HAL open science

Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation

van Hanh Nguyen, Catherine Matias

▶ **To cite this version:**

van Hanh Nguyen, Catherine Matias. Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation. ESAIM: Probability and Statistics, 2014, 18, pp.584-612. hal-00738555v2

HAL Id: hal-00738555

https://hal.science/hal-00738555v2

Submitted on 2 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation

Van Hanh Nguyen[1,2] and Catherine Matias[2]

April 2, 2013

1. Laboratoire de Mathématiques d'Orsay, Université Paris Sud, UMR CNRS 8628, Bâtiment 425, 91 405 Orsay Cedex, France. E-mail: nvanhanh@genopole.cnrs.fr
2. Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071- USC INRA, 23 bvd de France, 91 037 Évry, France. E-mail: catherine.matias@genopole.cnrs.fr

Abstract

In a multiple testing context, we consider a semiparametric mixture model with two components where one component is known and corresponds to the distribution of p-values under the null hypothesis and the other component f is nonparametric and stands for the distribution under the alternative hypothesis. Motivated by the issue of local false discovery rate estimation, we focus here on the estimation of the nonparametric unknown component f in the mixture, relying on a preliminary estimator of the unknown proportion θ of true null hypotheses. We propose and study the asymptotic properties of two different estimators for this unknown component. The first estimator is a randomly weighted kernel estimator. We establish an upper bound for its pointwise quadratic risk, exhibiting the classical nonparametric rate of convergence over a class of Hölder densities. To our knowledge, this is the first result establishing convergence as well as corresponding rate for the estimation of the unknown component in this nonparametric mixture. The second estimator is a maximum smoothed likelihood estimator. It is computed through an iterative algorithm, for which we establish a descent property. In addition, these estimators are used in a multiple testing procedure in order to estimate the local false discovery rate. Their respective performances are then compared on synthetic data.

Key words and phrases: False discovery rate; kernel estimation; local false discovery rate; maximum smoothed likelihood; multiple testing; p-values; semiparametric mixture model.

1 Introduction

In the framework of multiple testing problems (microarray analysis, neuro-imaging, etc), a mixture model with two populations is considered

$$\forall x \in \mathbb{R}^d, \quad g(x) = \theta\phi(x) + (1-\theta)f(x), \tag{1}$$

where θ is the unknown proportion of true null hypotheses, ϕ and f are the densities of the observations generated under the null and alternative hypotheses, respectively. More precisely, assume the test statistics are independent and identically distributed (iid) with a continuous distribution under the corresponding null hypotheses and we observe the p-values $X_1, X_2, \ldots, X_n$ associated with n independent tested hypotheses, then the density function ϕ is the uniform distribution on $[0,1]$ while the density function f is assumed unknown. The parameters of the model are (θ, f), where θ is a Euclidean parameter while f is an infinite-dimensional one and the model becomes

$$\forall x \in [0,1], \quad g(x) = \theta + (1-\theta) f(x). \tag{2}$$

In the following, we focus on model (2) that is slightly simpler than (1). A central problem in the multiple testing setup is the control of type I (*i.e.* false positive) and type II (*i.e.* false negative) errors. The most popular criterion regarding type I errors is the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995). To set up the notation, let H_i be the i-th (null) hypothesis. The outcome of testing n hypotheses simultaneously can be summarized as indicated in Table 1.

Table 1: Possible outcomes from testing n hypotheses $H_1, \ldots, H_n$.

	Accepts H_i	Rejects H_i	Total
H_i is true	TN	FP	n_0
H_i is false	FN	TP	n_1
Total	N	P	n

Benjamini and Hochberg (1995) define FDR as the expected proportion of rejections that are incorrect,

$$\mathrm{FDR} = \mathbb{E}\Big[\frac{\mathrm{FP}}{\max(\mathrm{P},1)}\Big] = \mathbb{E}\Big[\frac{\mathrm{FP}}{\mathrm{P}}\big|\mathrm{P}>0\Big]\mathbb{P}(\mathrm{P}>0).$$

They provide a multiple testing procedure that guarantees the bound FDR $\leq \alpha$, for a desired level α. Storey (2003) proposes to modify FDR so as to obtain a new criterion, the positive FDR (or pFDR), defined by

$$\mathrm{pFDR} = \mathbb{E}\Big[\frac{\mathrm{FP}}{\mathrm{P}}\big|\mathrm{P}>0\Big],$$

and argues that it is conceptually more sound than FDR. For microarray data for instance, there is a large value of the number of hypotheses n and the difference between pFDR and FDR is generally small as the extra factor $\mathbb{P}(P>0)$ is very close to 1 (see Liao et al., 2004). In a mixture context, the pFDR is given by

$$\mathrm{pFDR}(x) = \mathbb{P}(H_i \text{ being true } |X \leq x) = \frac{\theta\Phi(x)}{\theta\Phi(x) + (1-\theta)F(x)},$$

where Φ and F are the cumulative distribution functions (cdfs) for densities ϕ and f, respectively. (It is notationally convenient to consider events of the form $X \leq x$, but we could just as well consider tail areas to the right, two-tailed events, etc).

Efron et al. (2001) define the local false discovery rate (ℓFDR) to quantify the plausibility of a particular hypothesis being true, given its specific test statistic or p-value. In a mixture framework, the ℓFDR is the Bayes posterior probability

$$\ell\mathrm{FDR}(x) = \mathbb{P}(H_i \text{ being true } |X = x) = 1 - \frac{(1-\theta)f(x)}{\theta\phi(x) + (1-\theta)f(x)}. \tag{3}$$

In many multiple testing frameworks, we need information at the individual level about the probability for a given observation to be a false positive (Aubert et al., 2004). This motivates estimating the local false discovery rate ℓFDR. Moreover, the quantities pFDR and ℓFDR are analytically related by $\mathrm{pFDR}(x) = \mathbb{E}[\ell\mathrm{FDR}(X)|X \leq x]$. As a consequence (and recalling that the difference between pFDR and FDR is generally small), Robin et al. (2007) propose to estimate FDR by

$$\widehat{\mathrm{FDR}}(x_i) = \frac{1}{i}\sum_{j=1}^{i} \widehat{\ell\mathrm{FDR}}(x_j),$$

where $\widehat{\ell\mathrm{FDR}}$ is an estimator of ℓFDR and the observations $\{x_i\}$ are increasingly ordered. A natural strategy to estimate ℓFDR is to start by estimating both the proportion θ and either f or g. Another motivation for estimating the parameters in this mixture model comes from the works of Sun and Cai (2007; 2009), who develop adaptive compound decision rules for false discovery rate control. These rules are based on the estimation of the parameters in model (1) (dealing with z-scores) rather than model (2) (dealing with p-values). However, it appears that in some very specific cases (when the alternative is symmetric about the null), the oracle version of their procedure based on the p-values (and thus relying on estimators of the parameters in model (2)) may outperform the one based on model (1) (see Sun and Cai, 2007, for more details). In the following, we are thus interested in estimating parameters in model (2).

In a previous work (Nguyen and Matias, 2012), we discussed the estimation of the Euclidean part of the parameter θ in model (2). Thus, we will not consider further this point here. We rather focus on the estimation of the unknown density f, relying on a preliminary estimator of θ. We just mention that many estimators of θ have been proposed in the literature. One of the most well-known is the one proposed by Storey (2002), motivating its use in our simulations. Some of these estimators are proved to be consistent (under suitable model assumptions). Of course, we will need some specific properties of estimators $\hat{\theta}_n$ of θ to obtain rates of convergence of estimators of f. Besides, existence of estimators $\hat{\theta}_n$ satisfying those specific properties is a consequence of Nguyen and Matias (2012).

Now, different modeling assumptions on the marginal density f have been proposed in the literature. For instance, parametric models have been used with Beta distribution for the p-values (see for example Allison et al., 2002; Liao et al., 2004; Pounds and Morris, 2003) or Gaussian distribution of the probit transformation of the p-values (McLachlan et al., 2006). In the framework of nonparametric estimation, Strimmer (2008) proposed a modified Grenander density estimator for f, which has been initially suggested by Langaas et al. (2005). This approach requires monotonicity constraints on the density f. Other nonparametric approaches consist in relying on regularity assumptions on f. This is done for instance in Neuvial (2010), who is primarily interested in estimating θ under the assumption

that it is equal to $g(1)$. Relying on a kernel estimator of g, he derives nonparametric rates of convergence for θ. Another kernel estimator has been proposed by Robin et al. (2007), along with a multiple testing procedure, called `kerfdr`. This iterative algorithm is inspired by an expectation-maximization (`em`) procedure (Dempster et al., 1977). It is proved to be convergent as the number of iterations increases. However, it does not optimize any criterion and contrarily to the original `em` algorithm, it does not increase the observed data likelihood function. Besides, the asymptotic properties (with the number of hypotheses n) of the kernel estimator underlying Robin et al.'s approach have not been studied. Indeed, its iterative form prevents from obtaining any theoretical result on its convergence properties.

The first part of the present work focuses on the properties of a randomly weighted kernel estimator, which in essence, is very similar to the iterative approach proposed by Robin et al. (2007). Thus, this part may be viewed as a theoretical validation of `kerfdr` approach that gives some insights about the convergence properties (as the sample size increases) of this method. In particular, we establish that relying on a preliminary estimator of θ that roughly converges at parametric rate (see exact condition in Corollary 1), we obtain an estimator of the unknown density f that converges at the usual minimax nonparametric rate. To our knowledge, this is the first result establishing convergence as well as corresponding rate for the estimation of the unknown component in model (2). In a second part, we are interested in a new iterative algorithm for estimating the unknown density f, that aims at maximizing a smoothed likelihood. We refer to Paragraph 4.1 in Eggermont and LaRiccia (2001) for an interesting presentation of kernel estimators as maximum smoothed likelihood ones. Here, we base our approach on the work of Levine et al. (2011), who study a maximum smoothed likelihood estimator for multivariate mixtures. The main idea consists in introducing a nonlinear smoothing operator on the unknown component f as proposed in Eggermont and LaRiccia (1995). We prove that the resulting algorithm possesses a desirable descent property, just as an `em` algorithm does. We also show that it is competitive with respect to `kerfdr` algorithm, both when used to estimate f or ℓFDR.

The article is organized as follows. In Section 2, we start by describing different procedures to estimate f. We distinguish two types of procedures and first describe direct (non iterative) ones in Section 2.1. We mention a direct naive approach but the main procedure from this section is a randomly weighted kernel estimator. Then, we switch to iterative procedures (Section 2.2). The first one is not new: `kerfdr` has been proposed in Guedj et al. (2009); Robin et al. (2007). The second one, called `msl`, is new and adapted from the work of Levine et al. (2011) in a different context (multivariate mixtures). These iterative procedures are expected to be more accurate than direct ones, but their properties are in general more difficult to establish. As such, the direct randomly weighted kernel estimator from Section 2.1 may be viewed as a proxy for studying the convergence properties (with respect to f) of `kerfdr` procedure (properties that are unknown). Section 3 then gives the theoretical properties of the procedures described in Section 2. In particular, we establish (Theorem 1) an upper bound on the pointwise quadratic risk of the randomly weighted kernel procedure. Moreover, we prove that `msl` procedure possesses a descent property with respect to some criterion (Proposition 1). In Section 4, we rely on our different estimators to estimate both density f and the local false discovery rate ℓFDR. We present simulated experiments to compare their performances. All the proofs have been postponed to Section 5. Moreover, some of the more technical proofs have been further postponed to Appendix A.

2 Algorithmic procedures to estimate the density f

2.1 Direct procedures

Let us be given a preliminary estimator $\hat{\theta}_n$ of θ as well as a nonparametric estimator $\hat{g}_n$ of g. We propose here to rely on a kernel estimator of the density g

$$\hat{g}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) = \frac{1}{n}\sum_{i=1}^{n} K_{i,h}(x), \tag{4}$$

where K is a kernel (namely a real-valued integrable function such that $\int K(u)du = 1$), $h > 0$ is a bandwidth (both are to be chosen later) and

$$K_{i,h}(\cdot) = \frac{1}{h}K\Big(\frac{\cdot - X_i}{h}\Big). \tag{5}$$

Note that this estimator of g is consistent under appropriate assumptions.

A naive approach. From Equation (2), it is natural to propose to estimate f with

$$\hat{f}_n^{\text{naive}}(x) = \frac{\hat{g}_n(x) - \hat{\theta}_n}{1 - \hat{\theta}_n}\mathbf{1}_{\{\hat{\theta}_n \neq 1\}},$$

where 1_A is the indicator function of set A. This estimator has the same theoretical properties as the randomly weighted kernel estimator presented below. However, it is much worse in practice, as we shall see in the simulations of Section 4.

A randomly weighted kernel estimator. We now explain a natural construction for an estimator of f relying on a randomly weighted version of a kernel estimator of g. For any hypothesis, we introduce a (latent) random variable Z_i that equals 0 if the null hypothesis H_i is true and 1 otherwise,

$$\forall i = 1, \ldots, n \quad Z_i = \begin{cases} 0 & \text{if } H_i \text{ is true,} \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

Intuitively, it would be convenient to introduce a weight for each observation X_i, meant to select this observation only if it comes from f. Equivalently, the weights are used to select the indexes i such that $Z_i = 1$. Thus, a natural kernel estimate of f would be

$$f_1(x) = \frac{1}{h}\sum_{i=1}^{n} \frac{Z_i}{\sum_{k=1}^{n} Z_k} K\Big(\frac{x - X_i}{h}\Big) = \sum_{i=1}^{n} \frac{Z_i}{\sum_{k=1}^{n} Z_k} K_{i,h}(x), \ x \in [0,1].$$

However, f_1 is not an estimator and cannot be directly used since the random variables Z_i are not observed. A natural approach (initially proposed in Robin et al., 2007) is to replace them with their conditional expectation given the data $\{X_i\}_{1 \leq i \leq n}$, namely with the posterior probabilities $\tau(X_i) = \mathbb{E}(Z_i | X_i)$ defined by

$$\forall x \in [0,1], \ \tau(x) = \mathbb{E}(Z_i | X_i = x) = \frac{(1-\theta)f(x)}{g(x)} = 1 - \frac{\theta}{g(x)}. \tag{7}$$

This leads to the following definition

$$\forall x \in [0,1],\ f_2(x) = \sum_{i=1}^{n} \frac{\tau(X_i)}{\sum_{k=1}^{n} \tau(X_k)} K_{i,h}(x). \tag{8}$$

Once again, the weight $\tau_i = \tau(X_i)$ depends on the unknown parameters θ and f and thus f_2 is not an estimator but rather an oracle. To solve this problem, Robin et al. (2007) proposed an iterative approach, called `kerfdr` and discussed below, to approximate (8). For the moment, we propose to replace the posterior probabilities τ_i by direct (rather than iterative) estimators to obtain a randomly weighted kernel estimator of f. Specifically, we propose to estimate the posterior probability $\tau(x)$ by

$$\forall x \in [0,1],\ \hat{\tau}(x) = 1 - \frac{\hat{\theta}_n}{\hat{g}_n(x)}. \tag{9}$$

Then, by defining the weight

$$\hat{\tau}_i = \hat{\tau}(X_i) = 1 - \frac{\hat{\theta}_n}{\tilde{g}_n(X_i)},\ \text{where } \tilde{g}_n(X_i) = \frac{1}{(n-1)} \sum_{j \neq i}^{n} K_{j,h}(X_i), \tag{10}$$

we get a randomly weighted kernel estimator of the density f defined as

$$\forall x \in [0,1],\ \hat{f}_n^{\mathrm{rwk}}(x) = \sum_{i=1}^{n} \frac{\hat{\tau}_i}{\sum_{k=1}^{n} \hat{\tau}_k} K_{i,h}(x). \tag{11}$$

Note that it is not necessary to use the same kernel K in defining $\hat{g}_n$ and $\hat{f}_n^{\mathrm{rwk}}$, nor the same bandwidth h. In practice, we rely on the same kernel chosen with a compact support (to avoid boundary effects) and as we will see in Section 3, the bandwidths have to be chosen of the same order. Also note that the slight modification from $\hat{g}_n$ to $\tilde{g}_n$ in defining the weights (10) is minor and used in practice to reduce the bias of $\tilde{g}_n(X_i)$.

2.2 Iterative procedures

In this section, we still rely on a preliminary estimator $\hat{\theta}_n$ of θ. Two different procedures are described: `kerfdr` algorithm, proposed by Guedj et al. (2009); Robin et al. (2007) and a maximum smoothed likelihood `msl` estimator, inspired from the work of Levine et al. (2011) in the context of multivariate nonparametric mixtures. Both rely on an iterative randomly weighted kernel approach. The general form of these procedures is described by Algorithm 1. The main difference between the two procedures lies in the choice of the functions $\tilde{K}_{i,h}$ (that play the role of a kernel) and the way the weights are updated.

Note that the parameter θ is fixed throughout these iterative procedures. Indeed, as already noted by Robin et al. (2007), the solution $\theta = 0$ is a fixed point of a modified `kerfdr` algorithm where θ would be iteratively updated. This is also the case with the maximum smoothed likelihood procedure described below in the particular setup of model (2). This is why we keep θ fixed in both procedures. We now describe more explicitly the two procedures.

Algorithm 1: General structure of the iterative algorithms

`// Initialization;`
Set initial weights $\hat{\omega}_i^0 \sim \mathcal{U}\big([0,1]\big), i = 1, 2, \ldots, n$.

while $\max_i |\hat{\omega}_i^{(s)} - \hat{\omega}_i^{(s-1)}|/\hat{\omega}_i^{(s-1)} \geq \epsilon$ **do**

`// Update estimation of` f;
$\hat{f}^{(s)}(x_i) = \sum_j \hat{\omega}_j^{(s-1)} \tilde{K}_{j,h}(x_i) / \sum_k \hat{\omega}_k^{(s-1)}$

`// Update of weights;`
$\hat{\omega}_i^{(s)}$: depends on the procedure, see Equations (12) and (14)

$s \leftarrow s + 1$;

`// Return;`
$\hat{f}^{(s)}(\cdot) = \sum_i \hat{\omega}_i^{(s-1)} \tilde{K}_{i,h}(\cdot) / \sum_k \hat{\omega}_k^{(s-1)}$

`Kerfdr` **algorithm.** This procedure has been proposed by Guedj et al. (2009); Robin et al. (2007) as an approximation to the estimator suggested by (8). In this procedure, functions $\tilde{K}_{i,h}$ more simply denoted $K_{i,h}$ are defined through (5) where K is a kernel (namely $\int K(u)du = 1$) and following (7), the weights are updated as follows

$$\hat{\omega}_i^{(s)} = \frac{(1-\hat{\theta}_n)\hat{f}^{(s)}(x_i)}{\hat{\theta}_n + (1-\hat{\theta}_n)\hat{f}^{(s)}(x_i)}. \tag{12}$$

This algorithm has some `em` flavor (Dempster et al., 1977). Actually, updating the weights $\hat{\omega}_i^{(s)}$ is equivalent to `expectation`-step, and $\hat{f}^{(s)}(x)$ can be seen as an average of $\{K_{i,h}(x)\}_{1 \leq i \leq n}$ so that updating the estimator $\hat{f}$ may look like a `maximization`-step. However, as noted in Robin et al. (2007), the algorithm does not optimize any given criterion. Besides, it does not increase the observed data likelihood function.

The relation between $\hat{f}^{(s)}$ and $\hat{\omega}^{(s)}$ implies that the sequence $\{\hat{\omega}^{(s)}\}_{s \geq 0}$ satisfies $\hat{\omega}^{(s)} = \psi(\hat{\omega}^{(s-1)})$, where

$$\psi : [0,1]^n \backslash \{0\} \to [0,1]^n, \quad \psi_i(u) = \frac{\sum_i u_i b_{ij}}{\sum_i u_i b_{ij} + \sum_i u_i}, \quad \text{with} \quad b_{ij} = \frac{1-\hat{\theta}_n}{\hat{\theta}_n} \times \frac{K_{i,h}(x_j)}{\phi(x_j)}.$$

Thus, if the sequence $\{\hat{\omega}^{(s)}\}_{s \geq 0}$ is convergent, it has to converge towards a fixed point of ψ. Robin et al. (2007) prove that under some mild conditions, `kerfdr` estimator is self-consistent, meaning that as the number of iterations s increases, the sequence $\hat{f}^{(s)}$ converges towards the function

$$f_3(x) = \sum_{i=1}^{n} \frac{\hat{\omega}_i^*}{\sum_k \hat{\omega}_k^*} K_{i,h}(x),$$

where $\hat{\omega}_i^*$ is the (unique) limit of $\{\hat{\omega}_i^{(s)}\}_{s \geq 0}$. Note that contrarily to f_2, function f_3 is a randomly weighted kernel estimator of f. However, nothing is known about the convergence of f_3 nor $\hat{f}^{(s)}$ towards the true density f when the sample size n tends to infinity (while the

bandwidth $h = h_n$ tends to 0). Indeed, the weights $\{\hat{\omega}_i^{(s)}\}_{s\geq 0}$ used by the kernel estimator $\hat{f}^{(s)}$ form an iterative sequence. Thus it is very difficult to study the convergence properties of this weight sequence or of the corresponding estimator.

We thus propose another randomly weighted kernel estimator, whose weights are slightly different from those used in the construction of $\hat{f}^{(s)}$. More precisely, those weights are not defined iteratively but they mimic the sequence of weights $\{\hat{\omega}_i^{(s)}\}_{s\geq 0}$.

Maximum smoothed likelihood estimator. Following the lines of Levine et al. (2011), we construct an iterative estimator sequence of the density f that relies on the maximisation of a smoothed likelihood. Assume in the following that K is a positive and symmetric kernel on $\mathbb{R}$. We define its rescaled version as

$$K_h(x) = h^{-1}K(h^{-1}x).$$

We consider a linear smoothing operator $\mathcal{S} : \mathbb{L}_1([0,1]) \to \mathbb{L}_1([0,1])$ defined as

$$\mathcal{S}f(x) = \int_0^1 \frac{K_h(u-x)f(u)}{\int_0^1 K_h(s-u)ds}du, \text{ for all } x \in [0,1].$$

We remark that if f is a density on $[0,1]$ then $\mathcal{S}f$ is also a density on $[0,1]$. Let us consider a submodel of model (2) restricted to densities $f \in \mathcal{F}$ with

$$\mathcal{F} = \{\text{densities } f \text{ on } [0,1] \text{ such that } \log f \in \mathbb{L}_1([0,1])\}.$$

We denote by $\mathcal{S}^* : \mathbb{L}_1([0,1]) \to \mathbb{L}_1([0,1])$ the operator

$$\mathcal{S}^*f(x) = \frac{\int_0^1 K_h(u-x)f(u)du}{\int_0^1 K_h(s-x)ds}.$$

Note the difference between $\mathcal{S}$ and $\mathcal{S}^*$. The operator $\mathcal{S}^*$ is in fact the adjoint operator of $\mathcal{S}$. Here, we rely more specifically on the earlier work of Eggermont (1999) that takes into account the case where the density support ($[0,1]$ in our case) is different from the kernel support (usually $\mathbb{R}$). Indeed in this case, the normalisation terms introduce a difference between $\mathcal{S}$ and $\mathcal{S}^*$. Then for a density $f \in \mathcal{F}$, we approach it by a nonlinear smoothing operator $\mathcal{N}$ defined as

$$\mathcal{N}f(x) = \exp\{(\mathcal{S}^*(\log f))(x)\}, \quad x \in [0,1].$$

Note that $\mathcal{N}f$ is not necessarily a density. Now, the maximum smoothed likelihood procedure consists in applying Algorithm 1, relying on

$$\tilde{K}_{i,h}(x) = \frac{K_{i,h}(x)}{\int_0^1 K_{i,h}(s)ds}, \tag{13}$$

where $K_{i,h}$ is defined through (5) relying on a positive symmetric kernel K and

$$\hat{\omega}_i^{(s)} = \frac{(1-\hat{\theta}_n)\mathcal{N}\hat{f}^{(s)}(x_i)}{\hat{\theta}_n + (1-\hat{\theta}_n)\mathcal{N}\hat{f}^{(s)}(x_i)}. \tag{14}$$

In Section 3.2, we explain where these choices come from and why this procedure corresponds to a maximum smoothed likelihood approach. Let us remark that as in `kerfdr` algorithm, the sequence of weights $\{\hat{\omega}^{(s)}\}_{s\geq 0}$ also satisfies $\hat{\omega}^{(s)} = \varphi(\hat{\omega}^{(s-1)})$ for some specific function φ. Then, if the sequence $\{\hat{\omega}^{(s)}\}_{s\geq 0}$ is convergent, it must be convergent to a fixed point of φ. Existence and uniqueness of a fixed point for `msl` algorithm is explored below in Proposition 2.

In the following section, we thus establish theoretical properties of the procedures presented here. These are then further compared on simulated data in Section 4.

3 Mathematical properties of the algorithms

3.1 Randomly weighted kernel estimator

We provide below the convergence properties of the estimator $\hat{f}_n^{\text{rwk}}$ defined through (11). In fact, these naturally depend on the properties of the plug-in estimators $\hat{\theta}_n$ and $\hat{g}_n$. We are interested here in controlling the pointwise quadratic risk of $\hat{f}_n^{\text{rwk}}$. This is possible on a class of densities f that are regular enough. In the following, we denote by $\mathbb{P}_{\theta,f}$ and $\mathbb{E}_{\theta,f}$ the probability and corresponding expectation in the more specific model (2). Moreover, $\lfloor x \rfloor$ denotes the largest integer strictly smaller than x. Now, we recall that the order of a kernel is defined as its first nonzero moment (Tsybakov, 2009) and we recall below the definition of Hölder classes of functions.

Definition 1. *Fix $\beta > 0, L > 0$ and denote by $H(\beta, L)$ the set of functions $\psi : [0,1] \to \mathbb{R}$ that are l-times continuously differentiable on $[0,1]$ with $l = \lfloor \beta \rfloor$ and satisfy*

$$|\psi^{(l)}(x) - \psi^{(l)}(y)| \leq L|x-y|^{\beta - l}, \quad \forall x, y \in [0,1].$$

The set $H(\beta, L)$ is called the (β, L)-Hölder class of functions.

We denote by $\Sigma(\beta, L)$ the set

$$\Sigma(\beta, L) = \Big\{\psi : \psi \text{ is a density on } [0,1] \text{ and } \psi \in H(\beta, L)\Big\}.$$

According to the proof of Theorem 1.1 in Tsybakov (2009), we remark that

$$\sup_{\psi \in \Sigma(\beta, L)} \|\psi\|_\infty < +\infty.$$

In order to obtain the rate of convergence of $\hat{f}_n^{\text{rwk}}$ to f, we introduce the following assumptions

(A1) The kernel K is a right-continuous function.

(A2) K is of bounded variation.

(A3) The kernel K is of order $l = \lfloor \beta \rfloor$ and satisfies

$$\int K(u)du = 1, \ \int K^2(u)du < \infty, \text{ and } \int |u|^\beta |K(u)|du < \infty.$$

(B1) f is a uniformly continuous density function.

(C1) The bandwidth h is of order $\alpha n^{-1/(2\beta+1)}$, $\alpha > 0$.

Note that there exist kernels satisfying Assumptions **(A1)-(A3)** (see for instance Section 1.2.2 in Tsybakov, 2009). Note also that if $f \in \Sigma(\beta, L)$, it automatically satisfies Assumption **(B1)**.

Remark 1. *i) We first remark that if kernel K satisfies Assumptions **(A1)**, **(A2)** and if Assumptions **(B1)** and **(C1)** hold, then the kernel density estimator $\hat{g}_n$ defined by (4) converges uniformly almost surely to g (Wied and Weißbach, 2012). In other words*

$$\|\hat{g}_n - g\|_\infty \xrightarrow[n\to\infty]{a.s} 0.$$

*ii) If kernel K satisfies Assumption **(A3)** and if Assumption **(C1)** holds, then for all $n \geq 1$*

$$\sup_{x\in[0,1]} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_{\theta,f}(|\hat{g}_n(x) - g(x)|^2) \leq Cn^{\frac{-2\beta}{2\beta+1}},$$

where $C = C(\beta, L, \alpha, K)$ (see Theorem 1.1 in Tsybakov, 2009).

In the following theorem, we give the rate of convergence to zero of the pointwise quadratic risk of $\hat{f}_n^{\text{rwk}}$.

Theorem 1. *Assume that kernel K satisfies Assumptions **(A1)-(A3)** and $K \in \mathbb{L}_4(\mathbb{R})$. If $\hat{\theta}_n$ converges almost surely to θ and the bandwidth $h = \alpha n^{-1/(2\beta+1)}$ with $\alpha > 0$, then for any $\delta > 0$, the pointwise quadratic risk of $\hat{f}_n^{rwk}$ satisfies*

$$\sup_{x\in[0,1]} \sup_{\theta\in[\delta,1-\delta]} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_{\theta,f}(|\hat{f}_n^{rwk}(x) - f(x)|^2) \leq C_1 \sup_{\theta\in[\delta,1-\delta]} \sup_{f\in\Sigma(\beta,L)} \left[\mathbb{E}_{\theta,f}\left(|\hat{\theta}_n - \theta|\right)^4\right]^{\frac{1}{2}} + C_2 n^{\frac{-2\beta}{2\beta+1}},$$

where C_1, C_2 are two positive constants depending only on β, L, α, δ and K.

The proof of this theorem is postponed to Section 5.1. It works as follows: we first start by proving that the pointwise quadratic risk of f_2 (which is not an estimator) is of order $n^{-2\beta/(2\beta+1)}$. Then we compare estimator $\hat{f}_n^{\text{rwk}}$ with function f_2 to conclude the proof. We evidently obtain the following corollary from this theorem.

Corollary 1. *Under the assumptions of Theorem 1, if $\hat{\theta}_n$ is such that*

$$\limsup_{n\to+\infty} n^{\frac{2\beta}{2\beta+1}} \left[\mathbb{E}_{\theta,f}\left(|\hat{\theta}_n - \theta|\right)^4\right]^{\frac{1}{2}} < +\infty, \tag{15}$$

then for any fixed value (θ, f), there is some positive constant C such that

$$\sup_{x\in[0,1]} \mathbb{E}_{\theta,f}(|\hat{f}_n^{rwk}(x) - f(x)|^2) \leq Cn^{\frac{-2\beta}{2\beta+1}}.$$

Note that estimators $\hat{\theta}_n$ satisfying (15) exist. Indeed, relying on the same arguments as in the proofs of Propositions 2 or 3 in Nguyen and Matias (2012), we can prove that for instance histogram-based estimators or the estimator proposed by Celisse and Robin (2010) both satisfy that

$$\limsup_{n\to+\infty} n\left[\mathbb{E}_{\theta,f}\left(|\hat{\theta}_n-\theta|\right)^4\right]^{\frac{1}{2}} < +\infty.$$

Note also that the rate $n^{-\beta/(2\beta+1)}$ is the usual nonparametric minimax rate over the class $\Sigma(\beta, L)$ of Hölder densities in the case of direct observations. While we do not formally prove that this is also the case in undirect model (2), it is likely that the rate in this latter case is not faster as the problem is more difficult. A difficulty in establishing such a lower bound lies in the fact that when $\theta \in [\delta, 1-\delta]$ the direct model ($\theta = 0$) is not a submodel of (2). Anyway, such a lower bound would not be sufficient to conclude that estimator $\hat{f}_n^{\mathrm{rwk}}$ achieves the minimax rate. Indeed, the corollary states nothing about uniform convergence of $\hat{f}_n^{\mathrm{rwk}}(x)$ with respect to the parameter value (θ, f) since the convergence of the estimator $\hat{\theta}_n$ is not known to be uniform.

3.2 Maximum smoothed likelihood estimator

Let us now explain the motivations for considering an iterative procedure with functions $\tilde{K}_{i,h}$ and weights $\hat{\omega}_i^{(s)}$ respectively defined through (13) and (14). Instead of the classical log-likelihood, we follow the lines of Levine et al. (2011) and consider (the opposite of) a smoothed version of this log-likelihood as our criterion, namely

$$l_n(\theta, f) = \frac{-1}{n}\sum_{i=1}^{n} \log[\theta + (1-\theta)\mathcal{N}f(X_i)].$$

In this section, we denote by g_0 the true density of the observations X_i. For any fixed value of θ, up to the additive constant $\int_0^1 g_0(x)\log g_0(x)dx$, the smoothed log-likelihood $l_n(\theta, f)$ converges almost surely towards $l(\theta, f)$ defined as

$$l(\theta, f) := \int_0^1 g_0(x)\log\frac{g_0(x)}{\theta + (1-\theta)\mathcal{N}f(x)}dx.$$

This quantity may be viewed as a penalized Kullback-Leibler divergence between the true density g_0 and its smoothed approximation for parameters (θ, f). Indeed, let $D(a \mid b)$ denote the Kullback-Leibler divergence between (positive) measures a and b, defined as

$$D(a \mid b) = \int_0^1 \Big\{a(x)\log\frac{a(x)}{b(x)} + b(x) - a(x)\Big\}dx.$$

Note that in the above definition, a and b are not necessarily probability measures. Moreover it can be seen that we still have the property $D(a|b) \geq 0$ with equality if and only if $a = b$ (Eggermont, 1999). We now obtain

$$l(\theta, f) = D(g_0|\theta + (1-\theta)\mathcal{N}f) + (1-\theta)\big(1 - \int_0^1 \mathcal{N}f(x)dx\big).$$

The second term in the right-hand side of the above equation acts as a penalization term (Eggermont, 1999; Levine et al., 2011). Our goal is to construct an iterative sequence of estimators of f that possesses a descent property with respect to the criterion $l(\theta,\cdot)$, for fixed value θ. Indeed, as previously explained, θ has to remain fixed otherwise the following procedure gives a sequence $\{\theta^t\}$ that converges to 0. We start by describing such a procedure, relying on the knowledge of the parameters (thus an oracle procedure). Let us denote by $l_n(f)$ the smoothed log-likelihood $l_n(\theta,f)$ and by $l(f)$ the limit function $l(\theta,f)$. We want to construct a sequence of densities $\{f^t\}_{t\geq 0}$ such that

$$l(f^t)-l(f^{t+1})\geq cD(f^{t+1}\mid f^t)\geq 0, \tag{16}$$

where c is a positive constant depending on θ, the bandwidth h and the kernel K. We thus consider the difference

$$\begin{aligned} l(f^t)-l(f^{t+1}) &= \int_0^1 g_0(x)\log\frac{\theta+(1-\theta)\mathcal{N}f^{t+1}(x)}{\theta+(1-\theta)\mathcal{N}f^t(x)}dx \\ &= \int_0^1 g_0(x)\log\Big\{1-\omega_t(x)+\omega_t(x)\frac{\mathcal{N}f^{t+1}(x)}{\mathcal{N}f^t(x)}\Big\}dx, \end{aligned}$$

where

$$\omega_t(x)=\frac{(1-\theta)\mathcal{N}f^t(x)}{\theta+(1-\theta)\mathcal{N}f^t(x)}.$$

By the concavity of the logarithm function, we get that

$$\begin{aligned} l(f^t)-l(f^{t+1}) &\geq \int_0^1 g_0(x)\omega_t(x)\log\frac{\mathcal{N}f^{t+1}(x)}{\mathcal{N}f^t(x)}dx \\ &\geq \int_0^1 g_0(x)\omega_t(x)\Big[\mathcal{S}^*(\log f^{t+1})(x)-\mathcal{S}^*(\log f^t)(x)\Big]dx \\ &\geq \int_0^1 g_0(x)\omega_t(x)\Big(\int_0^1 K_h(s-x)ds\Big)^{-1}\Big(\int_0^1 K_h(u-x)\log\frac{f^{t+1}(u)}{f^t(u)}du\Big)dx \\ &\geq \int_0^1\Big(\int_0^1\frac{g_0(x)\omega_t(x)K_h(u-x)}{\int_0^1 K_h(s-x)ds}dx\Big)\log\frac{f^{t+1}(u)}{f^t(u)}du. \end{aligned} \tag{17}$$

Let us define

$$\alpha_t=\frac{1}{\int_0^1\omega_t(u)g_0(u)du}\ \text{and}\ f^{t+1}(x)=\alpha_t\int_0^1\frac{K_h(u-x)\omega_t(u)g_0(u)}{\int_0^1 K_h(s-u)ds}du, \tag{18}$$

then f^{t+1} is a density function on $[0,1]$ and

$$l(f^t)-l(f^{t+1})\geq\frac{1}{\alpha_t}D(f^{t+1}\mid f^t).$$

With the same arguments as in the proof of following Proposition 1, we can show that α_t^{-1} is lower bounded by a positive constant c depending on θ,h and K. The sequence $\{f^t\}_{t\geq 0}$ thus satisfies property (16). However, we stress that it is an oracle as it depends on the knowledge of the true density g_0 that is unknown. Now, the estimator sequence $\{\hat{f}^{(t)}\}_{t\geq 0}$ defined through Equations (13), (14) and Algorithm 1 is exactly the Monte Carlo approximation of $\{f^t\}_{t\geq 0}$. We prove in the next proposition that it also satisfies the descent property (16).

Proposition 1. *For any initial value of the weights $\hat{\omega}_0 \in (0,1)^n$, the sequence of estimators $\{\hat{f}^{(t)}\}_{t\geq 0}$ defined through* (13), (14) *and Algorithm 1 satisfies*

$$l_n(\hat{f}^{(t)}) - l_n(\hat{f}^{(t+1)}) \geq cD(\hat{f}^{(t+1)} \mid \hat{f}^{(t)}) \geq 0,$$

where c is a positive constant depending on θ, the bandwidth h and the kernel K.

To conclude this section, we study the behavior of the limiting criterion l. Let us introduce the set

$$\mathcal{B} = \{\mathcal{S}\varphi; \varphi \text{ density on } [0,1]\}.$$

Proposition 2. *The criterion l has a unique minimum $f^\star$ on $\mathcal{B}$. Moreover, if there exists a constant L depending on h such that for all $x, y \in [-1,1]$*

$$|K_h(x) - K_h(y)| \leq L|x-y|,$$

then the sequence of densities $\{f^t\}_{t\geq 0}$ converges uniformly to $f^\star$.

Note that the previous assumption may be satisfied by many different kernels. For instance, if K is the density of the standard normal distribution, then this assumption is satisfied with

$$L = \frac{1}{h^2\sqrt{2\pi}} e^{-1/2}.$$

As a consequence and since l_n is lower bounded, the sequence $\{\hat{f}^{(t)}\}_{t\geq 0}$ converges to a local minimum of l_n as t increases. Moreover, we recall that as the sample size n increases, the criterion l_n converges (up to a constant) to l. Thus, the outcome of Algorithm 1 that relies on Equations (13) and (14) is an approximation of the minimizer $f^\star$ of l.

4 Estimation of local false discovery rate and simulation study

4.1 Estimation of local false discovery rate

In this section, we study the estimation of local false discovery rate (ℓFDR) by using the previously introduced estimators of the density f and compare these different approaches on simulated data. Let us recall definition (3) of the local false discovery rate

$$\ell\text{FDR}(x) = \mathbb{P}(H_i \text{ being true } |X = x) = \frac{\theta}{\theta + (1-\theta)f(x)}, \quad x \in [0,1].$$

For a given estimator $\hat{\theta}$ of the proportion θ and an estimator $\hat{f}$ of the density f, we obtain a natural estimator of the local false discovery rate for observation x_i

$$\widehat{\ell\text{FDR}}(x_i) = \frac{\hat{\theta}}{\hat{\theta} + (1-\hat{\theta})\hat{f}(x_i)}. \tag{19}$$

Let us now denote by $\hat{f}_{\text{rwk}}$ the randomly weighted kernel estimator of f constructed in Section 2.1, by $\hat{f}_{\text{kerfdr}}$ the estimator of f presented in Algorithm 1 and by $\hat{f}_{\text{msl}}$ the maximum smoothed likelihood estimator of f presented in Algorithm 1. Note that $\hat{f}_{\text{kerfdr}}$ is available through the R package `kerfdr`. We also let $\widehat{\ell\text{FDR}}_m, m \in \{\text{rwk}, \text{kerfdr}, \text{msl}\}$ be the estimators

of ℓFDR induced by a plug-in of estimators $\hat{f}_m$ in (19) and $\widehat{\ell\mathrm{FDR}}_{st}$ be the estimator of ℓFDR computed by the method of Strimmer (2008). We compute the root mean squared error (RMSE) between the estimates and the true values

$$\mathrm{RMSE}_m = \frac{1}{S}\sum_{s=1}^{S}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\widehat{\ell\mathrm{FDR}}_m^{(s)}(x_i) - \ell\mathrm{FDR}(x_i)\}^2},$$

for $m \in \{\mathrm{rwk}, \mathrm{kerfdr}, \mathrm{msl}, \mathrm{st}\}$ and where $s = 1, \ldots, S$ denotes the simulation index (S being the total number of repeats). We also compare $\mathbb{L}^2$-norms between $\hat{f}_m$ and f for $m \in \{\mathrm{rwk}, \mathrm{kerfdr}, \mathrm{msl}\}$, relying on the root mean integrated squared error

$$\mathrm{RMISE}_m = \frac{1}{S}\sum_{s=1}^{S}\sqrt{\int_0^1 [\hat{f}_m^{(s)}(u) - f(u)]^2 du}.$$

The quality of the estimates provided by method m is measured by the mean RMSE_m or RMISE_m: the smaller these quantities, the better the performances of the method.

We mention that we also tested the naive method described in Section 2.1 and the results were bad. In order to present clear figures, we have chosen not to show those.

4.2 Simulation study

In this section, we give an illustration of the previous results on some simulated experiments. We simulate sets of p-values according to the mixture model (2). We consider three different cases for the alternative distribution f and two different values for the proportion: $\theta = 0.65$ and 0.85. In the first case, we simulate p-values under the alternative with distribution

$$f(x) = \rho\Big(1 - x\Big)^{\rho-1}\mathbf{1}_{[0,1]}(x),$$

where $\rho = 4$, as proposed in Celisse and Robin (2010). In the second case, the p-value corresponds to the statistic T which has a mixture distribution $\theta\mathcal{N}(0,1) + (1-\theta)\mathcal{N}(\mu,1)$, with $\mu = 2$. In the third case, the p-value corresponds to the statistic T which has a mixture density $\theta(1/2)\exp\{-|t|\} + (1-\theta)(1/2)\exp\{-|t-\mu|\}$, with $\mu = 1$. The p-values densities obtained with those three models are given in Figure 1 for $\theta = 0.65$.

For each of the $3 \times 2 = 6$ configurations, we generate $S = 100$ samples of size $n \in \{500, 1000, 2000, 5000\}$. In these experiments, we choose to consider the estimator of θ initially proposed by Schweder and Spjøtvoll (1982), namely

$$\hat{\theta} = \frac{\#\{X_i > \lambda; i = 1, \ldots, n\}}{n(1-\lambda)},$$

with parameter value λ optimally chosen by bootstrap method, as recommended by Storey (2002). The kernel is chosen with compact support, for example the triangular kernel or the rectangular kernel. The bandwidth is selected according to a rule of thumb due to (Silverman, 1986, Section 3.4.2),

$$h = 0.9\min\Big\{SD, \frac{IQR}{1.34}\Big\}n^{-1/5},$$

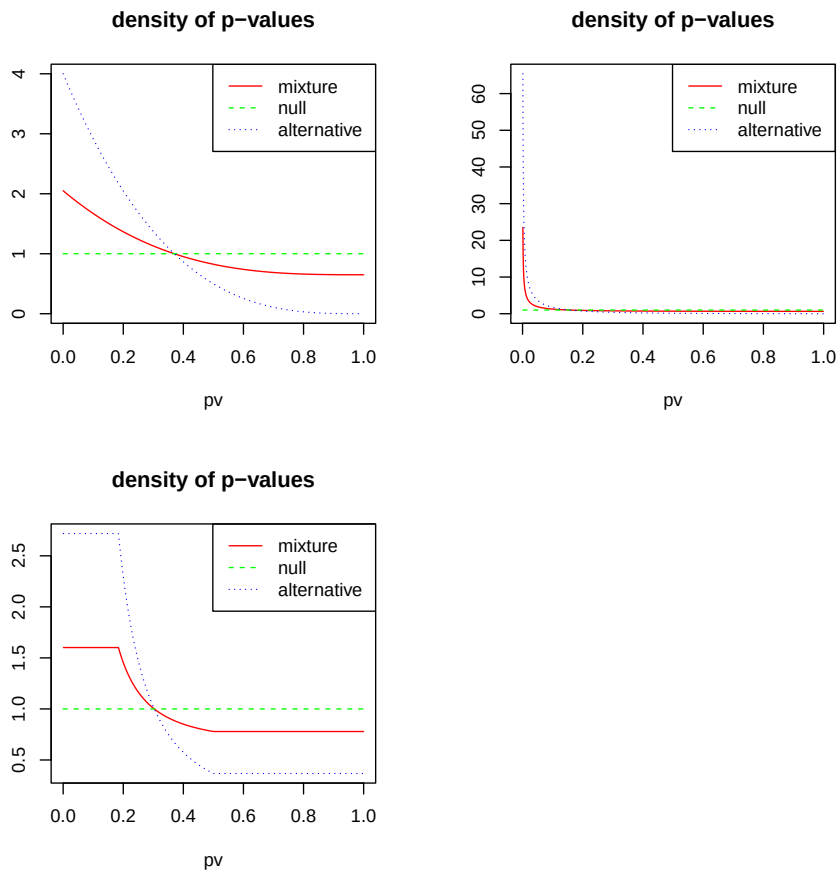

Figure 1: Densities of the p-values in the three different models, with $\theta = 0.65$. Top left: first model, top right: second model, bottom left: third model.

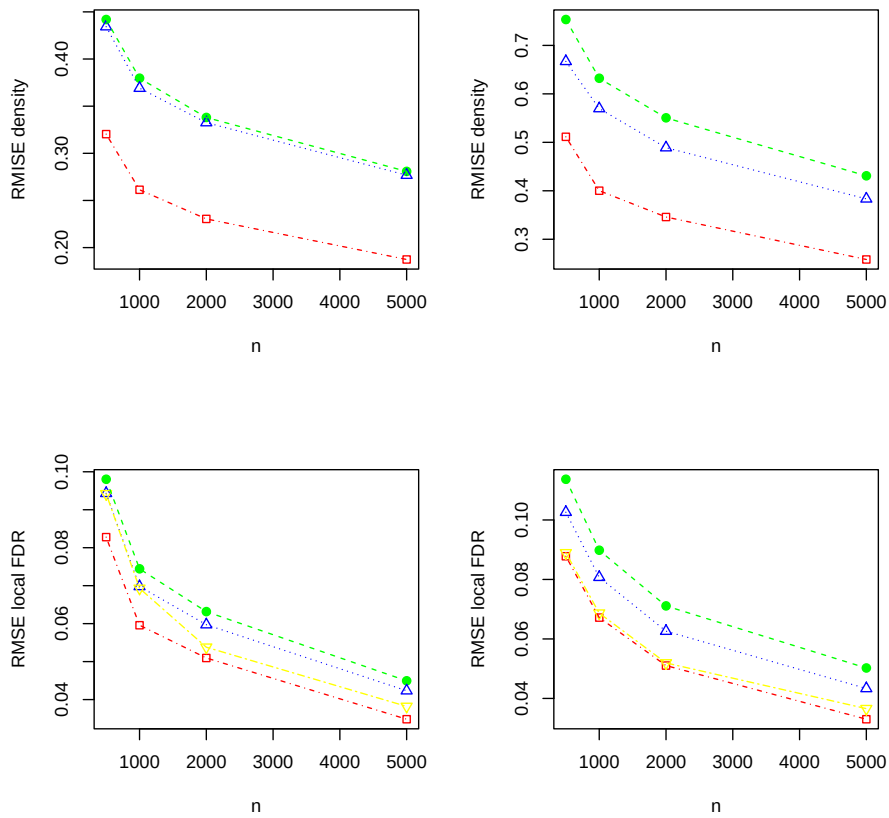

Figure 2: RMISE (for density f) and RMSE (for ℓFDR) in the first model as a function of n. Methods: "$\bullet$" = rwk, "$\triangle$" = `kerfdr`, "$\square$" = msl, "$\triangledown$" = st (only for ℓFDR). Left: $\theta = 0.65$, right: $\theta = 0.85$.

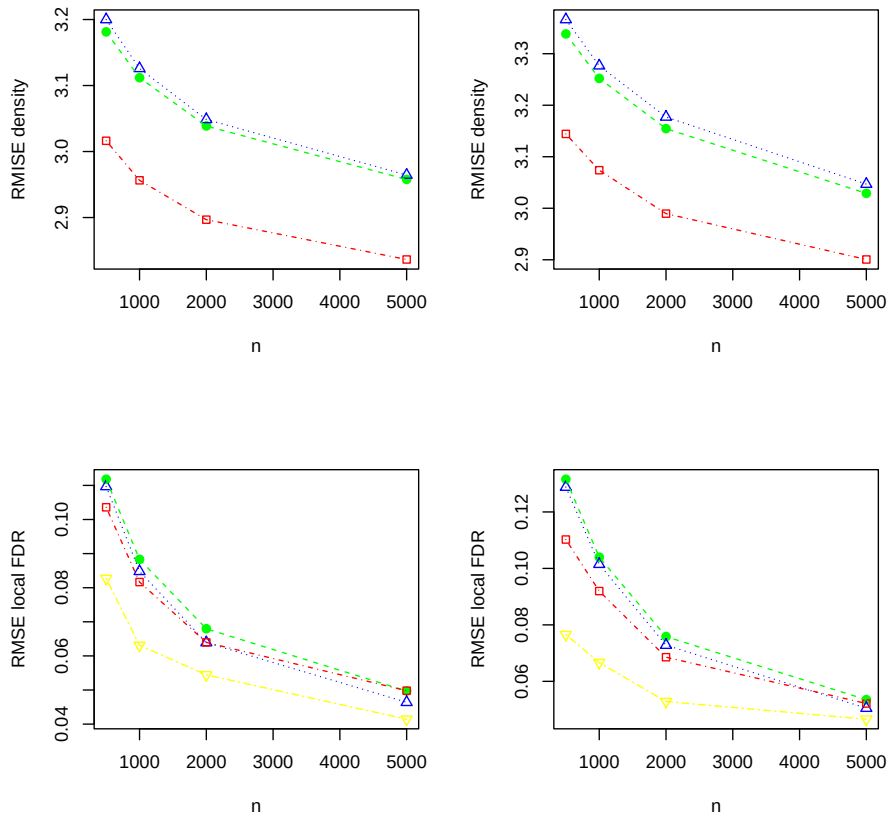

Figure 3: RMISE (for density f) and RMSE (for ℓFDR) in the second model as a function of n. Methods: "$\bullet$" = rwk, "$\triangle$" = `kerfdr`, "$\square$" = msl, "$\triangledown$" = st (only for ℓFDR). Left: $\theta = 0.65$, right: $\theta = 0.85$.

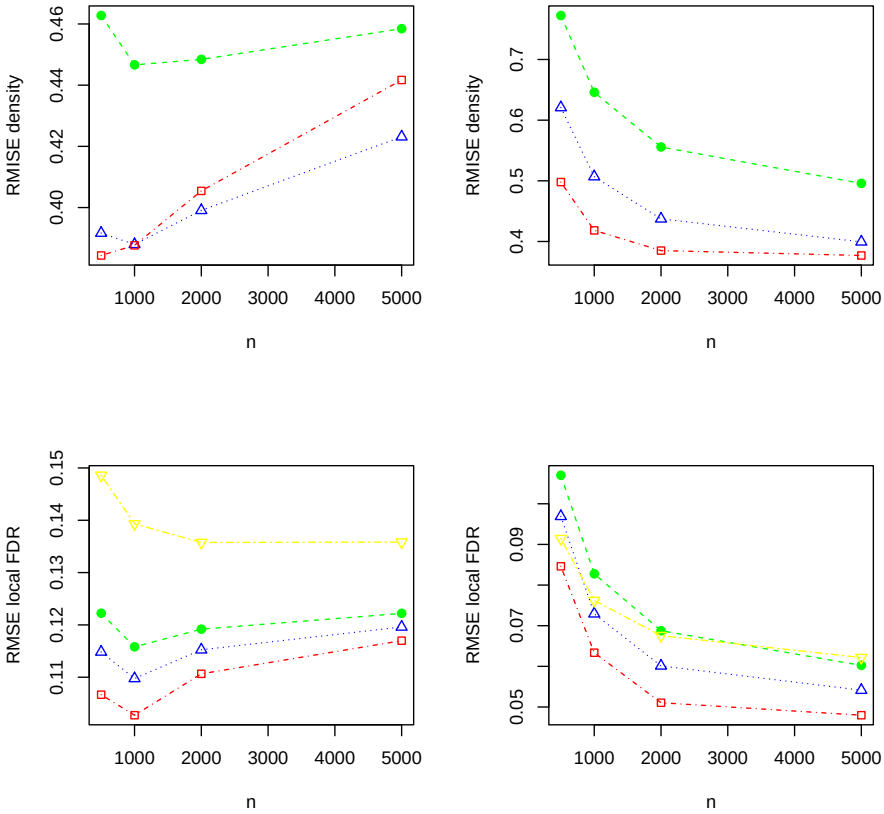

Figure 4: RMISE (for density f) and RMSE (for ℓFDR) in the third model as a function of n. Methods: "$\bullet$" = rwk, "$\triangle$" = `kerfdr`, "$\square$" = msl, "$\triangledown$" = st (only for ℓFDR). Left: $\theta = 0.65$, right: $\theta = 0.85$.

where SD and IQR are respectively the standard deviation and interquartile range of the data values. Figures 2, 3 and 4 show the RMISEs and the RMSEs for the six configurations and the four different methods.

We first comment the results on the estimation of f (top half of each figure). Except for model 2, the RMISEs obtained are small for all the three procedures. Model 2 exhibits a rather high RMISEs and this may be explained by the fact that density f is not bounded near 0 in this case. We note that the methods `rwk` and `kerfdr` have very similar performances, except in the third model where `kerfdr` seems to slightly outperform `rwk`. Let us recall that we introduced this latter method only as a way of approaching the theoretical performances of `kerfdr` method. Now, in five out of the six configurations, `msl` outperforms the two other methods (`rwk, kerfdr`).

Then, we switch to comparing the methods with respect to estimation of ℓFDR (bottom half of each figure). First, note that the four methods exhibit small RMSEs with respect to ℓFDR and are thus efficient for estimating this quantity. We also note that `rwk` tends to have lower performances than `kerfdr,msl`. Now, `msl` tends to slightly outperform `kerfdr`. Thus `msl` appears as a competitive method for ℓFDR estimation. The comparison with Strimmer (2008)'s approach is more difficult: for model 1, the method compares with `msl`, while it outperforms all the methods in model 2 and is outperformed by `msl` in model 3.

As a conclusion, we claim that `msl` is a competitive method for estimating both the alternative density f and the ℓFDR.

5 Proofs

5.1 Proof of Theorem 1

The proof works as follows: we first start by proving that the pointwise quadratic risk of function f_2 defined by (8) is order of $n^{-2\beta/(2\beta+1)}$ in the following proposition. Then we compare the estimator $\hat{f}_n^{\text{rwk}}$ with the function f_2 to conclude the proof. To simplify notation, we abbreviate $\hat{f}_n^{\text{rwk}}$ to $\hat{f}_n$.

We shall need the following two lemmas. The proof of the first one may be found for instance in Proposition 1.2 in Tsybakov (2009). The second one is known as Bochner's lemma and is a classical result in kernel density estimation. Therefore its proof is omitted.

Lemma 1. ***(Proposition 1.2 in Tsybakov (2009)).*** *Let p be a density in $\Sigma(\beta, L)$ and K a kernel function of order $l = \lfloor \beta \rfloor$ such that*

$$\int_{\mathbb{R}} |u|^{\beta} |K(u)| du < \infty.$$

Then there exists a positive constant C_3 depending only on β, L and K such that for all $x_0 \in \mathbb{R}$,

$$\Big| \int_{\mathbb{R}} K(u) \big[p(x_0 + uh) - p(x_0) \big] du \Big| \leq C_3 h^{\beta}, \quad \forall h > 0.$$

Lemma 2. ***(Bochner's lemma).*** *Let g be a bounded function on $\mathbb{R}$, continuous in a neighborhood of $x_0 \in \mathbb{R}$ and Q a function which satisfies*

$$\int_{\mathbb{R}} |Q(x)| dx < \infty.$$

Then, we have

$$\lim_{h\to 0}\frac{1}{h}\int_{\mathbb{R}} Q\Big(\frac{x-x_0}{h}\Big)g(x)dx = g(x_0)\int_{\mathbb{R}} Q(x)dx.$$

Now, we come to the first step in the proof.

Proposition 3. *Assume that kernel K satisfies Assumption **(A3)** and bandwidth $h = \alpha n^{-1/(2\beta+1)}$, with $\alpha > 0$. Then the pointwise quadratic risk of function f_2, defined by* (8) *and depending on (θ, f), satisfies*

$$\sup_{x\in[0,1]}\ \sup_{\theta\in[\delta,1-\delta]}\ \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_{\theta,f}(|f_2(x)-f(x)|^2) \le C_4 n^{\frac{-2\beta}{2\beta+1}},$$

where C_4 is a positive constant depending only on β, L, α, δ and K.

Proof of Proposition 3. Let us denote by

$$S_n = \sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}.$$

The pointwise quadratic risk of f_2 can be written as the sum of a bias term and a variance term

$$\mathbb{E}_{\theta,f}(|f_2(x)-f(x)|^2) = [\mathbb{E}_{\theta,f}(f_2(x)) - f(x)]^2 + \mathbb{V}\mathrm{ar}_{\theta,f}[f_2(x)].$$

Let us first study the bias term. According to (8) and the definition (7) of the weights, we have

$$\begin{aligned}
\mathbb{E}_{\theta,f}[f_2(x)] &= \frac{n}{h}\mathbb{E}_{\theta,f}\left[\tau_1 K\Big(\frac{x-X_1}{h}\Big)\Big(\sum_{k=1}^{n}\tau_k\Big)^{-1}\right]\\
&= \frac{n}{h}\mathbb{E}_{\theta,f}\left[\frac{f(X_1)}{g(X_1)}K\Big(\frac{x-X_1}{h}\Big)S_n^{-1}\right]\\
&= \frac{n}{h}\int_0^1 f(t)K\Big(\frac{x-t}{h}\Big)\mathbb{E}_{\theta,f}\left[\Big(\frac{f(t)}{g(t)}+S_{n-1}\Big)^{-1}\right]dt\\
&= n\int_{-x/h}^{(1-x)/h} K(t)f(x+th)\mathbb{E}_{\theta,f}\left[\Big(\frac{f(x+th)}{g(x+th)}+S_{n-1}\Big)^{-1}\right]dt. \qquad (20)
\end{aligned}$$

Since the functions f and g are related by the equation $g(t) = \theta + (1-\theta)f(t)$ for all $t \in [0,1]$, the ratio $f(t)/g(t)$ is well defined and satisfies

$$0 \le \frac{f(t)}{g(t)} \le \frac{1}{1-\theta} \le \delta^{-1}, \quad \forall t\in[0,1], \text{ and } \forall\theta\in[\delta,1-\delta].$$

Then for all $t \in [-x/h, (1-x)/h]$, we get

$$\frac{1}{S_{n-1}+\delta^{-1}} \le \left(\frac{f(x+th)}{g(x+th)}+S_{n-1}\right)^{-1} \le \frac{1}{S_{n-1}},$$

where the bounds are uniform with respect to t.
By combining this inequality with (20), we obtain

$$n\left(\int_{-x/h}^{(1-x)/h} K(t)f(x+th)dt\right)\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-1}+\delta^{-1}}\Big) \leq \mathbb{E}_{\theta,f}\big[f_2(x)\big]$$

$$\text{and} \quad \mathbb{E}_{\theta,f}\big[f_2(x)\big] \leq n\left(\int_{-x/h}^{(1-x)/h} K(t)f(x+th)dt\right)\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-1}}\Big).$$

Then, we apply the following lemma, whose proof is postponed to Appendix A.1.

Lemma 3. *There exist some positive constants c_1, c_2, c_3, c_4 (depending on δ) such that for n large enough,*

$$\mathbb{E}_{\theta,f}\Big(\frac{1}{S_n}\Big) \leq \frac{1}{n} + \frac{c_1}{n^2}, \tag{21}$$

$$\mathbb{E}_{\theta,f}\Big(\frac{1}{S_n^2}\Big) \leq \frac{c_2}{n^2}, \tag{22}$$

$$\mathbb{E}_{\theta,f}\Big(\frac{1}{S_n+2\delta^{-1}}\Big) \geq \frac{1}{n} - \frac{c_3}{n^2}, \tag{23}$$

$$\textit{and} \quad \mathbb{E}_{\theta,f}\Big(\frac{1}{S_n^2}\Big) - \mathbb{E}^2_{\theta,f}\Big(\frac{1}{\delta^{-1}+S_n}\Big) \leq \frac{c_4}{n^3}. \tag{24}$$

Relying on Inequalities (21) and (23), we have for n large enough

$$\int_{-x/h}^{(1-x)/h} K(t)f(x+th)dt - \frac{c_3}{n} \leq \mathbb{E}_{\theta,f}\big[f_2(x)\big] \leq \int_{-x/h}^{(1-x)/h} K(t)f(x+th)dt + \frac{c_1}{n}.$$

Since $f(x+th)=0$ for all $t \notin [-x/h, (1-x)/h]$, we may write

$$\int_{-x/h}^{(1-x)/h} K(t)f(x+th)dt = \int_{\mathbb{R}} K(t)f(x+th)dt.$$

Thus, the bias of $f_2(x)$ satisfies

$$|b(x)| = |\mathbb{E}_{\theta,f}\big[f_2(x)\big] - f(x)| \leq \int_{\mathbb{R}} K(t)|f(x+th)-f(x)|dt + \frac{c_5}{n}.$$

By using Lemma 1 and the choice of bandwidth h, we obtain that

$$b^2(x) \leq C_5 h^{2\beta},$$

where $C_5 = C_5(\beta, L, K)$. Let us study now the variance term of $f_2(x)$. We have

$$\mathbb{V}\mathrm{ar}_{\theta,f}\big[f_2(x)\big] = \frac{1}{h^2}\big[n\mathbb{V}\mathrm{ar}_{\theta,f}(Y_1) + n(n-1)\mathbb{C}\mathrm{ov}_{\theta,f}(Y_1,Y_2)\big], \tag{25}$$

where

$$Y_i = \frac{f(X_i)}{g(X_i)}K\Big(\frac{x-X_i}{h}\Big)S_n^{-1}.$$

The variance of Y_1 is bounded by its second moment and

$$\mathbb{E}_{\theta,f}(Y_1^2) = \mathbb{E}_{\theta,f}\left[\Big(\frac{f(X_1)}{g(X_1)}\Big)^2 K^2\Big(\frac{x-X_1}{h}\Big)S_n^{-2}\right] = \int_0^1 \frac{f^2(t)}{g(t)}K^2\Big(\frac{x-t}{h}\Big)\mathbb{E}_{\theta,f}\Big[\Big(\frac{f(t)}{g(t)}+S_{n-1}\Big)^{-2}\Big]dt.$$

Now, recalling that $0 \le f/g \le \delta^{-1}$ and using Inequality (22) of Lemma 3, we get

$$\begin{aligned}\mathbb{E}_{\theta,f}(Y_1^2) &\le h\Big(\int_{-x/h}^{(1-x)/h}\frac{f^2(x+th)}{g(x+th)}K^2(t)dt\Big)\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-1}^2}\Big)\\ &\le h\delta^{-1}\sup_{f\in\Sigma(\beta,L)}\|f\|_\infty\Big(\int K^2(t)dt\Big)\frac{c_2}{n^2} \le \frac{C_6 h}{n^2}. \qquad (26)\end{aligned}$$

We now study the covariance of Y_1 and Y_2

$$\begin{aligned}&\mathbb{C}\mathrm{ov}_{\theta,f}(Y_1,Y_2) = \mathbb{E}_{\theta,f}(Y_1Y_2) - \mathbb{E}_{\theta,f}^2(Y_1)\\ =\ &\mathbb{E}_{\theta,f}\left[\frac{f(X_1)f(X_2)}{g(X_1)g(X_2)}K\Big(\frac{x-X_1}{h}\Big)K\Big(\frac{x-X_2}{h}\Big)S_n^{-2}\right] - \mathbb{E}_{\theta,f}^2\left[\frac{f(X_1)}{g(X_1)}K\Big(\frac{x-X_1}{h}\Big)S_n^{-1}\right]\\ =\ &\int_{[0,1]^2} f(t)f(u)K\Big(\frac{x-t}{h}\Big)K\Big(\frac{x-u}{h}\Big)\mathbb{E}_{\theta,f}\left[\Big(\frac{f(t)}{g(t)}+\frac{f(u)}{g(u)}+S_{n-2}\Big)^{-2}\right]dtdu\\ &-\left(\int_0^1 f(t)K\Big(\frac{x-t}{h}\Big)\mathbb{E}_{\theta,f}\left[\Big(\frac{f(t)}{g(t)}+S_{n-1}\Big)^{-1}\right]dt\right)^2\\ =\ &\int_{[0,1]^2} f(t)f(u)K\Big(\frac{x-t}{h}\Big)K\Big(\frac{x-u}{h}\Big)A(t,u)dtdu,\end{aligned}$$

where

$$\begin{aligned}A(t,u) &= \mathbb{E}_{\theta,f}\left[\Big(\frac{f(t)}{g(t)}+\frac{f(u)}{g(u)}+S_{n-2}\Big)^{-2}\right] - \mathbb{E}_{\theta,f}\left[\Big(\frac{f(t)}{g(t)}+S_{n-1}\Big)^{-1}\right]\mathbb{E}_{\theta,f}\left[\Big(\frac{f(u)}{g(u)}+S_{n-1}\Big)^{-1}\right]\\ &\le \mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-2}^2}\Big) - \mathbb{E}_{\theta,f}^2\Big(\frac{1}{2\delta^{-1}+S_{n-2}}\Big).\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Y_1,Y_2) &\le \int_{[0,1]^2} f(t)f(u)K\Big(\frac{x-t}{h}\Big)K\Big(\frac{x-u}{h}\Big)\left[\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-2}^2}\Big) - \mathbb{E}_{\theta,f}^2\Big(\frac{1}{2\delta^{-1}+S_{n-2}}\Big)\right]dtdu\\ &\le h^2\left(\int_{\mathbb{R}} f(x+th)K(t)dt\right)^2\left[\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-2}^2}\Big) - \mathbb{E}_{\theta,f}^2\Big(\frac{1}{2\delta^{-1}+S_{n-2}}\Big)\right]\\ &\le C_7h^2\left[\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-2}^2}\Big) - \mathbb{E}_{\theta,f}^2\Big(\frac{1}{2\delta^{-1}+S_{n-2}}\Big)\right].\end{aligned}$$

According to Inequality (24) of Lemma 3, we have

$$\mathbb{E}_{\theta,f}\Big(\frac{1}{S_{n-2}^2}\Big) - \mathbb{E}_{\theta,f}^2\Big(\frac{1}{2\delta^{-1}+S_{n-2}}\Big) \le \frac{c_4}{n^3},$$

hence

$$\mathbb{C}\mathrm{ov}_{\theta,f}(Y_1,Y_2) \leq \frac{C_8 h^2}{n^3}. \tag{27}$$

By returning to Equality (25) and combining with (26) and (27), we obtain

$$\mathbb{V}\mathrm{ar}_{\theta,f}\big[f_2(x)\big] \leq \frac{1}{h^2}\left[\frac{C_6 h}{n} + n(n-1)h^2\frac{C_8 h^2}{n^3}\right] \leq \frac{C_9}{nh}.$$

Thus, as the bandwidth h is of order $n^{-1/(2\beta+1)}$, the pointwise quadratic risk of $f_2(x)$ satisfies

$$\mathbb{E}_{\theta,f}(|f_2(x)-f(x)|^2) \leq C_4 n^{\frac{-2\beta}{2\beta+1}}.$$

□

Proof of Theorem 1. First, the pointwise quadratic risk of $\hat{f}_n(x)$ is bounded in the following way

$$\mathbb{E}_{\theta,f}(|\hat{f}_n(x)-f(x)|^2) \leq 2\mathbb{E}_{\theta,f}(|f_2(x)-f(x)|^2) + 2\mathbb{E}_{\theta,f}(|\hat{f}_n(x)-f_2(x)|^2). \tag{28}$$

According to Proposition 3, we have

$$\mathbb{E}_{\theta,f}(|f_2(x)-f(x)|^2) \leq C_4 n^{\frac{-2\beta}{2\beta+1}}, \tag{29}$$

and it remains to study the second term appearing in the right-hand side of (28). We write

$$\begin{aligned}
\hat{f}_n(x)-f_2(x) &= \frac{1}{h}\sum_{i=1}^n\left(\frac{\hat{\tau}_i}{\sum_k\hat{\tau}_k}-\frac{\tau_i}{\sum_k\tau_k}\right)K\left(\frac{x-X_i}{h}\right)\\
&= \frac{1}{h}\sum_{i=1}^n\frac{\hat{\tau}_i-\tau_i}{\sum_k\hat{\tau}_k}K\left(\frac{x-X_i}{h}\right)+\frac{1}{h}\sum_{i=1}^n\tau_i\left(\frac{1}{\sum_k\hat{\tau}_k}-\frac{1}{\sum_k\tau_k}\right)K\left(\frac{x-X_i}{h}\right)\\
&= \frac{n}{\sum_k\hat{\tau}_k}\times\frac{1}{nh}\sum_{i=1}^n(\hat{\tau}_i-\tau_i)K\left(\frac{x-X_i}{h}\right)\\
&\quad+\frac{n^2}{\sum_k\hat{\tau}_k\sum_k\tau_k}\times\frac{\sum_k(\tau_k-\hat{\tau}_k)}{n}\times\frac{1}{nh}\sum_{i=1}^n\tau_i K\left(\frac{x-X_i}{h}\right).
\end{aligned}$$

Moreover, recalling the definition of the weights (10), we have for all $1\leq i\leq n$,

$$\hat{\tau}_i-\tau_i = \frac{\hat{\theta}_n}{\tilde{g}_n(X_i)}-\frac{\theta}{g(X_i)} = \hat{\theta}_n\Big[\frac{1}{\tilde{g}_n(X_i)}-\frac{1}{g(X_i)}\Big]+\frac{1}{g(X_i)}(\hat{\theta}_n-\theta),$$

and thus get

$$\begin{aligned}
\hat{f}_n(x)-f_2(x) &= \frac{n\hat{\theta}_n}{\sum_k\hat{\tau}_k}\times\frac{1}{nh}\sum_{i=1}^n\Big[\frac{1}{\tilde{g}_n(X_i)}-\frac{1}{g(X_i)}\Big]K\left(\frac{x-X_i}{h}\right)\\
&\quad+\frac{n(\hat{\theta}_n-\theta)}{\sum_k\hat{\tau}_k}\times\frac{1}{nh}\sum_{i=1}^n\frac{1}{g(X_i)}K\left(\frac{x-X_i}{h}\right)\\
&\quad+\frac{n^2\hat{\theta}_n}{\sum_k\hat{\tau}_k\sum_k\tau_k}\times\frac{1}{n}\sum_k\Big[\frac{1}{\tilde{g}_n(X_k)}-\frac{1}{g(X_k)}\Big]\times\frac{1}{nh}\sum_{i=1}^n\tau_i K\left(\frac{x-X_i}{h}\right)\\
&\quad+\frac{n^2(\hat{\theta}_n-\theta)}{\sum_k\hat{\tau}_k\sum_k\tau_k}\times\frac{1}{n}\sum_k\frac{1}{g(X_k)}\times\frac{1}{nh}\sum_{i=1}^n\tau_i K\left(\frac{x-X_i}{h}\right).
\end{aligned} \tag{30}$$

Let us control the different terms appearing in this latter equality. We first remark that for all i,

$$0 \leq \tau_i \leq 1 \text{ and } \frac{1}{g(X_i)} \leq \frac{1}{\theta} \leq \delta^{-1}. \tag{31}$$

Since by assumption $\hat{\theta}_n \xrightarrow[n\to\infty]{as} \theta \in [0,1]$, for n large enough we also get $|\hat{\theta}_n| < 3/2$, a.s. According to the law of large numbers and $\mathbb{E}_{\theta,f}(\tau_1) = 1-\theta$, we also obtain that for n large enough

$$\frac{\delta}{2} \leq \frac{1-\theta}{2} \leq \frac{1}{n}\sum_{i=1}^{n}\tau_i \leq \frac{3(1-\theta)}{2} \leq \frac{3(1-\delta)}{2} \quad \text{a.s.} \tag{32}$$

Moreover, by using a Taylor expansion of the function $u \mapsto 1/u$ with an integral form of the remainder term, we have for all i,

$$\Big|\frac{1}{\tilde{g}_n(X_i)} - \frac{1}{g(X_i)}\Big| = \frac{|\tilde{g}_n(X_i) - g(X_i)|}{g^2(X_i)} \int_0^1 \left(1 + s\frac{\tilde{g}_n(X_i) - g(X_i)}{g(X_i)}\right)^{-2} ds.$$

Since convergence of $\hat{g}_n$ to g is valid pointwise and in $\mathbb{L}_\infty$ norm (see Remark 1), and since $\tilde{g}_n$ is a slight modification of $\hat{g}_n$, we have almost surely, for n large enough and for all $s \in [0,1]$ and all $x \in [0,1]$,

$$1 + s\frac{\tilde{g}_n(x) - g(x)}{g(x)} \geq 1 - s\frac{\|\hat{g}_n - g\|_\infty}{\theta} \geq 1 - \frac{s}{2} > 0.$$

Hence, for all $x \in [0,1]$ and large enough n,

$$\int_0^1 \left(1 + s\frac{\tilde{g}_n(x) - g(x)}{g(x)}\right)^{-2} ds \leq \int_0^1 \frac{4ds}{(2-s)^2} = 2,$$

and we obtain

$$\Big|\frac{1}{\tilde{g}_n(X_i)} - \frac{1}{g(X_i)}\Big| \leq 2\delta^{-2}|\tilde{g}_n(X_i) - g(X_i)| \quad \text{a.s.} \tag{33}$$

We also use the following lemma, whose proof is postponed to Appendix A.2.

Lemma 4. *For large enough n, we have*

$$\frac{n}{|\sum_k \hat{\tau}_k|} \leq c_7 \quad a.s. \tag{34}$$

By returning to Equality (30) and combining with (31), (32), (33) and (34), we obtain

$$\begin{aligned} |\hat{f}_n(x) - f_2(x)|^2 \;\leq\; & c_8\left(\frac{1}{nh}\sum_{i=1}^{n}|\tilde{g}_n(X_i) - g(X_i)| \times \Big|K\Big(\frac{x-X_i}{h}\Big)\Big|\right)^2 \\ & + c_9|\hat{\theta}_n - \theta|^2\left(\frac{1}{nh}\sum_{i=1}^{n}\Big|K\Big(\frac{x-X_i}{h}\Big)\Big|\right)^2 \\ & + c_{10}\left(\frac{1}{n}\sum_{i=1}^{n}|\tilde{g}_n(X_i) - g(X_i)|\right)^2\left(\frac{1}{nh}\sum_{i=1}^{n}K\Big(\frac{x-X_i}{h}\Big)\right)^2 \quad \text{a.s.} \end{aligned} \tag{35}$$

We now successively control the expectations T_1, T_2 and T_3 of the three terms appearing in this upper-bound. For the first term, we have

$$\begin{aligned}
T_1 &= \mathbb{E}_{\theta,f}\left[\left(\frac{1}{nh}\sum_{i=1}^{n}|\tilde{g}_n(X_i)-g(X_i)|\times\left|K\Big(\frac{x-X_i}{h}\Big)\right|\right)^2\right] \\
&= \mathbb{E}_{\theta,f}\left[\frac{1}{n^2h^2}\sum_{i,j=1}^{n}|\tilde{g}_n(X_i)-g(X_i)||\tilde{g}_n(X_j)-g(X_j)|\times\left|K\Big(\frac{x-X_i}{h}\Big)K\Big(\frac{x-X_j}{h}\Big)\right|\right] \\
&= \frac{1}{nh}\mathbb{E}_{\theta,f}\left[\frac{1}{h}|\tilde{g}_n(X_1)-g(X_1)|^2K^2\Big(\frac{x-X_1}{h}\Big)\right] \\
&\quad+\frac{n-1}{n}\mathbb{E}_{\theta,f}\left[\frac{1}{h^2}|\tilde{g}_n(X_1)-g(X_1)||\tilde{g}_n(X_2)-g(X_2)|\times\left|K\Big(\frac{x-X_1}{h}\Big)K\Big(\frac{x-X_2}{h}\Big)\right|\right].
\end{aligned}$$

Now,

$$\begin{aligned}
T_{11} &= \mathbb{E}_{\theta,f}\left[\frac{1}{h}|\tilde{g}_n(X_1)-g(X_1)|^2K^2\Big(\frac{x-X_1}{h}\Big)\right] \\
&= \int_0^1 \mathbb{E}_{\theta,f}\left(|\hat{g}_{n-1}(t)-g(t)|^2\right)K^2\Big(\frac{x-t}{h}\Big)\frac{g(t)}{h}dt \quad \text{(according to definition (10))} \\
&\leq C_{10}n^{\frac{-2\beta}{2\beta+1}}\int_0^1 K^2\Big(\frac{x-t}{h}\Big)\frac{g(t)}{h}dt \quad \text{(according to Remark 1)} \\
&\leq C_{11}n^{\frac{-2\beta}{2\beta+1}} \text{ (according to Lemma 2)}, \qquad (36)
\end{aligned}$$

and in the same way

$$\begin{aligned}
T_{12} =& \mathbb{E}_{\theta,f}\left[\frac{1}{h^2}|\tilde{g}_n(X_1)-g(X_1)||\tilde{g}_n(X_2)-g(X_2)|K\Big(\frac{x-X_1}{h}\Big)K\Big(\frac{x-X_2}{h}\Big)\right] \\
=& \int_0^1\int_0^1 \mathbb{E}_{\theta,f}\Big[\Big|\frac{n-2}{n-1}\hat{g}_{n-2}(t)-g(t)+\frac{1}{(n-1)h}K\Big(\frac{t-s}{h}\Big)\Big| \\
&\times\Big|\frac{n-2}{n-1}\hat{g}_{n-2}(s)-g(s)+\frac{1}{(n-1)h}K\Big(\frac{s-t}{h}\Big)\Big|\Big]\Big|K\Big(\frac{x-t}{h}\Big)K\Big(\frac{x-s}{h}\Big)\Big|\frac{g(t)g(s)}{h^2}dtds.
\end{aligned}$$

This last term is upper-bound by

$$\begin{aligned}
T_{12} \leq& \int_0^1\int_0^1 \mathbb{E}_{\theta,f}\Big[\Big(|\hat{g}_{n-2}(t)-g(t)|+\frac{1}{n-1}g(t)+\frac{1}{(n-1)h}\Big|K\Big(\frac{t-s}{h}\Big)\Big|\Big) \\
&\times\Big(|\hat{g}_{n-2}(s)-g(s)|+\frac{1}{n-1}g(s)+\frac{1}{(n-1)h}\Big|K\Big(\frac{s-t}{h}\Big)\Big|\Big)\Big] \\
&\times\Big|K\Big(\frac{x-t}{h}\Big)K\Big(\frac{x-s}{h}\Big)\Big|\frac{g(t)g(s)}{h^2}dtds \\
\leq& \int_0^1\int_0^1\left\{\mathbb{E}_{\theta,f}^{1/2}\left[|\hat{g}_{n-2}(t)-g(t)|^2\right]\mathbb{E}_{\theta,f}^{1/2}\left[|\hat{g}_{n-2}(s)-g(s)|^2\right]+o\Big(\frac{1}{nh}\Big)\right\} \\
&\times\Big|K\Big(\frac{x-t}{h}\Big)K\Big(\frac{x-s}{h}\Big)\Big|\frac{g(t)g(s)}{h^2}dtds.
\end{aligned}$$

According to Remark 1, we have

$$T_{12} \leq C_{12} n^{\frac{-2\beta}{2\beta+1}} \left[\int_0^1 \left| K\Big(\frac{x-t}{h}\Big) \right| \frac{g(t)}{h} dt \right]^2 \leq C_{13} n^{\frac{-2\beta}{2\beta+1}} \text{ (according to Lemma 2).} \tag{37}$$

Thus we get that

$$T_1 = \mathbb{E}_{\theta,f} \left[\left(\frac{1}{nh} \sum_{i=1}^n |\tilde{g}_n(X_i) - g(X_i)| \left| K\Big(\frac{x-X_i}{h}\Big) \right| \right)^2 \right] \leq C_{14} n^{\frac{-2\beta}{2\beta+1}}. \tag{38}$$

For the second term in the right hand side of (35), we have

$$\begin{aligned} T_2 &= \mathbb{E}_{\theta,f} \left[|\hat{\theta}_n - \theta|^2 \left(\frac{1}{nh} \sum_{i=1}^n \left| K\Big(\frac{x-X_i}{h}\Big) \right| \right)^2 \right] \\ &\leq \mathbb{E}_{\theta,f}^{1/2} \left[|\hat{\theta}_n - \theta|^4 \right] \mathbb{E}_{\theta,f}^{1/2} \left[\left(\frac{1}{nh} \sum_{i=1}^n \left| K\Big(\frac{x-X_i}{h}\Big) \right| \right)^4 \right]. \end{aligned}$$

The proof of the following lemma is postponed to Appendix A.3.

Lemma 5. *There exist some positive constant C_{15} such that*

$$\mathbb{E}_{\theta,f} \left[\left(\frac{1}{nh} \sum_{i=1}^n \left| K\Big(\frac{x-X_i}{h}\Big) \right| \right)^4 \right] \leq C_{15}. \tag{39}$$

This lemma entails that

$$T_2 \leq C_{15} \left[\mathbb{E}_{\theta,f} \left(|\hat{\theta}_n - \theta|^4 \right) \right]^{\frac{1}{2}}. \tag{40}$$

Now, we turn to the third term in the right hand side of (35). We have

$$\begin{aligned} T_3 &= \mathbb{E}_{\theta,f} \left[\left(\frac{1}{n} \sum_{i=1}^n |\tilde{g}_n(X_i) - g(X_i)| \right)^2 \left(\frac{1}{nh} \sum_{i=1}^n \left| K\Big(\frac{x-X_i}{h}\Big) \right| \right)^2 \right] \\ &= \mathbb{E}_{\theta,f} \left[\frac{1}{n^4 h^2} \sum_{i,j,k,l=1}^n |\tilde{g}_n(X_i) - g(X_i)| |\tilde{g}_n(X_j) - g(X_j)| \left| K\Big(\frac{x-X_k}{h}\Big) K\Big(\frac{x-X_l}{h}\Big) \right| \right]. \end{aligned}$$

By using the same arguments as for obtaining (36) and (37), we can get that

$$T_3 \leq C_{16} n^{\frac{-2\beta}{2\beta+1}}. \tag{41}$$

According to (38), (40) and (41), we may conclude

$$\mathbb{E}_{\theta,f}(|\hat{f}_n(x) - f_2(x)|^2) \leq C_{15} \left[\mathbb{E}_{\theta,f} \left(|\hat{\theta}_n - \theta|^4 \right) \right]^{\frac{1}{2}} + C_{17} n^{\frac{-2\beta}{2\beta+1}}. \tag{42}$$

By returning to Inequality (28) and combining it with (29) and (42), we achieve that

$$\mathbb{E}_{\theta,f}(|\hat{f}_n(x) - f(x)|^2) \leq C_1 \left[\mathbb{E}_{\theta,f} \left(|\hat{\theta}_n - \theta|^4 \right) \right]^{\frac{1}{2}} + C_2 n^{\frac{-2\beta}{2\beta+1}}.$$

□

5.2 Other proofs

Proof of Proposition 1. By using the same arguments as for obtaining (17), we can get that

$$l_n(\hat{f}^{(t)}) - l_n(\hat{f}^{(t+1)}) \geq \frac{1}{n}\sum_{k=1}^{n} \hat{\omega}_k^{(t)} D(\hat{f}^{(t+1)} \mid \hat{f}^{(t)}).$$

Let us now denote by

$$m = \inf_{x\in[-1,1]} K_h(x) \text{ and } M = \sup_{x\in[-1,1]} K_h(x),$$

then m and M are two positive constants depending on the bandwidth h and the kernel K. We note that for all $x \in [0,1]$,

$$m \leq \int_0^1 K_h(u-x)du \leq \min(M,1).$$

Thus, for all $t \geq 1$, the estimate $\hat{f}^{(t)}$ is lower bounded by m. Since the operator $\mathcal{N}$ is increasing, it follows that $\mathcal{N}\hat{f}^{(t)}$ is also lower bounded by m. Now the function

$$x \mapsto \frac{(1-\theta)x}{\theta + (1-\theta)x}$$

is increasing, so that we finally obtain

$$\hat{\omega}_k^{(t)} = \frac{(1-\theta)\mathcal{N}\hat{f}^{(t)}(X_k)}{\theta + (1-\theta)\mathcal{N}\hat{f}^{(t)}(X_k)} \geq \frac{(1-\theta)m}{\theta + (1-\theta)m} = c.$$

This concludes the proof. □

Proof of Proposition 2. We start by stating a lemma, whose proof is postponed to Appendix A.4.

Lemma 6. *The function $l : \mathcal{B} \to \mathbb{R}$ is continuous with respect to the topology induced by uniform convergence on the set of functions defined on* $[0,1]$.

First, for all $f \in \mathcal{B}$, we remark that $m \leq f(\cdot) \leq M/m$. Thus, $\mathcal{N}(f)$ and $l(f)$ are well-defined for $f \in \mathcal{B}$. Moreover, it is easy to see that $l(f)$ is bounded below on $\mathcal{B}$. According to the definition (18) of the sequence $\{f^t\}_{t\geq 0}$, every function f^t belongs to $\mathcal{B}$. As a consequence, we obtain that the sequence $\{l(f^t)\}_{t\geq 0}$ is decreasing and lower bounded, thus it is convergent and the sequence $\{f^t\}_{t\geq 0}$ converges (simply) to a local minimum of l.

Now, it is easy to see that l is a strictly convex function on the convex set $\mathcal{B}$ (relying on Eggermont (1999)). Existence and uniqueness of the minimum $f^\star$ of l in $\mathcal{B}$ thus follows, as well as the simple convergence of the iterative sequence $\{f^t\}_{t\geq 0}$ to this unique minimum.

For all $x, y \in [0,1]$ and for all t, we have

$$\begin{aligned}
|f^t(x) - f^t(y)| &= \frac{1}{\int_0^1 \omega_t(u)g_0(u)du}\left|\int_0^1 \frac{[K_h(u-x) - K_h(u-y)]\omega_t(u)g_0(u)}{\int_0^1 K_h(s-u)ds}du\right| \\
&\leq \frac{1}{\int_0^1 \omega_t(u)g_0(u)du}\int_0^1 \frac{|K_h(u-x) - K_h(u-y)|\omega_t(u)g_0(u)}{m}du \\
&\leq \frac{L}{m}|x-y|,
\end{aligned}$$

so that the sequence $\{f^t\}$ is uniformly bounded and equicontinuous. Relying on Arzelà-Ascoli theorem, there exists a subsequence $\{f^{t_k}\}$ of $\{f^t\}$ which converges uniformly to some limit. However, this uniform limit must be the simple limit of the sequence, namely the minimum $f^\star$ of l. Now, uniqueness of the uniform limit value of the sequence $\{f^t\}_{t\geq 0}$ entails its convergence.

□

A Proofs of technical lemmas

A.1 Proof of Lemma 3

Proof. We first show (22). According to the law of large numbers, since $\mathbb{E}_{\theta,f}\big(f(X_1)/g(X_1)\big) = 1$, we have

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)} \xrightarrow[n\to\infty]{as} 1. \tag{43}$$

Hence

$$\frac{n^2}{S_n^2} = \big(\frac{S_n}{n}\big)^{-2} \xrightarrow[n\to\infty]{as} 1.$$

By the dominated convergence theorem, there exists a constant $c_2 > 0$ such that for n large enough

$$\mathbb{E}_{\theta,f}\big[\frac{1}{S_n^2}\big] = \frac{1}{n^2}\mathbb{E}_{\theta,f}\big[\frac{n^2}{S_n^2}\big] \leq \frac{c_2}{n^2},$$

establishing (22). Let us now prove (21). By using a Taylor's expansion, we have

$$\frac{1}{S_n} = \frac{1}{n}\times\frac{1}{1+(\frac{S_n}{n}-1)} = \frac{1}{n}\left[2-\frac{S_n}{n}+\Big(\frac{S_n}{n}-1\Big)^2\frac{1}{(1+\gamma_n(\frac{S_n}{n}-1))^3}\right],$$

where $\gamma_n \in]0,1[$ depends on S_n. Combining this with (43), we obtain

$$\frac{1}{(1+\gamma_n(\frac{S_n}{n}-1))^3} \xrightarrow[n\to\infty]{as} 1.$$

Thus, there exist some positive constants c, c' such that for n large enough,

$$\frac{1}{n}\Big[2-\frac{S_n}{n}+c'\big(\frac{S_n}{n}-1\big)^2\Big] \leq \frac{1}{S_n} \leq \frac{1}{n}\Big[2-\frac{S_n}{n}+c\big(\frac{S_n}{n}-1\big)^2\Big] \quad \text{a.s.} \tag{44}$$

This implies in particular that

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n}\Big] \leq \frac{1}{n}\Big[2-\frac{\mathbb{E}_{\theta,f}[S_n]}{n}+c\mathbb{E}_{\theta,f}\big[(\frac{S_n}{n}-1)^2\big]\Big] = \frac{1}{n}+\frac{c}{n}\mathbb{E}_{\theta,f}\Big[(\frac{S_n}{n}-1)^2\Big].$$

In addition,

$$\mathbb{E}_{\theta,f}\Big[(\frac{S_n}{n}-1)^2\Big] = \mathbb{V}\mathrm{ar}\Big(\frac{S_n}{n}\Big) = \frac{1}{n}\mathbb{V}\mathrm{ar}\left(\frac{f(X_1)}{g(X_1)}\right).$$

Remember that the ratio f/g is bounded (by δ^{-1}) and thus has finite variance. Hence, there exists a positive constant c_1 such that for n large enough

$$\mathbb{E}_{\theta,f}\big[\frac{1}{S_n}\big] \leq \frac{1}{n} + \frac{c_1}{n^2}.$$

We now prove (23). By using again a Taylor expansion, we have

$$\frac{1}{S_n+\delta^{-1}} = \frac{1}{S_n} \times \frac{1}{1+1/(\delta S_n)} = \frac{1}{S_n} - \frac{1}{\delta S_n^2} \times \frac{1}{[1+\beta_n/(\delta S_n)]^2},$$

where $\beta_n \in]0,1[$ depends on S_n. We also have

$$\frac{1}{[1+\beta_n/(\delta S_n)]^2} \xrightarrow[n\to\infty]{as} 1.$$

Thus, there exists a positive constant c'' such that for n large enough

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n+\delta^{-1}}\Big] = \mathbb{E}_{\theta,f}\Big[\frac{1}{S_n} - \frac{1}{\delta S_n^2} \times \frac{1}{[1+\beta_n/(\delta S_n)]^2}\Big] \geq \mathbb{E}_{\theta,f}\Big[\frac{1}{S_n}\Big] - \mathbb{E}_{\theta,f}\Big[\frac{c''}{S_n^2}\Big] \quad \text{a.s.}$$

According to (44), we have

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n}\Big] \geq \frac{1}{n}\left[2 - \frac{\mathbb{E}_{\theta,f}[S_n]}{n} + c'\mathbb{E}_{\theta,f}\Big[(\frac{S_n}{n}-1)^2\Big]\right] = \frac{1}{n} + \frac{c'}{n^2}\mathbb{V}\text{ar}\left(\frac{f(X_1)}{g(X_1)}\right),$$

and it is proved above that

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n^2}\Big] \leq \frac{c_2}{n^2}.$$

Thus we obtain Inequality (23), namely

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n+\delta^{-1}}\Big] \geq \frac{1}{n} - \frac{c_3}{n^2}.$$

Finally, we show (24). In the same way as we proved (23) above, we have for large enough n,

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n+2\delta^{-1}}\Big] \geq \frac{1}{n} - \frac{c_3'}{n^2} > 0$$

and thus

$$\mathbb{E}^2_{\theta,f}\Big[\frac{1}{S_n+2\delta^{-1}}\Big] \geq \frac{1}{n^2}\left(1 - \frac{2c_3'}{n} + \frac{c_3'^2}{n^2}\right) \geq \frac{1}{n^2}\Big(1 - \frac{2c_3'}{n}\Big). \tag{45}$$

According to Inequality (44) (containing only positive terms for n large enough), we have

$$\begin{aligned}\frac{1}{S_n^2} &\leq \frac{1}{n^2}\left[4 + \frac{S_n^2}{n^2} + c^2\Big(\frac{S_n}{n}-1\Big)^4 - 4\frac{S_n}{n} + 4c\Big(\frac{S_n}{n}-1\Big)^2 - 2c\frac{S_n}{n}\Big(\frac{S_n}{n}-1\Big)^2\right] \quad \text{(as)}\\ &\leq \frac{1}{n^2}\left[4 + \frac{S_n^2}{n^2} + c^2\Big(\frac{S_n}{n}-1\Big)^4 - 4\frac{S_n}{n} + 4c\Big(\frac{S_n}{n}-1\Big)^2\right] \quad \text{a.s.}\end{aligned}$$

Since

$$\mathbb{E}_{\theta,f}[S_n] = n, \quad \mathbb{E}_{\theta,f}[S_n^2] = n\mathbb{V}\text{ar}\left(\frac{f(X_1)}{g(X_1)}\right) + n^2 \quad \text{and} \quad \mathbb{E}_{\theta,f}\left[\Big(\frac{S_n}{n}-1\Big)^2\right] = \frac{1}{n}\mathbb{V}\text{ar}\left(\frac{f(X_1)}{g(X_1)}\right),$$

we have

$$\begin{aligned}
\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n^2}\Big] &\leq \frac{1}{n^2}\left[4+\frac{\mathbb{E}_{\theta,f}[S_n^2]}{n^2}+c^2\mathbb{E}_{\theta,f}\Big[\Big(\frac{S_n}{n}-1\Big)^4\Big]-4\frac{\mathbb{E}_{\theta,f}[S_n]}{n}+4c\mathbb{E}_{\theta,f}\Big[\Big(\frac{S_n}{n}-1\Big)^2\Big]\right] \\
&\leq \frac{1}{n^2}\left[4+\frac{1}{n}\mathbb{V}\mathrm{ar}\left(\frac{f(X_1)}{g(X_1)}\right)+1+c^2\mathbb{E}_{\theta,f}\Big[\Big(\frac{S_n}{n}-1\Big)^4\Big]-4+\frac{4c}{n}\mathbb{V}\mathrm{ar}\left(\frac{f(X_1)}{g(X_1)}\right)\right] \\
&\leq \frac{1}{n^2}\left[1+\frac{C_4}{n}+c^2\mathbb{E}_{\theta,f}\Big[\Big(\frac{S_n}{n}-1\Big)^4\Big]\right]. \qquad (46)
\end{aligned}$$

Combining (45) and (46), we get that

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n^2}\Big]-\mathbb{E}^2_{\theta,f}\Big[\frac{1}{S_n+2\delta^{-1}}\Big]\leq\frac{C}{n^3}+\frac{c^2}{n^2}\mathbb{E}_{\theta,f}\Big[\Big(\frac{S_n}{n}-1\Big)^4\Big]. \qquad (47)$$

We now upper-bound the quantity $\mathbb{E}_{\theta,f}\big[(\frac{S_n}{n}-1)^4\big]$. Let us denote by

$$U_i=\frac{f(X_i)}{g(X_i)}-1.$$

We have

$$\begin{aligned}
\Big(\frac{S_n}{n}-1\Big)^4 &= \frac{1}{n^4}\left(\sum_{i=1}^n U_i\right)^4=\frac{1}{n^4}\sum_{i=1}^n U_i^4+\frac{1}{n^4}\sum_{i\neq j}^n U_i^3U_j+ \\
&\quad +\frac{1}{n^4}\sum_{i\neq j}^n U_i^2U_j^2+\frac{1}{n^4}\sum_{i\neq j\neq k}^n U_i^2U_jU_k+\frac{1}{n^4}\sum_{i\neq j\neq k\neq l}^n U_iU_jU_kU_l.
\end{aligned}$$

Since the random variables U_i are iid with mean zero, we obtain

$$\mathbb{E}_{\theta,f}\left[\Big(\frac{S_n}{n}-1\Big)^4\right]=\frac{1}{n^4}\left[n\mathbb{E}_{\theta,f}(U_1^4)+n(n-1)\mathbb{E}_{\theta,f}(U_1^2U_2^2)\right]=O\Big(\frac{1}{n^2}\Big). \qquad (48)$$

Finally, according to (47) and (48) we have

$$\mathbb{E}_{\theta,f}\Big[\frac{1}{S_n^2}\Big]-\mathbb{E}^2_{\theta,f}\Big[\frac{1}{S_n+2\delta^{-1}}\Big]=O\Big(\frac{1}{n^3}\Big).$$

□

A.2 Proof of Lemma 4

Proof. We write

$$\frac{1}{\sum_k\hat\tau_k}=\frac{1}{\sum_k\tau_k+\sum_k(\hat\tau_k-\tau_k)}=\frac{1}{\sum_k\tau_k}-\frac{\sum_k(\hat\tau_k-\tau_k)}{(\sum_k\tau_k)^2}\times\int_0^1\left(1+s\frac{\sum_k(\hat\tau_k-\tau_k)}{\sum_k\tau_k}\right)^{-2}ds.$$

Let us establish that $\|\hat\tau-\tau\|_{\infty,[0,1]}=\sup_{x\in[0,1]}|\hat\tau(x)-\tau(x)|$ converges almost surely to zero. Indeed,

$$\hat\tau(x)-\tau(x)=(\theta-\hat\theta_n)\frac{1}{g(x)}+\hat\theta_n\left(\frac{1}{g(x)}-\frac{1}{\tilde g_n(x)}\right)$$

and using the same argument as for establishing (33), we get that for n large enough and for all $x \in [0,1]$,

$$|\hat{\tau}(x) - \tau(x)| \leq \frac{|\hat{\theta}_n - \theta|}{\theta} + 2|\hat{\theta}_n| \frac{\|\hat{g}_n - g\|_\infty}{\theta^2} \leq \delta^{-1}|\hat{\theta}_n - \theta| + 2\delta^{-2}\|\hat{g}_n - g\|_\infty.$$

By using consistency of $\hat{\theta}_n$ and Remark 1, we obtain that $\|\hat{\tau} - \tau\|_{\infty,[0,1]}$ converges almost surely to zero. Now,

$$\begin{aligned} \forall s \in [0,1], \quad 1 + s\frac{\sum_k(\hat{\tau}_k - \tau_k)}{\sum_k \tau_k} &\geq 1 - s\frac{n\|\hat{\tau}_k - \tau_k\|_{\infty,[0,1]}}{\sum_k \tau_k} \\ &\geq 1 - s\frac{2\|\hat{\tau}_k - \tau_k\|_{\infty,[0,1]}}{\theta} \geq 1 - \frac{s}{2} > 0 \quad \text{a.s.} \end{aligned}$$

We obtain that

$$\begin{aligned} \frac{n}{|\sum_k \hat{\tau}_k|} &\leq \frac{n}{\sum_k \tau_k} + \frac{n\sum_k |\hat{\tau}_k - \tau_k|}{(\sum_k \tau_k)^2} \times \int_0^1 \left(1 + s\frac{\sum_k(\hat{\tau}_k - \tau_k)}{\sum_k \tau_k}\right)^{-2} ds \\ &\leq \frac{n}{\sum_k \tau_k} + \frac{n^2\|\hat{\tau} - \tau\|_{\infty,[0,1]}}{(\sum_k \tau_k)^2} \times \int_0^1 \left(1 - \frac{s}{2}\right)^{-2} ds \\ &\leq \frac{2}{1-\theta} + \frac{8\|\hat{\tau} - \tau\|_{\infty,[0,1]}}{(1-\theta)^2} \leq c_7 \quad \text{a.s.} \end{aligned}$$

□

A.3 Proof of Lemma 5

Proof. In order to prove (39), let us consider iid random variables $U_1, \ldots, U_n$ defined as

$$U_i = \left|K\left(\frac{x - X_i}{h}\right)\right|.$$

For all $1 \leq p \leq 4$, we have

$$\mathbb{E}_{\theta,f}(U_i^p) = \int \left|K^p\left(\frac{x-t}{h}\right)\right| g(t)dt = h\int \left|K^p(t)\right| g(x+th)dt \leq C_{15}h.$$

We then write

$$\left(\frac{1}{nh}\sum_{i=1}^n \left|K\left(\frac{x - X_i}{h}\right)\right|\right)^4 = \frac{1}{n^4h^4}\left(\sum_i U_i\right)^4, \tag{49}$$

where

$$\left(\sum_i U_i\right)^4 = \sum_i U_i^4 + \sum_{i\neq j} U_i^3 U_j + \sum_{i\neq j} U_i^2 U_j^2 + \sum_{i\neq j\neq k} U_i^2 U_j U_k + \sum_{i\neq j\neq k\neq l} U_i U_j U_k U_l.$$

And for all choice of the bandwidth $h > 0$ such that $nh \to \infty$,

$$\begin{aligned}
&\mathbb{E}_{\theta,f}\Big[\Big(\sum_i U_i\Big)^4\Big] \\
=&n\mathbb{E}_{\theta,f}(U_1^4) + n(n-1)\mathbb{E}_{\theta,f}(U_1^3U_2) + n(n-1)\mathbb{E}_{\theta,f}(U_1^2U_2^2)+ \\
&+ n(n-1)(n-2)\mathbb{E}_{\theta,f}(U_1^2U_2U_3) + n(n-1)(n-2)(n-3)\mathbb{E}_{\theta,f}(U_1U_2U_3U_4) \\
=&n\mathbb{E}_{\theta,f}(U_1^4) + n(n-1)\mathbb{E}_{\theta,f}(U_1^3)\mathbb{E}_{\theta,f}(U_1) + n(n-1)\mathbb{E}^2_{\theta,f}(U_1^2)+ \\
&+ n(n-1)(n-2)\mathbb{E}_{\theta,f}(U_1^2)\mathbb{E}^2_{\theta,f}(U_1) + n(n-1)(n-2)(n-3)\mathbb{E}^4_{\theta,f}(U_1) \\
\leq&C_{15}n^4h^4. \qquad (50)
\end{aligned}$$

According to (49) and (50) we obtain the result. □

A.4 Proof of Lemma 6

Proof. Let f be a function in $\mathcal{B}$ and $\{f_n\}$ be a sequence of densities on $[0,1]$ such that $\|f_n - f\|_\infty \xrightarrow[n\to\infty]{} 0$. Let us recall that every $f \in \mathcal{B}$ satisfies the bounds $m \leq f \leq M/m$. We have

$$\begin{aligned}
|\, l(f_n) - l(f)\, | &= \Big|\int_0^1 g_0(x)\log\frac{\theta + (1-\theta)\mathcal{N}f(x)}{\theta + (1-\theta)\mathcal{N}f_n(x)}dx\Big| \\
&\leq \int_0^1 g_0(x)\Big|\log\Big\{1 + \frac{(1-\theta)[\mathcal{N}f_n(x) - \mathcal{N}f(x)]}{\theta + (1-\theta)\mathcal{N}f_n(x)}\Big\}\Big|dx,
\end{aligned}$$

and

$$\begin{aligned}
|\, \mathcal{N}f_n(x) - \mathcal{N}f(x)\, | &= \mathcal{N}f(x)\Big|\exp\frac{\int_0^1 K_h(u-x)[\log f_n(u) - \log f(u)]du}{\int_0^1 K_h(s-x)ds} - 1\Big| \\
&\leq \frac{M}{m}\Big|\exp\frac{\int_0^1 K_h(u-x)[\log f_n(u) - \log f(u)]du}{\int_0^1 K_h(s-x)ds} - 1\Big|.
\end{aligned}$$

For $|x| < \epsilon$ small enough, we have $|\log(1+x)| \leq 2|x|$ and $|\exp(x) - 1| \leq 2|x|$. Combining with the fact that f is bounded, we get that

$$\begin{aligned}
\Big|\int_0^1 K_h(u-x)[\log f_n(u) - \log f(u)]du\,| &\leq \int_0^1 K_h(u-x)\Big|\log\big\{1 + \frac{f_n(u) - f(u)}{f(u)}\big\}\Big|du \\
&\leq 2\|f_n - f\|_\infty
\end{aligned}$$

and thus

$$\|\mathcal{N}f_n - \mathcal{N}f\|_\infty \leq \frac{4M}{m^2}\|f_n - f\|_\infty.$$

We finally obtain

$$|\, l(f_n) - l(f)\, | \leq C\|f_n - f\|_\infty,$$

where C is a constant depending on h, K and θ. □

References

Allison, D. B., G. L. Gadbury, M. Heo, J. R. Fernández, C.-K. Lee, T. A. Prolla, and R. Weindruch (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal. 39*(1), 1–20.

Aubert, J., A. Bar-Hen, J.-J. Daudin, and S. Robin (2004). Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics 5*(1), 125.

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

Celisse, A. and S. Robin (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference 140*(11), 3132–3147.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B 39*(1), 1–38.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*(456), 1151–1160.

Eggermont, P. and V. LaRiccia (1995). Maximum smoothed likelihood density estimation for inverse problems. *Ann. Stat. 23*(1), 199–220.

Eggermont, P. and V. LaRiccia (2001). *Maximum penalized likelihood estimation. Vol. 1: Density estimation.* Springer Series in Statistics. New York, NY: Springer.

Eggermont, P. P. B. (1999). Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. *Applied Mathematics & Optimization 39*, 75–91.

Guedj, M., S. Robin, A. Celisse, and G. Nuel (2009). Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics 10*(1), 84.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (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.

Levine, M., D. R. Hunter, and D. Chauveau (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika 98*(2), 403–416.

Liao, J., Y. Lin, Z. E. Selvanayagam, and W. J. Shih (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics 20*(16), 2694–2701.

McLachlan, G., R. Bean, and L. B.-T. Jones (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics 22*(13), 1608–1615.

Neuvial, P. (2010). Intrinsic bounds and false discovery rate control in multiple testing problems. Technical report, arXiv:1003.0747.

Nguyen, V. and C. Matias (2012). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. Technical report, arXiv:1205.4097.

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

Robin, S., A. Bar-Hen, J.-J. Daudin, and L. Pierre (2007). A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput. Statist. Data Anal. 51*(12), 5483–5493.

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika 69*(3), 493–502.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. London: Chapman & Hall.

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. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist. 31*(6), 2013–2035.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics 9*(1), 303.

Sun, W. and T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc. 102*(479), 901–912.

Sun, W. and T. Cai (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 393–424.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. New York, NY: Springer.

Wied, D. and R. Weißbach (2012). Consistency of the kernel density estimator: a survey. *Statistical Papers 53*, 1–21.