Electronic Journal of Statistics
Vol. 8 (2014) 2879–2904
ISSN: 1935-7524
DOI: 10.1214/14-EJS976

Adaptive density estimation in deconvolution problems with unknown error distribution

Johanna Kappus

Institut für Mathematik, Universität Rostock
18051 Rostock, Germany
e-mail: johanna.kappus@uni-rostock.de

and

Gwennaëlle Mabon

CREST – ENSAE
3 avenue Pierre Larousse
92245 Malakoff, France
&
MAP5*, Université Paris Descartes*
45 rue des Saints-Pères
75006 Paris, France
e-mail: gwennaelle.mabon@ensae.fr

Abstract: We investigate the data driven choice of the cutoff parameter in density deconvolution problems with unknown error distribution. To make the target density identifiable, one has to assume that some additional information on the noise is available. We consider two different models: the framework where some additional sample of the pure noise is available, as well as the model of repeated measurements, where the contaminated random variables of interest can be observed repeatedly, with independent errors. We introduce spectral cutoff estimators and present upper risk bounds. The focus of this work lies on the optimal choice of the bandwidth by penalization strategies, leading to non-asymptotic oracle bounds.

MSC 2010 subject classifications: Primary 62G07; secondary 62G99.
Keywords and phrases: Adaptive estimation, deconvolution, density estimation, mean squared risk, nonparametric methods, replicate observations.

Received December 2013.

Contents

1 Introduction 2880
2 Statistical model, estimation procedure and risk bounds 2882
2.1 Notations 2882
2.2 Statistical model and estimators 2882
2.3 Upper risk bounds 2885

3 Data driven bandwidth selection and oracle bounds 2886
4 Illustrations 2888
4.1 Practical estimation procedure 2888
4.2 Comparison with [11] and influence of M in the NS-model . . . 2889
4.3 Illustrations in the RD-model 2891
4.4 Comparison with a kernel estimator 2891
5 Concluding remarks 2893
6 Proofs 2895
6.1 Preliminaries 2895
6.2 A technical auxiliary result 2896
6.3 Proof of the oracle bounds 2901
Acknowledgments 2902
References 2902

1. Introduction

This paper addresses the problem of the adaptive bandwidth selection via penalization in deconvolution problems with unknown error distribution. We study in parallel two different models. In both models, we assume that the random variables are real-valued.

Model 1. The random variable X of interest is perturbed by some additional additive error ε, independent of X, so the empirically accessible quantity is $Y = X + \varepsilon$. The distribution of the noise is assumed to be unknown. One observes n independent copies of Y:

$$Y_j = X_j + \varepsilon_j,\ j = 1, \dots, n. \tag{1}$$

In addition it is assumed that a sample $(\varepsilon_{-j})_{j=1,\dots,M}$ of the pure noise, independent of the Y_j, is available.

Model 2. The noisy random variable X can be measured repeatedly, with independent errors. The observations are then of the form,

$$Y_{j,k} = X_j + \varepsilon_{j,k},\ j = 1, \dots, n \quad \text{and} \quad k = 2. \tag{2}$$

All the X_j are assumed to be independent and identically distributed and all the $\varepsilon_{j,k}$ are independent and identically distributed and independent of the X_j. Again, it is assumed that the distribution of the $\varepsilon_{j,k}$ is unknown. In addition, the error terms are assumed to be centered.

Density deconvolution is a classical topic in nonparametric statistics and a large amount of literature on this subject has been published since the late 80s. Rates of convergence and their optimality have been studied, for example, in [9, 32, 33, 21] and [20]. For the study of sharp asymptotic optimality, see [5, 7, 8]. Adaptive estimation for deconvolution problems has then been investigated by [31], who apply wavelet techniques, by [13], who consider the adaptive bandwidth selection for projection estimators and by [6] for linear functionals. We can also cite [12] in a multivariate setting.

However, the above mentioned papers have been working under the assumption that the distribution of the errors is perfectly known, which is clearly not realistic in most fields of application. The systematic study of deconvolution with unknown error distribution in presence of an additional noise sample, which corresponds to Model 1, dates back to the late 90s. For the study of convergence rates, see [30, 22] or [27].

The rigorous study of adaptive procedures in a deconvolution model with unknown errors has only recently been addressed. We are aware of the work by [11], by [23] who consider a model of circular deconvolution and by [17], who deal with adaptive quantile estimation via Lespki's method.

In comparison to the classical deconvolution model with known errors, the research on the model of repeated measurements has not been so intense. For the more general model of repeated measurements with skew error densities, rates of convergence have been studied in [25] and recently been improved in [10]. Consistent estimation under minimal a priori assumptions is investigated in [28]. For repeated measurements with symmetric error density, we refer to [18] and [16]. This model has various applications in economics, see, for example, [4], but also in a medical context, see [16].

In the last mentioned paper, as well as in [18], practical strategies for the adaptive bandwidth selection have been proposed, but a theoretical justification is not given which is a motivation for the rigorous study presented in this paper. Essential tools for our approach rely on considerations presented in [24]. There are also many common points with recent contributions by [17]. In comparison to the last mentioned authors, the main difference lies in the fact that their approach is minimax and asymptotic whereas we are interested in non-asymptotic oracle bounds, thus following the model selection paradigma in the sense of [2] and [26].

Model 1 corresponds, in many respects, to the situation which has been investigated in [11]. Let us clarify the essential differences: in a deconvolution model with estimated characteristic function of the errors, the risk bounds are determined by the size M of the noise sample, as well as the number n of observations of Y. For sample sizes $M \geq n$ the risk bounds correspond to the model with known error distribution. However, for $M < n$ the bound on the risk gets worse. The approach by [11] is tailored for the case where M is, by a polynomial factor, larger than n, and cannot be extended to $M < n$. The additional considerations presented in the present work allow to handle the case of small noise samples. This seems to be of some practical relevance when one turns away from the classical measurement error model and regards, in some context of physics or biology, ε as a signal overlying some other signal X. In such a framework, the size of the noise sample will be determined by some extraneous influence and the assumption that $M > n$ may fail to hold true.

We want to emphasize another important difference between our reasoning and the arguments given in [11], but also in [17]. The last mentioned papers do always work under the standing assumption that the interesting density and error density belong to certain prescribed classes of functions. More precisely, it is assumed therein that the characteristic functions have an exponential or polynomial decay behavior. We are able to dispose completely of any of such semi-parametric assumptions.

This is motivated by arguments given in [2]. From a model selection point of view, rather than considering a family of parameter sets and aiming at building an estimator which is simultaneously asymptotically minimax, the target is to find the best estimator within a collection leading to non-asymptotic oracle inequalities. It is argued in [2] that these considerations make sense without specifying any particular family of parameter sets. From this point of view, it is desirable to avoid, as far as possible, any a priori parametric assumptions and provide a fully general treatment.

This paper is organized as follows. In Section 2, we fix the notation and assumptions, introduce the estimators and present upper risk bounds. In Section 3, the data driven choice of the cutoff parameter is investigated. We introduce penalized criteria and derive non asymptotic oracle bounds for the corresponding estimators. In Section 4, we present some data examples to illustrate the practical performance of our estimator. All proofs are postponed to Section 6.

2. Statistical model, estimation procedure and risk bounds

In the present section we fix the statistical model and assumptions, introduce the estimators and recall, for the readers convenience, the non asymptotic risk bounds which have been presented in earlier publications on the subject. We start by introducing some notation which will be used throughout the rest of the text.

2.1. Notations

For two real numbers a and b, $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. For two functions φ, $\psi : \mathbb{R} \to \mathbb{C}$ belonging to $\mathbb{L}^1(\mathbb{R}) \cap \mathbb{L}^2(\mathbb{R})$, $\|\varphi\|$ denotes the $\mathbb{L}^2$-norm of φ, defined by $\|\varphi\|^2 = \int_{\mathbb{R}} |\varphi(x)|^2 \,\mathrm{d}x$, and $\langle \varphi, \psi \rangle$ the scalar product between φ and ψ, defined by $\langle \varphi, \psi \rangle = \int_{\mathbb{R}} \varphi(x)\overline{\psi(x)} \,\mathrm{d}x$. The Fourier transform φ^* is defined by

$$\varphi^*(x) = \int e^{ixu} \varphi(u) \,\mathrm{d}u.$$

Besides, if φ^* belongs to $\mathbb{L}^1(\mathbb{R}) \cap \mathbb{L}^2(\mathbb{R})$, then the function φ is the inverse Fourier transform of φ^* and can be written $\varphi(x) = 1/(2\pi) \int e^{-ixu} \varphi^*(u) \,\mathrm{d}u$. The convolution product $*$ is defined as $(\varphi * \psi)(x) = \int \varphi(x-u)\psi(u) \,\mathrm{d}u$. Lastly, we abbreviate Model 1 by NS-model, where NS stands for noise sample, and Model 2 by RD-model for replicate data.

2.2. Statistical model and estimators

In the situation of Model 1 and Model 2 defined in the introductory part, the target is to recover the density f of X from the data. In the sequel, we limit our considerations to the case where the number k of repeated measurements is equal to two. This setting easily generalizes to a higher dimensional model. However, for sake of clarity, we omit the details.

Throughout the paper, we make the following assumptions:

(A1) X and ε have square integrable densities f and f_ε w.r.t. Lebesgue measure.
(A2) $\forall x \in \mathbb{R}$, $f_\varepsilon^*(x) \neq 0$.
(A3) f_ε is symmetric around zero.

Let f_Y denote the density of $X + \varepsilon$. By independence of X and ε, $f_Y = f * f_\varepsilon$. Under **(A2)**, we have the equality $f^* = f_Y^*/f_\varepsilon^*$. An estimator $\hat{f}_Y^*$ of f_Y can be calculated from the data. If f_ε^* is known, an unbiased plug-in-estimator of f^* is then given by $\hat{f}_Y^*/f_\varepsilon^*$. The inverse Fourier transform is then applied to get an estimate of f. However, since neither $\hat{f}_Y^*$ nor $1/f_\varepsilon^*$ are integrable, it is necessary to apply some regularization technique, for example, a spectral cutoff estimator. In this particular case, the estimator of f would be $1/(2\pi)\int_{|u|\leq \pi m} e^{-iux}\hat{f}_Y^*(u)/f_\varepsilon^*(u)\,\mathrm{d}u$. We can notice that this estimator corresponds both to a kernel estimator built with a sinc kernel ([5]) or to a projection type estimator as in [13].

In the present case, the error distribution is assumed to be unknown. To make the problem identifiable, some additional information on the noise is required. In the NS-model, we introduce the empirical characteristic function of ε,

$$\forall u \in \mathbb{R},\ \hat{f}_{\varepsilon,\mathrm{NS}}^*(u) = \frac{1}{M}\sum_{j=1}^{M} e^{iu\varepsilon_{-j}}.$$

In the same way, f_Y^* is estimated by its empirical version

$$\forall u \in \mathbb{R},\ \hat{f}_{Y,\mathrm{NS}}^*(u) = \frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}. \tag{3}$$

Secondly the identification of f_ε^* is also possible in the model of repeated observations. Under the symmetry assumption **(A3)**, we have the following equalities

$$\begin{aligned}\forall u \in \mathbb{R},\ \mathbb{E}\left[e^{iu(Y_{n+j,1}-Y_{n+j,2})}\right] &= \mathbb{E}\left[e^{iu(\varepsilon_{n+j,1}-\varepsilon_{n+j,2})}\right] = \left|\mathbb{E}\left[e^{iu\varepsilon_{n+j,1}}\right]\right|^2 \\ &= \left(\mathbb{E}\left[e^{iu\varepsilon_{n+j,1}}\right]\right)^2 = \left(f_\varepsilon^*(u)\right)^2. \end{aligned} \tag{4}$$

When **(A3)** is violated, the model is much more complicated and requires a completely different approach since f_ε^* is not a real positive function anymore. For further discussion, see [10]. Formula (4) suggests to define the following unbiased estimator of $(f_\varepsilon^*)^2$:

$$\forall u \in \mathbb{R},\ \widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u) = \frac{1}{n}\sum_{j=1}^{n}\cos\left(u(Y_{j,1}-Y_{j,2})\right). \tag{5}$$

Moreover, an unbiased estimator of f_Y^* is given by

$$\forall u \in \mathbb{R},\ \hat{f}^*_{Y,\mathrm{RD}}(u) = \frac{1}{2n}\sum_{j=1}^{n}\left(e^{iuY_{j,1}} + e^{iuY_{j,2}}\right). \tag{6}$$

When considering the empirical characteristic functions, one has to be careful about the fact that the $Y_{j,k}$ are not independent of each other, nor independent of the $\varepsilon_{j,k}$.

Finally, one has to pay attention to the fact that small values of the empirical characteristic function in the denominator lead to unfavorable effects and a bad performance of the estimator. This is an immediate consequence of the fact that a reasonable estimation of the ratio $1/f_\varepsilon^*$ is impossible as soon as the denominator is smaller than the standard deviation. This phenomenon has been investigated in [30]. This entails the necessity to consider a regularized version of the empirical characteristic function in the denominator. [18] propose a ridge-parameter approach. However, this requires a careful discussion of the choice of the ridge parameter. For this reason, we prefer the completely data driven approach proposed in [30], which has also been applied by [11] for the NS-model and by [16] for the RD-model. This approach corresponds to the following estimators of the ratio:

$$\frac{1}{\tilde{f}^*_{\varepsilon,\mathrm{NS}}(x)} = \frac{\mathbb{1}\left\{|\hat{f}^*_{\varepsilon,\mathrm{NS}}(x)| \geq M^{-1/2}\right\}}{\hat{f}^*_{\varepsilon,\mathrm{NS}}(x)} \quad \text{and} \quad \frac{1}{\tilde{f}^*_{\varepsilon,\mathrm{RD}}(x)} := \frac{\mathbb{1}\left\{\widehat{f^{*2}_{\varepsilon,\mathrm{RD}}}(x) \geq n^{-1/2}\right\}}{\sqrt{\widehat{f^{*2}_{\varepsilon,\mathrm{RD}}}(x)}}.$$

These definitions lead to defining the empirical versions of f^* as follows:

$$\check{f}^*_{\mathrm{NS}}(u) := \frac{\hat{f}^*_{Y,\mathrm{NS}}(u)}{\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)} \quad \text{and} \quad \check{f}^*_{\mathrm{RD}}(u) := \frac{\hat{f}^*_{Y,\mathrm{RD}}(u)}{\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)}.$$

Finally, the objects to be estimated are characteristic functions, so the absolute values are bounded by 1. For this reason, one should prefer the regularized versions

$$\hat{f}^*_{\mathrm{NS}}(u) := \frac{\check{f}^*_{\mathrm{NS}}(u)}{\max\{|\check{f}^*_{\mathrm{NS}}(u)|, 1\}} \quad \text{and} \quad \hat{f}^*_{\mathrm{RD}}(u) := \frac{\check{f}^*_{\mathrm{RD}}(u)}{\max\{|\check{f}^*_{\mathrm{RD}}(u)|, 1\}}.$$

Given any positive, real valued m, the above considerations lead to defining the spectral cutoff estimators of f as follows:

$$\hat{f}_{m,\mathrm{NS}}(x) := \frac{1}{2\pi}\int e^{-iux}\hat{f}^*_{m,\mathrm{NS}}(u)\,\mathrm{d}u \quad \text{and} \quad \hat{f}_{m,\mathrm{RD}}(x) := \frac{1}{2\pi}\int e^{-iux}\hat{f}^*_{m,\mathrm{RD}}(u)\,\mathrm{d}u, \tag{7}$$

with

$$\hat{f}^*_{m,\mathrm{NS}}(u) := \hat{f}^*_{\mathrm{NS}}(u)\mathbb{1}_{[-\pi m,\pi m]}(u) \quad \text{and} \quad \hat{f}^*_{m,\mathrm{RD}}(u) := \hat{f}^*_{\mathrm{RD}}(u)\mathbb{1}_{[-\pi m,\pi m]}(u).$$

Moreover, we use the notation

$$f_m(x) := \frac{1}{2\pi}\int e^{-iux}f^*_m(u)\,\mathrm{d}u \quad \text{and} \quad f^*_m(u) := f^*(u)\mathbb{1}_{[-\pi m,\pi m]}(u).$$

2.3. Upper risk bounds

The following non-asymptotic risk bounds are valid for the estimators defined in the preceding section.

Proposition 2.1.

*(i) In presence of an additional noise sample, under **(A1)–(A2)**, there exists a universal positive constant C such that*

$$\mathbb{E}\left\|f-\hat{f}_{m,\mathrm{NS}}\right\|^2 \leq 2\left\|f-f_m\right\|^2 + C\left(\frac{1}{n}\int_{-\pi m}^{\pi m}\frac{1}{|f_\varepsilon^*(u)|^2}\,\mathrm{d}u + \frac{1}{M}\int_{-\pi m}^{\pi m}\frac{|f^*(u)|^2}{|f_\varepsilon^*(u)|^2}\,\mathrm{d}u\right). \tag{8}$$

*(ii) In the model of repeated measurements, under **(A1)–(A3)**, there exists a universal positive constant C' such that*

$$\mathbb{E}\left\|f-\hat{f}_{m,\mathrm{RD}}\right\|^2 \leq 2\left\|f-f_m\right\|^2 + \frac{C'}{n}\int_{-\pi m}^{\pi m}\frac{1}{|f_\varepsilon^*(u)|^2}\,\mathrm{d}u + \frac{C'}{n}\int_{-\pi m}^{\pi m}\frac{|f^*(u)|^2}{|f_\varepsilon^*(u)|^2(|f_\varepsilon^*(u)|^2\vee n^{-1/2})}\,\mathrm{d}u. \tag{9}$$

Remark 1. The first two terms on the right-hand side of Equations (8) and (9) correspond to the usual terms when the distribution of the errors is known: the squared bias term and a bound on the variance. The last term is due to the estimation of f_ε^* which depends on the considered model. These bounds have already been established in the literature on deconvolution, see [30, 18] or [11, 16]. We can hence omit the proof.

In view of the rates of convergence, we observe the following: consider first the NS-model. If $M \geq n$, there is no loss in the rate in comparison to a deconvolution problem with known f_ε. However, a loss in the rate may occur for $M < n$. More precisely, if the ratio M/n is small, in comparison to f^*/f_ε^*, the estimator does not achieve the rates of convergence which are known to be optimal for deconvolution with known error distribution. This is intuitive, since f can only be identified through f_ε and there is no hope to estimate f with high precision when the information on f_ε is not reliable. For a detailed discussion and minimax lower bounds, we refer to [30]. Next, consider the RD-model. From Equation (9), one derives immediately, that there is no loss in the rate, in comparison to deconvolution with known error distribution, if the decay of f^* outbalances the decay of f_ε^*. If this is no longer true, the following holds: the smoother f_ε is, in comparison to f, the worse are the resulting rates of convergence. It can be shown that a loss in the rate is unavoidable in this context (Alexander Meister, personal communication), but to the best of our knowledge, minimax lower bounds have not been published for this particular case.

3. Data driven bandwidth selection and oracle bounds

The goal of this section is to provide a strategy for the optimal data driven choice of the smoothing parameter. Given a collection $\mathcal{M}$ of cutoff parameters, which may vary with M and n, the bandwidth $\widehat{m}$ should ideally outbalance the bias and variance term displayed in Equations (8) and (9). This trade-off is easier to realize when the variance term is assumed to be known, which is often the case in the literature on model selection. In a deconvolution problem with perfectly known error distribution $\widehat{m}$ should mimic the oracle choice

$$m_{th} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ -\|f_m\|^2 + \frac{1}{n}\frac{1}{2\pi}\int_{-\pi m}^{\pi m} \frac{(1-|f_Y^*(u)|^2)}{|f_\varepsilon^*(u)|^2}\,\mathrm{d}u \right\}.$$

In the present framework, the considerations are even more involved since the characteristic function in the denominator is unknown and the variance is hence not feasible to actually compute. Following the model selection paradigm, see [3, 2] or [26], we select $\widehat{m}$ as the minimizer of a penalized criterion such that

$$\widehat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ -\|\hat{f}_m\|^2 + \widehat{\mathrm{pen}}(m) \right\}.$$

The penalty term should be chosen large enough to annihilate the fluctuation of $\hat{f}_m$ around its target, for all m in the model collection $\mathcal{M}$ simultaneously, but on the other hand, should ideally be as close as possible to the variance term in order to preserve the non-asymptotic risk bounds. In a model selection problem with known variance, the penalty term is deterministic, which is no longer the case in the present situation.

Before introducing the stochastic penalty terms, we shall need the following definitions: for $\delta > 0$ and $u \in \mathbb{R}$,

$$w(u) := (\log(e + |u|))^{-\frac{1}{2}-\delta}.$$

Moreover

$$k_N(u) := N^{-1/2}(\log N)^{1/2} w(u)^{-1},\ N = n, M.$$

The weight function w has been introduced in [29] and the considerations presented in that paper, combined with ideas given in [24] play an important role for our arguments. Since the penalty terms will involve an empirical version of the characteristic function in the denominator, the oracle inequalities depend on a precise control of the deviation of $\hat{f}_\varepsilon^*$ from f_ε, simultaneously on the real line. It is shown in [29] that the distance between both object, weighted by w, is simultaneously small on the real axes. In the penalty terms, there will hence occur a loss of logarithmic order, in comparison to the variance term.

Let us now introduce the stochastic penalty terms. In the NS-model,

$$\widehat{\mathrm{pen}}_{\mathrm{NS}}(m) := \widehat{\mathrm{pen}}_{1,\mathrm{NS}}(m) + \widehat{\mathrm{pen}}_{2,\mathrm{NS}}(m),$$

$$:= \frac{16}{3\pi}\int_{-\pi m}^{\pi m} \frac{\tau_Y^2 k_n^2(u)}{|\tilde{f}^*_{\varepsilon,\text{NS}}(u)|^2}\,\mathrm{d}u + \frac{2}{3\pi}\int_{-\pi m}^{\pi m} \frac{\tau_\varepsilon^2 k_M^2(u)|\hat{f}^*_{Y,\text{NS}(u)}|^2}{|\tilde{f}^*_{\varepsilon,\text{NS}}(u)|^4}\,\mathrm{d}u, \quad (10)$$

with $\tau_Y = \tau_\varepsilon = \sqrt{6}$.

In the RD-model,

$$\widehat{\text{pen}}_{\text{RD}}(m) := \widehat{\text{pen}}_{1,\text{RD}}(m) + \widehat{\text{pen}}_{2,\text{RD}}(m),$$
$$:= \frac{16}{3\pi}\int_{-\pi m}^{\pi m} \frac{\tau_Y^2 k_n^2(u)}{|\tilde{f}^*_{\varepsilon,\text{RD}}(u)|^2}\,\mathrm{d}u + \frac{4}{9\pi}\int_{-\pi m}^{\pi m} \frac{\tau_\varepsilon^2 k_n^2(u)|\hat{f}^*_{Y,\text{RD}(u)}|^2}{|\tilde{f}^*_{\varepsilon,\text{RD}}(u)|^6}\,\mathrm{d}u, \quad (11)$$

with τ_Y as previously and $\tau_\varepsilon = \sqrt{3}$.

The cutoff parameters are selected as the minimizers of the following penalized criteria.

$$\widehat{m}_{\text{NS}} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{ -\left\|\hat{f}_{m,\text{NS}}\right\|^2 + \widehat{\text{pen}}_{\text{NS}}(m) \right\} \quad (12)$$

$$\widehat{m}_{\text{RD}} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{ -\left\|\hat{f}_{m,\text{RD}}\right\|^2 + \widehat{\text{pen}}_{\text{RD}}(m) \right\}. \quad (13)$$

Before formulating the oracle bound for the corresponding estimators, let us introduce the deterministic counterparts of the stochastic penalty terms:

$$\text{pen}_{\text{NS}}(m) := \text{pen}_{1,\text{NS}}(m) + \text{pen}_{2,\text{NS}}(m)$$
$$:= \frac{16}{3\pi}\int_{-\pi m}^{\pi m} \frac{\tau_Y^2 k_n^2(u)}{|f^*_\varepsilon(u)|^2}\,\mathrm{d}u + \frac{2}{3\pi}\int_{-\pi m}^{\pi m} \frac{\tau_\varepsilon^2 k_M^2(u)|f^*(u)|^2}{|f^*_\varepsilon(u)|^4}\,\mathrm{d}u \quad (14)$$

and

$$\text{pen}_{\text{RD}}(m) := \text{pen}_{1,\text{RD}}(m) + \text{pen}_{2,\text{RD}}(m)$$
$$:= \frac{16}{3\pi}\int_{-\pi m}^{\pi m} \frac{\tau_Y^2 k_n^2(u)}{|f^*_\varepsilon(u)|^2}\,\mathrm{d}u + \frac{4}{9\pi}\int_{-\pi m}^{\pi m} \frac{\tau_\varepsilon^2 k_n^2(u)|f^*(u)|^2}{|f^*_\varepsilon(u)|^4}\,\mathrm{d}u. \quad (15)$$

We are now ready to formulate the oracle bounds and hence the main result of this paper:

Theorem 3.1.

*(i) Let $\mathcal{M}$ be a collection of cutoff parameters, with $\max\mathcal{M} \le \sqrt{M\wedge n}$. Assume that (**A1**) and (**A2**) are satisfied. Let $\widehat{m}_{\text{NS}}$ be defined by (12) and $\hat{f}_{\widehat{m}_{\text{NS}},\text{NS}}$ according to (7). Then there exists a universal positive constant C^{ad} and a positive constant C depending on the particular choice of τ and δ, but not on any of the underlying distributions, such that*

$$\mathbb{E}\|f - \hat{f}_{\widehat{m}_{\text{NS}},\text{NS}}\|^2 \le C^{ad} \inf_{m\in\mathcal{M}} \left\{ \|f - f_m\|^2 + \text{pen}_{\text{NS}}(m) \right\} + \frac{C}{M\wedge n}. \quad (16)$$

(ii) Let $\mathcal{M}$ be a collection of cutoff parameters with $\max \mathcal{M} \leq \sqrt{n}$*. Assume that **(A1)–(A3)** are satisfied. Let $\widehat{m}_{\mathrm{RD}}$ and $\hat{f}_{\widehat{m}_{\mathrm{RD}},\mathrm{RD}}$ be defined according to (13) and (7). Then there is a universal positive constant C^{ad} and a positive constant C depending on the choice of γ and δ, but not on the underlying distributions, such that*

$$\mathbb{E}\|f - \hat{f}_{\widehat{m}_{\mathrm{RD}},\mathrm{RD}}\|^2 \leq C^{ad} \inf_{m\in\mathcal{M}} \left\{\|f - f_m\|^2 + \mathrm{pen}_{\mathrm{RD}}(m)\right\} + \frac{C}{n}. \tag{17}$$

Remark 2. It is remarkable that we are able to establish non-parametric oracle bounds which make sense without specifying any particular semi-parametric model. Related problems are frequently discussed under specific a priori assumptions on the decay behavior of f^* and f_ε^*, see for example [13] or [11]. In the present work, we can completely dispose of any such assumptions, so our approach is as general as possible.

Another interesting point about our considerations is the following: the only assumption imposed on the collection $\mathcal{M}$ of cutoff parameters is some upper bound on the largest index. No further specification is necessary and we may work with an arbitrarily fine grid, allowing good approximation properties. This is a consequence of the fact that our proofs rely on one sole application of the Talagrand inequality. Additional applications of the Bernstein inequality and sums over $\mathcal{M}$ are not required.

Finally, it is worth emphasizing that, to the best of our knowledge, the non-asymptotic oracle bounds for the RD-model are completely new and the same is true for the bounds in the NS model with $M < n$.

We have considered the problem from a non-asymptotic perspective, but Theorem 3.1 entails, from the asymptotic and minimax point of view, the following observation: in those cases where minimax rates of convergence are known, the procedure achieves, up to a logarithmic loss, automatic adaptation over prescribed non-parametric function classes, typically Sobolev-spaces or classes of super-smooth functions.

4. Illustrations

4.1. *Practical estimation procedure*

Let us describe first the adaptive procedures as it is implemented for the both models:

- ▷ For $m \in \mathcal{M} = \{m_1, \ldots, m_n\}$, compute $-\|\hat{f}_m\|^2 + \widehat{\mathrm{pen}}(m)$.
- ▷ Choose $\widehat{m}$ such as $\widehat{m} = \operatorname{argmin}_{m\in\mathcal{M}}\{-\|\hat{f}_m\|^2 + \widehat{\mathrm{pen}}(m)\}$.
- ▷ Compute $\hat{f}_{\widehat{m}}(x) = \frac{1}{2\pi}\int_{-\pi\widehat{m}}^{\pi\widehat{m}} e^{-ixu}\frac{\hat{f}_Y^*(u)}{\tilde{f}_\varepsilon^*(u)}du$.

As often in model selection methods, the values of the constants in the penalty, here denoted by τ_Y and τ_ε, which are obtained from the theory are too large in practice. Therefore a calibration step is required and done: for a small set of densities and different sample sizes the mean integrated squared error (MISE) is computed in order to determine admissible range for the values of the constants (see [1] for a description of this step). Finally, the penalties are chosen according to Equations (10) and (11) with τ_Y replaced by 0.6 and τ_ε by 0.3 for the NS-model and RD-model. Besides we consider the model collection $\mathcal{M} = \{m = k/10,\ 1 \leq k \leq 25\}$. In practice, one can take k in a much larger set and propose larger values for m; the first selected value can be followed by another run of the estimation algorithm with a thinner grid of proposals around the selected value. We limited the set here because the proposed values seemed adequate and allowed less time-consuming repeated experiments.

In the sequel we also use the notations $\hat{r}^{or}$ and $\hat{r}^{ad}$ define as follows

$$\hat{r}^{or} = \min_{m\in\mathcal{M}} \hat{\mathbb{E}}\|f - \hat{f}_m\|^2 \quad \text{and} \quad \hat{r}^{ad} = \hat{\mathbb{E}}\|f - \hat{f}_{\widehat{m}}\|^2,$$

where $\hat{\mathbb{E}}$ is the approximation of theoretical expectation computed via Monte-Carlo repetitions.

The whole implementation is conducted using R software. The integrated squared error (ISE) $\|f - \hat{f}_{\hat{m}}\|^2$ is computed via a standard approximation and discretization (over 300 points) of the integral on an interval of $\mathbb{R}$ denoted by I. Then the MISE $\mathbb{E}\|f - \hat{f}_{\hat{m}}\|^2$ is computed as the empirical mean of the approximated ISE over 500 simulation samples.

4.2. Comparison with [11] and influence of M in the NS-model

We compute different estimators of the signal for different values of M and consider different signal densities and two noises. Following [13], we study the following densities on the interval I:

- ▷ Laplace distribution, $f(x) = e^{-\sqrt{2}|x|}/\sqrt{2}$, $I = [-5, 5]$.
- ▷ Mixed Gamma distribution: $X = W/\sqrt{5.48}$, with $W \sim 0.4\Gamma(5, 1) + 0.6\Gamma(13, 1)$, $I = [-1.5, 26]$.
- ▷ Cauchy distribution: $f(x) = (\pi(1 + x^2))^{-1}$, $I = [-10, 10]$.
- ▷ Standard Gaussian distribution, $I = [-4, 4]$.

All the densities are normalized with unit variance except the Cauchy density.

We consider the two following noise densities with same variance 1/10.

- ▷ **Gaussian noise**: $f_\varepsilon(x) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma_\varepsilon^2})$, $f_\varepsilon^*(x) = \exp(-\frac{\sigma_\varepsilon^2 x^2}{2})$.
- ▷ **Laplace noise**: $f_\varepsilon(x) = \frac{1}{2\sigma_\varepsilon} \exp(-\frac{|x|}{\sigma_\varepsilon})$, $f_\varepsilon^*(x) = \frac{1}{1+\sigma_\varepsilon^2 x^2}$.

We want to study the influence of the relationship of n and M on the estimation of f in the NS-model. We then consider different values of n and values of $M = \sqrt{n}$ and $M = n$.

Results. The results of the simulations are given in Table 1. Table 1 illustrates the case where we can recover a preliminary sample of the noise ε (our so-called NS-model). First we see that the risk decreases when the sample size increases. Likewise the risk increases when the variance increases. Secondly the results are very close to those of [11]. Nevertheless our procedure is equivalent or better.

Table 1

Values of approximated MISE $\mathbb{E}(\|f - \hat{f}_{\hat{m}_{\mathrm{NS}},\mathrm{NS}}\|^2) \times 100$ *averaged over 500 samples with a Laplace and Gaussian noise of variance* $1/10$, $\hat{r}^{ad}$ *is the risk of the adaptive estimator,* $\hat{r}^{or}$ *is the risk of the oracle estimator,* M *is the size of the noise and* n *is the size of the* Y_i*'s*

			$n = 100$		$n = 250$		$n = 500$	
f		f_ε	Lap.	Gaus.	Lap.	Gaus.	Lap.	Gaus.
Laplace	$M = \lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	2.779	2.714	1.790	1.933	1.304	1.403
		$\hat{r}^{or}$	(2.223)	(1.370)	(1.342)	(0.828)	(0.988)	(0.544)
	$M = n$	$\hat{r}^{ad}$	2.606	2.620	1.543	1.626	1.136	1.339
		$\hat{r}^{or}$	(1.848)	(1.357)	(1.107)	(0.786)	(0.796)	(0.539)
Mixed Gamma	$M = \lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	0.787	0.777	0.377	0.385	0.247	0.247
		$\hat{r}^{or}$	(0.671)	(0.710)	(0.331)	(0.337)	(0.205)	(0.209)
	$M = n$	$\hat{r}^{ad}$	0.751	0.725	0.360	0.365	0.232	0.234
		$\hat{r}^{or}$	(0.642)	(0.682)	(0.325)	(0.319)	(0.194)	(0.197)
Cauchy	$M = \lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	0.891	0.888	0.433	0.465	0.311	0.325
		$\hat{r}^{or}$	(0.721)	(0.731)	(0.386)	(0.418)	(0.361)	(0.275)
	$M = n$	$\hat{r}^{ad}$	0.817	0.806	0.402	0.416	0.285	0.301
		$\hat{r}^{or}$	(0.687)	(0.687)	(0.361)	(0.379)	(0.232)	(0.254)
Gaussian	$M = \lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	0.774	0.777	0.450	0.446	0.275	0.303
		$\hat{r}^{or}$	(0.657)	(0.655)	(0.388)	(0.415)	(0.190)	(0.211)
	$M = n$	$\hat{r}^{ad}$	0.666	0.644	0.406	0.392	0.247	0.272
		$\hat{r}^{or}$	(0.512)	(0.540)	(0.345)	(0.325)	(0.127)	(0.126)

The main improvement is that our procedure has better performances when the size of the preliminary sample is small. Moreover we show that the procedure does not need big M since we reach the same performances as [11] for $M = n$ when they take $M = n^2$. Indeed if we consider the mixed Gamma distribution for $n = M = 500$, our risk is 0.232 with a Laplace noise, while [11] with $M = n^2$ have 0.382. We can make the same remarks for the other distributions (except for the Laplace distribution) with a sample size 100 or 250 and with the two noise distributions. For the Laplace distribution, the results of [11] are better but they do not outperform ours. Nonetheless, in our model, the risk decreases more rapidly when M and n increase.

Effect of the variance

We also test how our procedure behaves when the variance is increased. In Table 2, we present the results of simulations where the variance takes the values $1/4$, $1/2$ and 1. We only report the case of a Gaussian error distribution since the results of the Laplace error are very similar. Moreover a Gaussian distribution is a case less favorable. Indeed its Fourier transform decays exponentially and it is known to imply possibly slower asymptotic rates. Thus it is more difficult to recover the target density f.

Results. The results are reported in Table 2. As before the risk decreases when n and M increase. Similarly when we increase the contamination of the variable of interest by increasing the variance of the error distribution, the risk increases. The procedure performs still well since the adaptive risk is close to the oracle risk around twice bigger.

TABLE 2

Values of approximated MISE $\mathbb{E}(\|f-\hat{f}_{\hat{m}_{\mathrm{NS}},\mathrm{NS}}\|^2)\times 100$ averaged over 500 samples with a Gaussian noise, $\hat{r}^{ad}$ is the risk of the adaptive estimator, $\hat{r}^{or}$ is the risk of the oracle estimator, M is the size of the noise and n is the size of the Y_i's

				$\sigma_\varepsilon^2 = 1/4$	
f	M	n	100	250	500
Mixed Gamma	$\lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	1.767	0.889	0.471
		$\hat{r}^{or}$	(0.808)	(0.473)	(0.299)
	n	$\hat{r}^{ad}$	1.231	0.620	0.387
		$\hat{r}^{or}$	(0.791)	(0.439)	(0.265)
Gaussian	$\lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	1.663	0.920	0.543
		$\hat{r}^{or}$	(0.913)	(0.531)	(0.397)
	n	$\hat{r}^{ad}$	1.245	0.669	0.522
		$\hat{r}^{or}$	(0.578)	(0.280)	(0.180)
				$\sigma_\varepsilon^2 = 1/2$	
Mixed Gamma	$\lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	2.170	1.229	0.673
		$\hat{r}^{or}$	(0.940)	(0.652)	(0.456)
Gaussian	$\lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	1.951	1.185	0.867
		$\hat{r}^{or}$	(1.489)	(0.988)	(0.747)
	n	$\hat{r}^{ad}$	1.884	1.053	0.690
				$\sigma_\varepsilon^2 = 1$	
Mixed Gamma	$\lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	2.798	1.973	1.232
		$\hat{r}^{or}$	(1.042)	(0.865)	(0.692)
	n	$\hat{r}^{ad}$	1.594	1.135	0.951
		$\hat{r}^{or}$	(0.979)	(0.776)	(0.585)
Gaussian	$\lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	3.163	2.139	1.898
		$\hat{r}^{or}$	(2.621)	(1.988)	(1.542)
	n	$\hat{r}^{ad}$	2.458	1.432	0.982
		$\hat{r}^{or}$	(1.400)	(0.848)	(0.613)

4.3. Illustrations in the RD-model

For this model, we use the same signal and error distributions as described in 4.2. We consider different values of n: 200 and 2000, with variance $1/10$ and $1/2$.

Results. The results are reported in Table 3. We note that the values of the MISE are very close for both error distributions. Moreover the adaptive risk is also close to the oracle risk. It is multiplied by approximately 1.5 when the variance of the error distribution is $1/10$ and 2 when the variance equals $1/2$. Again the risk decreases when n and M increase. And the risk increases when the variance increases.

4.4. Comparison with a kernel estimator

Recently, some papers as [14] have investigated the necessity of inversion in statistical inverse problem. The idea is to compare the performances of a simple kernel estimator directly applied to the data Y_i with the adaptive estimator of the NS-model. This model allows us to choose a non symmetric noise which makes the estimation with a kernel estimator more difficult. That is why we

Table 3
Values of approximated MISE $\mathbb{E}(\|f - \hat{f}_{\hat{m}_{\mathrm{RD}},\mathrm{RD}}\|^2) \times 100$ averaged over 500 samples, $\hat{r}^{ad}$ is the risk of the adaptive estimator, $\hat{r}^{or}$ is the risk of the oracle estimator, M is the size of the noise and n is the size of the Y_i's

		$\sigma_\varepsilon^2 = 1/10$			
		$n = 200$		$n = 2000$	
f	f_ε	Lap.	Gaus.	Lap.	Gaus.
Laplace	$\hat{r}^{ad}$	1.706	1.683	0.655	0.725
	$\hat{r}^{or}$	(1.163)	(1.216)	(0.369)	(0.438)
Mixed	$\hat{r}^{ad}$	0.558	0.548	0.075	0.076
Gamma	$\hat{r}^{or}$	(0.381)	(0.391)	(0.061)	(0.060)
Cauchy	$\hat{r}^{ad}$	0.535	0.595	0.127	0.127
	$\hat{r}^{or}$	(0.358)	(0.377)	(0.068)	(0.069)
Gaussian	$\hat{r}^{ad}$	0.338	0.335	0.045	0.050
	$\hat{r}^{or}$	(0.231)	(0.235)	(0.034)	(0.035)
		$\sigma_\varepsilon^2 = 1/2$			
Laplace	$\hat{r}^{ad}$	3.901	3.539	1.938	2.007
	$\hat{r}^{or}$	(2.904)	(2.710)	(1.234)	(1.536)
Mixed	$\hat{r}^{ad}$	0.857	0.799	0.248	0.226
Gamma	$\hat{r}^{or}$	(0.583)	(0.550)	(0.144)	(0.133)
Cauchy	$\hat{r}^{ad}$	1.172	1.027	0.312	0.242
	$\hat{r}^{or}$	(0.599)	(0.617)	(0.181)	(0.174)
Gaussian	$\hat{r}^{ad}$	0.703	0.586	0168	0.155
	$\hat{r}^{or}$	(0.441)	(0.399)	(0.092)	(0.076)

compare our estimator with a kernel in two contexts: one with a symmetric noise and another one with an asymmetric noise.

First, for the symmetric noise, we take the results of Table 2 with variance 1/2 for a Gaussian noise with $M = n$ and compute the kernel estimator for this design.

Secondly we compute the estimator of the signal in the NS-model with $M = \lfloor \sqrt{n} \rfloor$ with an asymmetric noise. We consider two densities defined in the beginning of this section: mixed Gamma and Gaussian. We choose the error distribution as follows

▷ **Gamma noise**: $f_\varepsilon(x) = x^{\alpha-1}\frac{\beta^\alpha e^{-\beta x}}{\Gamma(\alpha)}\mathbb{1}_{x\geq 0}$

with parameters $\alpha = 1$ and $\beta = 2$. The variance of the error distribution is then 1/4. For the kernel estimator, we use the function *density* of R with a Gaussian kernel where the bandwidth is selected by cross-validation.

Results. The results are reported in Tables 4 and 5. In Table 4, we see that the results are very close. For the mixed Gamma of the kernel the results are a little bit better but the risk of our estimator decreases more rapidly. For the Gaussian distribution our results are better and the risk decreases also more rapidly when the sample size increases.

For the non symmetric noise the results are reported in Table 5. We see that for the mixed Gamma, the kernel estimator performs unexpectedly well. For $n = 100$, the risks are practically the same as those of the adaptive estimator. When

Table 4

Comparison of the method with the results of a Gaussian kernel estimator with a symmetric noise (MISE$\times 100$ averaged over 500 simulations), $\hat{r}^{ker}$ is the risk of the kernel estimator, $\hat{r}^{ad}$ is the risk of the adaptive estimator, $\hat{r}^{or}$ is the risk of the oracle estimator, M is the size of the noise and n is the size of the Y_i's

f			$n = 100$	$n = 250$	$n = 500$
Mixed		$\hat{r}^{ker}$	0.849	0.648	0.526
Gamma	$M = n$	$\hat{r}^{ad}$	1.483	0.835	0.554
		$\hat{r}^{or}$	(0.895)	(0.581)	(0.386)
Gaussian		$\hat{r}^{ker}$	1.673	1.333	1.145
	$M = n$	$\hat{r}^{ad}$	1.884	1.053	0.690
		$\hat{r}^{or}$	(0.832)	(0.444)	(0.292)

Table 5

Comparison of the method with the results of a Gaussian kernel estimator with an asymmetric noise (MISE$\times 100$ averaged over 500 simulations), $\hat{r}^{ker}$ is the risk of the kernel estimator, $\hat{r}^{ad}$ is the risk of the adaptive estimator, $\hat{r}^{or}$ is the risk of the oracle estimator, M is the size of the noise and n is the size of the Y_i's

f			$n = 100$	$n = 250$	$n = 500$
Mixed		$\hat{r}^{ker}$	1.086	0.881	0.785
Gamma	$M = \lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	1.063	0.573	0.332
		$\hat{r}^{or}$	(0.714)	(0.428)	(0.258)
Gaussian		$\hat{r}^{ker}$	3.181	3.052	2.807
	$M = \lfloor\sqrt{n}\rfloor$	$\hat{r}^{ad}$	1.364	0.677	0.449
		$\hat{r}^{or}$	(0.844)	(0.449)	(0.329)

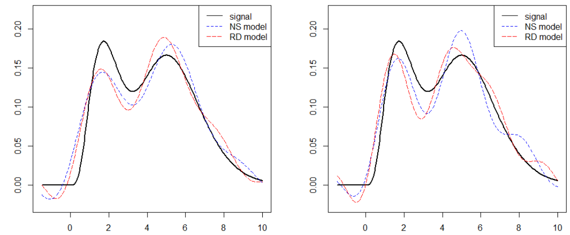
n increases the adaptive estimator performs better but the kernel estimator still gives satisfying results.

On the other hand, the results of the Gaussian distribution illustrate well the importance of inversion in statistical inverse problem. Indeed the risk with a kernel estimator is around 3.10^{-2} for the diverse values of n while for the adaptive estimator the risk is divided by 3. More precisely for $n = 100$ the risk of the kernel estimator is multiplied by 2 compared to the adaptive estimator, by 5 for $n = 250$ and by 7 for $n = 500$. So when the error distribution is unknown and the signal to noise ratio (named *s2n*) is not too large, our procedure is worthy of interest. When the *s2n* gets larger, we only expect the deconvolution procedure not to deteriorate the results compared to direct estimation. It has been checked to be true for known noise density by [14] see section 4.7 and 4.9. As in practice the value of the *s2n* is unknown, our procedure is always recommended.

Figures 1 and 2 illustrate the estimation of a mixed Gamma using both penalized estimators in the cases for $n = M = 200$ and $n = M = 500$. The estimation is made with additional Gaussian and Laplace noises with a variance of $1/10$. The bimodal specificity of the density is well described. Moreover the precision increases with the sample size.

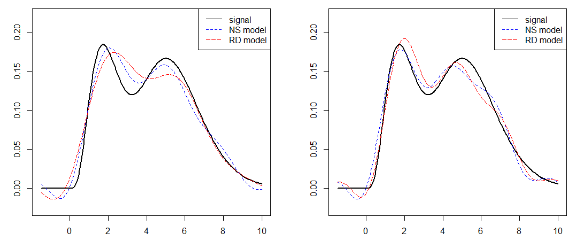
5. Concluding remarks

This paper deals with adaptive deconvolution estimation of a density when the noise density is unknown. We have considered two cases: one where a preliminary



(a) Laplace noise (b) Gaussian noise

Fig 1. *Estimations for $n = M = 200$ of a mixed Gamma (bold black line), in blue dashed line for the* NS *estimator and in red longdashed for the* RD *estimator.*



(a) Laplace noise (b) Gaussian noise

Fig 2. *Estimations for $n = M = 500$ of a mixed Gamma (bold black line), in blue dashed line for the* NS *estimator and in red longdashed for the* RD *estimator.*

sample of the noise can be observed and another one where the variable of interest X can be observed repeatedly with independent errors. For both models, we have proposed a theoretical adaptive procedure which automatically makes a data driven bias-variance compromise. Moreover it allows us to not specify rates of convergence since they are mechanically reached. This procedure enables us to treat the problem of adaptive estimation in repeated observation model which is completely new. The estimation procedure relies on the independence of the estimators of the characteristic functions of f_Y^* and f_ε^*. Its advantage is to be very general under weak assumptions. Indeed, that procedure takes into account cases where there can be small number of replications which matches realistic applications as in medicine or economics. Besides of its theoretical properties, our procedure has showed good performances in simulation.

At last we think that our procedure can be extended to density estimation of a random effect in linear mixed-effects model. Indeed we are aware of the work of [15] who proposed an adaptive procedure based on deconvolution methods in the unknown-error case which is not optimal and [19] who used a Lepski's method in the known-error case. In that model, the noise can also be recovered by successive difference similarly to the repeated model but the characteristic function of the noise would be raised to a greater power. We may then propose in the same spirit an adaptive procedure for the random effect in linear mixed-effects model.

6. Proofs

6.1. Preliminaries

We start by restating, for the reader's convenience, the following version of Talagrand's inequality:

Lemma 6.1. *Let I be some countable index set. For each $i \in I$, let $Z_1^{(i)}, \dots, Z_n^{(i)}$ be independent and identically distributed random variables with values in $[-1,1]$, defined on the same probability space. Let $v^2 := \sup_{i\in I} \mathbb{V}\mathrm{ar}[Z_1^{(i)}]$ and $S_n^{(i)} := 1/n \sum_{j=1}^n Z_j^{(i)}$. Then there are universal positive constants c_1 and c_2 such that for any $\kappa > 0$,*

$$\mathbb{P}\left[\left\{\sup_{i\in I} |S_n^{(i)}| \leq \frac{3}{2}\mathbb{E}\left[\sup_{i\in I} |S_n^{(i)}|\right] + \kappa\right\}\right] \leq 2\exp\left(-\left(\frac{n\kappa^2}{c_1 v^2 + c_2\kappa}\right)\right).$$

A proof of this result is given in [26], see page 170. It follows from the arguments presented therein that for any $r, s > 0$ with $\frac{1}{s} + \frac{1}{r} = 1$, the constants can be chosen $c_1 = 2s^2$ and $c_2 = 6r$.

The following result will be essential for the theoretical justification of the adaptive procedure. The proof was given in [24]. It uses a fundamental Lemma shown in [29], combined with Lemma 6.1.

Lemma 6.2. *Let $Z_1, \dots, Z_n$ be i.i.d. random variables. Assume that $\mathbb{E}|Z_1| \leq m_Z$ for some $m_Z > 0$. Let $\hat{f}_Z^*$ denote the empirical characteristic function. For some $\delta > 0$, let*

$$w(u) = (\log(e + |u|))^{-1/2-\delta}.$$

Assume that $\tau > 2\sqrt{p}$. Then there exists a positive constant C depending on m_Z and the choice of γ and δ such that for arbitrary $n \in \mathbb{N}$:

$$\mathbb{P}\left[\exists u \in \mathbb{R} : |\tilde{f}_Z^*(u) - f_Z^*(u)| > \tau\,(\log(n))^{1/2}\, w(u)^{-1} n^{-1/2}\right] \leq C n^{-p}. \tag{18}$$

Remark 3. In the situation of the preceding Lemma: if the Z_j are assumed to have a symmetric distribution and the empirical characteristic function is defined to be $\hat{f}_Z^*(u) = 1/n \sum_{j=1}^n \cos(uZ_j)$, it is enough to assume $\tau > \sqrt{2p}$ to obtain (18).

The next result has been formulated and proved in [30].

Lemma 6.3. *Let $Z_1, \dots, Z_n$ be i.i.d. random variables. Let f_Z^* be the true and $\hat{f}_Z^*$ the empirical characteristic function. Moreover let $1/\tilde{f}_Z^*(u) = \mathbb{1}\{|\hat{f}_Z^*(u)| \geq n^{-1/2}\}/\hat{f}_Z^*(u)$. Then for arbitrary $p \in \mathbb{N}$, there exists a positive constant C depending only on p such that*

$$\mathbb{E}\left[\left|\frac{1}{\tilde{f}_Z^*(u)} - \frac{1}{f_Z^*(u)}\right|^p\right] \leq C\left(\frac{1}{|f_Z^*(u)|^p} \wedge \frac{n^{-\frac{p}{2}}}{|f_Z^*(u)|^{2p}}\right). \tag{19}$$

Remark 4. In the preceding lemma, one should keep in mind that in the RD-model, $Y_{j,1} - Y_{j,2} = \varepsilon_{j,1} - \varepsilon_{j,2}$ plays the role of Z_j.

Lemma 6.3 yields the following useful corollary, which allows to compare the stochastic penalty terms to their deterministic counterparts.

Corollary 6.4. *Let the penalty terms be defined according to Equations* (10) *and* (11). *Then for some* $C > 0$,

$$\mathbb{E}\left[\widehat{\mathrm{pen}}_{\mathrm{NS}}(m)\right] \leq C\mathrm{pen}_{\mathrm{NS}}(m) \quad \textit{and} \quad \mathbb{E}\left[\widehat{\mathrm{pen}}_{\mathrm{RD}}(m)\right] \leq C\mathrm{pen}_{\mathrm{RD}}(m).$$

Proof. It is enough to consider $\widehat{\mathrm{pen}}_{\mathrm{RD}}$. Equation (19) gives immediately for some $C > 0$,

$$\mathbb{E}\left[\int_{-\pi m}^{\pi m} \frac{\tau_Y^2 k_n^2(u)}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u\right] \leq C \int_{-\pi m}^{\pi m} \frac{\tau_Y^2 k_n^2(u)}{|f^*_\varepsilon(u)|^2}\,\mathrm{d}u.$$

The Cauchy-Schwarz inequality and the estimate $1/|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2 \leq n^{1/2}$ imply

$$\begin{aligned}
&\mathbb{E}\left[\int_{-\pi m}^{\pi m} |\hat{f}^*_{Y,\mathrm{RD}}(u)|^2 \frac{\tau_\varepsilon^2 k_n^2(u)}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^6}\,\mathrm{d}u\right] \\
&\leq 2\int_{-\pi m}^{\pi m} \mathbb{E}^{1/2}\left[\left|\hat{f}^*_{Y,\mathrm{RD}}(u) - f^*_{Y,\mathrm{RD}}(u)\right|^4\right] \mathbb{E}^{1/2}\left[\frac{\tau_\varepsilon^2 k_n^2(u)}{\left|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)\right|^{12}}\right]\mathrm{d}u \\
&\quad + 2\int_{-\pi m}^{\pi m} \mathbb{E}\left[\frac{\tau_\varepsilon^2 k_n^2(u)\,|f^*_Y(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^6}\right]\mathrm{d}u \\
&\leq 2\int_{-\pi m}^{\pi m} \frac{\tau_\varepsilon^2 k_n^2(u)}{|f^*_\varepsilon(u)|^2}\,\mathrm{d}u + 2\int_{-\pi m}^{\pi m} \frac{\tau_\varepsilon^2 k_n^2(u)|f^*(u)|^2}{|f^*_\varepsilon(u)|^2\left(|f^*_\varepsilon(u)|^2 \vee n^{-1/2}\right)}\,\mathrm{d}u.
\end{aligned}$$

This completes the proof for the RD-model. The arguments for the NS-model are the same, line for line, so we omit the details. The only difference lies in the fact that, in the definition of $\hat{f}^*_{\varepsilon,\mathrm{NS}}$ and $\tilde{f}^*_{\varepsilon,\mathrm{NS}}$, M plays now the role of n as defined in Lemma 6.3. ☐

6.2. A technical auxiliary result

In the sequel, for arbitrary $k > m$, we use the notation

$$\widehat{\mathrm{pen}}_{\mathrm{NS}}(m,k) := \widehat{\mathrm{pen}}_{\mathrm{NS}}(k) - \widehat{\mathrm{pen}}_{\mathrm{NS}}(m) \quad \text{and} \quad \widehat{\mathrm{pen}}_{\mathrm{RD}}(m,k) := \widehat{\mathrm{pen}}_{\mathrm{RD}}(k) - \widehat{\mathrm{pen}}_{\mathrm{RD}}(m),$$

as well as

$$\begin{aligned}
\widehat{\mathrm{pen}}_{\ell,\mathrm{NS}}(m,k) &:= \widehat{\mathrm{pen}}_{\ell,\mathrm{NS}}(k) - \widehat{\mathrm{pen}}_{\ell,\mathrm{NS}}(m) \\
\widehat{\mathrm{pen}}_{\ell,\mathrm{RD}}(m,k) &:= \widehat{\mathrm{pen}}_{\ell,\mathrm{RD}}(k) - \widehat{\mathrm{pen}}_{\ell,\mathrm{RD}}(m), \quad \ell = 1, 2.
\end{aligned}$$

Moreover,

$$A(m,k) := \{u \in \mathbb{R} : m \leq |u| \leq k\}.$$

The proof of Theorem 3.1 relies on the following auxiliary result which is, in turn, a consequence of Lemma 6.2.

Proposition 6.5.

(i) In the model of repeated measurements, there exists an universal positive constant C such that

$$\mathbb{E}\left[\sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}}\left\{\left\|\hat{f}_{k,\mathrm{RD}}-\hat{f}_{m,\mathrm{RD}}\right\|^2-6\left\|f_k-f_m\right\|^2-\frac{3}{4}\widehat{\mathrm{pen}}_{\mathrm{RD}}(m,k)\right\}_+\right]\leq\frac{C}{n}.$$

(ii) If an additional sample of the pure noise is available, for some universal positive constant C,

$$\mathbb{E}\left[\sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}}\left\{\left\|\hat{f}_{k,\mathrm{NS}}-\hat{f}_{m,\mathrm{NS}}\right\|^2-6\left\|f_k-f_m\right\|^2-\frac{3}{4}\widehat{\mathrm{pen}}_{\mathrm{NS}}(m,k)\right\}_+\right]\leq\frac{C}{n\wedge M}.$$

Proof.

(i) Let us introduce the favorable events

$$\mathcal{E}_{Y,\mathrm{RD}}:=\left\{\forall u\in\mathbb{R}:|\hat{f}^*_{Y,\mathrm{RD}}(u)-f^*_Y(u)|\leq\tau_Y k_n(u)\right\}.$$
$$\mathcal{E}_{\varepsilon,\mathrm{RD}}:=\left\{\forall u\in\mathbb{R}:|\widehat{f^{*2}_{\varepsilon,\mathrm{RD}}}(u)-f^{*2}_\varepsilon(u)|\leq\tau_\varepsilon k_n(u)\right\}.$$

Applying Parseval's equality we can estimate

$$\begin{aligned}\|\hat{f}_{k,\mathrm{RD}}-\hat{f}_{m,\mathrm{RD}}\|^2&=\frac{1}{2\pi}\|\hat{f}^*_{k,\mathrm{RD}}-\hat{f}^*_{m,\mathrm{RD}}\|^2=\frac{1}{2\pi}\int_{A(m,k)}\frac{|\hat{f}^*_{Y,\mathrm{RD}}(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u\\&\leq\frac{1}{\pi}\int_{A(m,k)}\frac{|\hat{f}^*_{Y,\mathrm{RD}}(u)-f^*_Y(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u+\frac{1}{\pi}\int_{A(m,k)}\frac{|f^*_Y(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u.\end{aligned}\tag{20}$$

We start by dealing with the first summand appearing in the last line of (20). The definition of $\mathcal{E}_{Y,\mathrm{RD}}$ and the fact that

$$\frac{1}{\pi}\int_{A(m,k)}\frac{\tau_Y^2k_n^2(u)}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u\leq\frac{3}{8}\widehat{\mathrm{pen}}_{1,\mathrm{RD}}(m,k)$$

immediately imply the following inequality

$$\begin{aligned}&\sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}}\left\{\frac{1}{\pi}\int_{A(m,k)}\frac{|\hat{f}^*_{Y,\mathrm{RD}}(u)-f^*_Y(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u-\frac{3}{8}\widehat{\mathrm{pen}}_{1,\mathrm{RD}}(m,k))\right\}_+\mathbb{1}_{\mathcal{E}_{Y,\mathrm{RD}}}\\&\leq\sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}}\left\{\frac{1}{\pi}\int_{A(m,k)}\frac{\tau_Y^2k_n^2(u)}{|\tilde{f}^*_{\varepsilon,\mathrm{RD}}(u)|^2}\,\mathrm{d}u-\frac{3}{8}\widehat{\mathrm{pen}}_{1,\mathrm{RD}}(m,k)\right\}_+\mathbb{1}_{\mathcal{E}_{Y,\mathrm{RD}}}=0.\end{aligned}$$

Consider now the second summand in the last line of (20). Recall that f_ε^* and $\widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}$ are real valued and (formally, with $c/\infty := 0$),

$$|\tilde{f}_{\varepsilon,\mathrm{RD}}(u)|^2 = \begin{cases} \widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u), & \text{if } \widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u) \geq n^{-1/2} \\ \infty, & \text{else.} \end{cases}$$

This entails the series of inequalities

$$\begin{aligned}
&\frac{1}{\pi}\int_{A(m,k)} \frac{|f_Y^*(u)|^2}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2}\,\mathrm{d}u - 6\|f_k - f_m\|^2 \\
&= \frac{1}{\pi}\int_{A(m,k)} \frac{|f_Y^*(u)|^2}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2}\,\mathrm{d}u - \frac{3}{\pi}\|f_k^* - f_m^*\|^2 \\
&= \frac{1}{\pi}\int_{A(m,k)} |f_Y^*(u)|^2 \left(\frac{1}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2} - \frac{3}{|f_\varepsilon^*(u)|^2}\right)\mathrm{d}u \\
&\leq \frac{1}{\pi}\int_{A(m,k)} |f_Y^*(u)|^2 \left(\frac{1}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2} - \frac{3}{|f_\varepsilon^*(u)|^2}\right)\mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}}\,\mathrm{d}u \\
&\leq \frac{2}{3\pi}\int_{A(m,k)} |\hat{f}_{Y,\mathrm{RD}}^*(u)|^2 \frac{f_\varepsilon^{*2}(u) - 3\widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u)}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^4}\mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}}\,\mathrm{d}u \\
&\quad + \frac{2}{\pi}\int_{A(m,k)} |f_Y^*(u) - \hat{f}_{Y,\mathrm{RD}}^*(u)|^2 \frac{f_\varepsilon^{*2}(u) - 3\widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u)}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2 |f_\varepsilon^*(u)|^2}\mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}}\,\mathrm{d}u.
\end{aligned} \tag{21}$$

For the expression appearing in the last line of formula (21), we observe that

$$\begin{aligned}
&\frac{2}{\pi}\int_{A(m,k)} |f_Y^*(u) - \hat{f}_{Y,\mathrm{RD}}^*(u)|^2 \frac{f_\varepsilon^{*2}(u) - 3\widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u)}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2 |f_\varepsilon^*(u)|^2}\mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}}\,\mathrm{d}u \\
&\leq \frac{2}{\pi}\int_{A(m,k)} |f_Y^*(u) - \hat{f}_{Y,\mathrm{RD}}^*(u)|^2 \frac{\left|f_\varepsilon^*(u)\right|^2}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2 |f_\varepsilon^*(u)|^2}\mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}}\,\mathrm{d}u \\
&= \frac{2}{\pi}\int_{A(m,k)} \frac{|f_Y^*(u) - \hat{f}_{Y,\mathrm{RD}}^*(u)|^2}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2}\,\mathrm{d}u.
\end{aligned}$$

The definition of $\widehat{\mathrm{pen}}_{1,\mathrm{RD}}$ and $\mathcal{E}_{Y,\mathrm{RD}}$ readily imply that

$$\sup_{\substack{k \geq m \\ k,m \in \mathcal{M}}} \left\{\frac{2}{\pi}\int_{A(m,k)} \frac{|f_Y^*(u) - \hat{f}_{Y,\mathrm{RD}}^*(u)|^2}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2}\,\mathrm{d}u - \frac{3}{8}\widehat{\mathrm{pen}}_{1,\mathrm{RD}}(m,k)\right\}_+ \mathbb{1}_{\mathcal{E}_{Y,\mathrm{RD}}} = 0.$$

Consider now the second to last line in Equation (21). We observe that $3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2 \leq |f_\varepsilon^*(u)|^2$ implies $|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2 = \widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u) \leq (1/2)|f_\varepsilon^*(u)^2 - \widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u)|$.

From this, we derive that

$$\frac{f_\varepsilon^*(u) - 3\widehat{f_{\varepsilon,\mathrm{RD}}^{*2}}(u)}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^4} \mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}} \leq \frac{1}{2} \frac{\left|f_\varepsilon^*(u)^2 - \tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)^2\right|^2}{|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^6}.$$

Using this and the definition of $\mathcal{E}_{\varepsilon,\mathrm{RD}}$ and $\widehat{\mathrm{pen}}_{2,\mathrm{RD}}$, we arrive at

$$\begin{aligned}
&\sup_{\substack{k \geq m\\ k,m \in \mathcal{M}}} \left\{ \int_{A(m,k)} \frac{|\hat{f}_{Y,\mathrm{RD}}^*(u)|^2 |f_\varepsilon^*(u)^2 - \tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)^2|}{3/2\pi |\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^4} \mathbb{1}_{\left\{|f_\varepsilon^*(u)|^2 \geq 3|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^2\right\}} \,\mathrm{d}u \right.\\
&\qquad \left. - \frac{3}{4}\widehat{\mathrm{pen}}_{2,\mathrm{RD}}(m,k) \right\}_+ \mathbb{1}_{\mathcal{E}_{\varepsilon,\mathrm{RD}}}\\
&\leq \sup_{\substack{k \geq m\\ k,m \in \mathcal{M}}} \left\{ \int_{A(m,k)} |\hat{f}_{Y,\mathrm{RD}}^*(u)|^2 \frac{|f_\varepsilon^*(u)^2 - \tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)^2|^2}{3\pi |\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)|^6} \,\mathrm{d}u - \frac{3}{4}\widehat{\mathrm{pen}}_{2,\mathrm{RD}}(m,k) \right\}_+ \mathbb{1}_{\mathcal{E}_{\varepsilon,\mathrm{RD}}}\\
&\leq \sup_{\substack{k \geq m\\ k,m \in \mathcal{M}}} \left\{ \int_{A(m,k)} |\hat{f}_{Y,\mathrm{RD}}^*(u)|^2 \frac{\tau_\varepsilon^2 k_n^2(u)}{3\pi \left|\tilde{f}_{\varepsilon,\mathrm{RD}}^*(u)\right|^6} \,\mathrm{d}u - \frac{3}{4}\widehat{\mathrm{pen}}_{2,\mathrm{RD}}(m,k) \right\}_+ \mathbb{1}_{\mathcal{E}_{\varepsilon,\mathrm{RD}}} = 0.
\end{aligned}$$

Putting the above together, we have shown

$$\mathbb{E}\left[\sup_{\substack{k \geq m\\ k,m \in \mathcal{M}}} \left\{ \|\hat{f}_{k,\mathrm{RD}} - \hat{f}_{m,\mathrm{RD}}\|^2 - 6\|f_k - f_m\|^2 - \frac{3}{4}\widehat{\mathrm{pen}}_{\mathrm{RD}}(m,k) \right\}_+ \mathbb{1}_{\mathcal{E}_{Y,\mathrm{RD}} \cap \mathcal{E}_{\varepsilon,\mathrm{RD}}} \right] = 0.$$

There remains to consider the exceptional set $\mathcal{E}_{Y,\mathrm{RD}}^c \cup \mathcal{E}_{\varepsilon,\mathrm{RD}}^c$. Using the fact that for arbitrary m, the absolute value of $\hat{f}_{m,\mathrm{RD}}^*$ is, by definition, bounded by 1, as well as the fact that $\max \mathcal{M} \leq \sqrt{n}$, we can estimate

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{k \geq m\\ k,m \in \mathcal{M}}} \left\{ \|\hat{f}_{k,\mathrm{RD}}^* - \hat{f}_{m,\mathrm{RD}}^*\|^2 - 6\|f_k^* - f_m^*\|^2 - \frac{3}{4}\widehat{\mathrm{pen}}_{\mathrm{RD}}(m,k) \right\}_+ \mathbb{1}_{(\mathcal{E}_{\varepsilon,\mathrm{RD}} \cap \mathcal{E}_{Y,\mathrm{RD}})^c} \right]\\
&\leq \mathbb{E}\left[\sup_{\substack{k \geq m\\ k,m \in \mathcal{M}}} \|\hat{f}_{k,\mathrm{RD}}^* - \hat{f}_{m,\mathrm{RD}}^*\|^2 \mathbb{1}_{(\mathcal{E}_{\varepsilon,\mathrm{RD}} \cap \mathcal{E}_{Y,\mathrm{RD}})^c} \right]\\
&\leq \sqrt{n}\,\mathbb{P}\left((\mathcal{E}_{\varepsilon,\mathrm{RD}} \cap \mathcal{E}_{Y,\mathrm{RD}})^c\right) \leq \sqrt{n}\left(\mathbb{P}\left(\mathcal{E}_{Y,\mathrm{RD}}^c\right) + \mathbb{P}\left(\mathcal{E}_{\varepsilon,\mathrm{RD}}^c\right)\right).
\end{aligned}$$

Thanks to Lemma 6.2, it holds that

$$\sqrt{n}\left(\mathbb{P}\left(\mathcal{E}_{Y,\mathrm{RD}}^c\right) + \mathbb{P}\left(\mathcal{E}_{\varepsilon,\mathrm{RD}}^c\right)\right) \leq \frac{C}{n}. \tag{22}$$

This completes the proof of part *(i)*.

(ii) The proof of the second part uses essentially the same arguments as the proof of the first part, so we content ourselves with sketching the important steps.

$$\mathcal{E}_{Y,\mathrm{NS}} := \left\{ \forall u \in \mathbb{R} : |\hat{f}^*_{Y,\mathrm{NS}}(u) - f^*_Y(u)| \leq \tau_Y k_n(u) \right\}$$
$$\mathcal{E}_{\varepsilon,\mathrm{NS}} := \left\{ \forall u \in \mathbb{R} : |\hat{f}^*_{\varepsilon,\mathrm{NS}}(u) - f^*_\varepsilon(u)| \leq \tau_\varepsilon k_M(u) \right\}.$$

In analogy with (20) and (21), we find that

$$\begin{aligned}
&\|\hat{f}^*_{k,\mathrm{NS}} - \hat{f}^*_{m,\mathrm{NS}}\|^2 - 6\|f^*_k - f^*_m\|^2 \\
&\leq \frac{2}{\pi} \int_{A(m,k)} \frac{|\hat{f}^*_{Y,\mathrm{NS}}(u) - f^*_Y(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^2} \,\mathrm{d}u \\
&\quad + \frac{1}{3\pi} \int_{A(m,k)} |\hat{f}^*_{Y,\mathrm{RD}}(u)|^2 \frac{|f^*_\varepsilon(u)|^2 - 3|\hat{f}^*_{\varepsilon,\mathrm{NS}}(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^4} \,\mathrm{d}u.
\end{aligned}$$

Again, on $\mathcal{E}_{Y,\mathrm{NS}}$,

$$\sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}} \left\{ \frac{2}{\pi} \int_{A(m,k)} \frac{|\hat{f}^*_{Y,\mathrm{NS}}(u) - f^*_Y(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^2} \,\mathrm{d}u - \frac{3}{4} \widehat{\mathrm{pen}}_{1,\mathrm{NS}}(m,k) \right\}_+ = 0.$$

Using the fact that for $a > 0$, $x, y \in \mathbb{C}$, $|x+y|^2 \leq (1+a)|x|^2 + (1+1/a)|y|^2$ holds, we find that

$$\begin{aligned}
&\frac{1}{3\pi} \int_{A(m,k)} |\hat{f}^*_{Y,\mathrm{RD}}(u)|^2 \frac{|f^*_\varepsilon(u)|^2 - 3|\hat{f}^*_{\varepsilon,\mathrm{NS}}(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^4} \,\mathrm{d}u \\
&\leq \frac{1}{2\pi} \int_{A(m,k)} |\hat{f}^*_{Y,\mathrm{RD}}(u)|^2 \frac{|f^*_\varepsilon(u) - \hat{f}^*_{\varepsilon,\mathrm{NS}}(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^4} \,\mathrm{d}u.
\end{aligned}$$

Consequently, on $\mathcal{E}_{\varepsilon,\mathrm{NS}}$,

$$\begin{aligned}
&\sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}} \left\{ \frac{1}{3\pi} \int_{A(m,k)} |\hat{f}^*_{Y,\mathrm{RD}}(u)|^2 \frac{|f^*_\varepsilon(u)|^2 - 3|\hat{f}^*_{\varepsilon,\mathrm{NS}}(u)|^2}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^4} \,\mathrm{d}u - \frac{3}{4} \widehat{\mathrm{pen}}_{2,\mathrm{NS}}(m,k) \right\}_+ \\
&\leq \sup_{\substack{k\geq m\\ k,m\in\mathcal{M}}} \left\{ \frac{1}{2\pi} \int_{A(m,k)} |\hat{f}^*_{Y,\mathrm{RD}}(u)|^2 \frac{\tau_\varepsilon^2 k_M^2(u)}{|\tilde{f}^*_{\varepsilon,\mathrm{NS}}(u)|^4} \,\mathrm{d}u - \frac{3}{4} \widehat{\mathrm{pen}}_{2,\mathrm{NS}}(m,k) \right\}_+ = 0.
\end{aligned}$$

Finally, we find that

$$\mathbb{P}\left[\mathcal{E}^c_{Y,\mathrm{NS}} \cup \mathcal{E}^c_{\varepsilon,\mathrm{NS}}\right] \leq C(n^{-3/2} + M^{-3/2}),$$

leading to the desired bound on the complement set. □

6.3. Proof of the oracle bounds

Proposition 6.5 was the essential technical step towards proving the oracle bounds. It shows how the penalty terms control the fluctuation of $\|\hat{f}^*_m\|^2$ around its target, be it in the model of repeated measurements or in the deconvolution model with additional noise sample. Once this result is fixed, the considerations for both models run exactly along the same line. For this reason, we may henceforth drop all subscripts relative to one particular model.

Proof of Theorem 3.1

We denote by $m^\star$ the oracle cutoff,

$$m^\star = \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ -\|\hat{f}_m\|^2 + \mathrm{pen}(m) \right\}.$$

We have

$$\|f - \hat{f}_{\hat{m}}\|^2 \leq 2\|f - \hat{f}_{m^\star}\|^2 + 2\|\hat{f}_{m^\star} - \hat{f}_{\hat{m}}\|^2. \tag{23}$$

Taking expectation and applying Proposition 2.1 gives for the first summand on the right hand side

$$\mathbb{E}\left[\|f - \hat{f}_{m^\star}\|^2\right] \leq 2\|f - f_{m^\star}\|^2 + \mathrm{pen}(m^\star),$$

since the variance term is a fortiori bounded from above by the penalty term. So there remains to consider the second summand on the right hand side of (23).

• Consider first the set $\mathcal{G} = \{\widehat{m} \leq m^\star\}$. Let us notice that on $\mathcal{G}$

$$\left\|\hat{f}_{m^\star} - \hat{f}_{\widehat{m}}\right\|^2 \mathbb{1}_{\mathcal{G}} = \left(\left\|\hat{f}_{m^\star}\right\|^2 - \left\|\hat{f}_{\widehat{m}}\right\|^2 \right) \mathbb{1}_{\mathcal{G}}.$$

Besides according to the definition of $\hat{m}$, one has the following inequality:

$$-\left\|\hat{f}_{\widehat{m}}\right\|^2 + \widehat{\mathrm{pen}}(\widehat{m}) \leq -\left\|\hat{f}_{m^\star}\right\|^2 + \widehat{\mathrm{pen}}(m^\star), \tag{24}$$

which implies

$$-\left\|\hat{f}_{\hat{m}}\right\|^2 \leq -\left\|\hat{f}_{m^\star}\right\|^2 + \widehat{\mathrm{pen}}(m^\star).$$

Thus

$$\left\|\hat{f}_{m^\star} - \hat{f}_{\hat{m}}\right\|^2 \mathbb{1}_{\mathcal{G}} = \left(\left\|\hat{f}_{m^\star}\right\|^2 - \left\|\hat{f}_{\hat{m}}\right\|^2 \right) \mathbb{1}_{\mathcal{G}} \leq \widehat{\mathrm{pen}}(m^\star).$$

Taking expectation and applying Corollary 6.4 yields for some positive constant C,

$$\begin{aligned}\mathbb{E}\left[\left\|f - \hat{f}_{\hat{m}}\right\|^2 \mathbb{1}_{\mathcal{G}}\right] &\leq 2\mathbb{E}\left\|f - \hat{f}_{m^\star}\right\|^2 + 2\mathbb{E}\left[\widehat{\mathrm{pen}}(m^\star)\right] \\ &\leq 2\left\|f - f_{m^\star}\right\|^2 + 2C\mathrm{pen}(m^\star).\end{aligned}$$

We have thus proved the desired result on $\mathcal{G}$:

$$\mathbb{E}\left[\|f-\hat{f}_{\hat{m}}\|^2\mathbb{1}_{\mathcal{G}}\right] \le C \inf_{m\in\mathcal{M}} \left\{\|f-f_m\|^2 + \mathrm{pen}(m)\right\}. \tag{25}$$

• Next, we consider the set $\mathcal{G}^c = \{\widehat{m} > m^\star\}$. It holds that

$$\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2 \mathbb{1}_{\mathcal{G}^c} = 4\Big(\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2 - \frac{3}{4}\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2\Big)\mathbb{1}_{\mathcal{G}^c}.$$

We realize that, by definition of $\hat{f}_{\hat{m}}$, see (24), on $\mathcal{G}^c$,

$$\begin{aligned}
-\frac{3}{4}\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2 &= \frac{3}{4}\left(\left\|\hat{f}_{m^\star}\right\|^2 - \left\|\hat{f}_{\hat{m}}\right\|^2\right)\\
&\le \frac{3}{4}\left(\left\|\hat{f}_{m^\star}\right\|^2 - \left\|\hat{f}_{m^\star}\right\|^2 + \widehat{\mathrm{pen}}(m^\star) - \widehat{\mathrm{pen}}(\hat{m})\right)\\
&= -\frac{3}{4}\widehat{\mathrm{pen}}(m^\star, \hat{m}).
\end{aligned}$$

It follows from there that

$$\begin{aligned}
&\Big(\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2 - \frac{3}{4}\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2\Big)\mathbb{1}_{\mathcal{G}^c}\\
&\le \sup_{\substack{k\ge m^\star\\ k\in\mathcal{M}}}\left\{\left\|\hat{f}_k - \hat{f}_{m^\star}\right\|^2 - 6\left\|f_k - f_{m^\star}\right\|^2 - \frac{3}{4}\widehat{\mathrm{pen}}(m^\star, k)\right\}_+ + 6\sup_{k\ge m^\star}\|f_k - f_{m^\star}\|^2.
\end{aligned}$$

Taking expectation and applying Proposition 6.5, as well as the monotonicity of the bias term, we conclude that

$$\begin{aligned}
\mathbb{E}\left[\|f-\hat{f}_{\hat{m}}\|^2\mathbb{1}_{\mathcal{G}^c}\right] &\le 2\mathbb{E}\left[\left\|f - \hat{f}_{m^\star}\right\|^2\right] + 2\mathbb{E}\left[\left\|\hat{f}_{\hat{m}} - \hat{f}_{m^\star}\right\|^2\mathbb{1}_{\mathcal{G}^c}\right]\\
&\le C\mathbb{E}\left[\left\|f - \hat{f}_{m^\star}\right\|^2\right] + \frac{C}{N},
\end{aligned}$$

with $N = n$ in the RD-model and $N = M\wedge n$ in the NS-model.

This is the desired oracle bound for $\mathcal{G}^c$. □

Acknowledgments

The authors would like to thank F. Comte for her advice, suggestions and her understanding support during this work. They are also grateful to two anonymous referees for detailed and instructive remarks and comments which helped them to substantially improve this paper.

References

[1] Baudry, J., Maugis, C., and Michel, B. (2012). Slope heuristics: Overview and implementation. *Statistics and Computing*, 22(2):455–470. MR2865029

[2] Birgé, L. (1999). An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics: Festschrift for Willem R. van Zwet*, pages 113–133. IMS Lecture Notes-Monograph Series. MR1836557

[3] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York. MR1462939

[4] Bonhomme, S. and Robin, J.-M. (2010). Generalized non-parametric deconvolution with an application to earnings dynamics. *Review of Economic Studies*, 77(2):491–533. MR2650495

[5] Butucea, C. (2004). Deconvolution of supersmooth densities with smooth noise. *The Canadian Journal of Statistics*, 32(2):181–192. MR2064400

[6] Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98. MR2546799

[7] Butucea, C. and Tsybakov, A. (2008a). Sharp optimality in density deconvolution with dominating bias I. *Theory Proba. Appl.*, 52(1):24–39. MR2354572

[8] Butucea, C. and Tsybakov, A. (2008b). Sharp optimality in density deconvolution with dominating bias II. *Theory Proba. Appl.*, 52(2):237–249. MR2742504

[9] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186. MR0997599

[10] Comte, F. and Kappus, J. (2014). Density deconvolution from repeated measurements without symmetry assumption on the errors. Preprint hal-01010409.

[11] Comte, F. and Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *Journal of the Royal Statistical Society: Series B*, 73:601–627. MR2853732

[12] Comte, F. and Lacour, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. H. Poincaré Probab. Statist.*, 49(2):569–609. MR3088382

[13] Comte, F., Rozenholc, Y., and Taupin, M.-L. (2006). Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, 3(34):431–452. MR2328553

[14] Comte, F., Rozenholc, Y., and Taupin, M.-L. (2007). Finite sample penalization in adaptive density deconvolution. *Journal of Statistical Computation and Simulation*, 77(11):977–1000. MR2416478

[15] Comte, F. and Samson, A. (2012). Nonparametric estimation of random-effects densities in linear mixed-effects model. *Journal of Nonparametric Statistics*, 24(4):951–975. MR2995486

[16] Comte, F., Samson, A., and Stirnemann, J. (2012). Deconvolution estimation of onset of pregnancy with replicate observations. *Scandinavian Journal of Statistics*, 41(2):325–345. MR3207174

[17] Dattner, I., Reiss, M., and Trabs, M. (2013). Adaptive quantile estimation in deconvolution with unknown error distribution. arXiv:1303.1698. MR3069892

[18] Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, 36(2):665–685. MR2396811

[19] Dion, C. (2013). New strategies for nonparametric estimation in a linear mixed model. *Journal of Statistical Planning and Inference*, 150:30–48. MR3206719

[20] Efromovich, S. (1997). Density estimation for the case of supersmooth measurement errors. *Journal of the American Statistical Association*, 92:526–535. MR1467846

[21] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272. MR1126324

[22] Johannes, J. (2009). Deconvolution with unknown error distribution. *The Annals of Statistics*, 37(5a):2301–2323. MR2543693

[23] Johannes, J. and Schwarz, M. (2013). Adaptive circular deconvolution by model selection under unknown error distribution. *Bernoulli*, 19(5A):1576–1611. MR3129026

[24] Kappus, J. (2014). Adaptive nonparametric estimation for Lévy processes observed at low frequency. *Stochastic Processes and Their Applications*, 124:730–758. MR3131312

[25] Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65:139–165. MR1625869

[26] Massart, P. (2003). *Concentration Inequalities and Model Selection*. Number 1896 in Lecture Notes in Mathematics. Springer. MR2319879

[27] Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics. Springer. MR2768576

[28] Neumann, M. (2007). Deconvolution from panel data with unknown error distribution. *Journal of Multivariate Analysis*, 98(10):1955–1968. MR2396948

[29] Neumann, M. and Reiss, M. (2009). Nonparametric estimation for Lévy processes from low frequency observations. *Bernoulli*, 15(1):223–248. MR2546805

[30] Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7(4):307–330. MR1460203

[31] Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics*, 27(6):2033–2053. MR1765627

[32] Stefanski, L. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics and Probability Letters*, 9(3):229–235. MR1045189

[33] Stefanski, S. and Carroll, R. (1990). Deconvoluting kernel density estimators. *Statistics*, 21(2):169–184. MR1054861