



**HAL**  
open science

# Semi-parametric mixture models and applications to multiple testing

van Hanh Nguyen

► **To cite this version:**

van Hanh Nguyen. Semi-parametric mixture models and applications to multiple testing. General Mathematics [math.GM]. Université Paris Sud - Paris XI, 2013. English. NNT: 2013PA112196 . tel-00987035

**HAL Id: tel-00987035**

**<https://theses.hal.science/tel-00987035v1>**

Submitted on 12 May 2014

**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.



UNIVERSITÉ PARIS-SUD  
FACULTÉ DES SCIENCES D'ORSAY

## THÈSE

*présentée pour obtenir*

LE GRADE DE DOCTEUR EN SCIENCES  
DE L'UNIVERSITÉ PARIS XI

Spécialité : Mathématiques

*par*

Van Hanh NGUYEN

*Sujet :*

**MODÈLES DE MÉLANGE SEMI-PARAMÉTRIQUES ET  
APPLICATIONS AUX TESTS MULTIPLES.**

Rapporteurs : M. Gilles BLANCHARD  
M. Stéphane ROBIN

Soutenue le 01 Octobre 2013 devant la Commission d'examen :

Mme. CRISTINA BUTUCEA (Présidente du jury)  
M. ALAIN CELISSE (Examineur)  
Mme. ÉLISABETH GASSIAT (Co-directrice de thèse)  
Mme. CATHERINE MATIAS (Co-directrice de thèse)  
M. STÉPHANE ROBIN (Rapporteur)  
M. ETIENNE ROQUAIN (Examineur)

# Remerciements

Je voudrais tout d'abord remercier grandement mes deux directrices de thèse : Catherine MATIAS, pour son encadrement scientifique de qualité qui m'a permis de m'épanouir dans mes travaux de recherche, aussi pour sa grande gentillesse et sympathie qu'elle m'a apporté pendant ces trois années, ainsi que Elisabeth GASSIAT, pour ses multiples conseils et pour la confiance qu'elle m'a accordé en acceptant d'encadrer ce travail doctoral.

Je remercie également à Gilles BLANCHARD et Stéphane ROBIN, pour avoir accepté de rapporter sur ma thèse. Leurs conseils sur le manuscrit m'ont beaucoup aidé à améliorer ma thèse.

Mes remerciements vont également à Cristina BUTUCEA, Alain CELISSE et Etienne ROQUAIN qui me font l'honneur d'accepter d'être membres de Jury de ma soutenance.

Je voudrais remercier très chaleureusement tous les membres du laboratoire Statistique et Génome. Je remercie particulièrement Christophe AMBROISE et Michèle ILBERT pour leur gentillesse et leur aide. Je remercie également Marie-Luce TAUPIN pour sa co-direction de mon stage de Master 2. Il m'est impossible d'oublier les trois années passées au laboratoire.

Je voudrais remercier le Ministère de l'enseignement supérieur et de la recherche de la France qui a financé cette thèse en m'accordant un poste d'allocataire de recherche à l'Université Paris-Sud. Je remercie également David HARARI, les membres du département de Mathématiques d'Orsay et les personnels du secrétariat de l'Université Paris-Sud.

Enfin, je désire remercier ma famille, mes amis qui m'ont encouragé continuellement pendant toutes mes études en France.

## Résumé

Dans un contexte de test multiple, nous considérons un modèle de mélange semiparamétrique avec deux composantes. Une composante est supposée connue et correspond à la distribution des  $p$ -valeurs sous hypothèse nulle avec probabilité a priori  $\theta$ . L'autre composante  $f$  est nonparamétrique et représente la distribution des  $p$ -valeurs sous l'hypothèse alternative. Le problème d'estimer les paramètres  $\theta$  et  $f$  du modèle apparaît dans les procédures de contrôle du taux de faux positifs ("false discovery rate" ou FDR). Dans la première partie de cette dissertation, nous étudions l'estimation de la proportion  $\theta$ . Nous discutons de résultats d'efficacité asymptotique et établissons que deux cas différents arrivent suivant que  $f$  s'annule ou non surtout un intervalle non-vide. Dans le premier cas (annulation surtout un intervalle), nous présentons des estimateurs qui convergent à la vitesse paramétrique, calculons la variance asymptotique optimale et conjecturons qu'aucun estimateur n'est asymptotiquement efficace (*i.e* atteint la variance asymptotique optimale). Dans le deuxième cas, nous prouvons que le risque quadratique de n'importe quel estimateur ne converge pas à la vitesse paramétrique. Dans la deuxième partie de la dissertation, nous nous concentrons sur l'estimation de la composante inconnue nonparamétrique  $f$  dans le mélange, en comptant sur un estimateur préliminaire de  $\theta$ . Nous proposons et étudions les propriétés asymptotiques de deux estimateurs différents pour cette composante inconnue. Le premier estimateur est un estimateur à noyau avec poids aléatoires. Nous établissons une borne supérieure pour son risque quadratique ponctuel, en montrant une vitesse de convergence nonparamétrique classique sur une classe de Hölder. Le deuxième estimateur est un estimateur du maximum de vraisemblance régularisée. Il est calculé par un algorithme itératif, pour lequel nous établissons une propriété de décroissance d'un critère. De plus, ces estimateurs sont utilisés dans une procédure de test multiple pour estimer le taux local de faux positifs ("local false discovery rate" ou  $\ell$ FDR).

## Abstract

In a multiple testing context, we consider a semiparametric mixture model with two components. One component is assumed to be known and corresponds to the distribution of  $p$ -values under the null hypothesis with prior probability  $\theta$ . The other component  $f$  is nonparametric and stands for the distribution under the alternative hypothesis. The problem of estimating the parameters  $\theta$  and  $f$  of the model appears from the false discovery rate control procedures. In the first part of this dissertation, we study the estimation of the proportion  $\theta$ . 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 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 parametric rate. In the second part of the dissertation, we focus on the estimation of the nonparametric unknown component  $f$  in the mixture, relying on a preliminary estimator of  $\theta$ . 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. 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.

# Contents

<b>1</b>	<b>General Introduction</b>	<b>8</b>
1.1	Multiple testing framework	8
1.1.1	Multiple testing problem	8
1.1.2	An example of multiple testing	9
1.1.3	$P$ -value and $z$ -value of test	10
1.1.4	Mixture model in multiple testing setup	11
1.1.5	Multiple testing procedure	12
1.1.6	Type I and II error rates	12
1.2	Type I error rate control procedures	13
1.2.1	FWER control procedures	14
1.2.2	FDR control procedures	15
1.3	The FDR estimation approach	18
1.3.1	Estimation of pFDR and FDR	18
1.3.2	Connection between FDR estimation and FDR control	19
1.4	Local false discovery rate	19
1.5	Semiparametric inference	21
1.5.1	Tangent sets and efficient influence function	21
1.5.2	Asymptotically efficient estimator	23
1.5.3	Expressions for semiparametric models in a strict sense	26
1.5.4	One-step estimator method	27
1.5.5	The infinite bound case	29
1.6	Organization	29
<b>2</b>	<b>Estimation of the proportion of true null hypotheses</b>	<b>31</b>
2.1	Introduction	31

2.2	Lower bounds for the quadratic risk and efficiency	35
2.3	Upper bounds for the quadratic risk and efficiency (when $\delta > 0$ )	38
2.3.1	A histogram based estimator	38
2.3.2	Celisse and Robin [2010]'s procedure	39
2.3.3	One-step estimators	41
2.4	Simulations	43
2.5	Proofs of main results	46
2.5.1	Proof of Proposition 2.1	46
2.5.2	Proof of Theorem 2.1	48
2.5.3	Proofs from Sections 2.3.1 and 2.3.3	52
2.5.4	Proof of Theorem 2.3	55
2.6	Proofs of technical lemmas	60
2.6.1	Proof of Lemma 2.1	60
2.6.2	Proof of Lemma 2.2	61
2.6.3	Proof of Lemma 2.3	62
<b>3</b>	<b>Estimation of the density of the alternative</b>	<b>64</b>
3.1	Introduction	65
3.2	Algorithmic procedures to estimate the density $f$	69
3.2.1	Direct procedures	69
3.2.2	Iterative procedures	70
3.3	Mathematical properties of the algorithms	73
3.3.1	Randomly weighted kernel estimator	73
3.3.2	Maximum smoothed likelihood estimator	76
3.4	Estimation of local false discovery rate and simulation study	79
3.4.1	Estimation of local false discovery rate	79
3.4.2	Simulation study	80

## CONTENTS

---

3.5	Proofs of main results . . . . .	85
3.5.1	Proof of Theorem 3.1 . . . . .	85
3.5.2	Other proofs . . . . .	93
3.6	Proofs of technical lemmas . . . . .	95
3.6.1	Proof of Lemma 3.3 . . . . .	95
3.6.2	Proof of Lemma 3.4 . . . . .	97
3.6.3	Proof of Lemma 3.5 . . . . .	98
3.6.4	Proof of Lemma 3.6 . . . . .	99
<b>4</b>	<b>Another semiparametric mixture model</b>	<b>101</b>
4.1	Identifiability . . . . .	101
4.2	Efficient information matrix for estimating $\theta$ . . . . .	102
4.3	Perspectives . . . . .	107
	<b>Bibliography</b>	<b>108</b>
	<b>Appendix</b>	<b>114</b>
<b>A</b>	<b>Adaptive estimation via Lepski's method</b>	<b>115</b>
A.1	Lepski's method . . . . .	115
A.2	Perspectives . . . . .	116



# General Introduction

---

This overview briefly describes the main components of this dissertation, including multiple testing framework, type I error rate control procedures, FDR estimation approach, local false discovery rate and semiparametric inference. The last concept is the central motivation of this dissertation. This introduction borrows some material from Roquain (2011), Storey (2002, 2004) and van der Vaart (1998, 2002).

## Contents

---

<b>1.1 Multiple testing framework</b>	<b>8</b>
<b>1.2 Type I error rate control procedures</b>	<b>13</b>
<b>1.3 The FDR estimation approach</b>	<b>18</b>
<b>1.4 Local false discovery rate</b>	<b>19</b>
<b>1.5 Semiparametric inference</b>	<b>21</b>
<b>1.6 Organization</b>	<b>29</b>

---

## 1.1 Multiple testing framework

### 1.1.1 Multiple testing problem

The problem of multiple testing has a long history in the statistics literature. Microarray analysis [Dudoit and van der Laan, 2008], astrophysics [Meinshausen and Rice, 2006] or neuroimaging [Turkheimer et al., 2001] are some areas in which multiple testing problems occur. We first recall the basic paradigm for single-hypothesis testing. We wish to test a null hypothesis  $H_0$  versus an alternative  $H_1$  based on a statistic  $X$ . For a given rejection region  $\Gamma$ , we reject  $H_0$  when  $X \in \Gamma$  and we accept  $H_0$  when  $X \notin \Gamma$ . A type I error occurs when the null hypothesis ( $H_0$ ) is true but is rejected; while a type II error occurs when the null hypothesis is false but is accepted. To choose  $\Gamma$ , the acceptable type I error is set at some level  $\alpha$ , then all rejection regions are considered that have a type I error that is less than or equal to  $\alpha$ . The one that has the lowest type II error is chosen. Therefore, the rejection region is sought with respect

## 1.1. MULTIPLE TESTING FRAMEWORK

---

to controlling the type I error. Precisely, we find a rejection region with nearly optimal power (power = 1 - type II error) while maintaining the desired  $\alpha$ -level type I error.

Now, for multiple-hypothesis testing, the situation becomes much more complicated. For instance, we test simultaneously  $n = 10,000$  null hypotheses, of which  $n_0 = 8,000$  are true nulls (level  $\alpha = 0.05$  for each test). This procedure makes on average  $n_0\alpha = 400$  false positives (type I errors). It seems unsuitable because it is likely to select a lot of false positives. And it becomes unclear how we should measure the overall error rate. A multiple testing procedure aims at correcting a priori the level of the single tests in order to obtain the “quantity” of false positives that is below a nominal level  $\alpha$ . The “quantity” of false positives is measured by using global type I error rates, as for instance the probability to make at least one type I error among all the hypotheses (family wise error rate, FWER) or the expected proportion of false positives among all rejected hypotheses (false discovery rate, FDR).

### 1.1.2 An example of multiple testing

In a microarray experiment, the level expressions of a set of genes are measured under two different experimental conditions and we aim at finding the genes that are differentially expressed between the two conditions. For instance, the genes come from tumor cells in the first experimental condition, while the genes come from healthy cells in the second, the differentially expressed genes may be involved in the development of this tumor and thus are genes of special interest. The problem of finding differentially expressed genes can be formalized as a particular case of a general two-sample multiple testing problem. Let us observe a couple of two independent samples

$$(Y^1, \dots, Y^{n_1}) \in \mathbb{R}^{n \times n_1} \text{ and } (Z^1, \dots, Z^{n_2}) \in \mathbb{R}^{n \times n_2},$$

where  $(Y^1, \dots, Y^{n_1})$  is a family of  $n_1$  iid copies of a random vector  $Y$  in  $\mathbb{R}^n$  and  $(Z^1, \dots, Z^{n_2})$  is a family of  $n_2$  iid copies of a random vector  $Z$  in  $\mathbb{R}^n$ . In the context of microarray data,  $Y_i^j$  (resp.  $Z_i^j$ ) is the expression level of the  $i$ -th gene for the  $j$ -th individual of the first (resp. second) experimental condition. Suppose that  $Y \sim \mathcal{N}(\mu_1, \Sigma_1)$  and  $Z \sim \mathcal{N}(\mu_2, \Sigma_2)$ , where  $\mu_1 = (\mu_{11}, \dots, \mu_{1n})$  and  $\mu_2 = (\mu_{21}, \dots, \mu_{2n})$  are mean vectors of  $\mathbb{R}^n$ ,  $\Sigma_1$  and  $\Sigma_2$  are diagonal covariance matrices. The index set  $\{1 \leq i \leq n : \mu_{1i} \neq \mu_{2i}\}$  corresponds to differentially expressed genes. Then we aim at testing simultaneously  $n$  hypotheses

$$H_{0,i} : \text{“}\mu_{1i} = \mu_{2i}\text{” against } H_{1,i} : \text{“}\mu_{1i} \neq \mu_{2i}\text{”}.$$

The individual test statistic is the classical two-sample  $t$ -statistic

$$X_i = \frac{\bar{Y}_i - \bar{Z}_i}{\hat{\sigma}_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } \hat{\sigma}_i^2 = \frac{(n_1 - 1)\hat{\sigma}_{1i}^2 + (n_2 - 1)\hat{\sigma}_{2i}^2}{n_1 + n_2 - 2},$$

and  $\bar{Y}_i, \hat{\sigma}_{1i}^2$  (resp.  $\bar{Z}_i, \hat{\sigma}_{2i}^2$ ) are the sample mean and the sample variance of the data  $\{Y_i^j\}_j$  (resp.  $\{Z_i^j\}_j$ ).

### 1.1.3 $P$ -value and $z$ -value of test

We define the  $p$ -value as the probability of observing something as extreme as or more extreme than the observed test statistic given that the null hypothesis is true. That is, we can consider the  $p$ -value as the minimum probability under the null that our test statistic is in the rejection region (i.e., the minimum type I error rate) over the set of nested rejection regions containing the observed test statistic. Formally, we can write the  $p$ -value [see [Lehmann, 1986](#)], corresponding to an observed test statistic  $X = x$  as

$$p\text{-value}(x) = \inf_{\{\Gamma: x \in \Gamma\}} \{\mathbb{P}(X \in \Gamma | H = 0)\},$$

where  $\{\Gamma : x \in \Gamma\}$  is a set of nested rejection regions that contain the observed test statistic  $x$ . Any  $p$ -value is stochastically bounded by a uniform distribution under the null, namely,

$$\mathbb{P}(p_i(X) \leq t | H = 0) \leq t, \text{ for all } t \in [0, 1]. \quad (1.1)$$

For example, when the rejection regions  $\Gamma$  are of the forms  $\{X \geq c\}$ , the  $p$ -value of  $X = x$  is

$$\begin{aligned} p\text{-value}(x) &= \inf_{\{c: x \geq c\}} \{\mathbb{P}(X \geq c | H = 0)\} \\ &= \mathbb{P}(X \geq x | H = 0) = 1 - G_0(x), \end{aligned}$$

where  $G_0$  is the cumulative distribution function (CDF) of test statistic  $X$  under null hypothesis. If the distribution of the statistic  $X_i$  is absolutely continuous, (1.1) holds with equality, that is, the  $p$ -values are exactly distributed like a uniform variable in  $[0, 1]$  when  $H_0$  is true.

**Remark 1.1.** *When we reject the null hypotheses on the basis of  $p$ -values, all rejection regions are of the form  $[0, \gamma]$  for some  $\gamma > 0$ .*

*Indeed, according to the definition of  $p$ -value, for two  $p$ -values  $p_1$  and  $p_2$ , the relation  $p_1 \leq p_2$  implies that the respective observed statistics  $x_1$  and  $x_2$  are such that  $x_2 \in \Gamma$  implies  $x_1 \in \Gamma$ . Therefore, whenever  $p_2$  is rejected,  $p_1$  should also be rejected.*

We now define a  $z$ -value of test as the probit transformation

$$Z = \text{probit}(P) = \Phi^{-1}(P),$$

where  $P$  is a  $p$ -value and  $\Phi$  is the CDF of the standard normal distribution.

#### 1.1.4 Mixture model in multiple testing setup

Suppose that we are testing  $n$  identical hypothesis tests  $H_1, \dots, H_n$  with observed statistics  $X_1, \dots, X_n$ . The identical tests mean that the same rejection region type is used for each test. We let  $H_i = 0$  when the null hypothesis  $i$  is true and  $H_i = 1$  otherwise. We denote by  $T_i = T(X_i)$  a transformation of test statistic, for example,  $T_i$  is  $p$ -value  $P_i$ ,  $z$ -value  $Z_i$ , local false discovery rate  $\ell\text{FDR}(X_i)$  (defined as below) or identical to test statistic  $X_i$ . We assume that the nulls  $T_i|H_i = 0$  and the alternatives  $T_i|H_i = 1$  are identically distributed with respective distribution functions  $G_0$  that is known and  $G_1$  that is unknown. Finally we assume that the  $H_i$  are Bernoulli random variables with an unknown probability  $\mathbb{P}(H_i = 0) = \theta$ . The marginal distribution of each  $T_i$  is thus a mixture

$$G(x) = \theta G_0(x) + (1 - \theta)G_1(x),$$

and we denote by  $g = \theta g_0 + (1 - \theta)g_1$  the corresponding probability density function (pdf) of  $T_i$  (if it exists). When we assume that the statistics  $X_i$  under the null hypotheses are continuous variables, the  $p$ -values under the null hypotheses follow the uniform distribution  $\mathcal{U}([0, 1])$  on interval  $[0, 1]$  and the marginal distribution of each  $p$ -value is

$$F(x) = \theta x + (1 - \theta)F_1(x), \text{ for } x \in [0, 1],$$

and we denote the corresponding pdf by  $f(x) = \theta \mathbf{1}_{[0,1]}(x) + (1 - \theta)f_1(x)$ , where  $f_1$  is an unknown pdf on  $[0, 1]$ . If we consider the transformation  $T_i$  as  $z$ -value  $Z_i$ , then

$$G_0(x) = \mathbb{P}_{H_0}(Z_i \leq x) = \mathbb{P}_{H_0}(P_i \leq \Phi(x)) = \Phi(x),$$

and

$$G_1(x) = \mathbb{P}_{H_1}(Z_i \leq x) = \mathbb{P}_{H_1}(P_i \leq \Phi(x)) = F_1(\Phi(x)).$$

Thus the pdf of  $Z_i$  is  $g(x) = \phi(x)[\theta + (1 - \theta)f_1(\Phi(x))]$ , for  $x \in \mathbb{R}$ , where  $\phi$  is the pdf of the standard normal distribution.

### 1.1.5 Multiple testing procedure

A multiple testing procedure (MTP) provides rejection regions, i.e., sets of values for each  $T_i$  that lead to the decision to reject the corresponding null hypothesis  $H_i$ . In other words, a MTP produces a random subset  $R$  of  $\{1, \dots, n\}$  that the indexes selected correspond to the rejected null hypotheses. A multiple testing setting includes the  $p$ -value family  $p = \{p_i, 1 \leq i \leq n\} \in [0, 1]^n$ . The multiple testing procedure based on  $p$  is defined as a set-valued function

$$R : p = (p_i)_{1 \leq i \leq n} \in [0, 1]^n \mapsto R(p) \subset \{1, \dots, n\},$$

taking as input an element of  $[0, 1]^n$  and returning a subset of  $\{1, \dots, n\}$ . The indexes selected by the procedure  $R(p)$  correspond to the rejected null hypotheses. When we focus on the case of the identical tests based on the  $p$ -value family, one procedure, called thresholding based procedure, is of the form  $R(p) = \{1 \leq i \leq n : p_i \leq t(p)\}$ , where the threshold  $t(\cdot) \in [0, 1]$  can depend on the data.

### 1.1.6 Type I and II error rates

To measure the quality of a multiple testing procedure, various error rates have been proposed in the literature. These rates evaluate the importance of the null hypotheses wrongly rejected, that is the number of false positives (FP). Two error measures that are the most commonly used in multiple-hypothesis testing are the family wise error rate (FWER) and the false discovery rate (FDR). Moreover, the false discovery proportion (FDP) is also a widely used type I error. The definitions of these rates are recalled in the following. First, the outcome of testing  $n$  hypotheses simultaneously can be summarized as indicated in Table 1.1.

Table 1.1: Possible outcomes from testing  $n$  hypotheses  $H_1, \dots, H_n$ .

	Accepts $H_i$	Rejects $H_i$	Total
$H_i$ is true	TN	FP	$n_0$
$H_i$ is false	FN	TP	$n_1$
Total	W	R	$n$

The family wise error rate (FWER) is defined as the probability to make at least one false positive among all the hypotheses,

$$\text{FWER} = \mathbb{P}(\text{FP} \geq 1).$$

The false discovery proportion (FDP) is defined as the proportion of false positives among the rejected hypotheses,

$$\text{FDP} = \frac{\text{FP}}{\max(\text{R}, 1)}.$$

Let us remark that the FDP is a random variable, it does not define an error rate. [Benjamini and Hochberg \[1995\]](#) define the false discovery rate (FDR) as the expectation of the FDP,

$$\text{FDR} = \mathbb{E}\left[\frac{\text{FP}}{\max(\text{R}, 1)}\right] = \mathbb{E}\left[\frac{\text{FP}}{\text{R}} \mid \text{R} > 0\right] \mathbb{P}(\text{R} > 0).$$

They provided sequential  $p$ -value methods to control this quantity. FDR offers a much less strict multiple-testing criterion over FWER and therefore leads to an increase in power. [Storey \[2003\]](#) proposes to modify FDR so as to obtain a new criterion, the positive FDR (or pFDR) defined by

$$\text{pFDR} = \mathbb{E}\left[\frac{\text{FP}}{\text{R}} \mid \text{R} > 0\right]$$

and argues that it is conceptually more sound than FDR. Indeed, when controlling FDR at level  $\alpha$ , and positive findings have occurred then FDR has really only been controlled at level  $\alpha/\mathbb{P}(\text{R} > 0)$ . This can be quite dangerous, and it is not the case for pFDR. Other similar measure includes the marginal FDR (mFDR) defined as

$$\text{mFDR} = \frac{\mathbb{E}(\text{FP})}{\mathbb{E}(\text{R})}.$$

Under weak conditions, [Genovese and Wasserman \[2002\]](#) showed that  $\text{mFDR} = \text{FDR} + O(n^{-1/2})$  and [Storey \[2003\]](#) proved that mFDR and pFDR are identical. An analog of FDR in terms of false negatives (type II errors) is the false nondiscovery rate (FNR), defined as

$$\text{FNR} = \mathbb{E}\left[\frac{\text{FN}}{\max(\text{W}, 1)}\right] = \mathbb{E}\left[\frac{\text{FN}}{\text{W}} \mid \text{W} > 0\right] \mathbb{P}(\text{W} > 0).$$

Similarly, we define the positive false nondiscovery rate (pFNR) as the conditional expectation

$$\text{pFNR} = \mathbb{E}\left[\frac{\text{FN}}{\text{W}} \mid \text{W} > 0\right].$$

## 1.2 Type I error rate control procedures

A multiple testing control procedure aims at finding a rejection region whose type I error rate is no larger than a certain level. There is well-defined relationship between two type I error

rate: the FDR and the FWER. To see this, we write

$$\begin{aligned} \mathbb{E}\left[\frac{\text{FP}}{\max(\text{R}, 1)}\right] &= \mathbb{E}\left[\frac{\text{FP}}{\text{R}} \mid \text{FP} \geq 1\right] \mathbb{P}[\text{FP} \geq 1] + 0 \cdot \mathbb{P}[\text{FP} = 0] \\ &\leq \mathbb{P}[\text{FP} \geq 1], \end{aligned}$$

then the FDR is less than or equal to the FWER. This implies that any procedure that controls the FWER will also control the FDR. The reverse, however, is not true. That is, control of the FDR does not generally imply control of the FWER.

### 1.2.1 FWER control procedures

[Hochberg and Tamhane \[1987\]](#) describe a variety of FWER-controlling methods, based on cut-off rules for ordered  $p$ -values. [Westfall and Young \[1993\]](#) provide resampling-based multiple testing procedures for controlling the FWER. We only present here some classical procedures to control the FWER. [Bonferroni \[1936\]](#)'s procedure is perhaps the best-known procedure in the multiple testing literature. It controls the FWER for arbitrary test statistics joint null distributions.

**[Bonferroni \[1936\]](#)'s procedure.** The Bonferroni procedure rejects any null hypothesis  $H_i$  with a  $p$ -value less than or equal to the common threshold  $t(p) = \alpha/n$ . That is, the set of rejected null hypotheses is  $R(p) = \{1 \leq i \leq n : p_i \leq \alpha/n\}$ . This procedure controls the FWER under arbitrary conditions. That is,

$$\begin{aligned} \text{FWER} &= \mathbb{P}(\exists \text{ a false positive}) = \mathbb{P}\left(\bigcup_{i: H_{0,i} \text{ is true}} \{p_i \leq t(p)\}\right) \\ &\leq \sum_{i: H_{0,i} \text{ is true}} \mathbb{P}(p_i \leq t(p)) \leq n_0 \frac{\alpha}{n} \leq \alpha. \end{aligned}$$

Closely related to [Bonferroni \[1936\]](#)'s procedure is [Šidák \[1967\]](#)'s procedure, which guarantees control of the FWER for the test statistics distributions that satisfy [Šidák's Inequality](#). It rejects any null hypothesis  $H_i$  with a  $p$ -value less than or equal to the common threshold  $t(p) = 1 - (1 - \alpha)^{1/n}$ . Since  $\alpha/n \leq 1 - (1 - \alpha)^{1/n}$ , [Šidák \[1967\]](#)'s procedure is thus more powerful than [Bonferroni \[1936\]](#)'s one. In other words, using [Šidák \[1967\]](#)'s procedure, we reject a larger number of hypotheses while controlling the same error rate, which leads to larger power. Besides, there are some other procedures that intend to control the family wise error rate and they are more powerful than [Bonferroni \[1936\]](#)'s procedure. Among those procedures, we can recall [Holm \[1979\]](#)'s procedure and [Hochberg \[1988\]](#)'s procedure.

### 1.2.2 FDR control procedures

A common criticism of multiple testing procedures designed to control the FWER is their lack of power, especially for large-scale testing problems such as those encountered in biomedical and genomic research. In many situations, control of the FWER can lead to unduly conservative procedures. In current areas of application of multiple testing procedures, such as gene expression studies based on microarray experiments, thousands of tests are performed simultaneously and a fairly large proportion of null hypotheses are expected to be false. In this context, Type I error rates based on the proportion of false positives among the rejected hypotheses (FDR) may be more appropriate than error rates based on the absolute number of Type I errors (FWER). [Benjamini and Hochberg \[1995\]](#) provided a linear step-up procedure (the BH procedure) which controls the FDR at a certain level  $\alpha$ .

**A linear step-up procedure (the BH procedure).** Consider testing  $H_1, \dots, H_n$  based on the corresponding  $p$ -values  $p_1, \dots, p_n$ ,

- Step 1: let  $p_{(1)} \leq \dots \leq p_{(n)}$  be the ordered  $p$ -values and denote by  $H_{(i)}$  the null hypothesis corresponding to  $p_{(i)}$ ,
- Step 2: calculate  $\hat{k} = \max\{1 \leq i \leq n : p_{(i)} \leq i\alpha/n\}$ ,
- Step 3: if  $\hat{k}$  exists then reject all  $H_{(i)}$  for  $i = 1, \dots, \hat{k}$ , otherwise reject nothing.

[Benjamini and Hochberg \[1995\]](#) prove that this procedure controls the FDR for independent test statistics. The subsequent article of [Benjamini and Yekutieli \[2001\]](#) establishes FDR control for test statistics with the positive dependence structure called *positive regression dependence from a subset*. Since [Benjamini and Hochberg \[1995\]](#)'s article, many authors have proposed a variety of multiple testing procedures for controlling the FDR. We first describe an adaptive linear step-up procedure which is proposed by [Benjamini and Hochberg \[2000\]](#).

**An adaptive linear step-up procedure.** Note that in fact, the BH procedure controls the FDR at level  $\theta\alpha$  under independence or positive dependence conditions, this suggests the use of the following adaptive procedure that depends on an estimator of  $\theta$ :

- Step 1: compute an estimator of  $\theta$  as  $\hat{\theta}_n$ ,
- Step 2: if  $\hat{\theta}_n = 0$ , reject all hypotheses; otherwise, test the hypotheses by using the BH



linear step-up procedure at level  $\alpha/\hat{\theta}_n$ . That is, reject all  $H^{(i)}$  for  $i = 1, \dots, \hat{l}$ , where

$$\hat{l} = \max\{i : p_{(i)} \leq \frac{i\alpha}{n\hat{\theta}_n}\}.$$

Now suppose that we take the most conservative estimate  $\hat{\theta}_n = 1$ , then

$$\hat{l} = \max\{i : p_{(i)} \leq \frac{i\alpha}{n}\} = \hat{k},$$

it means that the adaptive linear step-up procedure identifies with the BH linear step-up one in this case. Moreover, if we take a better estimator  $\hat{\theta}_n < 1$ , then  $\hat{l} > \hat{k}$ . In other words, using the adaptive linear step-up procedure, we reject a larger number of hypotheses while controlling the same error rate, which leads to larger power.

Since 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  $[\lambda, 1]$ , for  $\lambda$  not too small, should correspond to the true null hypotheses, Schweder and Spjøtvoll [1982] suggested a procedure to estimate  $\theta$ , that depends on the unspecified parameter  $\lambda$ . This estimator is equal to the proportion of  $p$ -values larger than this threshold  $\lambda$  divided by  $1 - \lambda$ , namely

$$\hat{\theta}_n(\lambda) = \frac{\#\{P_i > \lambda : 1 \leq i \leq n\}}{n(1 - \lambda)}. \quad (1.2)$$

Benjamini and Hochberg [2000] used this estimator to propose an adaptive linear step-up procedure. They also showed that this adaptive procedure has higher power than the BH one. And Storey et al. [2004] provided a proof that it controls FDR at a level  $\alpha$ . Note that  $\hat{\theta}_n(\lambda)$  is a conservative estimator of  $\theta$  (it means that  $\hat{\theta}_n(\lambda)$  overestimates  $\theta$ ). Moreover, small values of  $\lambda$  typically produce estimators with higher bias but lower variance, whereas large values of  $\lambda$  yield low bias and high variance estimators. There exist many methods to choose the value of  $\lambda$  and the most popular choice is to let  $\lambda = 1/2$ . Recently, Liang and Nettleton [2012] have summed up many existing adaptive procedures under two different strategies to select  $\lambda$ , the first one includes the adaptive procedures that use predetermined values of  $\lambda$  and the second one includes the dynamic adaptive procedures where the parameter  $\lambda$  is determined by data.

**A plug-in threshold procedure.** We now present here the FDR controlling method that is proposed by Genovese and Wasserman (2002, 2004). They consider the threshold

$$t(\theta, F) = \sup\{0 \leq t \leq 1 : \frac{\theta t}{F(t)} \leq \alpha\},$$

where we recall that  $F$  is the cumulative distribution function of each  $p$ -value  $P_i$ . Suppose that we reject the null hypotheses whenever the  $p$ -value is less than  $t(\theta, F)$ . This threshold depends on the unknown parameters  $\theta$  and  $F$ , so we call  $t(\theta, F)$  the oracle threshold. From [Genovese and Wasserman \[2002\]](#), it follows that, asymptotically, the FDR is less than  $\alpha$ . Moreover, if  $F$  is concave this threshold has the smallest asymptotic FNR among all procedures with FDR less than or equal to  $\alpha$  (cf. [Genovese and Wasserman \[2002\]](#)). The standard plug-in method is to estimate the functional  $t(\theta, F)$  by  $t(\hat{\theta}_n, \hat{F})$ , where  $\hat{\theta}_n$  and  $\hat{F}$  are estimators of  $\theta$  and  $F$ . We thus call any threshold of the form  $t(\hat{\theta}_n, \hat{F})$  a plug-in threshold. For instance, let  $\hat{F}_n$  be the empirical cumulative distribution function of  $P_1, P_2, \dots, P_n$ . [Genovese and Wasserman \[2004\]](#) showed that under weak conditions on  $\hat{\theta}_n$ , the thresholding procedure  $t(\hat{\theta}_n, \hat{F}_n)$  asymptotically controls FDR at a level  $\alpha$ .

**One-stage and two-stage adaptive procedures.** [Blanchard and Roquain \[2009\]](#) propose two FDR control procedures called one-stage and two-stage adaptive step-up procedures. In their one-stage procedure, they reject all null hypotheses for which  $p_i \leq p_{(k)}$ , where

$$k = \max \left\{ i : p_{(i)} \leq \min \left\{ \frac{(1-\lambda)i\alpha}{m-i+1}, \lambda \right\} \right\} = \max \left\{ i : p_{(i)} \leq \frac{i\alpha}{m} \min \left\{ \frac{(1-\lambda)m}{m-i+1}, \frac{\lambda m}{i\alpha} \right\} \right\},$$

for a fixed constant  $\lambda \in (0, 1)$ . They focus on the choice  $\lambda = \alpha$ , then this procedure can be viewed as an adaptive linear step-up procedure with  $\theta$ -estimator defined as

$$\hat{\theta}_n(i) = \max \left\{ \frac{m-i+1}{(1-\alpha)m}, \frac{i}{m} \right\}.$$

Their two-stage procedure is defined as an adaptive linear step-up procedure with  $\theta$ -estimator given by

$$\hat{\theta}_n^{BR}(\lambda) = \frac{m - R^{BR}(\lambda) + 1}{(1-\alpha)m},$$

where  $R^{BR}(\lambda)$  is the number of rejections that result from using the one-stage adaptive step-up procedure at level  $\lambda \in (0, 1)$ . These two procedures are proved to be competitive with previous existing ones under the assumption of independence of the  $p$ -values. Moreover, the authors propose some adaptive step-up procedures that have provably controlled FDR under positive dependence and unspecified dependence of the  $p$ -values, respectively (for more detail, we refer to [Blanchard and Roquain \[2009\]](#)).

## 1.3 The FDR estimation approach

### 1.3.1 Estimation of pFDR and FDR

Rather than searching for a  $p$ -value threshold that can guarantee FDR control at a specified level  $\alpha$ , Storey (2002, 2004) proposed to estimate the FDR for a fixed rejection region and provided a family of conservative point estimators. The following is Theorem 1 from Storey [2002]. It allows us to write pFDR in a very simple form that does not depend on  $n$ .

**Theorem 1.1.** *Suppose that  $n$  identical hypothesis tests are performed with the independent statistics  $X_1, \dots, X_n$  and rejection region  $\Gamma$ . Then*

$$pFDR(\Gamma) = \frac{\theta \mathbb{P}(X \in \Gamma | H = 0)}{\mathbb{P}(X \in \Gamma)} = \mathbb{P}(H = 0 | X \in \Gamma).$$

In terms of  $p$ -values, instead of denoting rejection regions by  $\Gamma$ , we denote them by  $\gamma$ , which refers to the interval  $[0, \gamma]$ . Then the pFDR can be written as

$$pFDR(\gamma) = \frac{\theta \mathbb{P}(P \leq \gamma | H = 0)}{\mathbb{P}(P \leq \gamma)} = \frac{\theta \gamma}{F(\gamma)},$$

where  $P$  is the random  $p$ -value resulting from any test. And the FDR can be computed as  $FDR(\gamma) = pFDR(\gamma) \mathbb{P}(R > 0)$ , where

$$\begin{aligned} \mathbb{P}(R > 0) &= 1 - \mathbb{P}(R = 0) = 1 - \mathbb{P}(\forall i, P_i > \gamma) \\ &= 1 - [1 - \mathbb{P}(P \leq \gamma)]^n = 1 - [1 - F(\gamma)]^n. \end{aligned}$$

Thus, pFDR and FDR are asymptotically equivalent for a fixed rejection region, precisely we have

$$pFDR(\gamma) - FDR(\gamma) = pFDR(\gamma)[1 - F(\gamma)]^n \xrightarrow[n \rightarrow \infty]{} 0.$$

It is then natural to use the same estimates for  $FDR(\gamma)$  and  $pFDR(\gamma)$ . For a given estimator  $\hat{\theta}_n$  of  $\theta$ , we estimate  $FDR(\gamma)$  by

$$\widehat{FDR}(\gamma) = \frac{\hat{\theta}_n \gamma}{\hat{F}_n(\gamma)},$$

where  $\hat{F}_n$  is the empirical distribution function of  $P_1, \dots, P_n$ . For example, Storey [2002] considers a conservative estimate of  $\theta$  that depends on the tuning parameter  $\lambda$  and is defined as (1.2), then he proposes an estimate of  $FDR(\gamma)$  as

$$\widehat{FDR}_\lambda(\gamma) = \frac{\hat{\theta}_n(\lambda) \gamma}{\hat{F}_n(\gamma)}.$$

Note that a good  $\theta$ -estimator is very important as a conservative  $\theta$ -estimator in general leads to a conservative FDR estimator, which can be used to control the FDR; this point was well illustrated through the work of [Storey \[2002\]](#) and [Storey et al. \[2004\]](#). We can also refer to [Benjamini et al. \[2006\]](#) for more detail on this point.

### 1.3.2 Connection between FDR estimation and FDR control

Most FDR research has focused on FDR control instead of FDR estimation. However, there is a connection between these two approaches. Let us first note that

$$\hat{l} = \max\{i : p_{(i)} \leq \frac{i\alpha}{n\hat{\theta}_n}\} = \max\{i : \frac{n\hat{\theta}_n p_{(i)}}{i} \leq \alpha\} = \max\{i : \widehat{FDR}(p_{(i)}) \leq \alpha\},$$

ie, the adaptive linear step-up procedure is equivalent to finding the largest  $p$ -value  $p_{(l)}$  such that  $\widehat{FDR}(p_{(l)}) \leq \alpha$ . The FDR estimation approach can be thus viewed as the “inverse problem” of the FDR control approach. For any function  $h$  defined on  $[0, 1]$ , let the step-up thresholding function be

$$t_\alpha(h) = \sup\{0 \leq t \leq 1 : h(t) \leq \alpha\}.$$

Then the threshold of the adaptive linear step-up procedure is exactly  $t_\alpha(\widehat{FDR})$ . Similarly, the oracle threshold of the plug-in threshold procedure can be written

$$t(\theta, F) = \sup\{0 \leq t \leq 1 : \frac{\theta t}{F(t)} \leq \alpha\} = \sup\{0 \leq t \leq 1 : \text{pFDR}(t) \leq \alpha\} = t_\alpha(\text{pFDR}).$$

The plug-in threshold procedure is thus identical to the adaptive linear step-up procedure when we apply a common estimate  $\hat{\theta}_n$ . We now present how a FDR estimation approach leads to a FDR control approach. Since  $t_\alpha(\widehat{FDR})$  is a random variable, we use the following notation

$$\text{FDR}\{t_\alpha(\widehat{FDR})\} := \mathbb{E}\left[\frac{\text{FP}\{t_\alpha(\widehat{FDR})\}}{\text{R}\{t_\alpha(\widehat{FDR})\}}\right].$$

[Storey et al. \[2004\]](#) and [Liang and Nettleton \[2012\]](#) proposed some FDR estimation approaches such that  $\text{FDR}\{t_\alpha(\widehat{FDR})\} \leq \alpha$ . Therefore, these thresholding procedures  $t_\alpha(\widehat{FDR})$  control the FDR at level  $\alpha$ .

## 1.4 Local false discovery rate

[Efron et al. \[2001\]](#) define the local false discovery rate ( $\ell\text{FDR}$ ) to quantify the plausibility of a particular hypothesis being true, given its specific test statistic or  $p$ -value. In a mixture

framework, the  $\ell\text{FDR}$  is the Bayes posterior probability

$$\ell\text{FDR}(x) = \mathbb{P}(H_i \text{ being true} \mid X = x) = 1 - \frac{(1 - \theta)g_1(x)}{\theta g_0(x) + (1 - \theta)g_1(x)}.$$

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  $\ell\text{FDR}$ . Moreover, another motivation for estimating the parameters in this mixture model comes from the works of Sun and Cai (2009, 2007), who develop adaptive compound decision rules for false discovery rate control. These rules are based on the estimation of the local false discovery rate  $\ell\text{FDR}$ . Let  $R$  be the set of ranked  $\widehat{\ell\text{FDR}}(x_i)$ :  $R = \{\widehat{\ell\text{FDR}}_{(1)}, \dots, \widehat{\ell\text{FDR}}_{(n)}\}$ . Sun and Cai [2007] proposed the following adaptive step-up procedure:

$$\text{Let } k = \max\left\{i : \frac{1}{i} \sum_{j=1}^i \widehat{\ell\text{FDR}}_{(j)} \leq \alpha\right\};$$

then reject all  $H_{(i)}, i = 1, \dots, k$ .

Sun and Cai [2007] showed that this procedure asymptotically attains the performance of an oracle procedure and in some simulation studies, it is more efficient than the conventional  $p$ -value-based methods, including the step-up procedure of Benjamini and Hochberg [1995] and the plug-in procedure of Genovese and Wasserman [2004]. Moreover, recall that  $z_i$  denotes the  $z$ -value and  $p_i$  denotes the  $p$ -value, we can write

$$\begin{aligned} \ell\text{FDR}(i) &:= \ell\text{FDR}(z_i) := \frac{\theta \phi(z_i)}{\phi(z_i)[\theta + (1 - \theta)f_1(\Phi(z_i))]} \\ &= \frac{\theta}{\theta + (1 - \theta)f_1(p_i)} := \ell\text{FDR}(p_i), \end{aligned}$$

thus this procedure is more adaptive than the BH adaptive procedure in the sense that it adapts to both the global feature  $p$ -value and  $z$ -value.

Let us note that  $\text{pFDR}$  and  $\ell\text{FDR}$  are analytically related by

$$\begin{aligned} \text{pFDR}(\gamma) &= \int_{-\infty}^{\gamma} \ell\text{FDR}(p) f(p) dp \left( \int_{-\infty}^{\gamma} f(p) dp \right)^{-1} \\ &= \mathbb{E}\{\ell\text{FDR}(P) \mid P \leq \gamma\}, \end{aligned}$$

then we can estimate  $\text{pFDR}$  or  $\text{FDR}$  by

$$\widehat{\text{FDR}}(p_{(i)}) = \frac{1}{i} \sum_{j=1}^i \widehat{\ell\text{FDR}}_{(j)}.$$

So that the above adaptive step-up procedure can be viewed as a plug-in threshold procedure  $t_\alpha(\widehat{\text{FDR}})$ . To conclude, let us stress that all FDR control procedures presented in Sections 1.2.2 and 1.4 can also be viewed as plug-in threshold procedures  $t_\alpha(\widehat{\text{FDR}})$  with suitable estimates  $\widehat{\text{FDR}}$ .

## 1.5 Semiparametric inference

In this section, we recall concepts from semiparametric theory. We follow the notation of Chapter 25 and more particularly Section 25.4 in [van der Vaart \[1998\]](#) and refer to this book for more details. Semiparametric models are statistical models in which the parameters are indexed by a finite-dimensional vector and an infinite-dimensional parameter. Precisely, a semiparametric model in a strict sense may have a natural parametrization  $(\theta, f) \mapsto \mathbb{P}_{\theta, f}$ , where  $\theta$  is a Euclidean parameter and  $f$  belongs to a nonparametric class of distributions. Here, we aim at estimating the value  $\psi(\mathbb{P}_{\theta, f}) = \theta$  and consider  $f$  as a nuisance parameter. We shall recall the theory of asymptotic efficiency for semiparametric models which is extended from parametric models.

### 1.5.1 Tangent sets and efficient influence function

We first recall the definition of tangent set in a general model. In this section, suppose that we observe a random sample  $X_1, X_2, \dots, X_n$  from a distribution  $\mathbb{P}$  which belongs to a set  $\mathcal{P}$  of probability measures on some measurable space  $(\mathcal{X}, \mathcal{A})$ . In particular, we consider a framework that is more general than the semiparametric one. We aim at estimating the value  $\psi(\mathbb{P})$  of a functional  $\psi : \mathcal{P} \rightarrow \mathbb{R}^k$ . Assume for simplicity that the parameter to be estimated is one-dimensional ( $k = 1$ ). In parametric models, we have a strict definition for the Fisher information for estimating the parameters. So, what can we say about the information for the semiparametric model  $\mathcal{P}$  for estimating  $\psi(\mathbb{P})$ ? For every smooth parametric submodel  $\mathcal{P}_0 \subset \mathcal{P}$  that contains the true distribution  $\mathbb{P}$ , we can calculate its Fisher information for estimating  $\psi(\mathbb{P})$ . Then the information for estimating  $\psi(\mathbb{P})$  for the whole model is not bigger than the information covered by each of these parametric submodels. So it is certainly not bigger than the infimum of the informations over all submodels. The information for  $\mathcal{P}$  is then simply defined as this infimum. It seems that in most situations, it suffices to consider one-dimensional submodels  $\mathcal{P}_0$ . We know that they should pass through the true distribution  $\mathbb{P}$  and be differentiable in quadratic mean at  $\mathbb{P}$  which we shall define now.

**Definition 1.1.** A differentiable path is a map  $t \mapsto \mathbb{P}_t$  from a neighbourhood  $[0, \epsilon)$  of 0 to  $\mathcal{P}$  with  $\mathbb{P}_0 = \mathbb{P}$  such that, for some measurable function  $g : \mathcal{X} \rightarrow \mathbb{R}$ ,

$$\int \left( \frac{d\mathbb{P}_t^{1/2} - d\mathbb{P}^{1/2}}{t} - \frac{1}{2}g d\mathbb{P}^{1/2} \right)^2 \rightarrow 0 \text{ as } t \rightarrow 0. \quad (1.3)$$

The parametric submodel  $\{\mathbb{P}_t : 0 \leq t < \epsilon\}$  is called differentiable in quadratic mean at  $\mathbb{P}$  and the function  $g$  is called the score function of the submodel  $\{\mathbb{P}_t : 0 \leq t < \epsilon\}$ .

Letting  $t \mapsto \mathbb{P}_t$  range over a collection of these submodels, we obtain a collection of score functions, which we call a tangent set of the model  $\mathcal{P}$  at  $\mathbb{P}$ . We denote this tangent set by  $\dot{\mathcal{P}}_{\mathbb{P}}$ . When we consider all possible differentiable paths  $t \mapsto \mathbb{P}_t$ , we obtain the maximal collection of score functions. This set is referred to as the maximal tangent set. A tangent set is usually a cone: if  $g \in \dot{\mathcal{P}}_{\mathbb{P}}$  and  $a \geq 0$ , then  $ag \in \dot{\mathcal{P}}_{\mathbb{P}}$ , since the path  $t \mapsto \mathbb{P}_{at}$  has score function  $ag$  when  $t \mapsto \mathbb{P}_t$  has score function  $g$ . Usually, we construct the submodels  $t \mapsto \mathbb{P}_t$  such that, for every  $x$ ,

$$g(x) = \left. \frac{\partial}{\partial t} \right|_{t=0} \log d\mathbb{P}_t(x).$$

This pointwise differentiability is not required by (1.3). Conversely, given this pointwise differentiability, we are not assured to have (1.3). We still need to be able to apply a convergence theorem for integrals to obtain this type of convergence in quadratic mean, such as the dominated convergence theorem of Lebesgue or the monotone convergence theorem, since we need to interchange limit and integration. The following lemma solves most examples as stated in [van der Vaart \[2002\]](#).

**Lemma 1.1.** If  $p_t$  is the density function of a probability distribution  $\mathbb{P}_t$  relative to a fixed measure  $\mu$  and  $t \mapsto \sqrt{p_t(x)}$  is continuously differentiable in a neighbourhood of 0 and  $t \mapsto \int p_t^2/p_t d\mu$ , where  $\dot{p}_t = \partial p_t / \partial t$ , is finite and continuous in this neighbourhood, then  $t \mapsto \mathbb{P}_t$  is a differentiable path.

The following lemma gives two fundamental but familiar properties of score functions. These are consequences of the differentiability in quadratic mean. We denote by  $\mathbb{L}_2(\mathbb{P})$  the space of measurable functions  $g : \mathcal{X} \rightarrow \mathbb{R}$  with  $\mathbb{P}g^2 = \int g^2 d\mathbb{P} < \infty$ , where almost surely equal functions are identified.

**Lemma 1.2.** Every score function belongs to the set  $\{g \in \mathbb{L}_2(\mathbb{P}) : \mathbb{P}g = 0\}$ .

From this lemma, we can conclude that a tangent set is a subset of the space  $\mathbb{L}_2(\mathbb{P})$  with mean zero. The tangent set is often a linear space, in which case we speak of a tangent space.

**Example (nonparametric model).** Suppose that  $\mathcal{P}$  consists of all probability distributions on the sample space. Let  $g$  be an arbitrary function such that  $g \in \mathbb{L}_2(\mathbb{P})$  and  $\mathbb{P}g = 0$ . We consider the submodel given by  $t \mapsto p_t(x) = c(t)k(tg(x))p_0(x)$  for a nonnegative function  $k$  with  $k(0) = k'(0) = 1$  and  $[c(t)]^{-1} = \int k(tg(x))p_0(x)dx$ . The function  $k(x) = 2(1 + \exp(-2x))^{-1}$  can be used for example. By a direct calculation or by using Lemma 1.1, we see that the path  $t \mapsto p_t(x)$  is differentiable and the corresponding score function is  $g$ . Then the maximal tangent set coincides with the space  $\{g \in \mathbb{L}_2(\mathbb{P}) : \mathbb{P}g = 0\}$ .

For defining the information for estimating  $\psi(\mathbb{P})$ , only those submodels  $t \mapsto \mathbb{P}_t$  along which the parameter  $t \mapsto \psi(\mathbb{P}_t)$  is differentiable in an appropriate sense are of interest. A minimal requirement is that the map  $t \mapsto \psi(\mathbb{P}_t)$  is differentiable at  $t = 0$ , but we need more. More precisely, a map  $\psi : \mathcal{P} \rightarrow \mathbb{R}$  is called differentiable at  $\mathbb{P}$  relative to a given tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$  if there exists a continuous linear map  $\dot{\psi}_{\mathbb{P}} : \mathbb{L}_2(\mathbb{P}) \rightarrow \mathbb{R}$  such that for every  $g \in \dot{\mathcal{P}}_{\mathbb{P}}$  and a submodel  $t \mapsto \mathbb{P}_t$  with score function  $g$ ,

$$\left. \frac{\partial \psi(\mathbb{P}_t)}{\partial t} \right|_{t=0} = \lim_{t \rightarrow 0} \frac{\psi(\mathbb{P}_t) - \psi(\mathbb{P})}{t} = \dot{\psi}_{\mathbb{P}}g.$$

The Riesz representation theorem for Hilbert spaces yields the existence of a measurable function  $\tilde{\psi}_{\mathbb{P}} : \mathcal{X} \rightarrow \mathbb{R}$  such that

$$\dot{\psi}_{\mathbb{P}}g = \langle \tilde{\psi}_{\mathbb{P}}, g \rangle_{\mathbb{L}_2(\mathbb{P})} = \int \tilde{\psi}_{\mathbb{P}}gd\mathbb{P}. \quad (1.4)$$

A function  $\tilde{\psi}_{\mathbb{P}}$  satisfying (1.4) is defined to be an influence function. The Riesz representation theorem assures uniqueness of the function  $\tilde{\psi}_{\mathbb{P}}$  when the inner product  $\langle \cdot, \cdot \rangle_{\mathbb{L}_2(\mathbb{P})}$  is specified for all functions of  $\mathbb{L}_2(\mathbb{P})$ . Here, only inner products of  $\tilde{\psi}_{\mathbb{P}}$  with elements  $g$  of the tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$  are specified and the tangent set does not span all of  $\mathbb{L}_2(\mathbb{P})$ , therefore the function  $\tilde{\psi}_{\mathbb{P}}$  is not uniquely defined by the functional  $\psi$  and the model  $\mathcal{P}$ . However, using the projection theorem of Hilbert spaces, we can construct a unique  $\tilde{\psi}_{\mathbb{P}}$  contained in  $\overline{\text{lin}}\dot{\mathcal{P}}_{\mathbb{P}}$ , the closure of the linear span of the tangent set. This function is called the efficient influence function. So, for further reference, when we write  $\tilde{\psi}_{\mathbb{P}}$ , we refer this to be the efficient influence function.

### 1.5.2 Asymptotically efficient estimator

To motivate the definition of information in the semiparametric setup, we first consider a differentiable parametric submodel  $t \mapsto \mathbb{P}_t$  with score function  $g$ . It is easy to show that the Fisher information in this parametric submodel is equal to the variance of the score function  $g$ , i.e.  $I = \mathbb{P}g^2 = \langle g, g \rangle_{\mathbb{P}}$ . Thus, the Cramér-Rao bound for estimating the function  $t \mapsto \psi(\mathbb{P}_t)$ ,



evaluated at  $t = 0$ , is

$$\frac{[\partial\psi(\mathbb{P}_t)/\partial t|_{t=0}]^2}{\mathbb{P}g^2} = \frac{\langle \tilde{\psi}_{\mathbb{P}}, g \rangle_{\mathbb{P}}^2}{\langle g, g \rangle_{\mathbb{P}}}.$$

We now present an important lemma.

**Lemma 1.3.** *Suppose that the functional  $\psi : \mathcal{P} \rightarrow \mathbb{R}$  is differentiable at  $\mathbb{P}$  relative to a tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$ . Then*

$$\sup_{g \in \text{lin}\dot{\mathcal{P}}_{\mathbb{P}}} \frac{\langle \tilde{\psi}_{\mathbb{P}}, g \rangle_{\mathbb{P}}^2}{\langle g, g \rangle_{\mathbb{P}}} = \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2.$$

Now the special meaning of the efficient influence function becomes clear. The squared norm  $\mathbb{P}\tilde{\psi}_{\mathbb{P}}^2$  of the efficient influence function  $\tilde{\psi}_{\mathbb{P}}$  plays the role of a smallest asymptotic variance an estimator for  $\psi(\mathbb{P})$  can have. We thus call the number  $\mathbb{P}\tilde{\psi}_{\mathbb{P}}^2$  the efficiency bound or the optimal variance.

For every function  $g$  in a given tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$ , we write  $\mathbb{P}_{t,g}$  for a corresponding submodel with score function  $g$  along which the function  $\psi$  is differentiable. The asymptotic minimax risk of an estimator sequence  $T_n$  (relative to the tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$ ), is defined as

$$\sup_I \liminf_{n \rightarrow \infty} \sup_{g \in I} \mathbb{P}_{1/\sqrt{n},g} [\sqrt{n}(T_n - \psi(\mathbb{P}_{1/\sqrt{n},g}))]^2,$$

where the first supremum is taken over all finite subsets  $I$  of the tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$ . We now state the local asymptotic minimax theorem that gives a lower bound of the asymptotic minimax risk of an arbitrary estimator  $T_n$  [see Theorem 25.21 in [van der Vaart, 1998](#)].

**Theorem 1.2. (Local Asymptotic Minimax, LAM).** *Let the function  $\psi : \mathcal{P} \rightarrow \mathbb{R}$  be differentiable at  $\mathbb{P}$  relative to the tangent cone  $\dot{\mathcal{P}}_{\mathbb{P}}$  with efficient influence function  $\tilde{\psi}_{\mathbb{P}}$ . If  $\dot{\mathcal{P}}_{\mathbb{P}}$  is a convex cone, then for any estimator sequence  $T_n$ ,*

$$\sup_I \liminf_{n \rightarrow \infty} \sup_{g \in I} \mathbb{P}_{1/\sqrt{n},g} [\sqrt{n}(T_n - \psi(\mathbb{P}_{1/\sqrt{n},g}))]^2 \geq \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2. \quad (1.5)$$

*The first supremum is taken over all finite subsets  $I$  of the tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$ .*

An estimator sequence  $T_n$  is called regular at  $\mathbb{P}$  for estimating  $\psi(\mathbb{P})$  (relative to  $\dot{\mathcal{P}}_{\mathbb{P}}$ ) if there exists a probability measure  $L$  such that

$$\sqrt{n}(T_n - \psi(\mathbb{P}_{1/\sqrt{n},g})) \overset{\mathbb{P}_{1/\sqrt{n},g}}{\rightsquigarrow} L, \text{ for every } g \in \dot{\mathcal{P}}_{\mathbb{P}},$$

where  $\overset{\mathbb{P}}{\rightsquigarrow}$  denotes convergence in distribution under  $\mathbb{P}$ . We now state the convolution theorem, that shows that the limit distribution  $L$  writes as the convolution between some unknown distribution and the centered Gaussian distribution  $N(0, \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2)$  [see Theorem 25.20 in [van der Vaart, 1998](#)].

**Theorem 1.3. (Convolution).** *Let the function  $\psi : \mathcal{P} \rightarrow \mathbb{R}$  be differentiable at  $\mathbb{P}$  relative to the tangent cone  $\dot{\mathcal{P}}_{\mathbb{P}}$  with efficient influence function  $\tilde{\psi}_{\mathbb{P}}$ . Then the asymptotic variance of every regular sequence of estimators is bounded below by  $\mathbb{P}\tilde{\psi}_{\mathbb{P}}^2$ . Furthermore, if  $\dot{\mathcal{P}}_{\mathbb{P}}$  is a convex cone, then every limit distribution  $L$  of a regular sequence of estimators can be written  $L = U + M$  where  $U \sim N(0, \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2)$  and  $M$  is some probability distribution independent of  $U$ .*

According to this theorem, we say that an estimator sequence is asymptotically efficient at  $\mathbb{P}$  (relative to the tangent set  $\dot{\mathcal{P}}_{\mathbb{P}}$ ) if it is regular at  $\mathbb{P}$  with limit distribution  $L = N(0, \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2)$ , in other words it is the best regular estimator. The definition of asymptotic efficiency is not absolute since it is defined relative to a given tangent set. In practice, we aim at finding a tangent set and an estimator sequence such that the tangent set is big enough and the estimator sequence efficient enough so that this estimator sequence is asymptotically efficient relative to this tangent set. We end this section on general efficiency theory with an interesting lemma.

**Lemma 1.4.** *Let the functional  $\psi : \mathcal{P} \rightarrow \mathbb{R}$  be differentiable at  $\mathbb{P}$  relative to the tangent cone  $\dot{\mathcal{P}}_{\mathbb{P}}$  with efficient influence function  $\tilde{\psi}_{\mathbb{P}}$ . A sequence of estimators  $T_n$  is regular at  $\mathbb{P}$  with limit distribution  $N(0, \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2)$  if and only if*

$$\sqrt{n}(T_n - \psi(\mathbb{P})) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\psi}_{\mathbb{P}}(X_i) + o_{\mathbb{P}}(1).$$

The nice thing about asymptotically efficient estimators is that they have interesting asymptotic properties and they are fully characterized by their efficient influence function. First, we note that  $T_n$  is consistent, i.e.,  $T_n \xrightarrow{\mathbb{P}} \psi(\mathbb{P})$ . In addition, by the central limit theorem and Slutsky's theorem, we obtain that

$$\sqrt{n}(T_n - \psi(\mathbb{P})) \overset{\mathbb{P}}{\rightsquigarrow} N(0, \mathbb{P}\tilde{\psi}_{\mathbb{P}}^2).$$

This means an asymptotically efficient estimator is asymptotically normal with asymptotic variance equal to the optimal variance. By Prohorov's theorem, the estimator  $T_n$  is also  $\sqrt{n}$ -consistent, i.e.,  $\sqrt{n}(T_n - \psi(\mathbb{P})) = O_{\mathbb{P}}(1)$ . We conclude that every asymptotically efficient estimator is a consistent, moreover a  $\sqrt{n}$ -consistent and asymptotically normal estimator.

### 1.5.3 Expressions for semiparametric models in a strict sense

We now focus our attention on semiparametric models in a strict sense,  $\mathcal{P} = \{\mathbb{P}_{\theta,f} : \theta \in \Theta, f \in \mathcal{F}\}$ , with  $\Theta \subset \mathbb{R}$  an open set and  $\mathcal{F}$  an arbitrary infinite dimension set of probability distributions. Our aim is to study the efficiency of an estimator  $T_n$  for  $\psi(\mathbb{P}_{\theta,f}) = \theta$ . Thus, we are looking for the efficient influence function  $\tilde{\psi}_{\theta,f}$  in this special setting. We will express the efficient influence function in terms of the efficient score function and the efficient information matrix. As submodels, we use paths of the form  $t \mapsto \mathbb{P}_{\theta+ta,f_t}$ , for given paths  $t \mapsto f_t$  in  $\mathcal{F}$  and  $a \in \mathbb{R}$ . The score functions for such submodels will typically have the form of a sum of partial derivatives with respect to the parametric component  $\theta$  and the nonparametric component  $f$ . If  $\dot{l}_{\theta,f}$  is the ordinary score function for  $\theta$  in the model where  $f$  is fixed (as we consider an ordinary parametric model), then we expect

$$\left. \frac{\partial}{\partial t} \right|_{t=0} \log d\mathbb{P}_{\theta+ta,f_t} = a\dot{l}_{\theta,f} + g.$$

The function  $g$  has the interpretation of a score function for  $f$  when  $\theta$  is fixed and typically runs through an infinite-dimensional set. We refer to this set as the tangent set for  $f$ , and denote it by  ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ . The functional  $\psi(\mathbb{P}_{\theta+ta,f_t}) = \theta + ta$  is certainly differentiable with respect to  $t$  in ordinary sense with derivative  $a$ . However, to be differentiable at  $\mathbb{P}_{\theta,f}$  relative to  $\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ , we need something more. By definition,  $\psi$  is differentiable relative to  $\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$  if and only if there exists a function  $\tilde{\psi}_{\theta,f}$  such that

$$a = \left. \frac{\partial}{\partial t} \right|_{t=0} \psi(\mathbb{P}_{\theta+ta,f_t}) = \langle \tilde{\psi}_{\theta,f}, a\dot{l}_{\theta,f} + g \rangle_{\mathbb{P}_{\theta,f}}, \quad \forall a \in \mathbb{R}, g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}.$$

By putting  $a = 0$ , we obtain that  $\langle \tilde{\psi}_{\theta,f}, g \rangle_{\mathbb{P}_{\theta,f}} = 0$  for all  $g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ . Thus,  $\tilde{\psi}_{\theta,f}$  must be orthogonal to the tangent set  ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$  for the nuisance parameter. In particular, the efficient influence function, which we denote again by  $\tilde{\psi}_{\theta,f}$ , is orthogonal to the nuisance tangent space  ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ . We shall state a lemma that gives an interesting form for the efficient influence function. Before doing that, we define the operator  $\Pi_{\theta,f} : \mathbb{L}_2(\mathbb{P}_{\theta,f}) \rightarrow \overline{\text{lin}}_f \dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$  to be the orthogonal projection onto the closure of the linear span of the nuisance tangent space in  $\mathbb{L}_2(\mathbb{P}_{\theta,f})$ . The function defined by  $\tilde{l}_{\theta,f} = \dot{l}_{\theta,f} - \Pi_{\theta,f} \dot{l}_{\theta,f}$  is called the efficient score function for  $\theta$  and its variance  $\tilde{I}_{\theta,f} = \mathbb{P}_{\theta,f} \tilde{l}_{\theta,f}^2$  is called the efficient information matrix for  $\theta$ .

**Lemma 1.5.** *Suppose that for every  $a \in \mathbb{R}$  and every  $g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$  there exists a path  $t \mapsto f_t$  in  $\mathcal{F}$  such that*

$$\int \left[ \frac{d\mathbb{P}_{\theta+ta,f_t}^{1/2} - d\mathbb{P}_{\theta,f}^{1/2}}{t} - \frac{1}{2}(a\dot{l}_{\theta,f} + g)d\mathbb{P}_{\theta,f}^{1/2} \right]^2 \rightarrow 0 \text{ as } t \rightarrow 0.$$

## 1.5. SEMIPARAMETRIC INFERENCE

---

If  $\tilde{I}_{\theta,f}$  is nonsingular, then the function  $\psi(\mathbb{P}_{\theta,f}) = \theta$  is differentiable at  $\mathbb{P}_{\theta,f}$  relative to the tangent set  $\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}} = \lim \dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}} = \{a\dot{l}_{\theta,f} + g : a \in \mathbb{R}, g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}\}$  with efficient influence function  $\tilde{\psi}_{\theta,f} = \tilde{I}_{\theta,f}^{-1}\tilde{l}_{\theta,f}$ .

As a consequence, we obtain a specialized version of Lemma 1.4. Suppose the nuisance tangent set  ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$  is a cone, then a sequence of estimators  $T_n$  is regular at  $\mathbb{P}_{\theta,f}$  with limiting distribution  $N(0, \mathbb{P}_{\theta,f}\tilde{\psi}_{\theta,f}^2)$  (asymptotically efficient) if and only if it satisfies

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{I}_{\theta,f}^{-1}\tilde{l}_{\theta,f}(X_i) + o_{\mathbb{P}_{\theta,f}}(1).$$

We first see that

$$\mathbb{P}_{\theta,f}\tilde{\psi}_{\theta,f}^2 = \left(\mathbb{P}_{\theta,f}\tilde{l}_{\theta,f}^2\right)^{-1} = \tilde{I}_{\theta,f}^{-1}.$$

The variance of the efficient influence function is equal to the inverse of the variance of the efficient score function or the inverse of the efficient information matrix. Thus, the reason why we call  $\tilde{l}_{\theta,f}$  the efficient score function and  $\tilde{I}_{\theta,f}$  the efficient information matrix is now clear. Secondly, under regularity conditions (see Chapters 5 and 8, [van der Vaart \[1998\]](#)), the maximum likelihood estimator  $\hat{\theta}_n$  in a parametric model satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\theta}^{-1}\dot{l}_{\theta}(X_i) + o_{\mathbb{P}_{\theta}}(1),$$

where  $I_{\theta}$  is the ordinary Fisher information matrix and  $\dot{l}_{\theta}$  is the ordinary score function. The only difference in a semiparametric model is that the ordinary score function  $\dot{l}_{\theta,f}$  is replaced by the efficient score function  $\tilde{l}_{\theta,f}$  and the Fisher information matrix  $I_{\theta,f}$  for  $\theta$  is replaced by the efficient information matrix  $\tilde{I}_{\theta,f}$ . A part of the score function for  $\theta$  can be accounted for by score functions for the nuisance parameter  $f$  when the nuisance parameter is unknown, a part of the information for  $\theta$  is lost. The orthogonal projection  $\Pi_{\theta,f}\dot{l}_{\theta,f}$  of the score function for  $\theta$  onto the nuisance tangent space  ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$  corresponds with this loss. When there is no nuisance parameter, there is no nuisance tangent space and thus no loss of information for estimating  $\theta$ .

### 1.5.4 One-step estimator method

In this section, we introduce the one-step method to construct an asymptotically efficient estimator, relying on a  $\sqrt{n}$ -consistent one [see [van der Vaart, 1998](#), Section 25.8]. Let  $\hat{\theta}_n$  be a  $\sqrt{n}$ -consistent estimator of  $\theta$ , then  $\hat{\theta}_n$  can be discretized on grids of mesh width  $n^{-1/2}$ . Suppose

that we are given a sequence of estimators  $\hat{l}_{n,\theta}(\cdot) = \hat{l}_{n,\theta}(\cdot; X_1, \dots, X_n)$  of the efficient score function  $\tilde{l}_{\theta,f}$ . Define with  $m = \lfloor n/2 \rfloor$ ,

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

Thus, for  $X_i$  ranging through each of the two halves of the sample, we use an estimator  $\hat{l}_{n,\theta,i}$  based on the other half of the sample. Then the one-step estimator is defined as

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

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  $\theta$ , starting at the initial guess  $\hat{\theta}_n$ . We now assume that, for every deterministic sequence  $\theta_n = \theta + O(n^{-1/2})$ , we have

$$\sqrt{n} \mathbb{P}_{\theta_n,f} \hat{l}_{n,\theta_n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\theta,f}} 0, \quad (1.6)$$

$$\mathbb{P}_{\theta_n,f} \|\hat{l}_{n,\theta_n} - \tilde{l}_{\theta_n,f}\|^2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\theta,f}} 0, \quad (1.7)$$

$$\int \|\tilde{l}_{\theta_n,f} d\mathbb{P}_{\theta_n,f}^{1/2} - \tilde{l}_{\theta,f} d\mathbb{P}_{\theta,f}^{1/2}\|^2 \xrightarrow[n \rightarrow \infty]{} 0. \quad (1.8)$$

Note that in the above 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  $\hat{l}$ .

**Theorem 1.4.** [Theorem 25.57 in [van der Vaart, 1998](#)] Suppose that the model  $\{\mathbb{P}_{\theta,f} : \theta \in \Theta\}$  is differentiable in quadratic mean with respect to  $\theta$  at  $(\theta, f)$ , let the efficient information matrix  $\tilde{I}_{\theta,f}$  be nonsingular. Assume that (1.6)- (1.8) hold. Then the one-step estimator  $\tilde{\theta}_n$  is asymptotically efficient at  $(\theta, f)$ .

This theorem reduces the problem of efficient estimation of  $\theta$  to estimation of the efficient score function. The estimator of the efficient score function must satisfy a “no-bias” (1.6) and a consistency (1.7) conditions. The consistency condition is usually easy to arrange, but the “no-bias” condition requires a convergence to zero of the bias at a rate faster than  $1/\sqrt{n}$ . If it fails, then the sequence  $\tilde{\theta}_n$  is not asymptotically efficient and may even converge at a slower rate than  $\sqrt{n}$ . The good news is that if an efficient estimator sequence exists, then it can always be constructed by the one-step method. In that sense the no-bias condition is necessary.

**Theorem 1.5.** *[Theorem 7.4 in [van der Vaart, 2002](#)] Suppose that the model  $\{\mathbb{P}_{\theta,f} : \theta \in \Theta\}$  is differentiable in quadratic mean with respect to  $\theta$  at  $(\theta, f)$ , let the efficient information matrix  $\tilde{I}_{\theta,f}$  be nonsingular, and assume that (1.8) holds. Then the existence of an asymptotically efficient estimator of  $\psi(\mathbb{P}_{\theta,f}) = \theta$  implies the existence of a sequence of estimators  $\hat{\lambda}_{n,\theta}$  satisfying (1.6) and (1.7).*

### 1.5.5 The infinite bound case

We end this section with an impossibility result due to [Chamberlain \[1986\]](#). Chamberlain showed that if the semiparametric efficiency bound (i.e., the variance of the efficient influence function  $\tilde{\psi}_{\mathbb{P}}$ ) is infinitely large (e.g.,  $\tilde{I}_{\theta,f} = \mathbb{P}_{\theta,f} \tilde{l}_{\theta,f}^2$  is singular), then no regular estimator exist. More precisely, if the efficient information matrix is singular, the variance of the efficient influence function is infinite and since this is a lower bound for the variance of any regular estimator, no regular estimator can exist. More details about this and other impossibility theorems can be found in [Newey \[1990\]](#).

## 1.6 Organization

Throughout this dissertation, we assume the test statistics are independent and identically distributed (iid) with a continuous distribution under the corresponding null or alternative hypotheses, then the  $p$ -values are iid and follow the uniform distribution  $\mathcal{U}([0, 1])$  in interval  $[0, 1]$  under the null hypotheses. The density  $g$  of the  $p$ -values is modeled by a two-component mixture with following expression

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

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). We recall that an adaptive linear step-up procedure or a plug-in threshold procedure requires an estimator of the parameter  $\theta$ . A good  $\theta$ -estimator is very important since a conservative  $\theta$ -estimator in general leads to a conservative FDR estimator, which can be used in FDR control procedures. Besides, the problem of estimating the component  $f$  appears from the estimation of the local false discovery rate, which is used in the adaptive step-up procedure of [Sun and Cai \[2007\]](#).

In the second chapter, we study the estimation of the proportion  $\theta$ . Firstly, let us note that many different estimators of  $\theta$  have been proposed in the literature but their rate of convergence

or asymptotic efficiency has only been partly studied. To our knowledge, there only exists some estimators that converge to  $\theta$  at nonparametric rate and it has not been investigated whether the parametric rate of convergence may be achieved by a consistent estimator of  $\theta$  in this semi-parametric setup. 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 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 parametric rate. We illustrate those results on simulated data. This chapter is a revised version that is submitted for publication in a journal of statistics.

Motivated by the issue of local false discovery rate estimation, in the third chapter, we focus on the estimation of the nonparametric unknown component  $f$  in the mixture, relying on a preliminary estimator of the unknown proportion  $\theta$  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. This chapter is accepted for publication in ESAIM: Probability and Statistics.

In the fourth chapter, we consider another mixture model that is useful to analyze gene expression data coming from microarray analysis. It is a mixture of two components where one component is assumed to be a known density with prior probability  $1-p$  and the other component is an unknown density that is assumed to be symmetric on  $\mathbb{R}$  with non-null location parameter  $\mu$ . This model has been introduced by [Bordes et al. \[2006\]](#). Here, we aim at computing the efficient information matrix for estimating the Euclidean parameter  $\theta = (p, \mu)$  and some ideas are proposed for future work.

# Estimation of the proportion of true null hypotheses

---

## Abstract

We consider the problem of estimating the proportion  $\theta$  of true null hypotheses in a multiple testing context. The setup is classically modeled through a semiparametric mixture with two components: a uniform distribution on interval  $[0, 1]$  with prior probability  $\theta$  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 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 parametric rate. We illustrate those results on simulated data.

## Contents

---

<b>2.1</b>	<b>Introduction</b>	<b>31</b>
<b>2.2</b>	<b>Lower bounds for the quadratic risk and efficiency</b>	<b>35</b>
<b>2.3</b>	<b>Upper bounds for the quadratic risk and efficiency (when <math>\delta &gt; 0</math>)</b>	<b>38</b>
<b>2.4</b>	<b>Simulations</b>	<b>43</b>
<b>2.5</b>	<b>Proofs of main results</b>	<b>46</b>
<b>2.6</b>	<b>Proofs of technical lemmas</b>	<b>60</b>

---

## 2.1 Introduction

The problem of estimating the proportion  $\theta$  of true null hypotheses is of interest in situation where several thousands of (independent) hypotheses can be tested simultaneously. One of the typical applications in which multiple testing problems occur is estimating the proportion of



genes that are not differentially expressed in deoxyribonucleic acid (DNA) microarray experiments [see for instance [Dudoit and van der Laan, 2008](#)]. Among other application domains, we mention astrophysics [[Meinshausen and Rice, 2006](#)] or neuroimaging [[Turkheimer et al., 2001](#)]. A reliable estimate of  $\theta$  is important when one wants to control multiple error rates, such as the false discovery rate (FDR) introduced by [Benjamini and Hochberg \[1995\]](#). In this work, we discuss asymptotic efficiency of estimators of the true proportion of null hypotheses. We stress that the asymptotic framework is particularly relevant in the above mentioned contexts where the number of tested hypotheses is huge.

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

$$\forall x \in [0, 1], \quad g(x) = \theta + (1 - \theta)f(x), \quad (2.1)$$

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).

Many different identifiability conditions on the parameter  $(\theta, f)$  in model (2.1) have been discussed in the literature. For example, [Genovese and Wasserman \[2004\]](#) introduce the concept of purity that corresponds to the case where the essential infimum of  $f$  on  $[0, 1]$  is zero. They prove that purity implies identifiability but not *vice versa*. [Langaas et al. \[2005\]](#) suppose that  $f$  is decreasing with  $f(1) = 0$  while [Neuvial \[2010\]](#) assumes that  $f$  is regular near  $x = 1$  with  $f(1) = 0$  and [Celisse and Robin \[2010\]](#) consider 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 2.5.1.

**Proposition 2.1.** *The parameter  $(\theta, f)$  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 to 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 and Wasserman \[2004\]](#). However, from an estimation perspective, the purity assumption is very weak and it is hopeless to obtain a reliable estimate of  $\theta$  based on the value of  $f$  at a unique value (or at a finite number of values). Since 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  $\delta$  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 and Robin \[2010\]](#). In the following, we explore asymptotic efficiency results for the estimation of  $\theta$  by assuming that the function  $f$  belongs to a set of densities (with respect to the Lebesgue measure  $\mu$ ) defined as

$$\begin{aligned} \mathcal{F}_\delta = \{f : [0, 1] \mapsto \mathbb{R}^+, \text{ continuously non increasing density, positive on } [0, 1 - \delta] \\ \text{and such that } f|_{[1-\delta, 1]} = 0\}. \end{aligned} \quad (2.2)$$

We establish that two different cases are to be distinguished:  $\delta$  is positive and  $\delta$  is equal to zero. In the first case, we obtain the existence of  $\sqrt{n}$ -consistent estimators of  $\theta$  that is to say estimators  $\hat{\theta}_n$  such that  $\sqrt{n}(\hat{\theta}_n - \theta)$  is bounded in probability (denoted by  $\sqrt{n}(\hat{\theta}_n - \theta) = O_{\mathbb{P}}(1)$ ). We exhibit such estimators and also compute the asymptotic optimal variance for this problem. Moreover, we conjecture that asymptotically efficient estimators (that is estimators asymptotically attaining this variance lower bound) do not exist. In the second case, while the existence of an estimator  $\hat{\theta}_n$  of  $\theta$  converging at parametric rate has not been established yet, we prove that if such a  $\sqrt{n}$ -consistent estimator of  $\theta$  exists, then the variance  $\text{Var}(\sqrt{n}\hat{\theta}_n)$  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.

Let us now discuss the different estimators of  $\theta$  proposed in the literature, starting with those assuming (implicitly or not) that  $f$  attains its minimum value on a whole interval. First, [Schweder and Spjøtvoll \[1982\]](#) suggested a procedure to estimate  $\theta$ , that 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  $\lambda$  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 and Spjøtvoll \[1982\]](#) nor

the one by Storey [2002]). Note that even if such an assumption were made, it would not solve the problem of choosing  $\lambda$  such that  $f$  attains its infimum on  $[\lambda, 1]$ . Adapting this procedure in order to end up with an estimate of the positive FDR (pFDR), Storey [2002] proposes a bootstrap strategy to pick  $\lambda$ . More precisely, his procedure minimizes the mean squared error for estimating the pFDR. Note that Genovese and Wasserman [2004] established that, for fixed value  $\lambda$  such that the cumulative distribution function (cdf)  $F$  of  $f$  satisfies  $F(\lambda) < 1$ , Storey's estimator converges at parametric rate and is asymptotically normal, but is also asymptotically biased: thus it does not converge to  $\theta$  at parametric rate. Some other choices of  $\lambda$  are, for instance, based on break point estimation [Turkheimer et al., 2001] or spline smoothing [Storey, 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 and 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  $\theta$  are constructed using nonparametric estimates  $\hat{g}$  of the density  $g$  and then estimate  $\theta$  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  $\theta$ . 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 were proven to be convergent to  $\theta$  at parametric rate. In Theorem 2.2 below, we prove that a very simple histogram based estimator possesses this property, while in Theorem 2.3, we establish that this is also true for the more elaborate procedure proposed by Celisse and 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  $\theta$  that asymptotically attains the optimal variance in this context. Besides, one might conjecture that such a procedure does not exist for regular models (see Section 2.3.3).

Other estimators of  $\theta$  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 nonparametric estimates of  $g$  and appear to inherit nonparametric rates of convergence. Langaas et al. [2005] derive estimators based on nonparametric maximum likelihood estimation of the  $p$ -value density, in two setups: decreasing and convex decreasing densities  $f$ . We mention that no theoretical properties of these estimators are given. Hengartner and Stark [1995] propose a very general finite sample confidence envelope for a monotone density.

Relying on this result and assuming moreover that the cdf  $G$  of  $g$  is concave and that  $g$  is Lipschitz in a neighborhood of  $x = 1$ , [Genovese and Wasserman \[2004\]](#) construct an estimator converging to  $g(1) = \theta$  at rate  $(\log n)^{1/3}n^{-1/3}$ . Under some regularity assumptions on  $f$  near  $x = 1$ , [Neuivial \[2010\]](#) establishes that by letting  $\lambda \rightarrow 1$ , [Storey's](#) estimator may be turned into a consistent estimator of  $\theta$ , with a nonparametric rate of convergence equal to  $n^{-k/(2k+1)}\eta_n$ , where  $\eta_n \rightarrow +\infty$  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, as is expected from the fact that the quadratic risk of any estimator of  $\theta$  cannot converge at parametric rate in this case (see [Theorem 2.1](#)).

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

The article is organized as follows. [Section 2.2](#) establishes lower bounds on the quadratic risk for the estimation of  $\theta$ , while [Section 2.3](#) explores corresponding upper bounds, *i.e.* the existence of  $\sqrt{n}$ -consistent estimators of  $\theta$  and the existence of asymptotically efficient estimators. [Section 2.4](#) illustrates our results relying on simulations. The proofs of the main results are postponed to [Section 2.5](#), while some technical lemmas are proved in [Section 2.6](#).

## 2.2 Lower bounds for the quadratic risk and efficiency

In this section, we give lower bounds for the quadratic risk of any estimator of  $\theta$ . 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 [\(2.2\)](#). Note that for any fixed value  $\delta \in [0, 1)$ , the condition stated in [Proposition 2.1](#) 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  $(\theta, f)$  is identifiable on  $(0, 1) \times \mathcal{F}_\delta$ .

We follow notation from 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 2.5.2](#) 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  $(\theta, f)$ . 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  $\theta$  (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

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

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

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

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  $L = N(0, \mathbb{P}_{\theta,f}(\tilde{\psi}_{\theta,f}^2))$ , 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

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

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

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

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

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

**Theorem 2.1.** *1) When  $\delta = 0$ , there is no regular estimator for  $\theta$  relative to the tangent set  $\dot{\mathcal{P}}_0$  and any estimator sequence  $\hat{\theta}_n$  has an infinite quadratic risk, namely*

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

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  $\hat{\theta}_n$ ,

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

where the first supremum is taken over all finite subsets  $E_\delta$  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

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{\delta} \mathbf{1}_{X_i \in [1-\delta,1]} + o_{\mathbb{P}_{\theta,f}}(n^{-1/2}). \quad (2.3)$$

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 (2.2) 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  $\sqrt{n}(\hat{\theta}_n - \theta) = O_{\mathbb{P}}(1)$  for some estimator  $\hat{\theta}_n$  but then  $\mathbb{V}\text{ar}(\sqrt{n}\hat{\theta}_n) \rightarrow +\infty$ . However, we note that the only rates of convergence obtained until now in this case are nonparametric ones.

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

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

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

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

In this sense, the oracle version of Storey's estimator that picks  $\lambda = 1 - \delta$  (namely choosing  $\lambda$  as the smallest value such that  $f$  vanishes on  $[\lambda, 1]$ ) is asymptotically efficient. Note also that  $\hat{\theta}^{\text{Storey}}(\lambda)$  automatically satisfies (2.3).

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

In this section, we investigate the existence of asymptotically efficient estimators for  $\theta$ , in the case where  $\delta > 0$ . We consider histogram based estimators of  $\theta$  where a nonparametric histogram estimator  $\hat{g}$  of  $g$  is combined with an interval selection that aims at picking an interval where  $g$  is equal to  $\theta$ . We start by establishing the existence of  $\sqrt{n}$ -consistent estimators: a simple histogram based procedure is studied in Section 2.3.1 while a more elaborate one is the object of Section 2.3.2. Finally in Section 2.3.3 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 generalised to the case where  $f$  vanishes on a non-empty interval included in  $[0, 1]$  and is monotone outside this interval.

### 2.3.1 A histogram based estimator

Let  $\hat{g}_I$  be a histogram estimator corresponding to a partition  $I = (I_k)_{1, \dots, D}$  of  $[0, 1]$ , defined by

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

where  $n_k = \text{card}\{i : X_i \in I_k\}$  is the number of observations in  $I_k$  and  $|I_k|$  is the width of interval  $I_k$ . We estimate  $\theta$  by the minimal value of  $\hat{g}_I$ , that is

$$\hat{\theta}_{I,n} = \min_{1 \leq k \leq D} \frac{n_k}{n|I_k|} = \frac{n_{\hat{k}_n}}{n|I_{\hat{k}_n}|}, \quad (2.4)$$

where we let

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

Note that histogram estimators are natural nonparametric 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 mean squared error of the order  $1/n$ . The proof of this result appears in Section 2.5.3.

### 2.3. UPPER BOUNDS FOR THE QUADRATIC RISK AND EFFICIENCY (WHEN $\delta > 0$ )

---

**Theorem 2.2.** 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  $\theta$ ,
- ii)  $\limsup_{n \rightarrow \infty} n\mathbb{E}[(\hat{\theta}_{I,n} - \theta)^2] < +\infty$ .

Note that since  $\hat{\theta}_{I,n}$  has the mean squared error 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, an 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  $n_{\hat{I}}/(n|\hat{I}|)$ . 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.

#### 2.3.2 Celisse and Robin [2010]'s procedure

We recall here the procedure for estimating  $\theta$  that is presented in Celisse and 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 2.4), 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 since not many estimators of  $\theta$  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

$$\mathcal{I} = \bigcup_{m_{min} \leq m \leq m_{max}} \mathcal{I}_{2^m}. \quad (2.5)$$



### 2.3. UPPER BOUNDS FOR THE QUADRATIC RISK AND EFFICIENCY (WHEN $\delta > 0$ )

---

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*  $\hat{g}_{I^*}$  is obtained through

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

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

$$\hat{R}_p(I) = \frac{2n-p}{(n-1)(n-p)} \sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n}\right)^2. \quad (2.6)$$

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

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

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

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

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

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

**Remark 2.1.** *In our procedure, we consider the set of natural partitions defined by (2.5), while [Celisse and Robin \[2010\]](#) use the one defined by*

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

### 2.3. UPPER BOUNDS FOR THE QUADRATIC RISK AND EFFICIENCY (WHEN $\delta > 0$ )

---

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  $2 \leq k + 2 \leq l \leq M$ ,

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

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

In [Celisse and Robin \[2010\]](#), the authors only establish 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 and 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

$$\forall I \in \mathcal{I}, \quad 8s_{11}s_{21} - 2s_{11}^2 + 8s_{32} - 10s_{21}^2 - 4s_{22} \neq 0, \quad s_{21} - s_{22} - s_{32} + 3s_{11} \neq 0. \quad (2.7)$$

This technical condition is used in [Celisse and 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 2.5.4](#).

**Theorem 2.3.** *Suppose that  $f$  satisfies the technical condition (2.7) 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  $\theta$ ,
- ii)  $\hat{\theta}_n^{CR}$  is  $\sqrt{n}$ -consistent, i.e.  $\sqrt{n}(\hat{\theta}_n^{CR} - \theta) = O_{\mathbb{P}}(1)$ ,
- iii) If  $p$  is fixed then  $\limsup_{n \rightarrow \infty} n\mathbb{E}[(\hat{\theta}_n^{CR} - \theta)^2] < +\infty$ .

Here again, asymptotic normality of  $\hat{\theta}_n^{CR}$  or the exact value of its asymptotic variance are 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.

#### 2.3.3 One-step estimators

In this section, we introduce the one-step method to construct an asymptotically efficient estimator, relying on a  $\sqrt{n}$ -consistent one [see [van der Vaart, 1998](#), Section 25.8]. Here again,

### 2.3. UPPER BOUNDS FOR THE QUADRATIC RISK AND EFFICIENCY (WHEN $\delta > 0$ )

---

we use terminology from semiparametric theory. Let  $\hat{\theta}_n$  be a  $\sqrt{n}$ -consistent estimator of  $\theta$ , then  $\hat{\theta}_n$  can be discretized on grids of mesh width  $n^{-1/2}$ . Suppose that we are given a sequence of estimators  $\hat{l}_{n,\theta}(\cdot) = \hat{l}_{n,\theta}(\cdot; X_1, \dots, X_n)$  of the efficient score function  $\tilde{l}_{\theta,f}$  (an expression of the efficient score function in our context is given in Section 2.5.2). Define with  $m = \lfloor n/2 \rfloor$ ,

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

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

$$\sqrt{n} \mathbb{P}_{\theta_n, f} \hat{l}_{n,\theta_n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\theta, f}} 0, \quad (2.8)$$

$$\mathbb{P}_{\theta_n, f} \|\hat{l}_{n,\theta_n} - \tilde{l}_{\theta_n, f}\|^2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}_{\theta, f}} 0, \quad (2.9)$$

$$\int \|\tilde{l}_{\theta_n, f} d\mathbb{P}_{\theta_n, f}^{1/2} - \tilde{l}_{\theta, f} d\mathbb{P}_{\theta, f}^{1/2}\|^2 \xrightarrow[n \rightarrow \infty]{} 0. \quad (2.10)$$

Note that in the above 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 above assumptions, the one-step estimator defined as

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

is asymptotically efficient at  $(\theta, f)$  [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  $\theta$ , 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  $\theta$  and its implications in model  $\mathcal{P}_\delta$ .

Under condition (2.10), it is shown in Theorem 7.4 from van der Vaart [2002] that the existence of an asymptotically efficient sequence of estimators of  $\theta$  implies the existence of a sequence of estimators  $\hat{l}_{n,\theta}$  of  $\tilde{l}_{\theta, f}$  satisfying (2.8) and (2.9). 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 (2.10) holds. Then, the estimator  $\hat{l}_{n,\theta}$  of the efficient score function  $\tilde{l}_{\theta, f}$  must satisfy both a "no-bias" (2.8) and a consistency (2.9) 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 2.5.3.

**Proposition 2.2.** *The existence of an asymptotically efficient sequence of estimators of  $\theta$  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 (2.8) and (2.9). 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  $\delta$ , then this condition is equivalent to  $\sqrt{n}(\hat{\delta}_n - \delta) = o_{\mathbb{P}}(1)$ .*

Let us now explain the consequences of this result. The proposition states that efficient estimators of  $\theta$  exist if and only if estimators of  $\tilde{l}_{\theta,f}$  that satisfy (2.8) and (2.9) can be constructed. As there is no general method to estimate an efficient score function, such an estimator should rely on the specific expression (2.15). Though 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 (2.15), it is likely to be the case. Then, existence of efficient estimators of  $\theta$  is equivalent to existence of estimators of  $\delta$  that converge at faster than parametric rate. Note that this is possible for irregular models [see Chapter 6 in [Ibragimov and Hasminskii, 1981](#), for more details]. However, for regular models, such estimators cannot be constructed and one might conjecture that efficient estimators of  $\theta$  do not exist in regular models.

## 2.4 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  $\theta$  previously discussed. We choose to compare three different estimators: the histogram based estimator  $\hat{\theta}_{I,n}$  defined in Section 2.3.1 through (2.4), the more elaborate histogram based estimator  $\hat{\theta}_n^{CR}$  proposed in [Celisse and 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  $\theta$  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 and 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 labeled

## 2.4. SIMULATIONS

---

as described in Table 2.1, 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 2.1. For each estimator  $\hat{\theta}_n$  of  $\theta$ , we compare the quantity  $n\mathbb{E}[(\hat{\theta}_n - \theta)^2]$  with the optimal variance  $\theta(\delta^{-1} - \theta)$  when this bound exists. Equivalently, we compare the logarithm of mean squared error,  $\log(\text{MSE}) = \log\mathbb{E}[(\hat{\theta}_n - \theta)^2]$  for each estimator  $\hat{\theta}_n$  with  $-\log(n) + \log[\theta(\delta^{-1} - \theta)]$ . When  $\delta = 0$ , we only compare the slope of the line induced by  $\log(\text{MSE})$  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.

$(s, \theta)$	$\delta = 0.3$	$\delta = 0$
(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)	( $d_1$ )	( $d_2$ )

Table 2.1: Labels of the 8 models with different parameter values.

The results are presented in Figure 2.2 for the case  $\delta > 0$  and Figure 2.3 for the case  $\delta = 0$ . First, we note that in both cases ( $\delta > 0$  and  $\delta = 0$ ), Langaas et al.'s estimator  $\hat{\theta}_n^L$  has nonparametric rate of convergence (null slope) and performs badly compared to  $\hat{\theta}_{I,n}$  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 behaviors 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 are thus robust to the assumption that  $f$  vanishes on an interval in the convex setup. The results

## 2.4. SIMULATIONS

---

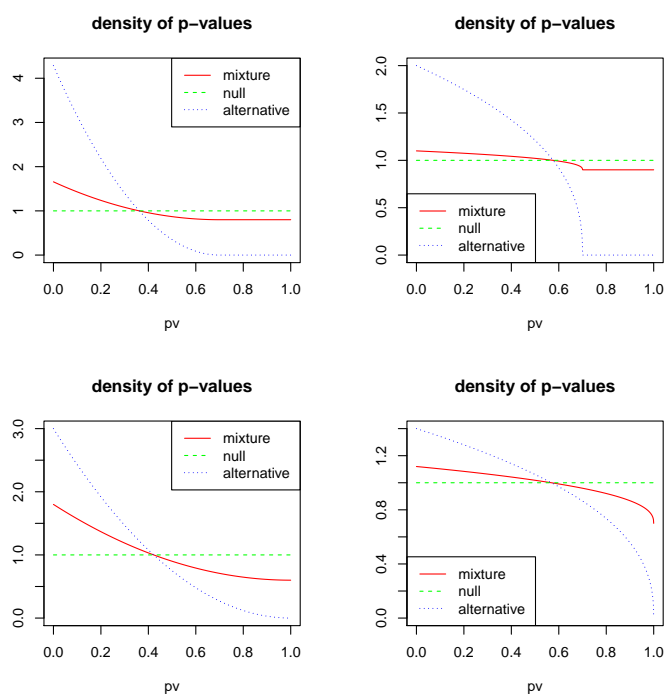


Figure 2.1: Density function of the  $p$ -values. Top left: model ( $b_1$ ); top right: model ( $d_1$ ); bottom left: model ( $a_2$ ); bottom right: model ( $c_2$ ).

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 nonparametric 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 parametric rate of convergence and almost optimal asymptotic variance.

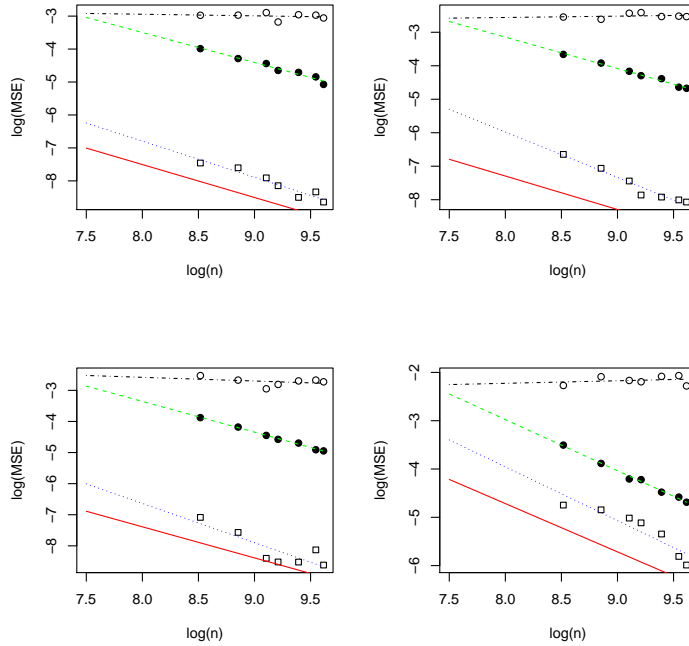


Figure 2.2: Logarithm of the mean squared error as a function of  $\log(n)$  and corresponding linear regression for  $\hat{\theta}_n^L$  ( $\circ$  and black line, respectively),  $\hat{\theta}_n^{CR}$  ( $\square$  and blue line, respectively) and  $\hat{\theta}_{I,n}$  ( $\bullet$  and green line, 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  $y = -\log(n) + \log[\theta(\delta^{-1} - \theta)]$ .

## 2.5 Proofs of main results

### 2.5.1 Proof of Proposition 2.1

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  $\theta$  and  $f$  are identifiable on the set  $(0, 1) \times \mathcal{F}$  by contradiction.

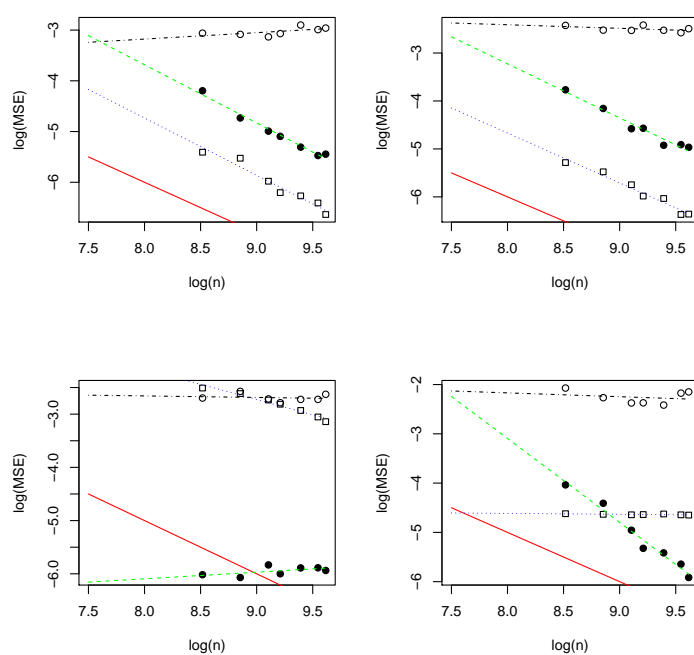


Figure 2.3: Logarithm of the mean squared error as a function of  $\log(n)$  and corresponding linear regression for  $\hat{\theta}_n^L$  ( $\circ$  and black line, respectively),  $\hat{\theta}_n^{CR}$  ( $\square$  and blue line, respectively) and  $\hat{\theta}_{I,n}$  ( $\bullet$  and green line, 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  $c$ .



Suppose that there exist  $(\theta_1, f_1)$  and  $(\theta_2, f_2) \in \mathcal{F}$ ,  $(\theta_1, f_1) \neq (\theta_2, f_2)$  such that

$$\theta_1 + (1 - \theta_1)f_1(x) = \theta_2 + (1 - \theta_2)f_2(x), \text{ for all } x \in [0, 1]. \quad (2.11)$$

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

$$\theta_1 + (1 - \theta_1)f_1(x) = \theta_2 + (1 - \theta_2)(c + (1 - c)f_1(x)), \text{ for all } x \in [0, 1]. \quad (2.12)$$

From (2.11) and (2.12), 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.

Necessity: Suppose that the parameters  $\theta$  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  $\theta$  and  $f$  are not identifiable on the set  $(0, 1) \times \mathcal{F}$ .

### 2.5.2 Proof of Theorem 2.1

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

$$t \mapsto f_t(x) = c(t)k(th_0(x))f(x), \quad (2.13)$$

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 normalising 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 (2.13) 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

$${}_f\dot{\mathcal{P}}_\delta = \left\{ h = \frac{(1 - \theta)fh_0}{\theta + (1 - \theta)f}; h_0 \text{ is continuous and non increasing on } [0, 1 - \delta) \text{ with } \int fh_0 = 0 \right\}.$$

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 (2.13). We remark that if  $\dot{l}_{\theta, f}$  is the ordinary score function for

## 2.5. PROOFS OF MAIN RESULTS

---

$\theta$  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  $a\dot{l}_{\theta,f} + h$  is a score function for  $(\theta, f)$  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  $(\theta, f)$  is given by the linear span

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

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

$$\dot{l}_{\theta,f}(x) = \frac{\partial}{\partial \theta} \log[\theta + (1 - \theta)f(x)] = \frac{1 - f(x)}{\theta + (1 - \theta)f(x)}. \quad (2.14)$$

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

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

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

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

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  $\tilde{I}_{\theta,f}^{-1}\tilde{l}_{\theta,f}$  [see Lemma 25.25 in [van der Vaart, 1998](#)]. The following proposition provides expressions for these quantities in our setup.

**Proposition 2.3.** *The efficient score function  $\tilde{l}_{\theta,f}$  and the efficient information  $\tilde{I}_{\theta,f}$  for estimating  $\theta$  in model  $\mathcal{P}_\delta$  are given by*

$$\tilde{l}_{\theta,f}(x) = \frac{1}{\theta} - \frac{1}{\theta(1 - \theta\delta)}\mathbf{1}_{[0,1-\delta)}(x) \quad \text{and} \quad \tilde{I}_{\theta,f} = \frac{\delta}{\theta(1 - \theta\delta)}, \quad (2.15)$$

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 2.3.* The ordinary score function  $\dot{l}_{\theta,f}$  can be written as

$$\begin{aligned} \dot{l}_{\theta,f}(x) &= \frac{\partial}{\partial \theta} \log[\theta + (1-\theta)f(x)] \\ &= \left( \frac{1-f(x)}{\theta + (1-\theta)f(x)} + \frac{\delta}{1-\theta\delta} \right) \mathbf{1}_{[0,1-\delta)}(x) + \frac{1}{\theta} \mathbf{1}_{[1-\delta,1]}(x) - \frac{\delta}{1-\theta\delta} \mathbf{1}_{[0,1-\delta)}(x). \end{aligned} \quad (2.16)$$

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 (2.16), namely

$$\Pi_{\theta,f} \dot{l}_{\theta,f}(x) = \left( \frac{1-f(x)}{\theta + (1-\theta)f(x)} + \frac{\delta}{1-\theta\delta} \right) \mathbf{1}_{[0,1-\delta)}(x), \quad (2.17)$$

and then the efficient score function for  $\theta$  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

$$\begin{aligned} h_0(x) &= -\left( \frac{1-f(x)}{(1-\theta)f(x)} + \frac{\delta}{1-\theta\delta} \times \frac{\theta + (1-\theta)f(x)}{(1-\theta)f(x)} \right) \mathbf{1}_{[0,1-\delta)}(x) \\ &= \frac{1}{(1-\theta)(1-\theta\delta)} \left( 1 - \delta - \frac{1}{f(x)} \right) \mathbf{1}_{[0,1-\delta)}(x). \end{aligned}$$

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

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

Hence

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

Now, to conclude the proof of (2.17), it is necessary to establish that the second term in the right hand side of (2.16) is orthogonal to the closure of the linear span of  ${}_f\dot{\mathcal{P}}_\delta$ , namely

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

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

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

the scalar product between  $h$  and the remaining term in (2.16) is given by

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

This establishes (2.17). Let us now calculate the efficient information

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

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

$$\begin{aligned} \tilde{\psi}_{\theta,f}(x) &= \tilde{I}_{\theta,f}^{-1} \tilde{l}_{\theta,f}(x) \\ &= \frac{\theta(1-\theta\delta)}{\delta} \left( \frac{1}{\theta} \mathbf{1}_{[1-\delta,1]}(x) - \frac{\delta}{1-\theta\delta} \mathbf{1}_{[0,1-\delta)}(x) \right) \\ &= \frac{1-\theta\delta}{\delta} \mathbf{1}_{[1-\delta,1]}(x) - \theta \mathbf{1}_{[0,1-\delta)}(x) \\ &= \frac{1}{\delta} \mathbf{1}_{[1-\delta,1]}(x) - \theta, \end{aligned}$$

which concludes the proof. □

We are now ready to conclude the proof of Theorem 2.1.

*Proof of Theorem 2.1.* 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  $\theta$  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  $\dot{l}_{\theta,f} \notin {}_f\dot{\mathcal{P}}_{0,m} := \text{lin}(h_1, h_2, \dots, h_m)$ . We thus have

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

where  $E_0$  and  $F_0$  range through all finite subsets of the tangent sets  $\dot{\mathcal{P}}_0 = \text{lin}(\dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_0) = {}_f\dot{\mathcal{P}}_0$  and  $\text{lin}(\dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_{0,m}) = {}_f\dot{\mathcal{P}}_{0,m}$ , respectively. The efficient score function for  $\theta$  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  $\tilde{I}_{\theta,f,m} = \mathbb{P}_{\theta,f}(\tilde{l}_{\theta,f,m}^2)$  is non zero. Using Lemma 25.25 from [van der Vaart \[1998\]](#), the efficient influence function relative to the tangent set  $\text{lin}(\dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_{0,m})$  is  $\tilde{\psi}_{\theta,f,m} = \tilde{I}_{\theta,f,m}^{-1} \tilde{l}_{\theta,f,m}$ . So we can apply Theorem 25.21 from [van der Vaart \[1998\]](#) to obtain that

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

Since  $\tilde{I}_{\theta,f,m} \xrightarrow{m \rightarrow \infty} \tilde{I}_{\theta,f} = 0$ , we obtain the result. The second part of the proof concerning  $\delta > 0$  is an immediate consequence of Proposition 2.3 together with Theorem 25.21 and Lemma 25.23 in [van der Vaart \[1998\]](#).  $\square$

### 2.5.3 Proofs from Sections 2.3.1 and 2.3.3

*Proof of Theorem 2.2.* Let us denote by  $\mathcal{D} = \{1, 2, \dots, D\}$ ,  $\mathcal{D}_0 = \{k \in \mathcal{D} \text{ such that } I_k \subseteq [1 - \delta, 1]\}$  and  $\mathcal{D}_1 = \mathcal{D} \setminus \mathcal{D}_0 = \{k \in \mathcal{D} \text{ such that } I_k \not\subseteq [1 - \delta, 1]\}$ . 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  $\theta$ . Indeed, we can write that

$$\hat{\theta}_{I,n} = \theta + \sum_{k \in \mathcal{D}_0} \left( \frac{n_k}{n|I_k|} - \theta \right) \mathbf{1}\{\hat{k}_n = k\} + (\hat{\theta}_{I,n} - \theta) \mathbf{1}\{I_{\hat{k}_n} \not\subseteq [1 - \delta, 1]\}, \quad (2.18)$$

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 convergences

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

## 2.5. PROOFS OF MAIN RESULTS

---

As a consequence, we obtain that the second term in the right-hand side of (2.18) converges almost surely to zero, namely

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

The third term in the right-hand side of (2.18) also converges almost surely to zero. Indeed, we have

$$|\hat{\theta}_{I,n} - \theta| \mathbf{1}\{I_{\hat{k}_n} \notin [1 - \delta, 1]\} \leq \left( \max_{1 \leq k \leq D} \frac{1}{|I_k|} - \theta \right) \sum_{k \in \mathcal{D}_1} \mathbf{1}\{\hat{k}_n = k\}.$$

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

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

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

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

we obtain that

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

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

$$\mathbb{E}[(\sqrt{n}(\hat{\theta}_{I,n} - \theta))^2] = \sum_{k \in \mathcal{D}_0} \mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|} - \theta\right)\right)^2 \mathbf{1}_{\hat{k}_n = k}\right] + \sum_{k \in \mathcal{D}_1} \mathbb{E}\left[\left(\sqrt{n}\left(\frac{n_k}{n|I_k|} - \theta\right)\right)^2 \mathbf{1}_{\hat{k}_n = k}\right]. \quad (2.19)$$

The second term in the right-hand side of (2.19) is bounded by

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

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

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

For the first term in the right-hand side of (2.19), we apply Cauchy-Schwarz's inequality

$$\begin{aligned} \sum_{k \in \mathcal{D}_0} \mathbb{E} \left[ \left( \sqrt{n} \left( \frac{n_k}{n|I_k|} - \theta \right) \right)^2 \mathbf{1}_{\hat{k}_n = k} \right] &\leq \sqrt{\sum_{k \in \mathcal{D}_0} \mathbb{E} \left[ \left( \sqrt{n} \left( \frac{n_k}{n|I_k|} - \theta \right) \right)^4 \right]} \sqrt{\sum_{k \in \mathcal{D}_0} \mathbb{P}(\hat{k}_n = k)} \\ &\leq \sqrt{\sum_{k \in \mathcal{D}_0} \mathbb{E} \left[ \left( \sqrt{n} \left( \frac{n_k}{n|I_k|} - \theta \right) \right)^4 \right]}, \end{aligned} \quad (2.20)$$

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

$$\begin{aligned} \mathbb{E} \left[ \left( \sqrt{n} \left( \frac{n_k}{n|I_k|} - \theta \right) \right)^4 \right] &= \mathbb{E} \left[ \frac{1}{n^2} \left( \sum_{i=1}^n \left( \frac{1}{|I_k|} \mathbf{1}_{\{X_i \in I_k\}} - \theta \right) \right)^4 \right] \\ &= \frac{1}{n} \mathbb{E} \left[ \left( \frac{1}{|I_k|} \mathbf{1}_{\{X_1 \in I_k\}} - \theta \right)^4 \right] + \frac{n-1}{n} \mathbb{E}^2 \left[ \left( \frac{1}{|I_k|} \mathbf{1}_{\{X_1 \in I_k\}} - \theta \right)^2 \right] \\ &= \frac{\theta}{n} \left( \frac{1}{|I_k|^3} - \frac{4\theta}{|I_k|^2} + \frac{6\theta^2}{|I_k|} - 3\theta^3 \right) + \frac{n-1}{n} \sigma_k^4. \end{aligned} \quad (2.21)$$

Thus, we finally obtain that

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

□

*Proof of Proposition 2.2.* Let us first establish that condition (2.10) holds. In fact, with the notation  $p_{\theta,f} = \theta + (1-\theta)f$ , we have

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

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  $\theta$  is equivalent to the existence of a sequence of estimators  $\hat{l}_{n,\theta}$  satisfying (2.8) and (2.9).

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  $\delta$  in order to estimate  $\tilde{l}_{\theta,f}$ . Let  $\hat{\delta}_n$  be any given consistent (in probability) estimator of  $\delta$ . Let us examine condition (2.8) more closely. We have

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

Hence, the "no-bias" condition (2.8) is equivalent to the existence of an estimator  $\hat{\delta}_n$  of  $\delta$  that converges at a rate faster than  $1/\sqrt{n}$ , namely such that  $\sqrt{n}(\hat{\delta}_n - \delta) = o_{\mathbb{P}}(1)$ . With the same argument as in the previous calculation, the consistency condition (2.9) is satisfied as soon as the estimator  $\hat{\delta}_n$  converges in probability to  $\delta$ .  $\square$

#### 2.5.4 Proof of Theorem 2.3

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

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

We introduce three lemmas that are needed to prove Theorem 2.3. The proofs of these technical lemmas is further postponed to Section 2.6.



**Lemma 2.1.** *Let  $I = (I_k)_{k=1}^D$  be an arbitrary partition of  $[0, 1]$ . Then the variance term of the mean squared error of a histogram estimator  $\hat{g}_I$  is bounded by  $C/n$ , where  $C$  is a positive constant. In other words,*

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

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 mean squared error of a histogram estimator  $\hat{g}_I$  and its estimator.

**Lemma 2.2.** *Let  $I = (I_k)_{1, \dots, D}$  be an arbitrary partition of  $[0, 1]$ . Let  $p \in \{1, 2, \dots, n-1\}$  such that  $\lim_{n \rightarrow \infty} p/n < 1$ . Then we have the following results*

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

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

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  $I \not\trianglelefteq J$  otherwise.

**Lemma 2.3.** *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*

$$i) \text{ If } I \text{ is a subdivision of } I^{(N)}, \text{ then } L(I) = L(I^{(N)}).$$

$$ii) \text{ If } I \text{ is not a subdivision of } I^{(N)}, \text{ then } L(I) > L(I^{(N)}).$$

We are now ready to prove Theorem 2.3, starting by establishing point *i*). First, we remark that under condition (2.7), Celisse and Robin prove in their Proposition 2.1 that

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

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

$$\hat{\theta}_n^{CR} = \theta + \sum_{I=(N, \lambda) \trianglelefteq I^{(N)}} \left[ \frac{1}{n(1-\lambda)} \sum_{i=1}^n \mathbf{1}\{X_i \in [\lambda, 1]\} - \theta \right] \mathbf{1}\{\hat{\lambda} = \lambda\} + (\hat{\theta}_n^{CR} - \theta) \mathbf{1}_{\hat{\Lambda} \not\trianglelefteq I^{(N)}}, \quad (2.22)$$

where  $N = 2^{m_{max}}$  as in Lemma 2.3. For each partition  $I = (N, \lambda) \trianglelefteq I^{(N)}$ , we have  $[\lambda, 1] \subseteq \Lambda^*$ .

By applying the strong law of large numbers we get that

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

## 2.5. PROOFS OF MAIN RESULTS

---

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

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

Using Lemma 2.3, we have  $L(\hat{I}) > L(I^{(N)})$ . Let

$$\gamma = \min_{I \notin I^{(N)}} L(I) - L(I^{(N)}) > 0, \quad (2.23)$$

we obtain that

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

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

$$\begin{aligned} |\hat{\theta}_n^{CR} - \theta| \mathbf{1}_{\hat{I} \notin I^{(N)}} &\leq (N - \theta) \mathbf{1}\{\sup_{I \in \mathcal{I}} |\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}\} \\ &\leq (N - \theta) \sum_{I \in \mathcal{I}} \mathbf{1}\{|\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}\}. \end{aligned} \quad (2.24)$$

Since  $\forall I \in \mathcal{I}$ , we both have  $\hat{L}_{\hat{p}(I)} \xrightarrow[n \rightarrow \infty]{a.s.} L(I)$  and  $\hat{p}(I)/n \xrightarrow[n \rightarrow \infty]{a.s.} l_\infty(I) \in [0, 1)$  as well as the fact that  $\hat{R}_{\hat{p}}(I)$  (given by (2.6)) is a continuous function of  $p/n$ , we obtain  $\hat{L}_{\hat{p}(I)}(I) \xrightarrow[n \rightarrow \infty]{a.s.} L(I)$ .

Therefore,

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

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  $(\hat{\theta}_n^{CR} - \theta) \mathbf{1}_{\hat{I} \notin I^{(N)}} \xrightarrow{a.s.} 0$ . We finally get that  $\hat{\theta}_n^{CR} \xrightarrow{a.s.} \theta$ .

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

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

For each partition  $I = (N, \lambda) \trianglelefteq I^{(N)}$ , by applying the central limit theorem, we get that

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

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

$$\sum_{I=(N,\lambda)\leq I^{(N)}} \sqrt{n} \left[ \frac{1}{n(1-\lambda)} \sum_{i=1}^n \mathbf{1}_{X_i \in [\lambda, 1]} - \theta \right] \mathbf{1}_{\hat{\lambda}=\lambda} = O_{\mathbb{P}}(1). \quad (2.25)$$

We shall now prove that  $\sqrt{n}(\hat{\theta}_n^{CR} - \theta) \mathbf{1}_{\hat{I} \not\leq I^{(N)}} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0$ . In fact, according to (2.24), for all  $\epsilon > 0$ , we have

$$\begin{aligned} \mathbb{P}(\sqrt{n}|\hat{\theta}_n^{CR} - \theta| \mathbf{1}_{\hat{I} \not\leq I^{(N)}} > \epsilon) &\leq \mathbb{P}(\hat{I} \not\leq I^{(N)}) \\ &\leq \mathbb{P}(\sup_{I \in \mathcal{I}} |\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}) \\ &\leq \sum_{I \in \mathcal{I}} \mathbb{P}(|\hat{L}_{\hat{p}(I)}(I) - L(I)| \geq \frac{\gamma}{2}) \xrightarrow[n \rightarrow \infty]{} 0, \end{aligned}$$

where  $\gamma$  is defined by (2.23). Therefore,  $\sqrt{n}(\hat{\theta}_n^{CR} - \theta) \mathbf{1}_{\hat{I} \not\leq I^{(N)}} = o_{\mathbb{P}}(1)$ . We finally conclude that  $\sqrt{n}(\hat{\theta}_n^{CR} - \theta) = O_{\mathbb{P}}(1)$ .

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

$$\begin{aligned} \mathbb{E}[(\sqrt{n}(\hat{\theta}_n^{CR} - \theta))^2] &= \sum_{I=(N,\lambda)\leq I^{(N)}} \mathbb{E} \left[ \frac{1}{n} \left( \sum_{i=1}^n \left( \frac{1}{1-\lambda} \mathbf{1}_{\{X_i \in [\lambda, 1]\}} - \theta \right) \right)^2 \mathbf{1}_{\{\hat{\lambda}=\lambda\}} \right] \\ &\quad + \mathbb{E}[(\sqrt{n}(\hat{\theta}_n^{CR} - \theta))^2 \mathbf{1}_{\{\hat{I} \not\leq I^{(N)}\}}]. \end{aligned}$$

The first term of the above equation is bounded as in the proof of Proposition 2.2 (see inequalities (2.20) and (2.21))

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

The second term is bounded by

$$\begin{aligned} \mathbb{E}[(\sqrt{n}(\hat{\theta}_n^{CR} - \theta))^2 \mathbf{1}_{\{\hat{I} \not\leq I^{(N)}\}}] &\leq (N - \theta)^2 n \mathbb{P}(\hat{I} \not\leq I^{(N)}) \\ &\leq (N - \theta)^2 n \mathbb{P}(\sup_{I \in \mathcal{I}} |\hat{L}_p(I) - L(I)| \geq \frac{\gamma}{2}) \\ &\leq (N - \theta)^2 n \sum_{I \in \mathcal{I}} \mathbb{P}(|\hat{L}_p(I) - L(I)| \geq \frac{\gamma}{2}). \end{aligned}$$

For each partition  $I \in \mathcal{I}$ , according to the calculations in the proof of Lemma 2.1, we have

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

This leads to

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

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

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

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

## 2.6 Proofs of technical lemmas

### 2.6.1 Proof of Lemma 2.1

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

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

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

$$\begin{aligned} \|g - g_I\|_2^2 &= \inf_{(a_k)_k \in \mathbb{R}} \left[ \|g\|_2^2 - 2 \int_0^1 \left( \sum_k a_k \mathbf{1}_{I_k}(x) \right) g(x) dx + \int_0^1 \left( \sum_k a_k \mathbf{1}_{I_k}(x) \right)^2 dx \right] \\ &= \inf_{(a_k)_k \in \mathbb{R}} \left[ \|g\|_2^2 - 2 \sum_k a_k \alpha_k + \sum_k a_k^2 |I_k| \right] \\ &= \|g\|_2^2 - \sum_k \frac{\alpha_k^2}{|I_k|} = \|g\|_2^2 - s_{21}. \end{aligned} \tag{2.26}$$

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

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

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

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

Therefore,

$$\begin{aligned} \mathbb{E}[\|g - \hat{g}_I\|_2^2] &= \|g\|_2^2 + \sum_k \frac{n^2\alpha_k^2 + n\alpha_k(1 - \alpha_k)}{n^2 |I_k|} - 2 \sum_k \frac{n\alpha_k^2}{n |I_k|} \\ &= \|g\|_2^2 - s_{21} + \frac{1}{n}(s_{11} - s_{21}). \end{aligned} \tag{2.27}$$

Using (2.26) and (2.27), we obtain the desired result, namely

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

**2.6.2 Proof of Lemma 2.2**

i) Since

$$\lim_{n \rightarrow \infty} \frac{p}{n} < 1 \text{ and } \frac{n_k}{n} \xrightarrow[n \rightarrow \infty]{a.s.} \alpha_k, \text{ for all } k,$$

we obtain that

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

ii) By definition of  $R(I)$  and using (2.27), we have

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

This gives that

$$\begin{aligned} \sqrt{n}[\hat{R}_p(I) - R(I)] &= \sqrt{n} \left[ \frac{2n-p}{(n-1)(n-p)} \sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n}\right)^2 \right. \\ &\quad \left. + s_{21} - \frac{1}{n}(s_{11} - s_{21}) \right] \\ &= \frac{2n-p}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} [\sqrt{n}(\frac{n_k}{n} - \alpha_k)] + \frac{(2n-p)\sqrt{n}}{(n-1)(n-p)} s_{11} \\ &\quad - \frac{n(n-p+1)}{\sqrt{n}(n-1)(n-p)} \sum_k \frac{1}{|I_k|} [\sqrt{n}(\frac{n_k}{n} - \alpha_k)]^2 - \frac{(2n-p)\sqrt{n}}{(n-1)(n-p)} s_{21} \\ &\quad - \frac{2n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{\alpha_k}{|I_k|} [\sqrt{n}(\frac{n_k}{n} - \alpha_k)] - \frac{1}{\sqrt{n}}(s_{11} - s_{21}) \\ &= T_1 - \frac{2n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{\alpha_k}{|I_k|} [\sqrt{n}(\frac{n_k}{n} - \alpha_k)]. \end{aligned} \tag{2.28}$$

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

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

It thus follows that  $T_1 = o_{\mathbb{P}}(1)$ . We now consider the remaining term in (2.28). We have

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

Let us denote

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

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

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

By the central limit theorem, we obtain

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

Combining this with (2.28) implies that

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

It is easy to calculate that

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

Hence, we have

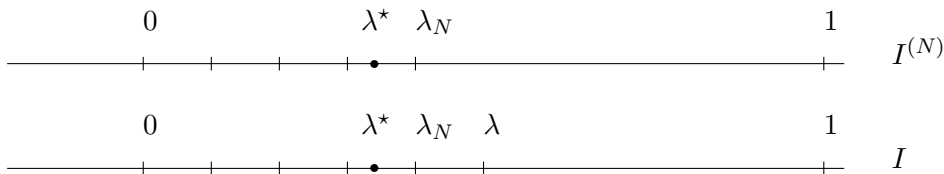
$$\sqrt{n}[\hat{L}_p(I) - L(I)] \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 4\sigma_I^2),$$

which completes the proof.

### 2.6.3 Proof of Lemma 2.3

i) Let us denote by  $\lambda^* = 1 - \delta$ . If  $I$  is a subdivision of  $I^{(N)}$ , then  $I = (N, \lambda)$  with  $[\lambda, 1] \subset [\lambda^*, 1]$ .

For example, we may have the following situation



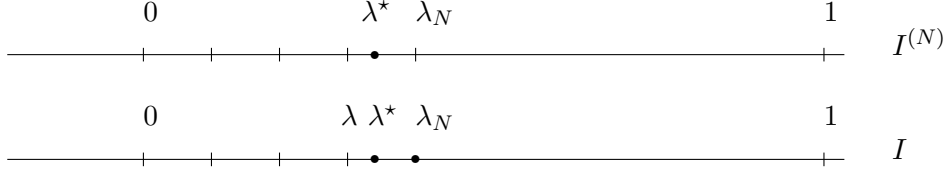
Since  $g$  is constant on the interval  $[\lambda^*, 1] \supset [\lambda_N, 1] \supset [\lambda, 1]$ , we have  $g_I = g_{I^{(N)}} = g$  on the interval  $[\lambda_N, 1]$ . This implies that  $\|g_I - g\|_2^2 = \|g_{I^{(N)}} - g\|_2^2$ .

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  $[\lambda, 1] \not\subset [\lambda_N, 1]$ . For example, we may have

## 2.6. PROOFS OF TECHNICAL LEMMAS

---



Since  $g_{I^{(N)}} = g$  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

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

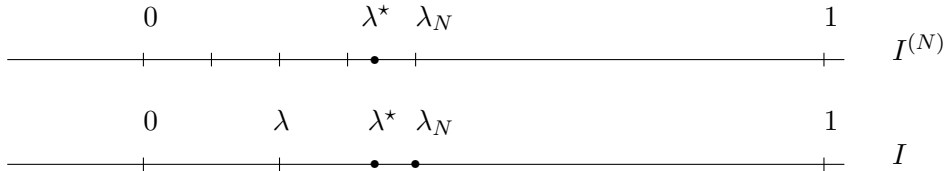
and

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

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 get 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.



# Estimation of the density of the alternative

## 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  $\theta$  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.

## Contents

<b>3.1</b>	<b>Introduction</b>	<b>65</b>
<b>3.2</b>	<b>Algorithmic procedures to estimate the density <math>f</math></b>	<b>69</b>
<b>3.3</b>	<b>Mathematical properties of the algorithms</b>	<b>73</b>
<b>3.4</b>	<b>Estimation of local false discovery rate and simulation study</b>	<b>79</b>
<b>3.5</b>	<b>Proofs of main results</b>	<b>85</b>
<b>3.6</b>	<b>Proofs of technical lemmas</b>	<b>95</b>

### 3.1 Introduction

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

$$\forall x \in \mathbb{R}^d, \quad g(x) = \theta\phi(x) + (1 - \theta)f(x), \quad (3.1)$$

where  $\theta$  is the unknown proportion of true null hypotheses,  $\phi$  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  $X_1, X_2, \dots, X_n$  associated with  $n$  independent tested hypotheses, then the density function  $\phi$  is the uniform distribution on  $[0, 1]$  while the density function  $f$  is assumed unknown. The parameters of the model are  $(\theta, f)$ , where  $\theta$  is a Euclidean parameter while  $f$  is an infinite-dimensional one and the model becomes

$$\forall x \in [0, 1], \quad g(x) = \theta + (1 - \theta)f(x). \quad (3.2)$$

In the following, we focus on model (3.2) that is slightly simpler than (3.1). 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 3.1.

Table 3.1: Possible outcomes from testing  $n$  hypotheses  $H_1, \dots, H_n$ .

	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,

$$\text{FDR} = \mathbb{E}\left[\frac{\text{FP}}{\max(\text{P}, 1)}\right] = \mathbb{E}\left[\frac{\text{FP}}{\text{P}} \mid \text{P} > 0\right] \mathbb{P}(\text{P} > 0).$$

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

$$\text{pFDR} = \mathbb{E}\left[\frac{\text{FP}}{\text{P}} \mid \text{P} > 0\right],$$

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

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

where  $\Phi$  and  $F$  are the cumulative distribution functions (cdfs) for densities  $\phi$  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 ( $\ell\text{FDR}$ ) to quantify the plausibility of a particular hypothesis being true, given its specific test statistic or  $p$ -value. In a mixture framework, the  $\ell\text{FDR}$  is the Bayes posterior probability

$$\ell\text{FDR}(x) = \mathbb{P}(H_i \text{ being true} \mid X = x) = 1 - \frac{(1 - \theta)f(x)}{\theta\phi(x) + (1 - \theta)f(x)}. \quad (3.3)$$

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  $\ell\text{FDR}$ . Moreover, the quantities pFDR and  $\ell\text{FDR}$  are analytically related by  $\text{pFDR}(x) = \mathbb{E}[\ell\text{FDR}(X) \mid X \leq x]$ . As a consequence (and recalling that the difference between pFDR and FDR is generally small), [Robin et al. \[2007\]](#) propose to estimate FDR by

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

where  $\widehat{\ell\text{FDR}}$  is an estimator of  $\ell\text{FDR}$  and the observations  $\{x_i\}$  are increasingly ordered. A natural strategy to estimate  $\ell\text{FDR}$  is to start by estimating both the proportion  $\theta$  and either  $f$  or  $g$ . Another motivation for estimating the parameters in this mixture model comes from the works of [Sun and Cai \(2009, 2007\)](#), who develop adaptive compound decision rules for false discovery rate control. These rules are based on the estimation of the parameters in model (3.1) (dealing with  $z$ -scores) rather than model (3.2) (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 (3.2)) may outperform the one based on model (3.1) [see [Sun and Cai, 2007](#), for more details]. In the following, we are thus interested in estimating parameters in model (3.2).

In a previous work [Nguyen and Matias, 2012], we discussed the estimation of the Euclidean part of the parameter  $\theta$  in model (3.2). 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  $\theta$ . We just mention that many estimators of  $\theta$  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  $\theta$  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 Nevial [2010], who is primarily interested in estimating  $\theta$  under the assumption that it is equal to  $g(1)$ . Relying on a kernel estimator of  $g$ , he derives nonparametric rates of convergence for  $\theta$ . 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  $\theta$  that roughly converges at

parametric rate (see exact condition in Corollary 3.1), 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 (3.2). 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  $\ell$ FDR.

The article is organized as follows. In Section 3.2, we start by describing different procedures to estimate  $f$ . We distinguish two types of procedures and first describe direct (non iterative) ones in Section 3.2.1. 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 3.2.2). 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 3.2.1 may be viewed as a proxy for studying the convergence properties (with respect to  $f$ ) of `kerfdr` procedure (properties that are unknown). Section 3.3 then gives the theoretical properties of the procedures described in Section 3.2. In particular, we establish (Theorem 3.1) 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 3.1). In Section 3.4, we rely on our different estimators to estimate both density  $f$  and the local false discovery rate  $\ell$ FDR. We present simulated experiments to compare their performances. All the proofs have been postponed to Section 3.5. Moreover, some of the more technical proofs have been further postponed to Section 3.6.

## 3.2 Algorithmic procedures to estimate the density $f$

### 3.2.1 Direct procedures

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

$$\hat{g}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_{i,h}(x), \quad (3.4)$$

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

$$K_{i,h}(\cdot) = \frac{1}{h} K\left(\frac{\cdot - X_i}{h}\right). \quad (3.5)$$

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

**A naive approach.** From Equation (3.2), 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  $\mathbf{1}_A$  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 3.4.

**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,

$$\forall i = 1, \dots, n \quad Z_i = \begin{cases} 0 & \text{if } H_i \text{ is true,} \\ 1 & \text{otherwise.} \end{cases} \quad (3.6)$$

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

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

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

$$\forall x \in [0, 1], \tau(x) = \mathbb{E}(Z_i|X_i = x) = \frac{(1 - \theta)f(x)}{g(x)} = 1 - \frac{\theta}{g(x)}. \quad (3.7)$$

This leads to the following definition

$$\forall x \in [0, 1], f_2(x) = \sum_{i=1}^n \frac{\tau(X_i)}{\sum_{k=1}^n \tau(X_k)} K_{i,h}(x). \quad (3.8)$$

Once again, the weight  $\tau_i = \tau(X_i)$  depends on the unknown parameters  $\theta$  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 (3.8). For the moment, we propose to replace the posterior probabilities  $\tau_i$  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

$$\forall x \in [0, 1], \hat{\tau}(x) = 1 - \frac{\hat{\theta}_n}{\hat{g}_n(x)}. \quad (3.9)$$

Then, by defining the weight

$$\hat{\tau}_i = \hat{\tau}(X_i) = 1 - \frac{\hat{\theta}_n}{\tilde{g}_n(X_i)}, \text{ where } \tilde{g}_n(X_i) = \frac{1}{(n-1)} \sum_{j \neq i}^n K_{j,h}(X_i), \quad (3.10)$$

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

$$\forall x \in [0, 1], \hat{f}_n^{\text{rwk}}(x) = \sum_{i=1}^n \frac{\hat{\tau}_i}{\sum_{k=1}^n \hat{\tau}_k} K_{i,h}(x). \quad (3.11)$$

Note that it is not necessary to use the same kernel  $K$  in defining  $\hat{g}_n$  and  $\hat{f}_n^{\text{rwk}}$ , 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 3.3, 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 (3.10) is minor and used in practice to reduce the bias of  $\tilde{g}_n(X_i)$ .

### 3.2.2 Iterative procedures

In this section, we still rely on a preliminary estimator  $\hat{\theta}_n$  of  $\theta$ . 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.

---

**Algorithm 1:** General structure of the iterative algorithms

---

```

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

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 (3.12) and (3.14)

     $s \leftarrow s + 1;$ 

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

---

Note that the parameter  $\theta$  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  $\theta$  would be iteratively updated. This is also the case with the maximum smoothed likelihood procedure described below in the particular setup of model (3.2). This is why we keep  $\theta$  fixed in both procedures. We now describe more explicitly the two procedures.

**Kerfdr algorithm.** This procedure has been proposed by Guedj et al. [2009], Robin et al. [2007] as an approximation to the estimator suggested by (3.8). In this procedure, functions  $\tilde{K}_{i,h}$  more simply denoted  $K_{i,h}$  are defined through (3.5) where  $K$  is a kernel (namely  $\int K(u)du = 1$ ) and following (3.7), the weights are updated as follows

$$\hat{\omega}_i^{(s)} = \frac{(1 - \hat{\theta}_n) \hat{f}^{(s)}(x_i)}{\hat{\theta}_n + (1 - \hat{\theta}_n) \hat{f}^{(s)}(x_i)}. \quad (3.12)$$

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

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

Thus, if the sequence  $\{\hat{\omega}^{(s)}\}_{s \geq 0}$  is convergent, it has to converge towards a fixed point of  $\psi$ . 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  $\mathcal{S} : \mathbb{L}_1([0, 1]) \rightarrow \mathbb{L}_1([0, 1])$  defined as

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

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 (3.2) 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  $\mathcal{S}^* : \mathbb{L}_1([0, 1]) \rightarrow \mathbb{L}_1([0, 1])$  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

$$\tilde{K}_{i,h}(x) = \frac{K_{i,h}(x)}{\int_0^1 K_{i,h}(s)ds}, \quad (3.13)$$

where  $K_{i,h}$  is defined through (3.5) relying on a positive symmetric kernel  $K$  and

$$\hat{\omega}_i^{(s)} = \frac{(1 - \hat{\theta}_n)\mathcal{N}\hat{f}^{(s)}(x_i)}{\hat{\theta}_n + (1 - \hat{\theta}_n)\mathcal{N}\hat{f}^{(s)}(x_i)}. \quad (3.14)$$

In Section 3.3.2, 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  $\varphi$ . Then, if the sequence  $\{\hat{\omega}^{(s)}\}_{s \geq 0}$  is convergent, it must be convergent to a fixed point of  $\varphi$ . Existence and uniqueness of a fixed point for `msl` algorithm is explored below in Proposition 3.2.

In the following section, we thus establish theoretical properties of the procedures presented here. These are then further compared on simulated data in Section 3.4.

### 3.3 Mathematical properties of the algorithms

#### 3.3.1 Randomly weighted kernel estimator

We provide below the convergence properties of the estimator  $\hat{f}_n^{\text{rwk}}$  defined through (3.11). 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 (3.2). 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 3.1.** Fix  $\beta > 0, L > 0$  and denote by  $H(\beta, L)$  the set of functions  $\psi : [0, 1] \rightarrow \mathbb{R}$  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  $(\beta, L)$ -Hölder class of functions.

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

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

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

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

(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 3.1.** *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 (3.4) converges uniformly almost surely to  $g$  [Wied and Weißbach, 2012]. In other words*

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

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

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

where  $C_1, C_2$  are two positive constants depending only on  $\beta, L, \alpha, \delta$  and  $K$ .

The proof of this theorem is postponed to Section 3.5.1. 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 3.1.** *Under the assumptions of Theorem 3.1, if  $\hat{\theta}_n$  is such that*

$$\limsup_{n \rightarrow +\infty} n^{\frac{2\beta}{2\beta+1}} \left[ \mathbb{E}_{\theta,f} \left( |\hat{\theta}_n - \theta| \right)^4 \right]^{\frac{1}{2}} < +\infty, \quad (3.15)$$

*then for any fixed value  $(\theta, f)$ , there is some positive constant  $C$  such that*

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

Note that estimators  $\hat{\theta}_n$  satisfying (3.15) exist. Indeed, relying on the same arguments as in the proofs of Theorems 2.2 or 2.3, we can prove that for instance histogram-based estimators or the estimator proposed by Celisse and Robin [2010] both satisfy that

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

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 (3.2), 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 (3.2). Anyway, such a lower bound would not be sufficient to conclude that estimator  $\hat{f}_n^{\text{rwk}}$  achieves the minimax rate. Indeed, the corollary states nothing about uniform convergence of  $\hat{f}_n^{\text{rwk}}(x)$  with respect to the parameter value  $(\theta, f)$  since the convergence of the estimator  $\hat{\theta}_n$  is not known to be uniform. Finally, note that the kernel estimator  $\hat{f}_n^{\text{rwk}}(x)$  also depends on the regularity  $\beta$  of the density  $f$ . In Appendix A, we apply Lepski's method to explore an adaptive kernel estimator of  $f$ .

### 3.3.2 Maximum smoothed likelihood estimator

Let us now explain the motivations for considering an iterative procedure with functions  $\tilde{K}_{i, \theta, h}$  and weights  $\hat{\omega}_i^{(s)}$  respectively defined through (3.13) and (3.14). 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  $\theta$ , 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  $(\theta, f)$ . Indeed, let  $D(a | b)$  denote the Kullback-Leibler divergence between (positive) measures  $a$  and  $b$ , defined as

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

### 3.3. MATHEMATICAL PROPERTIES OF THE ALGORITHMS

---

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

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

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  $\theta$ . Indeed, as previously explained,  $\theta$  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

$$l(f^t) - l(f^{t+1}) \geq cD(f^{t+1} | f^t) \geq 0, \quad (3.16)$$

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

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

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

$$\begin{aligned} l(f^t) - l(f^{t+1}) &\geq \int_0^1 g_0(x) \omega_t(x) \log \frac{\mathcal{N}f^{t+1}(x)}{\mathcal{N}f^t(x)} dx \\ &\geq \int_0^1 g_0(x) \omega_t(x) \left[ \mathcal{S}^*(\log f^{t+1})(x) - \mathcal{S}^*(\log f^t)(x) \right] dx \\ &\geq \int_0^1 g_0(x) \omega_t(x) \left( \int_0^1 K_h(s - x) ds \right)^{-1} \left( \int_0^1 K_h(u - x) \log \frac{f^{t+1}(u)}{f^t(u)} du \right) dx \\ &\geq \int_0^1 \left( \int_0^1 \frac{g_0(x) \omega_t(x) K_h(u - x)}{\int_0^1 K_h(s - x) ds} dx \right) \log \frac{f^{t+1}(u)}{f^t(u)} du. \end{aligned} \quad (3.17)$$

Let us define

$$\alpha_t = \frac{1}{\int_0^1 \omega_t(u) g_0(u) du} \quad \text{and} \quad f^{t+1}(x) = \alpha_t \int_0^1 \frac{K_h(u - x) \omega_t(u) g_0(u)}{\int_0^1 K_h(s - u) ds} du, \quad (3.18)$$

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

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

With the same arguments as in the proof of following Proposition 3.1, we can show that  $\alpha_t^{-1}$  is lower bounded by a positive constant  $c$  depending on  $\theta, h$  and  $K$ . The sequence  $\{f^t\}_{t \geq 0}$  thus satisfies property (3.16). 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 (3.13), (3.14) 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 (3.16).

**Proposition 3.1.** *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 (3.13), (3.14) and Algorithm 1 satisfies*

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

where  $c$  is a positive constant depending on  $\theta$ , 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 3.2.** *The criterion  $l$  has a unique minimum  $f^*$  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  $f^*$ .

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 (3.13) and (3.14) is an approximation of the minimizer  $f^*$  of  $l$ .

### 3.4 Estimation of local false discovery rate and simulation study

#### 3.4.1 Estimation of local false discovery rate

In this section, we study the estimation of local false discovery rate ( $\ell$ FDR) by using the previously introduced estimators of the density  $f$  and compare these different approaches on simulated data. Let us recall definition (3.3) of the local false discovery rate

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

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

$$\widehat{\ell\text{FDR}}(x_i) = \frac{\hat{\theta}}{\hat{\theta} + (1 - \hat{\theta})\hat{f}(x_i)}. \quad (3.19)$$

Let us now denote by  $\hat{f}_{\text{rwk}}$  the randomly weighted kernel estimator of  $f$  constructed in Section 3.2.1, 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  $\ell$ FDR induced by a plug-in of estimators  $\hat{f}_m$  in (3.19) and  $\widehat{\ell\text{FDR}}_{\text{st}}$  be the estimator of  $\ell$ FDR computed by the method of Strimmer [2008]. We compute the root mean squared error (RMSE) between the estimates and the true values

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

for  $m \in \{\text{rwk}, \text{kerfdr}, \text{msl}, \text{st}\}$  and where  $s = 1, \dots, S$  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  $m \in \{\text{rwk}, \text{kerfdr}, \text{msl}\}$ , relying on the root mean integrated squared error

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

The quality of the estimates provided by method  $m$  is measured by the mean  $\text{RMSE}_m$  or  $\text{RMISE}_m$ : the smaller these quantities, the better the performances of the method.

We mention that we also tested the naive method described in Section 3.2.1 and the results were bad. In order to present clear figures, we have chosen not to show those.



### 3.4.2 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 (3.2). 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

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

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 3.1](#) 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  $\theta$  initially proposed by [Schweder and Spjøtvoll \[1982\]](#), namely

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

with parameter value  $\lambda$  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 [\[Silverman, 1986, Section 3.4.2\]](#),

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

where  $SD$  and  $IQR$  are respectively the standard deviation and interquartile range of the data values. [Figures 3.2, 3.3 and 3.4](#) 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, `ms1` outperforms the two other methods (`rwk`, `kerfdr`).

### 3.4. ESTIMATION OF LOCAL FALSE DISCOVERY RATE AND SIMULATION STUDY

---

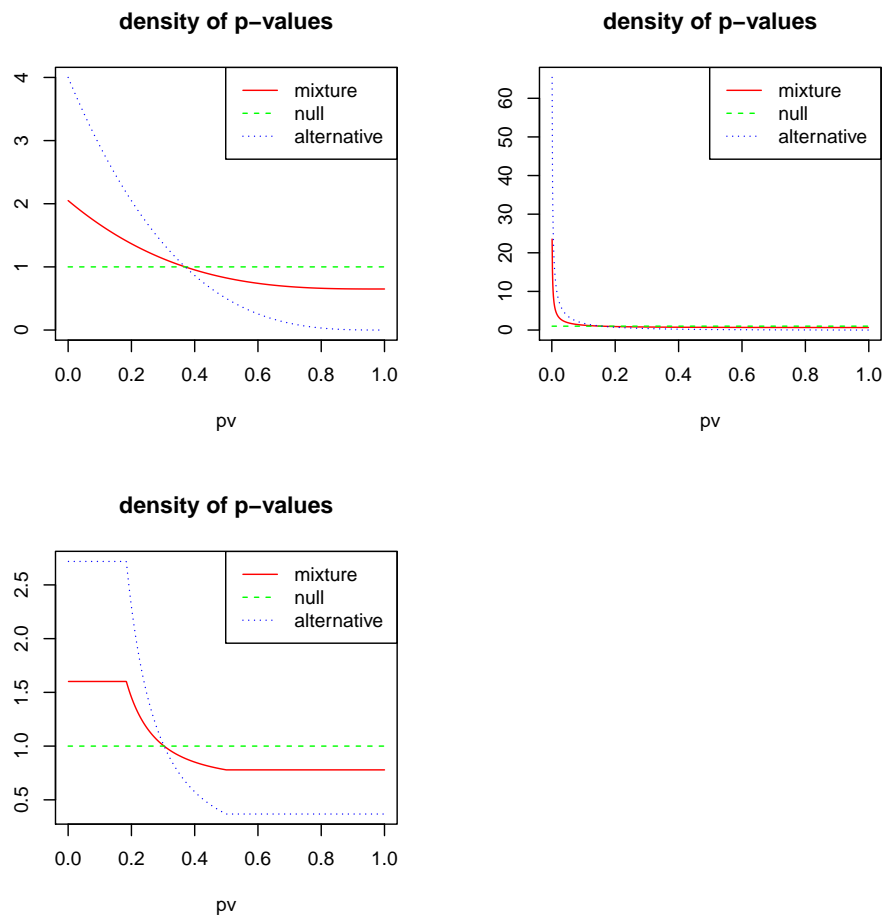


Figure 3.1: Densities of the  $p$ -values in the three different models, with  $\theta = 0.65$ . Top left: first model, top right: second model, bottom left: third model.

### 3.4. ESTIMATION OF LOCAL FALSE DISCOVERY RATE AND SIMULATION STUDY

---

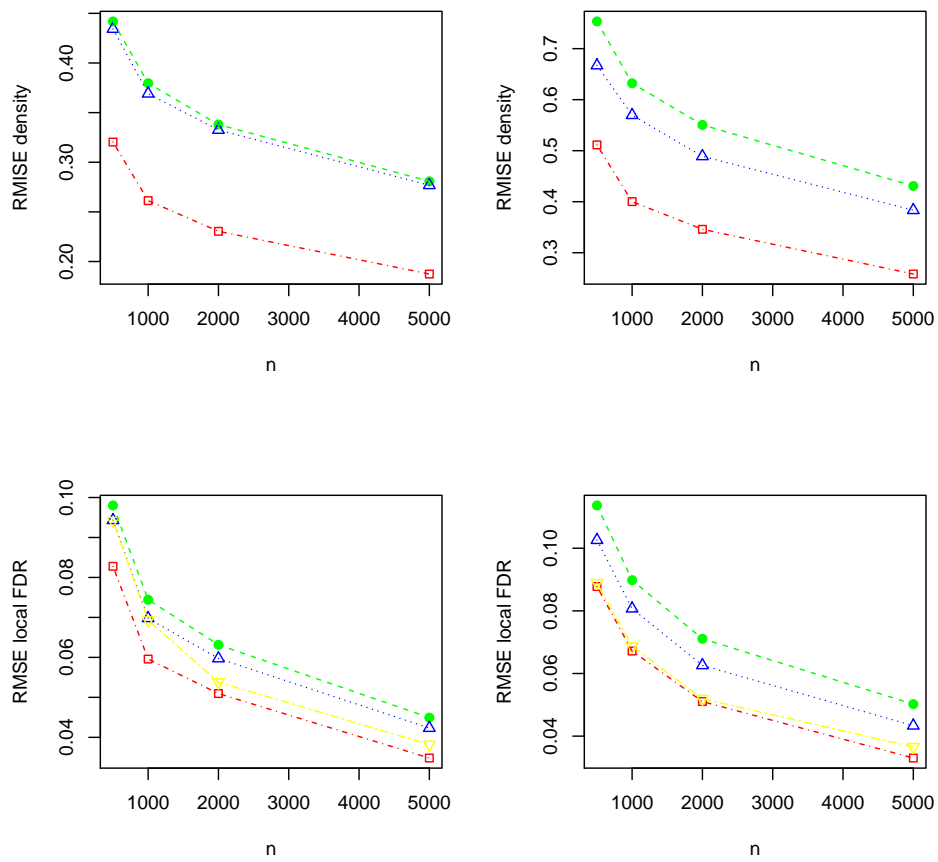


Figure 3.2: RMISE (for density  $f$ ) and RMSE (for  $\ell$ FDR) in the first model as a function of  $n$ . Methods: "●" = `rwk`, "△" = `kerfdr`, "□" = `msl`, "▽" = `st` (only for  $\ell$ FDR). Left:  $\theta = 0.65$ , right:  $\theta = 0.85$ .

### 3.4. ESTIMATION OF LOCAL FALSE DISCOVERY RATE AND SIMULATION STUDY

---

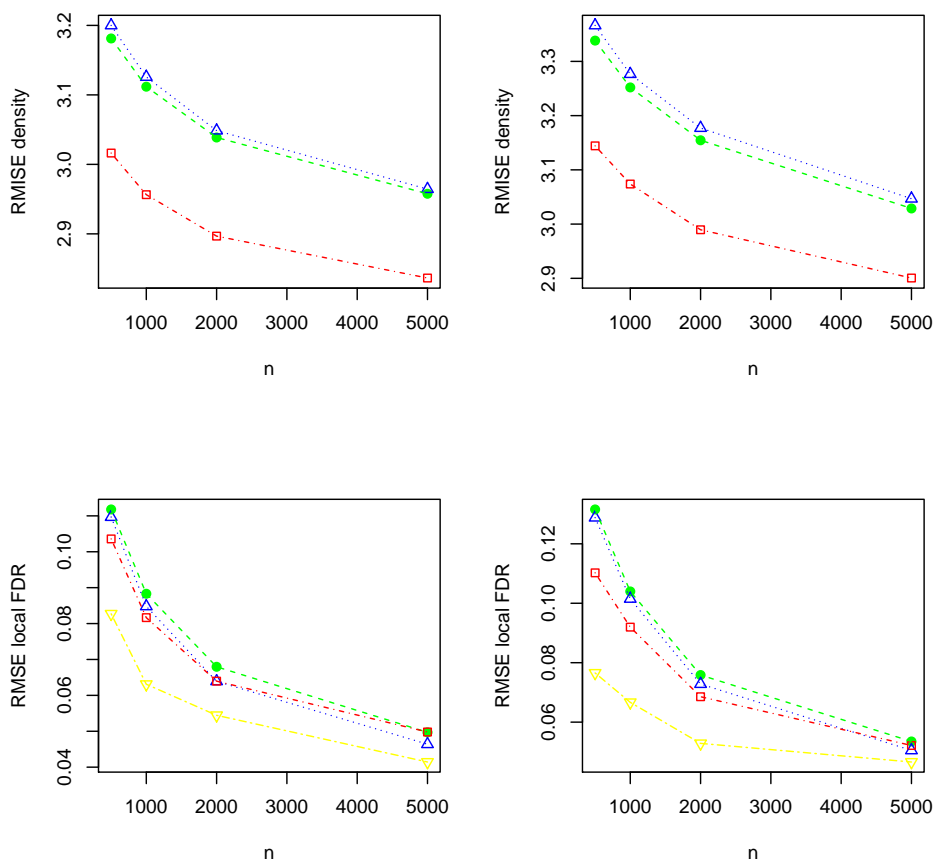


Figure 3.3: RMISE (for density  $f$ ) and RMSE (for  $\ell$ FDR) in the second model as a function of  $n$ . Methods: "●" = `rwk`, "△" = `kerfdr`, "□" = `msl`, "▽" = `st` (only for  $\ell$ FDR). Left:  $\theta = 0.65$ , right:  $\theta = 0.85$ .

### 3.4. ESTIMATION OF LOCAL FALSE DISCOVERY RATE AND SIMULATION STUDY

---

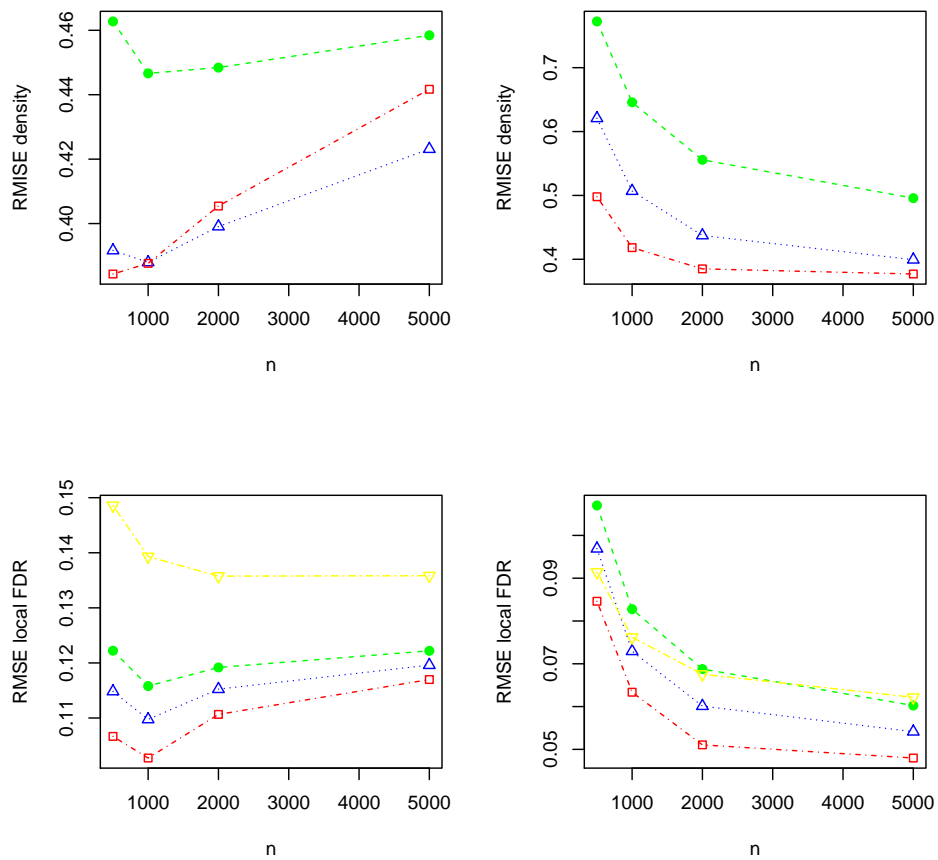


Figure 3.4: RMISE (for density  $f$ ) and RMSE (for  $\ell$ FDR) in the third model as a function of  $n$ . Methods: "●" = `rwk`, "△" = `kerfdr`, "□" = `msl`, "▽" = `st` (only for  $\ell$ FDR). Left:  $\theta = 0.65$ , right:  $\theta = 0.85$ .

Then, we switch to comparing the methods with respect to estimation of  $\ell$ FDR (bottom half of each figure). First, note that the four methods exhibit small RMSEs with respect to  $\ell$ FDR and are thus efficient for estimating this quantity. We also note that `rwk` tends to have lower performances than `kerfdr,ms1`. Now, `ms1` tends to slightly outperform `kerfdr`. Thus `ms1` appears as a competitive method for  $\ell$ FDR estimation. The comparison with [Strimmer \[2008\]](#)'s approach is more difficult: for model 1, the method compares with `ms1`, while it outperforms all the methods in model 2 and is outperformed by `ms1` in model 3.

As a conclusion, we claim that `ms1` is a competitive method for estimating both the alternative density  $f$  and the  $\ell$ FDR.

## 3.5 Proofs of main results

### 3.5.1 Proof of Theorem 3.1

The proof works as follows: we first start by proving that the pointwise quadratic risk of function  $f_2$  defined by (3.8) 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 3.1.** (*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  $\beta, L$  and  $K$  such that for all  $x_0 \in \mathbb{R}$ ,*

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

**Lemma 3.2.** (*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*

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

### 3.5. PROOFS OF MAIN RESULTS

---

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

**Proposition 3.3.** *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 (3.8) and depending on  $(\theta, f)$ , satisfies*

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

where  $C_4$  is a positive constant depending only on  $\beta, L, \alpha, \delta$  and  $K$ .

*Proof of Proposition 3.3.* 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

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

Let us first study the bias term. According to (3.8) and the definition (3.7) of the weights, we have

$$\begin{aligned} \mathbb{E}_{\theta, f}[f_2(x)] &= \frac{n}{h} \mathbb{E}_{\theta, f} \left[ \tau_1 K\left(\frac{x - X_1}{h}\right) \left(\sum_{k=1}^n \tau_k\right)^{-1} \right] \\ &= \frac{n}{h} \mathbb{E}_{\theta, f} \left[ \frac{f(X_1)}{g(X_1)} K\left(\frac{x - X_1}{h}\right) S_n^{-1} \right] \\ &= \frac{n}{h} \int_0^1 f(t) K\left(\frac{x-t}{h}\right) \mathbb{E}_{\theta, f} \left[ \left(\frac{f(t)}{g(t)} + S_{n-1}\right)^{-1} \right] dt \\ &= n \int_{-x/h}^{(1-x)/h} K(t) f(x+th) \mathbb{E}_{\theta, f} \left[ \left(\frac{f(x+th)}{g(x+th)} + S_{n-1}\right)^{-1} \right] dt. \end{aligned} \quad (3.20)$$

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

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

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

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

where the bounds are uniform with respect to  $t$ .

By combining this inequality with (3.20), we obtain

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

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

Then, we apply the following lemma, whose proof is postponed to Section 3.6.1.

**Lemma 3.3.** *There exist some positive constants  $c_1, c_2, c_3, c_4$  (depending on  $\delta$ ) such that for  $n$  large enough,*

$$\mathbb{E}_{\theta,f} \left( \frac{1}{S_n} \right) \leq \frac{1}{n} + \frac{c_1}{n^2}, \tag{3.21}$$

$$\mathbb{E}_{\theta,f} \left( \frac{1}{S_n^2} \right) \leq \frac{c_2}{n^2}, \tag{3.22}$$

$$\mathbb{E}_{\theta,f} \left( \frac{1}{S_n + 2\delta^{-1}} \right) \geq \frac{1}{n} - \frac{c_3}{n^2}, \tag{3.23}$$

$$\text{and } \mathbb{E}_{\theta,f} \left( \frac{1}{S_n^2} \right) - \mathbb{E}_{\theta,f}^2 \left( \frac{1}{\delta^{-1} + S_n} \right) \leq \frac{c_4}{n^3}. \tag{3.24}$$

Relying on Inequalities (3.21) and (3.23), we have for  $n$  large enough

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

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

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

By using Lemma 3.1 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

$$\text{Var}_{\theta,f} [f_2(x)] = \frac{1}{h^2} [n \text{Var}_{\theta,f} (Y_1) + n(n-1) \text{Cov}_{\theta,f} (Y_1, Y_2)], \tag{3.25}$$



### 3.5. PROOFS OF MAIN RESULTS

---

where

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

The variance of  $Y_1$  is bounded by its second moment and

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

Now, recalling that  $0 \leq f/g \leq \delta^{-1}$  and using Inequality (3.22) of Lemma 3.3, we get

$$\begin{aligned} \mathbb{E}_{\theta,f}(Y_1^2) &\leq h \left( \int_{-x/h}^{(1-x)/h} \frac{f^2(x+th)}{g(x+th)} K^2(t) dt \right) \mathbb{E}_{\theta,f} \left( \frac{1}{S_{n-1}^2} \right) \\ &\leq h \delta^{-1} \sup_{f \in \Sigma(\beta,L)} \|f\|_\infty \left( \int K^2(t) dt \right) \frac{c_2}{n^2} \leq \frac{C_6 h}{n^2}. \end{aligned} \quad (3.26)$$

We now study the covariance of  $Y_1$  and  $Y_2$

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

where

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

Hence

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &\leq \int_{[0,1]^2} f(t)f(u) K\left(\frac{x - t}{h}\right) K\left(\frac{x - u}{h}\right) \left[ \mathbb{E}_{\theta,f} \left( \frac{1}{S_{n-2}^2} \right) - \mathbb{E}_{\theta,f}^2 \left( \frac{1}{2\delta^{-1} + S_{n-2}} \right) \right] dt du \\ &\leq h^2 \left( \int_{\mathbb{R}} f(x+th) K(t) dt \right)^2 \left[ \mathbb{E}_{\theta,f} \left( \frac{1}{S_{n-2}^2} \right) - \mathbb{E}_{\theta,f}^2 \left( \frac{1}{2\delta^{-1} + S_{n-2}} \right) \right] \\ &\leq C_7 h^2 \left[ \mathbb{E}_{\theta,f} \left( \frac{1}{S_{n-2}^2} \right) - \mathbb{E}_{\theta,f}^2 \left( \frac{1}{2\delta^{-1} + S_{n-2}} \right) \right]. \end{aligned}$$

### 3.5. PROOFS OF MAIN RESULTS

---

According to Inequality (3.24) of Lemma 3.3, we have

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

hence

$$\text{Cov}_{\theta,f}(Y_1, Y_2) \leq \frac{C_8 h^2}{n^3}. \quad (3.27)$$

By returning to Equality (3.25) and combining with (3.26) and (3.27), we obtain

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

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

$$\mathbb{E}_{\theta,f}(|\hat{f}_n(x) - f(x)|^2) \leq 2\mathbb{E}_{\theta,f}(|f_2(x) - f(x)|^2) + 2\mathbb{E}_{\theta,f}(|\hat{f}_n(x) - f_2(x)|^2). \quad (3.28)$$

According to Proposition 3.3, we have

$$\mathbb{E}_{\theta,f}(|f_2(x) - f(x)|^2) \leq C_4 n^{\frac{-2\beta}{2\beta+1}}, \quad (3.29)$$

and it remains to study the second term appearing in the right-hand side of (3.28). 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 (3.10), we have for all  $1 \leq i \leq n$ ,

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

and thus get

$$\begin{aligned}
 \hat{f}_n(x) - f_2(x) &= \frac{n\hat{\theta}_n}{\sum_k \hat{\tau}_k} \times \frac{1}{nh} \sum_{i=1}^n \left[ \frac{1}{\tilde{g}_n(X_i)} - \frac{1}{g(X_i)} \right] K\left(\frac{x - X_i}{h}\right) \\
 &+ \frac{n(\hat{\theta}_n - \theta)}{\sum_k \hat{\tau}_k} \times \frac{1}{nh} \sum_{i=1}^n \frac{1}{g(X_i)} K\left(\frac{x - X_i}{h}\right) \\
 &+ \frac{n^2\hat{\theta}_n}{\sum_k \hat{\tau}_k \sum_k \tau_k} \times \frac{1}{n} \sum_k \left[ \frac{1}{\tilde{g}_n(X_k)} - \frac{1}{g(X_k)} \right] \times \frac{1}{nh} \sum_{i=1}^n \tau_i K\left(\frac{x - X_i}{h}\right) \\
 &+ \frac{n^2(\hat{\theta}_n - \theta)}{\sum_k \hat{\tau}_k \sum_k \tau_k} \times \frac{1}{n} \sum_k \frac{1}{g(X_k)} \times \frac{1}{nh} \sum_{i=1}^n \tau_i K\left(\frac{x - X_i}{h}\right). \tag{3.30}
 \end{aligned}$$

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

$$0 \leq \tau_i \leq 1 \text{ and } \frac{1}{g(X_i)} \leq \frac{1}{\theta} \leq \delta^{-1}. \tag{3.31}$$

Since by assumption  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{a.s.} \theta \in [0, 1]$ , 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

$$\frac{\delta}{2} \leq \frac{1 - \theta}{2} \leq \frac{1}{n} \sum_{i=1}^n \tau_i \leq \frac{3(1 - \theta)}{2} \leq \frac{3(1 - \delta)}{2} \quad \text{a.s.} \tag{3.32}$$

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  $i$ ,

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

Since convergence of  $\hat{g}_n$  to  $g$  is valid pointwise and in  $\mathbb{L}_\infty$  norm (see Remark 3.1), 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

$$\left| \frac{1}{\tilde{g}_n(X_i)} - \frac{1}{g(X_i)} \right| \leq 2\delta^{-2} |\tilde{g}_n(X_i) - g(X_i)| \quad \text{a.s.} \tag{3.33}$$

We also use the following lemma, whose proof is postponed to Section 3.6.2.

**Lemma 3.4.** *For large enough  $n$ , we have*

$$\frac{n}{|\sum_k \hat{\tau}_k|} \leq c_7 \quad a.s. \quad (3.34)$$

By returning to Equality (3.30) and combining with (3.31), (3.32), (3.33) and (3.34), we obtain

$$\begin{aligned} |\hat{f}_n(x) - f_2(x)|^2 &\leq c_8 \left( \frac{1}{nh} \sum_{i=1}^n |\tilde{g}_n(X_i) - g(X_i)| \times \left| K\left(\frac{x - X_i}{h}\right) \right| \right)^2 \\ &\quad + c_9 |\hat{\theta}_n - \theta|^2 \left( \frac{1}{nh} \sum_{i=1}^n \left| K\left(\frac{x - X_i}{h}\right) \right| \right)^2 \\ &\quad + c_{10} \left( \frac{1}{n} \sum_{i=1}^n |\tilde{g}_n(X_i) - g(X_i)| \right)^2 \left( \frac{1}{nh} \sum_{i=1}^n \left| K\left(\frac{x - X_i}{h}\right) \right| \right)^2 \quad a.s. \end{aligned} \quad (3.35)$$

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

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

Now,

$$\begin{aligned} T_{11} &= \mathbb{E}_{\theta, f} \left[ \frac{1}{h} |\tilde{g}_n(X_1) - g(X_1)|^2 K^2\left(\frac{x - X_1}{h}\right) \right] \\ &= \int_0^1 \mathbb{E}_{\theta, f} \left[ (|\hat{g}_{n-1}(t) - g(t)|^2) K^2\left(\frac{x - t}{h}\right) \frac{g(t)}{h} \right] dt \quad (\text{according to definition (3.10)}) \\ &\leq C_{10} n^{\frac{-2\beta}{2\beta+1}} \int_0^1 K^2\left(\frac{x - t}{h}\right) \frac{g(t)}{h} dt \quad (\text{according to Remark 3.1}) \\ &\leq C_{11} n^{\frac{-2\beta}{2\beta+1}} \quad (\text{according to Lemma 3.2}), \end{aligned} \quad (3.36)$$

and in the same way

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

This last term is upper-bound by

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

According to Remark 3.1, we have

$$T_{12} \leq C_{12} n^{\frac{-2\beta}{2\beta+1}} \left[ \int_0^1 \left| K\left(\frac{x-t}{h}\right) \right| \frac{g(t)}{h} dt \right]^2 \leq C_{13} n^{\frac{-2\beta}{2\beta+1}} \quad (\text{according to Lemma 3.2}). \quad (3.37)$$

Thus we get that

$$T_1 = \mathbb{E}_{\theta, f} \left[ \left( \frac{1}{nh} \sum_{i=1}^n |\tilde{g}_n(X_i) - g(X_i)| \left| K\left(\frac{x - X_i}{h}\right) \right| \right)^2 \right] \leq C_{14} n^{\frac{-2\beta}{2\beta+1}}. \quad (3.38)$$

For the second term in the right hand side of (3.35), we have

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

The proof of the following lemma is postponed to Section 3.6.3.

**Lemma 3.5.** *There exist some positive constant  $C_{15}$  such that*

$$\mathbb{E}_{\theta, f} \left[ \left( \frac{1}{nh} \sum_{i=1}^n \left| K\left(\frac{x - X_i}{h}\right) \right| \right)^4 \right] \leq C_{15}. \quad (3.39)$$

This lemma entails that

$$T_2 \leq C_{15} \left[ \mathbb{E}_{\theta, f} \left( |\hat{\theta}_n - \theta|^4 \right) \right]^{\frac{1}{2}}. \quad (3.40)$$

Now, we turn to the third term in the right hand side of (3.35). We have

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

By using the same arguments as for obtaining (3.36) and (3.37), we can get that

$$T_3 \leq C_{16} n^{\frac{-2\beta}{2\beta+1}}. \quad (3.41)$$

According to (3.38), (3.40) and (3.41), we may conclude

$$\mathbb{E}_{\theta, f} (|\hat{f}_n(x) - f_2(x)|^2) \leq C_{15} \left[ \mathbb{E}_{\theta, f} \left( |\hat{\theta}_n - \theta|^4 \right) \right]^{\frac{1}{2}} + C_{17} n^{\frac{-2\beta}{2\beta+1}}. \quad (3.42)$$

By returning to Inequality (3.28) and combining it with (3.29) and (3.42), we achieve that

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

□

### 3.5.2 Other proofs

*Proof of Proposition 3.1.* By using the same arguments as for obtaining (3.17), we can get that

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

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.  $\square$

*Proof of Proposition 3.2.* We start by stating a lemma, whose proof is postponed to Section 3.6.4.

**Lemma 3.6.** *The function  $l : \mathcal{B} \rightarrow \mathbb{R}$  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 (3.18) 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  $f^*$  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  $f^*$  of  $l$ . Now, uniqueness of the uniform limit value of the sequence  $\{f^t\}_{t \geq 0}$  entails its convergence.  $\square$

## 3.6 Proofs of technical lemmas

### 3.6.1 Proof of Lemma 3.3

*Proof.* We first show (3.22). According to the law of large numbers, since  $\mathbb{E}_{\theta,f}(f(X_1)/g(X_1)) = 1$ , we have

$$\frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} \xrightarrow[n \rightarrow \infty]{as} 1. \quad (3.43)$$

Hence

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

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

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

establishing (3.22). Let us now prove (3.21). By using a Taylor's expansion, we have

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

where  $\gamma_n \in ]0, 1[$  depends on  $S_n$ . Combining this with (3.43), we obtain

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

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

$$\frac{1}{n} \left[ 2 - \frac{S_n}{n} + c' \left(\frac{S_n}{n} - 1\right)^2 \right] \leq \frac{1}{S_n} \leq \frac{1}{n} \left[ 2 - \frac{S_n}{n} + c \left(\frac{S_n}{n} - 1\right)^2 \right] \quad \text{a.s.} \quad (3.44)$$

This implies in particular that

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

In addition,

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

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

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



### 3.6. PROOFS OF TECHNICAL LEMMAS

---

We now prove (3.23). 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

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

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

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

According to (3.44), we have

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

and it is proved above that

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

Thus we obtain Inequality (3.23), namely

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

Finally, we show (3.24). In the same way as we proved (3.23) above, we have for large enough  $n$ ,

$$\mathbb{E}_{\theta, f} \left[ \frac{1}{S_n + 2\delta^{-1}} \right] \geq \frac{1}{n} - \frac{c'_3}{n^2} > 0$$

and thus

$$\mathbb{E}_{\theta, f}^2 \left[ \frac{1}{S_n + 2\delta^{-1}} \right] \geq \frac{1}{n^2} \left( 1 - \frac{2c'_3}{n} + \frac{c'^2_3}{n^2} \right) \geq \frac{1}{n^2} \left( 1 - \frac{2c'_3}{n} \right). \quad (3.45)$$

According to Inequality (3.44) (containing only positive terms for  $n$  large enough), we have

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

Since

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

we have

$$\begin{aligned}
 \mathbb{E}_{\theta,f} \left[ \frac{1}{S_n^2} \right] &\leq \frac{1}{n^2} \left[ 4 + \frac{\mathbb{E}_{\theta,f}[S_n^2]}{n^2} + c^2 \mathbb{E}_{\theta,f} \left[ \left( \frac{S_n}{n} - 1 \right)^4 \right] - 4 \frac{\mathbb{E}_{\theta,f}[S_n]}{n} + 4c \mathbb{E}_{\theta,f} \left[ \left( \frac{S_n}{n} - 1 \right)^2 \right] \right] \\
 &\leq \frac{1}{n^2} \left[ 4 + \frac{1}{n} \text{Var} \left( \frac{f(X_1)}{g(X_1)} \right) + 1 + c^2 \mathbb{E}_{\theta,f} \left[ \left( \frac{S_n}{n} - 1 \right)^4 \right] - 4 + \frac{4c}{n} \text{Var} \left( \frac{f(X_1)}{g(X_1)} \right) \right] \\
 &\leq \frac{1}{n^2} \left[ 1 + \frac{C_4}{n} + c^2 \mathbb{E}_{\theta,f} \left[ \left( \frac{S_n}{n} - 1 \right)^4 \right] \right]. \tag{3.46}
 \end{aligned}$$

Combining (3.45) and (3.46), we get that

$$\mathbb{E}_{\theta,f} \left[ \frac{1}{S_n^2} \right] - \mathbb{E}_{\theta,f}^2 \left[ \frac{1}{S_n + 2\delta^{-1}} \right] \leq \frac{C}{n^3} + \frac{c^2}{n^2} \mathbb{E}_{\theta,f} \left[ \left( \frac{S_n}{n} - 1 \right)^4 \right]. \tag{3.47}$$

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

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

We have

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

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

$$\mathbb{E}_{\theta,f} \left[ \left( \frac{S_n}{n} - 1 \right)^4 \right] = \frac{1}{n^4} \left[ n \mathbb{E}_{\theta,f}(U_1^4) + n(n-1) \mathbb{E}_{\theta,f}(U_1^2 U_2^2) \right] = O\left( \frac{1}{n^2} \right). \tag{3.48}$$

Finally, according to (3.47) and (3.48) we have

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

□

### 3.6.2 Proof of Lemma 3.4

*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,

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

and using the same argument as for establishing (3.33), 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 3.1, 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

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

□

### 3.6.3 Proof of Lemma 3.5

*Proof.* In order to prove (3.39), let us consider iid random variables  $U_1, \dots, U_n$  defined as

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

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

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

We then write

$$\left( \frac{1}{nh} \sum_{i=1}^n \left| K \left( \frac{x - X_i}{h} \right) \right| \right)^4 = \frac{1}{n^4 h^4} \left( \sum_i U_i \right)^4, \quad (3.49)$$

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  $nh \rightarrow \infty$ ,

$$\begin{aligned}
 & \mathbb{E}_{\theta, f} \left[ \left( \sum_i U_i \right)^4 \right] \\
 &= n \mathbb{E}_{\theta, f}(U_1^4) + n(n-1) \mathbb{E}_{\theta, f}(U_1^3 U_2) + n(n-1) \mathbb{E}_{\theta, f}(U_1^2 U_2^2) + \\
 & \quad + n(n-1)(n-2) \mathbb{E}_{\theta, f}(U_1^2 U_2 U_3) + n(n-1)(n-2)(n-3) \mathbb{E}_{\theta, f}(U_1 U_2 U_3 U_4) \\
 &= n \mathbb{E}_{\theta, f}(U_1^4) + n(n-1) \mathbb{E}_{\theta, f}(U_1^3) \mathbb{E}_{\theta, f}(U_1) + n(n-1) \mathbb{E}_{\theta, f}^2(U_1^2) + \\
 & \quad + n(n-1)(n-2) \mathbb{E}_{\theta, f}(U_1^2) \mathbb{E}_{\theta, f}^2(U_1) + n(n-1)(n-2)(n-3) \mathbb{E}_{\theta, f}^4(U_1) \\
 & \leq C_{15} n^4 h^4.
 \end{aligned} \tag{3.50}$$

According to (3.49) and (3.50) we obtain the result.  $\square$

### 3.6.4 Proof of Lemma 3.6

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

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

and

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

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

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

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  $\theta$ .  $\square$

## Conclusion

In this thesis, we have considered a semiparametric mixture model in a multiple testing setup where several thousands of independent hypotheses can be tested simultaneously. We observe the  $p$ -values associated with  $n$  independent tested hypotheses. The Euclidean parameter of the model, denoted by  $\theta$  stands for the proportion of true null hypotheses and the nonparametric component  $f$  is the probability density function of the  $p$ -values under the alternative hypothesis. The problem of estimating the parameters of the model appears from the false discovery rate control procedures. We first have studied the estimation of the parameter  $\theta$  by supposing the nonincreasing assumption on  $f$  and distinguishing into two cases: models where  $f$  vanishes on a non-empty interval or not. In the first case, we obtain the existence of  $\sqrt{n}$ -consistent estimators of  $\theta$  that is to say estimators  $\hat{\theta}_n$  such that  $\sqrt{n}(\hat{\theta}_n - \theta)$  is bounded in probability (denoted by  $\sqrt{n}(\hat{\theta}_n - \theta) = O_{\mathbb{P}}(1)$ ). We exhibit such estimators and also compute the asymptotic optimal variance for this problem. Moreover, we conjecture that asymptotically efficient estimators (that are estimators asymptotically attaining this variance lower bound) do not exist in regular models. In the second case, we compute that the Fisher information for  $\theta$  is equal to zero, then there is no regular estimator for  $\theta$  and the quadratic risk of any estimator does not converge at parametric rate.

Secondly, we have studied the estimation of the nonparametric component  $f$  of the model, relying on a preliminary estimator of  $\theta$ . We describe different procedures to estimate  $f$ . The first one is a randomly weighted kernel estimator and the second one is an iterative algorithm for estimating  $f$ , that aims at maximizing a smoothed likelihood. The randomly weighted kernel procedure may be viewed as a theoretical validation of `kerfdr` approach and we establish an upper bound on its pointwise quadratic risk. Moreover, we prove that the resulting iterative algorithm possesses a desirable descent property, just as an `em` algorithm does. From some simulations to compare their performances, we claim that this iterative procedure is a competitive method for estimating both the alternative density  $f$  and the local false discovery rate  $\ell$ FDR.

# Another semiparametric mixture model

---

In this chapter, we consider another semiparametric mixture model that has proved to be useful to analyze gene expression data coming from microarray analysis. This model has been considered by many authors such as [Bordes et al. \[2006\]](#) and [Bordes and Vandekerkhove \[2010\]](#). They consider the following semiparametric mixture

$$g(x) = (1 - p)f_0(x) + pf(x - \mu), \quad \forall x \in \mathbb{R}, \quad (4.1)$$

where  $f_0$  is assumed to be a known density function and where the unknown parameters are the mixture proportion  $p \in (0, 1)$ , the non-null location parameter  $\mu$  and the density function  $f$ . The density  $f$  is supposed symmetric and has  $\mathbb{R}$ -support. We denote by  $\mathcal{F}$  the set of all symmetric densities with support equal to  $\mathbb{R}$ . [Bordes and Vandekerkhove \[2010\]](#) propose an asymptotically normal estimator of the unknown parameters under some mild assumptions. Here, we want to apply the semiparametric theory to study the asymptotic efficiency of estimators of the Euclidean parameters  $\theta = (p, \mu)$  of this mixture. We will suppress the subscripted  $\mathbb{R}$  on the integral sign from now on.

## 4.1 Identifiability

Note that model (4.1) is not identifiable in general, as it is shown in the following example (mentioned in [Bordes et al. \[2006\]](#))

$$(1 - p)f_0(x) + pf(x - \mu) = (1 - \frac{p}{2})f_0(x) + \frac{p}{2}f(x - 2\mu), \quad \forall x \in \mathbb{R},$$

where  $\mu \in \mathbb{R}^*$ ,  $f_0$  is a symmetric density function,  $p \in (0, 1)$  and  $f(x) = [f_0(x + \mu) + f_0(x - \mu)]/2$ . Then [Bordes et al. \[2006\]](#) give some sufficient conditions to obtain the identifiability of the parameters of model (4.1). We define the set  $\mathcal{F}_q = \{f \in \mathcal{F}; \int |x|^q f(x) dx < +\infty\}$  for  $q \geq 1$  and denote by  $\bar{f}_0$  the Fourier transform of the density  $f_0$ .

**Identifiability condition (I)** (Bordes et al. [2006]). The mixture model (4.1), with  $f_0 \in \mathcal{F}_3$  and  $\bar{f}_0 > 0$ , is identifiable if

$$(p, \mu, f) \in (0, 1) \times \mathbb{R}^* \times \mathcal{F}_3 \text{ and } m \neq m_0 + \frac{k \pm 2}{3k} \mu^2, \forall k \in \mathbb{N}^*,$$

where  $m_0$  and  $m$  are the second-order moments of  $f_0$  and  $f$  respectively.

**Example (Gaussian mixture).** We consider a Gaussian mixture

$$(1 - p)\mathcal{N}(0, 1) + p\mathcal{N}(\mu, \sigma^2), \quad (4.2)$$

where  $\mathcal{N}(\alpha, \beta)$  denotes the Gaussian distribution with mean  $\alpha$  and variance equal to  $\beta$ . Here, we have  $(f_0, f) \in \mathcal{F}_3^2$  and  $\bar{f}_0(x) = \exp(-x^2/2) > 0, \forall x \in \mathbb{R}$ , then the identifiability condition (I) is equivalent to

$$\sigma^2 \neq 1 + \frac{k \pm 2}{3k} \mu^2, \forall k \in \mathbb{N}^* \text{ or equivalently } \frac{2\mu^2}{3\sigma^2 - \mu^2 - 3} \notin \mathbb{Z}.$$

## 4.2 Efficient information matrix for estimating $\theta$

In this section, we calculate the efficient information matrix for estimating the parameter  $\theta = (p, \mu)$ . Firstly, let us denote

$$s(x) = \frac{g(x + \mu)g(-x + \mu)}{g(x + \mu) + g(-x + \mu)}, \quad (4.3)$$

and

$$A = 1 + \frac{(1 - p)^2}{4} \left[ \int s(x) dx \right]^{-1} \int \frac{[f_0(x + \mu) - f_0(-x + \mu)]^2}{g(x + \mu) + g(-x + \mu)} dx. \quad (4.4)$$

**Proposition 4.1.** *If we assume that  $f$  is continuously differentiable on  $\mathbb{R}$ , the efficient information matrix  $\tilde{I}_{\theta, f}$  for estimating the parameter  $\theta = (p, \mu)$  is given by its components  $a_{i, j} (1 \leq i, j \leq 2)$ , where*

$$a_{11} = \frac{A}{2} \int \frac{[f_0(-x + \mu) - f_0(x + \mu)]^2 dx}{g(x + \mu) + g(-x + \mu)},$$

$$a_{12} = a_{21} = Ap \int \frac{f'(x)[f_0(x + \mu) - f_0(-x + \mu)] dx}{g(x + \mu) + g(-x + \mu)},$$

and

$$a_{22} = 2p^2 \int \frac{[f'(x)]^2 dx}{g(x + \mu) + g(-x + \mu)} + \frac{p^2(1 - p)^2}{2} \left[ \int s(x) dx \right]^{-1} \times \left( \int \frac{f'(x)[f_0(x + \mu) - f_0(-x + \mu)] dx}{g(x + \mu) + g(-x + \mu)} \right)^2.$$

Moreover, if  $f'$  and  $f_0(\cdot + \mu) - f_0(-\cdot + \mu)$  are not linearly dependent functions, the matrix  $\tilde{I}_{\theta, f}$  is nonsingular.

*Proof of Proposition 4.1.* Firstly, the ordinary score function  $\dot{l}_{\theta, f}$  is computed as

$$\dot{l}(x) = \begin{pmatrix} \dot{l}^{(1)}(x) \\ \dot{l}^{(2)}(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial p} \log g(x) \\ \frac{\partial}{\partial \mu} \log g(x) \end{pmatrix} = \frac{1}{g(x)} \begin{pmatrix} f(x - \mu) - f_0(x) \\ -pf'(x - \mu) \end{pmatrix}. \quad (4.5)$$

For a symmetric function  $h$  such that  $\int h(x)f(x)dx = 0$ , let us consider a density path in  $\mathcal{F}$  defined by  $f_t(x) = c(t)k(th(x))f(x)$  for  $t > 0$ , where we recall that  $k(x) = 2(1 + \exp(-2x))^{-1}$  and  $c(t)$  is a normalizing constant. For every  $x$ , we have

$$\left. \frac{\partial}{\partial t} \right|_{t=0} \log [(1-p)f_0(x) + pf_t(x - \mu)] = \frac{pf(x - \mu)h(x - \mu)}{g(x)}.$$

Then, we obtain a tangent set for  $f$ , and denote it by  $\dot{\mathcal{P}}_f$ ,

$$\dot{\mathcal{P}}_f = \left\{ \frac{pf(x - \mu)h(x - \mu)}{g(x)}; h \text{ is a symmetric function and } \int h(x)f(x)dx = 0 \right\}.$$

Relying on definition (4.3) of function  $s$ , we define a constant vector as

$$C_{\theta, f} = \left[ \int s(x)dx \right]^{-1} \int s(x) [\dot{l}(x + \mu) + \dot{l}(-x + \mu)] dx,$$

and a function vector

$$h_0(x) = \frac{s(x)}{pf(x)} [\dot{l}(x + \mu) + \dot{l}(-x + \mu) - C_{\theta, f}]. \quad (4.6)$$

Note that the two coordinate functions of  $h_0$  are symmetric functions and

$$\begin{aligned} \int h_0(x)f(x)dx &= \frac{1}{p} \int s(x) [\dot{l}(x + \mu) + \dot{l}(-x + \mu) - C_{\theta, f}] dx \\ &= \frac{1}{p} \int s(x) [\dot{l}(x + \mu) + \dot{l}(-x + \mu)] dx - \frac{1}{p} C_{\theta, f} \int s(x) dx = 0. \end{aligned}$$

Hence we get that the two coordinate functions of

$$\frac{pf(x - \mu)h_0(x - \mu)}{g(x)} \text{ belong to } \dot{\mathcal{P}}_f. \quad (4.7)$$

In the following calculation, note that for every integrable function  $t$  on  $\mathbb{R}$ , we have

$$\int t(x)dx = \frac{1}{2} \int [t(x) + t(-x)] dx. \quad (4.8)$$

We now aim at proving that

$$\dot{l}(x) - \frac{pf(x - \mu)h_0(x - \mu)}{g(x)} \perp \overline{\text{lin}(\dot{\mathcal{P}}_f)}. \quad (4.9)$$



Indeed, for every score function

$$\frac{pf(x-\mu)h(x-\mu)}{g(x)} \in \dot{\mathcal{P}}_f,$$

we have (since  $fh_0$  is symmetric and combining with the property (4.8))

$$\begin{aligned} & \int \left[ \dot{l}(x) - \frac{pf(x-\mu)h_0(x-\mu)}{g(x)} \right] \frac{pf(x-\mu)h(x-\mu)}{g(x)} g(x) dx \\ &= p \int \left[ \dot{l}(x+\mu) - \frac{pf(x)h_0(x)}{g(x+\mu)} \right] f(x)h(x) dx \\ &= \frac{p}{2} \int \left[ \dot{l}(x+\mu) + \dot{l}(-x+\mu) - \frac{pf(x)h_0(x)}{g(x+\mu)} - \frac{pf(-x)h_0(-x)}{g(-x+\mu)} \right] f(x)h(x) dx \\ &= \frac{p}{2} \int \left[ \dot{l}(x+\mu) + \dot{l}(-x+\mu) - \frac{pf(x)h_0(x)}{s(x)} \right] f(x)h(x) dx \end{aligned} \quad (4.10)$$

According to definition (4.6) of function  $h_0$ , we have

$$\dot{l}(x+\mu) + \dot{l}(-x+\mu) - \frac{pf(x)h_0(x)}{s(x)} = C_{\theta,f},$$

and combining with the fact that  $\int h(x)f(x)dx = 0$ , we get that the integral (4.10) is equal to 0. This proves (4.9). According to (4.7) and (4.9), the efficient score function for  $\theta$  is equal to

$$\begin{aligned} \tilde{l}(x) &= \dot{l}(x) - \frac{pf(x-\mu)h_0(x-\mu)}{g(x)} \\ &= \dot{l}(x) - \frac{s(x-\mu)[\dot{l}(x) + \dot{l}(-x+2\mu) - C_{\theta,f}]}{g(x)}. \end{aligned}$$

Relying on definition (4.3) of function  $s$ , we have

$$\frac{s(x-\mu)}{g(x)} = \frac{g(-x+2\mu)}{g(x) + g(-x+2\mu)},$$

hence

$$\begin{aligned} \tilde{l}(x) &= \dot{l}(x) - \frac{g(-x+2\mu)[\dot{l}(x) + \dot{l}(-x+2\mu) - C_{\theta,f}]}{g(x) + g(-x+2\mu)} \\ &= \frac{\dot{l}(x)g(x) - \dot{l}(-x+2\mu)g(-x+2\mu) + C_{\theta,f}g(-x+2\mu)}{g(x) + g(-x+2\mu)}. \end{aligned}$$

The first coordinate of  $\tilde{l}$  is then simplified as (combining with Equation (4.5) and the symmetry of  $f$ )

$$\begin{aligned} \tilde{l}^{(1)}(x) &= \frac{\dot{l}^{(1)}(x)g(x) - \dot{l}^{(1)}(-x+2\mu)g(-x+2\mu) + C_{\theta,f}^{(1)}g(-x+2\mu)}{g(x) + g(-x+2\mu)} \\ &= \frac{f_0(-x+2\mu) - f_0(x) + C_{\theta,f}^{(1)}g(-x+2\mu)}{g(x) + g(-x+2\mu)}, \end{aligned}$$

where the constant  $C_{\theta,f}^{(1)}$  is computed as the following (combining with definition (4.3) of function  $s$ , Equation (4.5) and the property (4.8))

$$\begin{aligned} C_{\theta,f}^{(1)} \int s(x)dx &= \int s(x) [i^{(1)}(x+\mu) + i^{(1)}(-x+\mu)] dx \\ &= \int \frac{[f(x) - f_0(x+\mu)]g(-x+\mu) + [f(x) - f_0(-x+\mu)]g(x+\mu)}{g(x+\mu) + g(-x+\mu)} dx \end{aligned} \quad (4.11)$$

$$\begin{aligned} &= \int [f(x) - f_0(x+\mu)] dx + \int \frac{[f_0(x+\mu) - f_0(-x+\mu)]g(x+\mu)}{g(x+\mu) + g(-x+\mu)} dx \\ &= \int \frac{[f_0(x+\mu) - f_0(-x+\mu)] [(1-p)f_0(x+\mu) + pf(x)]}{g(x+\mu) + g(-x+\mu)} dx \\ &= (1-p) \int \frac{[f_0(x+\mu) - f_0(-x+\mu)] f_0(x+\mu)}{g(x+\mu) + g(-x+\mu)} dx \\ &= \frac{1-p}{2} \int \frac{[f_0(x+\mu) - f_0(-x+\mu)]^2}{g(x+\mu) + g(-x+\mu)} dx. \end{aligned} \quad (4.12)$$

And the second coordinate of  $\tilde{l}$  is simplified as (since  $f'$  is odd)

$$\begin{aligned} \tilde{l}^{(2)}(x) &= \frac{i^{(2)}(x)g(x) - i^{(2)}(-x+2\mu)g(-x+2\mu) + C_{\theta,f}^{(2)}g(-x+2\mu)}{g(x) + g(-x+2\mu)} \\ &= \frac{-2pf'(x-\mu) + C_{\theta,f}^{(2)}g(-x+2\mu)}{g(x) + g(-x+2\mu)}, \end{aligned}$$

where the constant  $C_{\theta,f}^{(2)}$  is also computed as the following

$$\begin{aligned} C_{\theta,f}^{(2)} \int s(x)dx &= \int s(x) [i^{(2)}(x+\mu) + i^{(2)}(-x+\mu)] dx \\ &= p \int \frac{g(x+\mu)g(-x+\mu)}{g(x+\mu) + g(-x+\mu)} \left[ \frac{-pf'(x)}{g(x+\mu)} + \frac{-pf'(-x)}{g(-x+\mu)} \right] dx \\ &= p \int \frac{f'(x)[g(x+\mu) - g(-x+\mu)]}{g(x+\mu) + g(-x+\mu)} dx. \end{aligned} \quad (4.13)$$

According to (4.1), we have

$$g(x+\mu) - g(-x+\mu) = (1-p)[f_0(x+\mu) - f_0(-x+\mu)], \quad (4.14)$$

hence

$$C_{\theta,f}^{(2)} \int s(x)dx = p(1-p) \int \frac{f'(x)[f_0(x+\mu) - f_0(-x+\mu)]}{g(x+\mu) + g(-x+\mu)} dx. \quad (4.15)$$

Let us now calculate the efficient information matrix

$$\tilde{I}_{\theta,f} = \mathbb{P}_{\theta,f}(\tilde{l} \tilde{l}^t) = \mathbb{P}_{\theta,f}(\tilde{l} \tilde{l}^t) = (a_{ij})_{i,j=1,2}.$$

The components  $a_{ij}$  are given by

$$\begin{aligned}
 a_{11} &= \mathbb{P}_{\theta,f}(\tilde{l}^{(1)}i^{(1)}) = \int \tilde{l}^{(1)}(x)i^{(1)}(x)g(x)dx \\
 &= \int \frac{f_0(-x+2\mu) - f_0(x) + C_{\theta,f}^{(1)}g(-x+2\mu)}{g(x) + g(-x+2\mu)} [f(x-\mu) - f_0(x)]dx \\
 &= \int \frac{f_0(-x+\mu) - f_0(x+\mu)}{g(x+\mu) + g(-x+\mu)} [f(x) - f_0(x+\mu)]dx \\
 &\quad + C_{\theta,f}^{(1)} \int \frac{[f(x) - f_0(x+\mu)]g(-x+\mu)dx}{g(x+\mu) + g(-x+\mu)} \\
 &= \frac{1}{2} \int \frac{f_0(-x+\mu) - f_0(x+\mu)}{g(x+\mu) + g(-x+\mu)} [f(x) - f_0(x+\mu) - f(-x) + f_0(-x+\mu)]dx \\
 &\quad + \frac{1}{2} C_{\theta,f}^{(1)} \int \frac{[f(x) - f_0(x+\mu)]g(-x+\mu) + [f(-x) - f_0(-x+\mu)]g(x+\mu)}{g(x+\mu) + g(-x+\mu)} dx.
 \end{aligned}$$

Since  $f$  is even and according to the calculation of (4.11), we have

$$\begin{aligned}
 a_{11} &= \frac{1}{2} \int \frac{[f_0(-x+\mu) - f_0(x+\mu)]^2 dx}{g(x+\mu) + g(-x+\mu)} \\
 &\quad + \frac{1-p}{4} C_{\theta,f}^{(1)} \int \frac{[f_0(x+\mu) - f_0(-x+\mu)]^2 dx}{g(x+\mu) + g(-x+\mu)} \\
 &= \frac{1}{2} \left( 1 + \frac{1-p}{2} C_{\theta,f}^{(1)} \right) \int \frac{[f_0(-x+\mu) - f_0(x+\mu)]^2 dx}{g(x+\mu) + g(-x+\mu)}.
 \end{aligned}$$

According to definition (4.4) of constant  $A$  and (4.12), we have

$$A = 1 + \frac{1-p}{2} C_{\theta,f}^{(1)},$$

hence

$$a_{11} = \frac{A}{2} \int \frac{[f_0(-x+\mu) - f_0(x+\mu)]^2 dx}{g(x+\mu) + g(-x+\mu)}.$$

The components  $a_{12}$  and  $a_{22}$  is calculated in the same way

$$\begin{aligned}
 a_{12} &= a_{21} = \mathbb{P}_{\theta,f}(\tilde{l}^{(1)}i^{(2)}) = \int \tilde{l}^{(1)}(x)i^{(2)}(x)g(x)dx \\
 &= \int \frac{f_0(-x+2\mu) - f_0(x) + C_{\theta,f}^{(1)}g(-x+2\mu)}{g(x) + g(-x+2\mu)} [-pf'(x-\mu)]dx \\
 &= p \int \frac{f'(x)[f_0(x+\mu) - f_0(-x+\mu)]dx}{g(x+\mu) + g(-x+\mu)} - pC_{\theta,f}^{(1)} \int \frac{f'(x)g(-x+\mu)}{g(x+\mu) + g(-x+\mu)} dx \\
 &= p \int \frac{f'(x)[f_0(x+\mu) - f_0(-x+\mu)]dx}{g(x+\mu) + g(-x+\mu)} + \frac{p}{2} C_{\theta,f}^{(1)} \int \frac{f'(x)[g(x+\mu) - g(-x+\mu)]}{g(x+\mu) + g(-x+\mu)} dx \\
 &= p \left( 1 + \frac{1-p}{2} C_{\theta,f}^{(1)} \right) \int \frac{f'(x)[f_0(-x+\mu) - f_0(x+\mu)]dx}{g(x+\mu) + g(-x+\mu)} \quad (\text{according to (4.14)}) \\
 &= pA \int \frac{f'(x)[f_0(-x+\mu) - f_0(x+\mu)]dx}{g(x+\mu) + g(-x+\mu)},
 \end{aligned}$$

and

$$\begin{aligned}
 a_{22} &= \mathbb{P}_{\theta,f}(\tilde{l}^{(2)}i^{(2)}) = \int \tilde{l}^{(2)}(x)i^{(2)}(x)g(x)dx \\
 &= \int \frac{2pf'(x-\mu) - C_{\theta,f}^{(2)}g(-x+2\mu)}{g(x)+g(-x+2\mu)} [pf'(x-\mu)]dx \\
 &= 2p^2 \int \frac{[f'(x)]^2 dx}{g(x+\mu)+g(-x+\mu)} - pC_{\theta,f}^{(2)} \int \frac{f'(x)g(-x+\mu)dx}{g(x+\mu)+g(-x+\mu)} \\
 &= 2p^2 \int \frac{[f'(x)]^2 dx}{g(x+\mu)+g(-x+\mu)} - \frac{p}{2} C_{\theta,f}^{(2)} \int \frac{f'(x)[g(-x+\mu)-g(x+\mu)]dx}{g(x+\mu)+g(-x+\mu)},
 \end{aligned}$$

According to (4.13) and (4.15), we get that

$$\begin{aligned}
 a_{22} &= 2p^2 \int \frac{[f'(x)]^2 dx}{g(x+\mu)+g(-x+\mu)} + \frac{p^2(1-p)^2}{2} \left[ \int s(x)dx \right]^{-1} \\
 &\quad \times \left( \int \frac{f'(x)[f_0(x+\mu)-f_0(-x+\mu)]dx}{g(x+\mu)+g(-x+\mu)} \right)^2.
 \end{aligned}$$

We now can deduce the determinant of the efficient information matrix  $\tilde{I}_{\theta,f}$  as the following

$$\begin{aligned}
 \det(\tilde{I}_{\theta,f}) &= p^2 A \left[ \int \frac{[f_0(-x+\mu)-f_0(x+\mu)]^2 dx}{g(x+\mu)+g(-x+\mu)} \int \frac{[f'(x)]^2 dx}{g(x+\mu)+g(-x+\mu)} \right. \\
 &\quad \left. - \left( \int \frac{f'(x)[f_0(x+\mu)-f_0(-x+\mu)]}{g(x+\mu)+g(-x+\mu)} \right)^2 \right]
 \end{aligned}$$

According to Cauchy-Schwarz inequality, we obtain that  $\det(\tilde{I}_{\theta,f}) \geq 0$  and the equality occurs if and only if  $f'$  and  $f_0(\cdot + \mu) - f_0(-\cdot + \mu)$  are linearly dependent functions.  $\square$

**Example (Gaussian mixture).** We consider the Gaussian mixture (4.2). Note that

$$f'(x) = -\frac{x}{\sigma^2\sqrt{2\pi}\sigma^2} \exp -\frac{x^2}{2\sigma^2},$$

and

$$f_0(x+\mu) - f_0(-x+\mu) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2} \left[ 1 - \exp(2\mu x) \right].$$

Then the functions  $f'$  and  $f_0(\cdot + \mu) - f_0(-\cdot + \mu)$  are not linearly dependent and the efficient information matrix  $\tilde{I}_{\theta,f}$  is nonsingular.

### 4.3 Perspectives

The previous section presents preliminary results on model (4.1). There are many issues that should be further studied. Firstly, let us note that [Bordes et al. \[2006\]](#) have only given

some sufficient conditions to obtain the identifiability of the parameters of model (4.1). Then, a perspective is to try to improve sufficient identifiability conditions or to study necessary and sufficient identifiability conditions. In the previous section, we have only calculated the efficient information matrix and [Bordes and Vandekerkhove \[2010\]](#) propose an asymptotically normal estimator of the unknown parameters. We will try to compare their asymptotic covariance matrix with the inverse of the efficient information matrix. Although these two matrices have complex forms, we can compare their performances on some simulated data. Moreover, we will try to apply the one-step method to find out an asymptotically efficient estimator of the Euclidean parameter  $\theta$  if it is possible.

# Bibliography

- David B. Allison, Gary L. Gadbury, Moonseong Heo, José R. Fernández, Cheol-Koo Lee, Tomas A. Prolla, and Richard Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.*, 39(1):1–20, 2002.
- J Aubert, A Bar-Hen, J-J Daudin, and S Robin. Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics*, 5(1):125, 2004.
- Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat. Ser.*, 25, 2000.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001.
- Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.
- Gilles Blanchard and Étienne Roquain. Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.*, 10:2837–2871, 2009.
- C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- L. Bordes and P. Vandekerkhove. Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Math. Methods Statist.*, 19(1):22–41, 2010.
- Laurent Bordes, Céline Delmas, and Pierre Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statist.*, 33(4):733–752, 2006.
- Per Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6(1):199, 2005.

## BIBLIOGRAPHY

---

- Alain Celisse and Stéphane Robin. Nonparametric density estimation by exact leave- $p$ -out cross-validation. *Comput. Statist. Data Anal.*, 52(5):2350–2368, 2008.
- Alain Celisse and Stéphane Robin. A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference*, 140(11):3132–3147, 2010.
- Gary Chamberlain. Asymptotic efficiency in semiparametric models with censoring. *J. Econometrics*, 32(2):189–218, 1986.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.
- Sandrine Dudoit and Mark J. van der Laan. *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. Springer, New York, 2008.
- Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160, 2001.
- P. P. B. Eggermont. Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. *Applied Mathematics & Optimization*, 39:75–91, 1999.
- P.P.B. Eggermont and V.N. LaRiccia. Maximum smoothed likelihood density estimation for inverse problems. *Ann. Stat.*, 23(1):199–220, 1995.
- P.P.B. Eggermont and V.N. LaRiccia. *Maximum penalized likelihood estimation. Vol. 1: Density estimation*. Springer Series in Statistics. New York, NY: Springer., 2001.
- Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):499–517, 2002.
- Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061, 2004.
- Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.
- Mickael Guedj, Stéphane Robin, Alain Celisse, and Gregory Nuel. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10(1):84, 2009.

## BIBLIOGRAPHY

---

- Nicolas W. Hengartner and Philip B. Stark. Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.*, 23(2):525–550, 1995.
- Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.
- Yosef Hochberg and Ajit C. Tamhane. *Multiple comparison procedures*. Wiley Series in Probability and Statistics. Wiley, 1 edition, 1987.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70, 1979.
- I. A. Ibragimov and R. Z. Hasminskii. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981. Asymptotic theory, Translated from the Russian by Samuel Kotz.
- Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. 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, 2005.
- E. L. Lehmann. *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition, 1986.
- Oleg Lepski. Lectures given for the paris-berlin seminar at garchy, 1998.
- M. Levine, D. R. Hunter, and D. Chauveau. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2):403–416, 2011.
- Kun Liang and Dan Nettleton. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(1):163–182, 2012.
- J.G. Liao, Yong Lin, Zachariah E. Selvanayagam, and Weichung Joe Shih. A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, 20(16):2694–2701, 2004.
- G.J. McLachlan, R.W. Bean, and L. Ben-Tovim Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22(13):1608–1615, 2006.



## BIBLIOGRAPHY

---

- Nicolai Meinshausen and Peter Bühlmann. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, 92(4): 893–907, 2005.
- Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1):373–393, 2006.
- Mathias O. Mosig, Ehud Lipkin, Galina Khutoreskaya, Elena Tchourzyna, Morris Soller, and Adam Friedmann. 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(4):1683–1698, 2001.
- Dan Nettleton, J.T.Gene Hwang, RicoA. Caldo, and RogerP. Wise. Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:337–356, 2006.
- Pierre Neuvial. Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. Technical report, arXiv:1003.0747, 2010.
- Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2): 99–135, 1990.
- V.H. Nguyen and C. Matias. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. Technical report, arXiv:1205.4097, 2012.
- Stan Pounds and Stephan W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.
- Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre. A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput. Statist. Data Anal.*, 51(12):5483–5493, 2007.
- Etienne Roquain. Type I error rate control for testing many hypotheses: a survey with proofs. *J. SFdS*, 152(2):3–38, 2011.
- T. Schweder and E. Spjøtvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.

## BIBLIOGRAPHY

---

- Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.*, 62:626–633, 1967.
- B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.
- John D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498, 2002.
- John D. Storey. The positive false discovery rate: a Bayesian interpretation and the  $q$ -value. *Ann. Statist.*, 31(6):2013–2035, 2003.
- John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205, 2004.
- Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303, 2008.
- Wenguang Sun and T. Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424, 2009.
- Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.*, 102(479):901–912, 2007.
- Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. New York, NY: Springer., 2009.
- F.E. Turkheimer, C.B. Smith, and K. Schmidt. Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage*, 13(5):920 – 930, 2001.
- A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.
- Aad van der Vaart. Semiparametric statistics. Bolthausen, Erwin et al., Lectures on probability theory and statistics. Ecole d’été de probabilités de Saint-Flour XXIX - 1999, Saint-Flour, France, July 8-24, 1999. Berlin: Springer. Lect. Notes Math. 1781, 331-457 (2002)., 2002.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment (Wiley Series in Probability and Statistics)*. Wiley, 1 edition, 1993.

Dominik Wied and Rafael Weißbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53:1–21, 2012.

# Adaptive estimation via Lepski's method

---

## A.1 Lepski's method

Lepski's method is a method for choosing a "best" estimator (in an appropriate sense) among a family of those, under suitable restrictions on this family. We recall here two selection procedures for choosing an adaptive estimator of a probability density function.

Let  $X$  be a random variable in  $D \subset \mathbb{R}$  having density function  $f$  with respect to the Lebesgue measure which is supposed to belong to a given set  $\mathcal{F}_\beta$  with an unknown parameter  $\beta$ . We want to estimate  $f$  on the basis of the i.i.d sample  $X_1, \dots, X_n$  drawn from  $f$ . The performance of any estimator  $\hat{f}_n$  is measured by the risk

$$R[\hat{f}; f] = \mathbb{E}_f[l^q(\hat{f}_n - f)],$$

where  $l(\cdot)$  is a semi-norm and  $q \geq 1$  is a given real number. If we choose the semi-norm  $l$  as the  $\mathbb{L}_p$ -norm  $l(g) = \|g\|_p$ , we come to the global estimation and the corresponding risk is given by

$$R[\hat{f}_n; f] = \mathbb{E}_f[\|\hat{f}_n - f\|_p^q], \quad p \in [1, +\infty].$$

If we aim at estimating function  $f$  at a fixed point  $x$ , then we come to the pointwise estimation and the semi-norm and the corresponding risk are defined as

$$l(g) = |g(x)| \quad \text{and} \quad R[\hat{f}_n; f] = \mathbb{E}_f[|\hat{f}_n(x) - f(x)|^q].$$

We then define the maximal rate  $R[\hat{f}_n; \beta]$  on  $\mathcal{F}_\beta$  of a given estimator  $\hat{f}_n$  and the minimax rate  $R_M[\beta]$  on  $\mathcal{F}_\beta$  as

$$R[\hat{f}_n; \beta] = \sup_{f \in \mathcal{F}_\beta} R[\hat{f}_n; f] = \sup_{f \in \mathcal{F}_\beta} \mathbb{E}_f[l(\hat{f}_n - f)]^q \quad \text{and} \quad R_M[\beta] = \inf_{\hat{f}_n} R[\hat{f}_n; \beta],$$

where the infimum is taken over all positive estimators. We say that two risks  $R$  and  $R'$  are equivalent and denote by  $R \asymp R'$  if  $\limsup_{n \rightarrow \infty} R/R' < +\infty$  and  $\limsup_{n \rightarrow \infty} R'/R < +\infty$ . An

estimator  $\hat{f}_n$  is called "rate asymptotically minimax" on  $\mathcal{F}_\beta$  if  $R[\hat{f}_n; \beta] \asymp R_M[\beta]$ . We now consider a family of parameter sets  $\{\mathcal{F}_\beta\}_{\beta \in \Xi}$ . Our goal is the following: from a family of rate asymptotically minimax estimators  $\{\hat{f}_{n,\beta}\}_{\beta \in \Xi}$ , how can we get an adaptive estimator over the family  $\{\mathcal{F}_\beta\}_{\beta \in \Xi}$ , i.e, construct a new estimator  $\tilde{f}_n$  independent of  $\beta$  which is simultaneously rate asymptotically minimax over all the set  $\mathcal{F}_\beta$ ? In other words, this estimator satisfies  $R[\tilde{f}_n; \beta] \asymp R_M[\beta]$  for all  $\beta \in \Xi$ . Lepski (1991) proposes a selection rule for choosing an adaptive estimator  $\hat{f}_{n,\hat{\beta}}$  to solve this problem. We simplify here his assumptions in the following way. His assumptions are essentially equivalent to

1.  $\Xi$  is a bounded subset of  $\mathbb{R}^+$ ;
2. the family  $\{\mathcal{F}_\beta\}_{\beta \in \Xi}$  is nondecreasing with respect to  $\beta$ ;
3. the minimax rates  $R_M[\beta]$  are continuous with respect to  $\beta$ ;
4. for each  $\beta \in \Xi$ , there is a rate asymptotically minimax estimator  $\hat{f}_{n,\beta}$  on  $\mathcal{F}_\beta$ ;
5. for  $n$  large enough and each  $\beta \in \Xi$ ,  $l^q(\hat{f}_{n,\beta} - f)$  is suitably concentrated around its expectation.

Lepski then chooses a suitable finite discretization  $\beta_1 < \dots < \beta_{N_n}$  of  $\Xi$  and defines  $\hat{\beta} = \beta_{\hat{j}}$ , where

$$\hat{j} = \inf \{j \leq N_n : l^q(\hat{f}_{n,\beta_j} - \hat{f}_{n,\beta_k}) \leq KR[\hat{f}_{n,\beta_k}; \beta_k], \forall k, j < k \leq N_n\},$$

for some given large enough constant  $K$ . He shows that the adaptive estimator  $\tilde{f}_n = \hat{f}_{n,\hat{\beta}}$  is simultaneously rate asymptotically minimax over all the sets  $\mathcal{F}_\beta$ . This method has been applied in various contexts and by many authors. Recently, Lepski has improved his method by relaxing the monotonicity assumptions (Lepski [1998]). In particular, he proposes a new general selection rule from a family of linear estimators. We can refer this method to Goldenshluger and Lepski [2011].

## A.2 Perspectives

In this section, we try to apply Lepski's method to propose an adaptive estimator of the density  $f$  in the mixture model (3.2) based on the  $p$ -values. Firstly, we recall the randomly weighted kernel estimator of  $f$  defined as (3.11). We define the bandwidth  $h_\beta$  depending on a parameter  $\beta$  given by

$$h_\beta = C_1 n^{\frac{-1}{2\beta+1}}, \quad \text{where } C_1 \text{ is a known positive constant.}$$

Let  $K$  be a kernel and  $\hat{\theta}_n$  be a given estimator of  $\theta$ , then, by defining the weight

$$\hat{\tau}_i = \hat{\tau}(X_i) = 1 - \frac{\hat{\theta}_n}{\tilde{g}_n(X_i)}, \text{ where } \tilde{g}_n(X_i) = \frac{1}{(n-1)h_\beta} \sum_{j \neq i}^n K\left(\frac{X_i - X_j}{h_\beta}\right),$$

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

$$\forall x \in [0, 1], \hat{f}_{n,\beta}(x) = \frac{1}{h_\beta} \sum_{i=1}^n \frac{\hat{\tau}_i}{\sum_{k=1}^n \hat{\tau}_k} K\left(\frac{x - X_i}{h_\beta}\right).$$

We suppose that the density  $f$  belongs to a Hölder class  $\Sigma(\beta, L)$ , where  $\beta$  is assumed to belong to a discrete set  $B_{N_n} = \{\beta_1, \beta_2, \dots, \beta_{N_n}\}$ . We evaluate the regularity  $\beta$  of the estimated density  $f$  and replace it into the kernel estimator  $\hat{f}_{n,\beta}$  in order to obtain an adaptive estimator. More precisely, let  $C_2 > 0$  be a sufficiently large constant and we define

$$\hat{\beta} = \max \left\{ \beta \in B_{N_n} : |\hat{f}_{n,\beta}(x_0) - \hat{f}_{n,\gamma}(x_0)| \leq C_2 \left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}, \forall \gamma < \beta, \gamma \in B_{N_n} \right\}.$$

Finally, we define

$$\hat{f}_n^*(x_0) = \hat{f}_{n,\hat{\beta}}(x_0).$$

We would like to prove that under some assumptions on the sequence  $\{N_n\}$ , the set  $B_{N_n}$ , the given estimator  $\hat{\theta}_n$  and the kernel  $K$ , we have

$$\limsup_{n \rightarrow \infty} \sup_{\beta \in B_{N_n}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_{\theta, f} \left[ \left(\frac{n}{\log n}\right)^{\frac{2\beta}{2\beta+1}} |\hat{f}_n^*(x_0) - f(x_0)|^2 \right] < +\infty.$$

In fact, note that when we study the quadratic risk of the kernel density estimators  $\hat{f}_{n,\beta}(x_0)$  (Section 3.5.1) and  $\hat{f}_n^*(x_0)$ , we encounter a situation that is more difficult than the case of classical kernel density estimators. It comes from the fact that the weights in kernel estimator  $\hat{f}_{n,\beta}(x_0)$  are random and dependent. The main problem for applying the previous strategy is that we would need to apply some exponential inequalities (Hoeffding's or Bernstein's inequalities), in a non-i.i.d framework. Nonetheless, we hope that we will improve our calculations or apply different approaches to solve this problem. Besides, we will try to see the performance of the adaptive estimator  $\hat{f}_n^*(x_0)$  on simulated data.