_{F_0}\liminf_{n\to\infty}\sup_{g\in F_0}\mathbb{E}_{P_{1/\sqrt{n},g}}\left[\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right)\right]^2,\end{aligned}$$

where E_0 and F_0 range through all finite subsets of the tangent sets $\dot{\mathcal{P}}_0 = \text{lin}\left(\dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_0\right) = {}_f\dot{\mathcal{P}}_0$ and $\text{lin}\left(\dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_{0,m}\right) = {}_f\dot{\mathcal{P}}_{0,m}$, respectively. The efficient score function for θ corresponding to the tangent set ${}_f\dot{\mathcal{P}}_{0,m}$ is

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

Moreover, the efficient information $\tilde{I}_{\theta,f,m} = \mathbb{P}_{\theta,f}\left(\tilde{l}^2_{\theta,f,m}\right)$ is non-zero. Using lemma 25.25 from van der Vaart (1998), the efficient influence function relative to the tangent set $\text{lin}\left(\dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_{0,m}\right)$ is $\tilde{\psi}_{\theta,f,m} = \tilde{I}^{-1}_{\theta,f,m}\tilde{l}_{\theta,f,m}$. So we can apply theorem 25.21 from van der Vaart (1998) to obtain that

$$\sup_{F_0}\liminf_{n\to\infty}\sup_{g\in F_0}\mathbb{E}_{P_{1/\sqrt{n},g}}\left[\sqrt{n}\left(\hat{\theta}_n - \psi\left(P_{1/\sqrt{n},g}\right)\right)\right]^2 \geq \tilde{I}^{-1}_{\theta,f,m}.$$

Because $\tilde{I}_{\theta,f,m} \xrightarrow[m\to\infty]{} \tilde{I}_{\theta,f} = 0$, we obtain the result. The second part of the proof concerning $\delta > 0$ is an immediate consequence of proposition 3 together with theorem 25.21 and lemma 25.23 in van der Vaart (1998). □

5.3. Proofs from Sections 3.1 and 3.3

Proof of theorem 2. Let us denote by $\mathcal{D} = \{1, 2, \cdots, D\}$ $\mathcal{D}_0 = \{k \in \mathcal{D}$ such that $I_k \subseteq [1-\delta, 1]\}$ and $\mathcal{D}_1 = \mathcal{D}\setminus\mathcal{D}_0 = \{k \in \mathcal{D}$ such that $I_k \nsubseteq [1-\delta, 1]\}$. We fix an integer $k_0 \in \mathcal{D}_0$. We start by proving that the estimator $\hat{\theta}_{I,n}$ converges almost surely to θ. Indeed, we can write that

$$\hat{\theta}_{I,n} = \theta + \sum_{k\in\mathcal{D}_0}\left(\frac{n_k}{n|I_k|} - \theta\right)\mathbf{1}\left\{\hat{k}_n = k\right\} + \left(\hat{\theta}_{I,n} - \theta\right)\mathbf{1}\left\{I_{\hat{k}_n} \nsubseteq [1-\delta, 1]\right\}, \tag{18}$$

where $\mathbf{1}\{A\}$ or $\mathbf{1}_A$ is used to denote the indicator function of set A. By using the strong law of large numbers, we have the almost sure convergence

$$\begin{aligned}
\forall k \in \mathcal{D}_0, &\quad \frac{n_k}{n|I_k|} \xrightarrow[n\to+\infty]{a.s.} \theta, \\
\forall k \in \mathcal{D}_1, &\quad \frac{n_k}{n|I_k|} \xrightarrow[n\to+\infty]{a.s.} \frac{\alpha_k}{|I_k|} = \frac{1}{|I_k|}\int_{I_k} g(u)du > \theta.
\end{aligned}$$

As a consequence, we obtain that the second term in the right-hand side of (18) converges almost surely to zero, namely,

$$\left| \sum_{k\in\mathcal{D}_0} \left(\frac{n_k}{n|I_k|} - \theta\right) \mathbf{1}\left\{\hat{k}_n = k\right\} \right| \le \sum_{k\in\mathcal{D}_0} \left|\frac{n_k}{n|I_k|} - \theta\right| \xrightarrow[n\to+\infty]{a.s.} 0.$$

The third term in the right-hand side of (18) also converges almost surely to zero. Indeed, we have

$$\left|\hat{\theta}_{I,n} - \theta\right| \mathbf{1}\left\{I_{\hat{k}_n} \not\subseteq [1-\delta, 1]\right\} \le \left(\max_{1\le k\le D} \frac{1}{|I_k|} - \theta\right) \sum_{k\in\mathcal{D}_1} \mathbf{1}\left\{\hat{k}_n = k\right\}.$$

For all $k \in \mathcal{D}_1$,

$$\begin{aligned}
\mathbf{1}\left\{\hat{k}_n = k\right\} &= \mathbf{1}\left\{\frac{n_k}{n|I_k|} \le \frac{n_j}{n|I_j|}, \forall j \in \mathcal{D}\right\} \\
&\le \mathbf{1}\left\{\frac{n_{k_0}}{n|I_{k_0}|} - \theta + \frac{\alpha_k}{|I_k|} - \frac{n_k}{n|I_k|} \ge \frac{\alpha_k}{|I_k|} - \theta\right\}.
\end{aligned}$$

Because $\epsilon_k = \alpha_k/|I_k| - \theta > 0$ and

$$\frac{n_{k_0}}{n|I_{k_0}|} - \theta + \frac{\alpha_k}{|I_k|} - \frac{n_k}{n|I_k|} \xrightarrow[n\to+\infty]{a.s.} 0,$$

we obtain that

$$\mathbf{1}\left\{\frac{n_{k_0}}{n|I_{k_0}|} - \theta + \frac{\alpha_k}{|I_k|} - \frac{n_k}{n|I_k|} \ge \epsilon_k\right\} \xrightarrow[n\to+\infty]{a.s.} 0,$$

which concludes the proof of the almost sure convergence of $\hat{\theta}_{I,n}$. We now prove the second statement of the proposition. We have

$$\begin{aligned}
\mathbb{E}\left[\left(\sqrt{n}\left(\hat{\theta}_{I,n} - \theta\right)\right)^2\right] &= \sum_{k\in\mathcal{D}_0} \mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|} - \theta\right)\right)^2 \mathbf{1}_{\hat{k}_n = k}\right] \\
&\quad + \sum_{k\in\mathcal{D}_1} \mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|} - \theta\right)\right)^2 \mathbf{1}_{\hat{k}_n = k}\right].
\end{aligned} \tag{19}$$

The second term in the right-hand side of (19) is bounded by

$$\sum_{k\in\mathcal{D}_1}\mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|}\theta\right)\right)^2\mathbf{1}_{\hat{k_n}=k}\right]\le\left(\max_{1\le k\le D}\frac{1}{|I_k|}-\theta\right)^2\sum_{k\in\mathcal{D}_1}n\mathbb{P}\left(\hat{k}_n=k\right),$$

where for all $k\in\mathcal{D}_1$, according to Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}\left(\hat{k}_n=k\right)&\le\mathbb{P}\left(\frac{n_k}{n|I_k|}\le\frac{n_{k_0}}{n|I_{k_0}|}\right)\\&\le\mathbb{P}\left[\sum_{i=1}^{n}\left(\frac{1}{|I_{k_0}|}\mathbf{1}\left\{X_i\in I_{k_0}\right\}-\theta+\frac{\alpha_k}{|I_k|}-\frac{1}{|I_k|}\mathbf{1}\left\{X_i\in I_k\right\}\right)\ge n\epsilon_k\right]\\&\le\exp\left[-2n\epsilon_k^2\left(\frac{1}{|I_k|}+\frac{1}{|I_{k_0}|}\right)^{-2}\right].\end{aligned}$$

For the first term in the right-hand side of (19), we apply Cauchy–Schwarz's inequality

$$\begin{aligned}\sum_{k\in\mathcal{D}_0}\mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|}-\theta\right)\right)^2\mathbf{1}_{\hat{k_n}=k}\right]&\le\sqrt{\sum_{k\in\mathcal{D}_0}\mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|}-\theta\right)\right)^4\right]}\sqrt{\sum_{k\in\mathcal{D}_0}\mathbb{P}\left(\hat{k}_n=k\right)}\\&\le\sqrt{\sum_{k\in\mathcal{D}_0}\mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|}-\theta\right)\right)^4\right]},\end{aligned}\tag{20}$$

where for all $k\in\mathcal{D}_0$,

$$\begin{aligned}&\mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|}-\theta\right)\right)^4\right]=\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\frac{1}{|I_k|}\mathbf{1}\{X_i\in I_k\}-\theta\right)\right)^4\right]\\&=\frac{1}{n}\mathbb{E}\left[\left(\frac{1}{|I_k|}\mathbf{1}\{X_1\in I_k\}-\theta\right)^4\right]+\frac{n-1}{n}\mathbb{E}^2\left[\left(\frac{1}{|I_k|}\mathbf{1}\{X_1\in I_k\}-\theta\right)^2\right]\\&=\frac{\theta}{n}\left(\frac{1}{|I_k|^3}-\frac{4\theta}{|I_k|^2}+\frac{6\theta^2}{|I_k|}-3\theta^3\right)+\frac{n-1}{n}\sigma_k^4.\end{aligned}\tag{21}$$

Thus, we finally obtain that

$$\begin{aligned}n\mathbb{E}\left[\left(\hat{\theta}_{I,n}-\theta\right)^2\right]&\le\sqrt{\sum_{k\in\mathcal{D}_0}\left[\frac{\theta}{n}\left(\frac{1}{|I_k|^3}-\frac{4\theta}{|I_k|^2}+\frac{6\theta^2}{|I_k|}-3\theta^3\right)+\frac{n-1}{n}\sigma_k^4\right]}+\\&\left(\max_{1\le k\le D}\frac{1}{|I_k|}-\theta\right)^2\sum_{k\in\mathcal{D}_1}n\exp\left[-2n\epsilon_k^2\left(\frac{1}{|I_k|}+\frac{1}{|I_{k_0}|}\right)^{-2}\right]\xrightarrow[n\to+\infty]{}\sqrt{\sum_{k\in\mathcal{D}_0}\sigma_k^4}.\end{aligned}$$

□

Proof of proposition 2. Let us first establish that condition (10) holds. In fact, with the notation $p_{\theta,f}=\theta+(1-\theta)f$, we have

$$\begin{aligned}\int\|\tilde{l}_{\theta_n,f}d\mathbb{P}_{\theta_n,f}^{1/2}-\tilde{l}_{\theta,f}d\mathbb{P}_{\theta,f}^{1/2}\|^2&=\int_0^1\left(\tilde{l}_{\theta_n,f}(x)\sqrt{p_{\theta_n,f}(x)}-\tilde{l}_{\theta,f}(x)\sqrt{p_{\theta,f}(x)}\right)^2dx\\&\le2\int_0^1\left(\tilde{l}_{\theta_n,f}(x)-\tilde{l}_{\theta,f}(x)\right)^2p_{\theta_n,f}(x)dx+2\int_0^1\tilde{l}_{\theta,f}^2(x)\left(\sqrt{p_{\theta_n,f}(x)}-\sqrt{p_{\theta,f}(x)}\right)^2dx\end{aligned}$$

$$\begin{aligned}
&\le 2\int_0^1 \left[\frac{1}{\theta_n}-\frac{1}{\theta}+\left(\frac{1}{\theta(1-\theta\delta)}-\frac{1}{\theta_n(1-\theta_n\delta)}\right)\mathbf{1}_{\{f(x)>0\}}\right]^2 p_{\theta_n,f}(x)dx\\
&\quad+2\int_0^1\left[\frac{1}{\theta}-\frac{1}{\theta(1-\theta\delta)}\mathbf{1}_{\{f(x)>0\}}\right]^2 \frac{(\theta_n-\theta)^2(1-f(x))^2}{\left(\sqrt{p_{\theta_n,f}(x)}+\sqrt{p_{\theta,f}(x)}\right)^2}dx\\
&\le 2\int_0^1(\theta_n-\theta)^2\left[\frac{1}{\theta\theta_n}+\frac{\delta(\theta+\theta_n)+1}{\theta\theta_n(1-\theta\delta)(1-\theta_n\delta)}\mathbf{1}_{\{f(x)>0\}}\right]^2 p_{\theta_n,f}(x)dx\\
&\quad+2\int_0^1(\theta_n-\theta)^2 2\left[\frac{1}{\theta^2}+\frac{1}{\theta^2(1-\theta)^2}\right]\frac{(1-f(x))^2}{\left(\sqrt{\theta_n}+\sqrt{\theta}\right)^2}dx\\
&\le(\theta_n-\theta)^2\left[\frac{C}{\theta^2}+\frac{C(1+2C\theta)}{\theta^2(1-\theta)^2}\right]^2+C(\theta_n-\theta)^2\left[\frac{1}{\theta^3}+\frac{1}{\theta^3(1-\theta)^2}\right]=O\left(\frac{1}{n}\right),
\end{aligned}$$

where C is some positive constant. Thus, according to theorem 7.4 from van der Vaart (2002), 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 (8) and (9).

Now in model $\mathcal{P}_\delta$, the efficient score function $\tilde{l}_{\theta,f}$ is given by

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

so that it is natural to estimate the parameter δ in order to estimate $\tilde{l}_{\theta,f}$. Let $\hat{\delta}_n$ be any given consistent (in probability) estimator of δ. Let us examine condition (8) more closely. We have

$$\begin{aligned}
\sqrt{n}\mathbb{P}_{\theta_n,f}\hat{l}_{n,\theta_n} &= \sqrt{n}\mathbb{P}_{\theta_n,f}\left(\hat{l}_{n,\theta_n}-\tilde{l}_{\theta_n,f}\right)\\
&= \sqrt{n}\int_0^1\frac{1}{\theta_n}\left[\frac{1}{1-\theta_n\hat{\delta}_n}\mathbf{1}_{[0,1-\hat{\delta}_n)}(x)-\frac{1}{1-\theta_n\delta}\mathbf{1}_{[0,1-\delta)}(x)\right]g_{\theta_n,f}(x)dx\\
&= \int_0^1\frac{\sqrt{n}}{\theta_n}\left[\left(\frac{1}{1-\theta_n\hat{\delta}_n}-\frac{1}{1-\theta_n\delta}\right)\mathbf{1}_{[0,1-\hat{\delta}_n)}(x)\right.\\
&\qquad\left.+\frac{1}{1-\theta_n\delta}\left(\mathbf{1}_{[0,1-\hat{\delta}_n)}(x)-\mathbf{1}_{[0,1-\delta)}(x)\right)\right]g_{\theta_n,f}(x)dx\\
&= \sqrt{n}\left(\hat{\delta}_n-\delta\right)\int_0^{1-\hat{\delta}_n}\frac{g_{\theta_n,f}(x)}{(1-\theta_n\delta)\left(1-\theta_n\hat{\delta}_n\right)}dx+\sqrt{n}\int_{1-\delta}^{1-\hat{\delta}_n}\frac{g_{\theta_n,f}(x)}{1-\theta_n\delta}dx\\
&= \sqrt{n}\left(\hat{\delta}_n-\delta\right)\left[\int_0^{1-\delta}\frac{g_{\theta,f}(x)}{(1-\theta\delta)^2}dx-\frac{g_{\theta,f}(1-\delta)}{1-\theta\delta}+o_{\mathbb{P}}(1)\right].
\end{aligned}$$

Hence, the ‘no-bias’ condition (8) is equivalent to the existence of an estimator $\hat{\delta}_n$ of δ that converges at a rate faster than $1/\sqrt{n}$, namely such that $\sqrt{n}\left(\hat{\delta}_n-\delta\right)=o_{\mathbb{P}}(1)$. With the same argument as in the previous calculation, the consistency condition (9) is satisfied as soon as the estimator $\hat{\delta}_n$ converges in probability to δ. □

5.4. *Proof of theorem 3*

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 MSE of a histogram estimator $\hat{g}_I$ can be written as the sum of a bias term and a variance term

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

We introduce three lemmas that are needed to prove theorem 3. The proofs of these technical lemmas are further postponed to Appendix A.

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

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

For any partition $I = (I_k)_{1,\dots,D}$ of $[0, 1]$, we let

$$L(I) = ||g_I - g||_2^2 \quad \text{and} \quad \hat{L}_p(I) = \hat{R}_p(I) + ||g||_2^2,$$

respectively, the bias term of the MSE of a histogram estimator $\hat{g}_I$ and its estimator.

Lemma 2. *Let* $I = (I_k)_{1,\dots,D}$ *be an arbitrary partition of* $[0, 1]$. *Let* $p \in \{1, 2, \dots, n-1\}$ *such that* $\lim_{n \overset{\infty}{\longrightarrow}} p/n < 1$. *Then we have the following results:*

$$(i)\ \hat{L}_p(I) \xrightarrow[n\to\infty]{a.s.} L(I)$$

$$(ii)\ \sqrt{n}\left(\hat{L}_p(I) - L(I)\right) = \sqrt{n}\left(\hat{R}_p(I) - R(I)\right) + \frac{1}{\sqrt{n}}(s_{11} - s_{21}) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, 4\left(s_{32} - s_{21}^2\right)\right).$$

Let I, J be two partitions in $\mathcal{I}$, then I is called a subdivision of J and we denote $I \trianglelefteq J$, if $\mathcal{F}_J \subset \mathcal{F}_I$ and $I \ntrianglelefteq J$ otherwise.

Lemma 3. *Suppose that function* f *belongs to* $\mathcal{F}_\delta$. *Let us consider* $m_{\max}$ *large enough such that* $\delta > 2^{1-m_{\max}}$. *Define* $N = 2^{m_{\max}}$ *and* $I^{(N)} = (N, \lambda_N) \in \mathcal{I}$ *with* $\lambda_N = \lceil N(1 - \delta \rceil / N$. *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 $L(I) > L(I^{(N)})$.

We are now ready to prove theorem 3, starting by establishing point i). First, we remark that under condition (7), Celisse and Robin proved in their proposition 2.1 that

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

Denoting by $\Lambda^\star = [1 - \delta, 1]$ and $\hat{\Lambda} = [\hat{\lambda}, 1]$, we may write

$$\begin{aligned}\hat{\theta}_n^{CR} = \theta + \sum_{I=(N,\lambda)\trianglelefteq I^{(N)}} \left[\frac{1}{n(1-\lambda)} \sum_{i=1}^n \mathbf{1}\{X_i \in [\lambda, 1]\} - \theta\right] \mathbf{1}\left\{\hat{\lambda} = \lambda\right\} \\ + \left(\hat{\theta}_n^{CR} - \theta\right) \mathbf{1}_{\hat{I} \ntrianglelefteq I^{(N)}},\end{aligned} \quad (22)$$

where $N = 2^{m_{\max}}$ as in lemma 3. For each partition $I = (N, \lambda) \trianglelefteq I^{(N)}$, we have $[\lambda, 1] \subseteq \Lambda^\star$. By applying the strong law of large numbers, we obtain that

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

Because 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

$$\left(\hat{\theta}_n^{CR}-\theta\right)\mathbf{1}_{\hat{I}\ntrianglelefteq I^{(N)}}\xrightarrow[n\to\infty]{a.s.}0.$$

Using lemma 3, we have $L\left(\hat{I}\right)>L(I^{(N)})$. Let

$$\gamma=\min_{I\ntrianglelefteq I^{(N)}}L(I)-L(I^{(N)})>0, \tag{23}$$

we obtain that

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

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

$$\begin{aligned}\left|\hat{\theta}_n^{CR}-\theta\right|\mathbf{1}_{\hat{I}\ntrianglelefteq I^{(N)}}&\leq(N-\theta)\mathbf{1}\left\{\sup_{I\in\mathcal{I}}\left|\hat{L}_{\hat{p}(I)}(I)-L(I)\right|\geq\frac{\gamma}{2}\right\}\\&\leq(N-\theta)\sum_{I\in\mathcal{I}}\mathbf{1}\left\{\left|\hat{L}_{\hat{p}(I)}(I)-L(I)\right|\geq\frac{\gamma}{2}\right\}.\end{aligned} \tag{24}$$

Because $\forall I\in\mathcal{I}$, we both have $\hat{L}_p(I)\xrightarrow[n\to\infty]{a.s.}L(I)$ and $\hat{p}(I)/n\xrightarrow[n\to\infty]{a.s.}l_\infty(I)\in[0,1)$ as well as the fact that $\hat{R}_p(I)$ (given by (6)) is a continuous function of p/n, we obtain $\hat{L}_{\hat{p}(I)}(I)\xrightarrow[n\to\infty]{a.s.}L(I)$. Therefore,

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

Indeed, if $X_n\xrightarrow{a.s.}X$, then $\forall\epsilon>0$, we have $\mathbf{1}\{|X_n-X|\geq\epsilon\}\xrightarrow{a.s.}0$. It thus follows that $\left(\hat{\theta}_n^{CR}-\theta\right)\mathbf{1}_{\hat{I}\ntrianglelefteq I^{(N)}}\xrightarrow{a.s.}0$. We finally obtain that $\hat{\theta}_n^{CR}\xrightarrow{a.s.}\theta$.

We now turn to point ii). We may write as previously

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n^{CR}-\theta\right)=&\sum_{I=(N,\lambda)\trianglelefteq I^{(N)}}\sqrt{n}\left[\frac{1}{n(1-\lambda)}\sum_{i=1}^{n}\mathbf{1}\{X_i\in[\lambda,1]\}-\theta\right]\mathbf{1}_{\{\hat{\lambda}=\lambda\}}\\&+\sqrt{n}\left(\hat{\theta}_n^{CR}-\theta\right)\mathbf{1}_{\{\hat{I}\ntrianglelefteq I^{(N)}\}}.\end{aligned}$$

For each partition $I = (N, \lambda) \trianglelefteq I^{(N)}$, by applying the CLT, we obtain that

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

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

$$\sum_{I=(N,\lambda)\trianglelefteq I^{(N)}} \sqrt{n}\left[\frac{1}{n(1-\lambda)}\sum_{i=1}^{n}\mathbf{1}_{X_i \in [\lambda,1]} - \theta\right]\mathbf{1}_{\hat{\lambda}=\lambda} = O_{\mathbb{P}}(1). \tag{25}$$

We shall now prove that $\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right)\mathbf{1}_{\hat{I} \ntrianglelefteq I^{(N)}} \xrightarrow[n\to\infty]{\mathbb{P}} 0$. In fact, according to (24), for all $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sqrt{n}\left|\hat{\theta}_n^{CR} - \theta\right|\mathbf{1}_{\hat{I} \ntrianglelefteq I^{(N)}} > \epsilon\right) &\le \mathbb{P}\left(\hat{I} \ntrianglelefteq I^{(N)}\right) \\
&\le \mathbb{P}\left(\sup_{I \in \mathcal{I}}\left|\hat{L}_{\hat{p}(I)}(I) - L(I)\right| \ge \frac{\gamma}{2}\right) \\
&\le \sum_{I \in \mathcal{I}} \mathbb{P}\left(\left|\hat{L}_{\hat{p}(I)}(I) - L(I)\right| \ge \frac{\gamma}{2}\right) \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

where γ is defined by (23). Therefore, $\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right)\mathbf{1}_{\hat{I} \ntrianglelefteq I^{(N)}} = o_{\mathbb{P}}(1)$. We finally conclude that $\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right) = O_{\mathbb{P}}(1)$.

We now prove the last statement $iii)$ of the proposition. We have

$$\begin{aligned}
\mathbb{E}\left[\left(\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right)\right)^2\right] = \sum_{I=(N,\lambda)\trianglelefteq I^{(N)}} &\mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n}\left(\frac{1}{1-\lambda}\mathbf{1}\{X_i \in [\lambda, 1]\} - \theta\right)\right)^2 \mathbf{1}_{\{\hat{\lambda}=\lambda\}}\right] \\
&+ \mathbb{E}\left[\left(\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right)\right)^2 \mathbf{1}\left\{\hat{I} \ntrianglelefteq I^{(N)}\right\}\right].
\end{aligned}$$

The first term of the aforementioned equation is bounded as in the proof of proposition 2 (see inequalities (20) and (21))

$$\begin{aligned}
&\sum_{I=(N,\lambda)\trianglelefteq I^{(N)}} \mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n}\left(\frac{1}{1-\lambda}\mathbf{1}\{X_i \in [\lambda, 1]\} - \theta\right)\right)^2 \mathbf{1}_{\{\hat{\lambda}=\lambda\}}\right] \\
&\le \sqrt{\sum_{I=(N,\lambda)\trianglelefteq I^{(N)}} \left[\frac{\theta}{n}\left(\frac{1}{(1-\lambda)^3} - \frac{4\theta}{(1-\lambda)^2} + \frac{6\theta^2}{1-\lambda} - 3\theta^3\right) + \frac{n-1}{n}\theta^2\left(\frac{1}{(1-\lambda)} - \theta\right)^2\right]}.
\end{aligned}$$

The second term is bounded by

$$\begin{aligned}
\mathbb{E}\left[\left(\sqrt{n}\left(\hat{\theta}_n^{CR} - \theta\right)\right)^2 \mathbf{1}\left\{\hat{I} \ntrianglelefteq I^{(N)}\right\}\right] &\le (N - \theta)^2 n \mathbb{P}\left(\hat{I} \ntrianglelefteq I^{(N)}\right) \\
&\le (N - \theta)^2 n \mathbb{P}\left(\sup_{I \in \mathcal{I}}\left|\hat{L}_p(I) - L(I)\right| \ge \frac{\gamma}{2}\right) \\
&\le (N - \theta)^2 n \sum_{I \in \mathcal{I}} \mathbb{P}\left(\left|\hat{L}_p(I) - L(I)\right| \ge \frac{\gamma}{2}\right).
\end{aligned}$$

For each partition $I \in \mathcal{I}$, according to the calculations in the proof of lemma 1, we have

$$\begin{aligned}
\hat{L}_p(I) - L(I) = {} & \frac{2n-p}{(n-1)(n-p)} \left\{ \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right) + s_{11} - s_{21} \right\} \\
& - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right)^2 \\
& - \frac{2n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{\alpha_k}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right).
\end{aligned}$$

This leads to

$$\begin{aligned}
\mathbb{P}\left(\left| \hat{L}_p(I) - L(I) \right| \geq \frac{\gamma}{2} \right) \leq {} & \mathbb{P}\left(\left| \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right) \right| \geq \frac{(n-1)(n-p)\gamma}{6(2n-p)} - |s_{21} - s_{11}| \right) \\
& + \mathbb{P}\left(\sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right)^2 \geq \frac{(n-1)(n-p)\gamma}{6n(n-p+1)} \right) \\
& + \mathbb{P}\left(\left| \sum_k \frac{\alpha_k}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right) \right| \geq \frac{(n-1)(n-p)\gamma}{12n(n-p+1)} \right).
\end{aligned}$$

According to Hoeffding's inequality, we have

$$\begin{aligned}
& \mathbb{P}\left(\left| \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right) \right| \geq \frac{(n-1)(n-p)\gamma}{6(2n-p)} - |s_{21} - s_{11}| \right) \\
& = \mathbb{P}\left(\left| \sum_{i=1}^n \sum_k \frac{1}{|I_k|} \left(\mathbf{1}\{X_i \in I_k\} - \alpha_k \right) \right| \geq \frac{n(n-1)(n-p)\gamma}{6(2n-p)} - n|s_{21} - s_{11}| \right) \\
& \leq 2 \exp\left[-2n \left(\sum_k \frac{1}{|I_k|} \right)^{-2} \left(\frac{(n-1)(n-p)\gamma}{6(2n-p)} - |s_{21} - s_{11}| \right)^2 \right],
\end{aligned}$$

as well as

$$\mathbb{P}\left(\left| \sum_k \frac{\alpha_k}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right) \right| \geq \frac{(n-1)(n-p)\gamma}{12n(n-p+1)} \right) \leq 2 \exp\left[-2n s_{11}^{-2} \left(\frac{(n-1)(n-p)\gamma}{12n(n-p+1)} \right)^2 \right],$$

and

$$\begin{aligned}
& \mathbb{P}\left(\left| \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n} - \alpha_k \right)^2 \right| \geq \frac{(n-1)(n-p)\gamma}{6n(n-p+1)} \right) \\
& \leq \sum_k \mathbb{P}\left(\left| \sum_{i=1}^n \left(\mathbf{1}\{X_i \in I_k\} - \alpha_k \right) \right|^2 \geq \frac{|I_k| n(n-1)(n-p)\gamma}{6D(n-p+1)} \right) \\
& \leq 2 \exp\left[-2 \left(\frac{|I_k|(n-1)(n-p)\gamma}{6D(n-p+1)} \right) \right].
\end{aligned}$$

Hence, we obtain that $n\mathbb{P}\left(\left| \hat{L}_p(I) - L(I) \right| \geq \frac{\gamma}{2} \right) \xrightarrow[n \to +\infty]{} 0$. Finally, we conclude that $\limsup_{n \to \infty} n\mathbb{E}\left[\left(\hat{\theta}_n^{CR} - \theta \right)^2 \right] < +\infty$.

Acknowledgements

The authors are grateful to Cyril Dalmasso, Elisabeth Gassiat and Pierre Neuvial for fruitful discussions concerning this work. They also would like to thank an anonymous referee whose remarks and suggestions greatly improved this work.

References

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**, 289–300.

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

Celisse, A. & Robin, S. (2008). Nonparametric density estimation by exact leave-p-out cross-validation. *Comput. Statist. Data Anal.* **52**, 2350–2368.

Celisse, A. & Robin, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference* **140**, 3132–3147.

Chamberlain, G. (1986). Asymptotic efficiency in semiparametric models with censoring. *J. Econometrics* **32**, 189–218.

Dudoit, S. & van der Laan, M. J. (2008). *Multiple testing procedures with applications to genomics*, Springer Series in Statistics, Springer, New York.

Genovese, C. & Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–1061.

Hengartner, N. W. & Stark, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23**, 525–550.

Ibragimov, I. A. & Has'minskiĭ, R. Z. (1981). *Statistical estimation*, Applications of Mathematics, vol. 16, Springer-Verlag, New York. Asymptotic theory, Translated from the Russian by Samuel Kotz.

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**, 555–572.

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**, 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**, 373–393.

Mosig, M. O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M. & Friedmann, A. (2001). A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* **157**, 1683–1698.

Nettleton, D., Hwang, J., Caldo, R. & Wise, R. (2006). Estimating the number of true null hypotheses from a histogram of p values. *J. Agric. Biol. Envir. S.* **11**, 337–356.

Neuvial, P. (2013). Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. *J. Mach. Learn. Res.* **14**, 1423–1459.

Schweder, T. & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 479–498.

Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (electronic).

Turkheimer, F. E., Smith, C. B. & Schmidt, K. (2001). Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage* **13**, 920–930.

van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics. Ecole d'été de probabilités de Saint-Flour XXIX – 1999, Saint-Flour, France, July 8–24, 1999* (eds Bolthausen, E. *et al.*), Lect. Notes Math. 1781 Springer, Berlin; 331–457.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Cambridge series in statistical and probabilistic mathematics, vol. 3, Cambridge University Press, Cambridge.

Received January 2013, in final form January 2014

Van Hanh Nguyen, Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071, 23 bvd de France, 91037 Évry, France.
E-mail: nvhanh@hua.edu.vn

Appendix A: Proofs of technical lemmas

A1. *Proof of lemma 1*

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

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

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

$$\begin{aligned}||g-g_I||_2^2 &= \inf_{(a_k)_k\in\mathbb{R}}\left[||g||_2^2-2\int_0^1\left(\sum_k a_k\mathbf{1}_{I_k}(x)\right)g(x)dx+\int_0^1\left(\sum_k a_k\mathbf{1}_{I_k}(x)\right)^2dx\right]\\ &= \inf_{(a_k)_k\in\mathbb{R}}\left[||g||_2^2-2\sum_k a_k\alpha_k+\sum_k a_k^2|I_k|\right]\\ &= ||g||_2^2-\sum_k\frac{\alpha_k^2}{|I_k|}=||g||_2^2-s_{21}.\end{aligned} \tag{26}$$

Let us now calculate the MSE of $\hat{g}_I$

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

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

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

Therefore,

$$\mathbb{E}\left[||g-\hat{g}_I||_2^2\right]=||g||_2^2-s_{21}+\frac{1}{n}(s_{11}-s_{21}). \tag{27}$$

Using (26) and (27), we obtain the desired result, namely

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

A2. *Proof of lemma 2*

(i) Because $\lim_{n\to\infty}\frac{p}{n}<1$ and $\frac{n_k}{n}\xrightarrow[n\to\infty]{a.s.}\alpha_k$, for all k, we obtain that

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

(ii) By definition of $R(I)$ and using (27), we have

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

This gives that

$$\begin{aligned}\sqrt{n}\left[\hat{R}_p(I) - R(I)\right] &= \frac{2n-p}{(n-1)(n-p)}\sum_k \frac{1}{|I_k|}\left[\sqrt{n}\left(\frac{n_k}{n} - \alpha_k\right)\right] + \frac{(2n-p)\sqrt{n}}{(n-1)(n-p)}s_{11} \\ &\quad - \frac{n(n-p+1)}{\sqrt{n}(n-1)(n-p)}\sum_k \frac{1}{|I_k|}\left[\sqrt{n}\left(\frac{n_k}{n} - \alpha_k\right)\right]^2 - \frac{(2n-p)\sqrt{n}}{(n-1)(n-p)}s_{21} \\ &\quad - \frac{2n(n-p+1)}{(n-1)(n-p)}\sum_k \frac{\alpha_k}{|I_k|}\left[\sqrt{n}\left(\frac{n_k}{n} - \alpha_k\right)\right] - \frac{1}{\sqrt{n}}(s_{11} - s_{21}) \\ &= T_1 - \frac{2n(n-p+1)}{(n-1)(n-p)}\sum_k \frac{\alpha_k}{|I_k|}\left[\sqrt{n}\left(\frac{n_k}{n} - \alpha_k\right)\right].\end{aligned} \tag{28}$$

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

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

It thus follows that $T_1 = o_{\mathbb{P}}(1)$. We now consider the remaining term in (28). We have

$$\sum_k \frac{\alpha_k}{|I_k|}\left[\sqrt{n}\left(\frac{n_k}{n} - \alpha_k\right)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\sum_k \frac{\alpha_k}{|I_k|}\mathbf{1}_{X_i \in I_k} - s_{21}\right).$$

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 i.i.d. centred with variance

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

By the CLT, we obtain

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

Combining this with (28) implies that

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

It is easy to calculate that

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

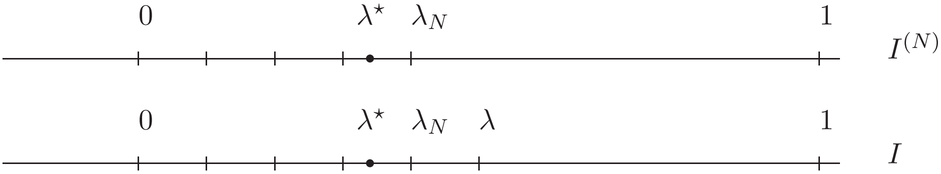
Hence, we have

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

which completes the proof.

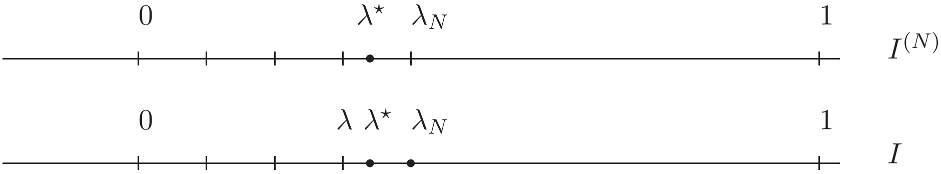
A3. *Proof of lemma 3*

(i) Let us denote by $\lambda^\star = 1 - \delta$. If I is a subdivision of $I^{(N)}$, then $I = (N, \lambda)$ with $[\lambda, 1] \subset [\lambda^\star, 1]$. For example, we may have the following situation:



Because g is constant on the interval $[\lambda^\star, 1] \supset [\lambda_N, 1] \supset [\lambda, 1]$, we have $g_I = g_{I^{(N)}} = g$ on the interval $[\lambda_N, 1]$. This implies that $||g_I - g||_2^2 = ||g_{I^{(N)}} - g||_2^2$.

(ii) If $I = (2^m, \lambda)$ is not a subdivision of $I^{(N)}$, then there are two cases to consider: If $m = m_{\max}$ then $[\lambda, 1] \not\subseteq [\lambda_N, 1]$. For example, we may have



Because $g_{I^{(N)}} = g$ on the interval $[\lambda_N, 1]$ and the two partitions I and $I^{(N)}$ restricted to the interval $[0, \lambda]$ are the same, we thus have

$$||g_I - g||_{2,[0,\lambda]}^2 = ||g_{I^{(N)}} - g||_{2,[0,\lambda]}^2,$$

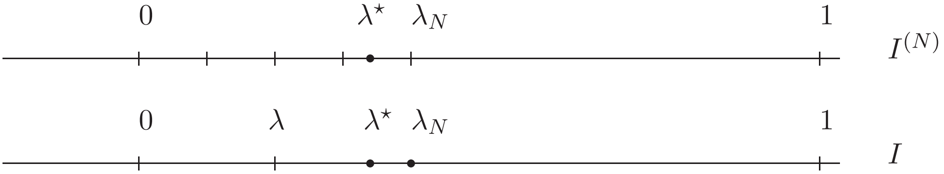
and

$$\begin{aligned} ||g_I - g||_2^2 - ||g_{I^{(N)}} - g||_2^2 &= ||g_I - g||_{2,[\lambda,1]}^2 - ||g_{I^{(N)}} - g||_{2,[\lambda,\lambda_N]}^2 \\ &= (\lambda_N - \lambda)(a - b)^2 + (1 - \lambda_N)(a - \theta)^2, \end{aligned}$$

where

$$a = \frac{1}{1-\lambda}\int_\lambda^1 g(x)dx, \quad b = \frac{1}{\lambda_N - \lambda}\int_\lambda^{\lambda_N} g(x)dx.$$

Using the assumption that $f \in \mathcal{F}_\delta$, we obtain that $L(I) > L(I^{(N)})$.
If $m < m_{\max}$, we may have, for example,



As before, we may show that

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

which completes the proof.