*The Annals of Statistics*
2018, Vol. 46, No. 6B, 3767–3804
https://doi.org/10.1214/17-AOS1675


# RHO-ESTIMATORS REVISITED: GENERAL THEORY AND APPLICATIONS

BY YANNICK BARAUD AND LUCIEN BIRGÉ

*Université Côte d'Azur and Sorbonne Universités*

Following Baraud, Birgé and Sart [*Invent. Math.* **207** (2017) 425–517], we pursue our attempt to design a robust universal estimator of the joint distribution of $n$ independent (but not necessarily i.i.d.) observations for an Hellinger-type loss. Given such observations with an unknown joint distribution $\mathbf{P}$ and a dominated model $\mathscr{Q}$ for $\mathbf{P}$, we build an estimator $\widehat{\mathbf{P}}$ based on $\mathscr{Q}$ (a $\rho$-estimator) and measure its risk by an Hellinger-type distance. When $\mathbf{P}$ does belong to the model, this risk is bounded by some quantity which relies on the local complexity of the model in a vicinity of $\mathbf{P}$. In most situations, this bound corresponds to the minimax risk over the model (up to a possible logarithmic factor). When $\mathbf{P}$ does not belong to the model, its risk involves an additional bias term proportional to the distance between $\mathbf{P}$ and $\mathscr{Q}$, whatever the true distribution $\mathbf{P}$. From this point of view, this new version of $\rho$-estimators improves upon the previous one described in Baraud, Birgé and Sart [*Invent. Math.* **207** (2017) 425–517] which required that $\mathbf{P}$ be absolutely continuous with respect to some known reference measure. Further additional improvements have been brought as compared to the former construction. In particular, it provides a very general treatment of the regression framework with random design as well as a computationally tractable procedure for aggregating estimators. We also give some conditions for the maximum likelihood estimator to be a $\rho$-estimator. Finally, we consider the situation where the statistician has at her or his disposal many different models and we build a penalized version of the $\rho$-estimator for model selection and adaptation purposes. In the regression setting, this penalized estimator not only allows one to estimate the regression function but also the distribution of the errors.

**1. Introduction.** In a previous paper, namely Baraud, Birgé and Sart (2017), we introduced a new class of estimators that we called $\rho$-estimators for estimating the distribution $\mathbf{P}$ of a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ with values in some measurable space $(\mathscr{X}, \mathscr{B})$ under the assumption that the $X_i$ are independent but not necessarily i.i.d. These estimators are based on density models, a *density model* being a family of densities $\mathbf{t}$ with respect to some reference measure $\boldsymbol{\mu}$ on $\mathscr{X}$. We also assumed that $\mathbf{P}$ was absolutely continuous with respect to $\boldsymbol{\mu}$ with density $\mathbf{s}$ and, following Le Cam (1973), we measured the performance of an estimator $\widehat{\mathbf{s}}$

Received June 2016; revised November 2017.
*MSC2010 subject classifications.* 62G35, 62G05, 62G07, 62G08, 62C20, 62F99.
*Key words and phrases.* $\rho$-estimation, robust estimation, density estimation, regression with random design, statistical models, maximum likelihood estimators, metric dimension, VC-classes.

of $\mathbf{s}$ in terms of $\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})$, where $\mathbf{h}$ is a Hellinger-type distance to be defined later. Originally, the motivations for this construction were to design an estimator $\widehat{\mathbf{s}}$ of $\mathbf{s}$ with the following properties.

— Given a density model $\mathbf{S}$, the estimator $\widehat{\mathbf{s}}$ should be nearly optimal over $\mathbf{S}$ from the minimax point of view, which means that it is possible to bound the risk of the estimator $\widehat{\mathbf{s}}$ over $\mathbf{S}$ from above by some quantity $CD_n(\mathbf{S})$ which is approximately of the order of the minimax risk over $\mathbf{S}$.

— Since in statistics we typically have uncomplete information about the true distribution of the observations, when we assume that $\mathbf{s}$ belongs to $\mathbf{S}$ nothing ever warrants that this is true. We may more reasonably expect that $\mathbf{s}$ is close to $\mathbf{S}$ which means that the model $\mathbf{S}$ is not exact but only approximate and that the quantity $\mathbf{h}(\mathbf{s}, \mathbf{S}) = \inf_{\mathbf{t}\in\mathbf{S}} \mathbf{h}(\mathbf{s}, \mathbf{t})$ might therefore be positive. In this case, we would like the risk of $\widehat{\mathbf{s}}$ to be bounded by $C'[D_n(\mathbf{S}) + \mathbf{h}^2(\mathbf{s}, \mathbf{S})]$ for some universal constant $C'$. In the case of $\rho$-estimators, the previous bound can actually be slightly refined and expressed in the following way. It is possible to define on $\mathbf{S}$ a positive function $R$ such that the risk of the $\rho$-estimator is not larger than $R(\mathbf{s})$, with $R(\mathbf{s}) \leq CD_n(\mathbf{S})$ if $\mathbf{s}$ belongs to the model $\mathbf{S}$ and not larger than $C' \inf_{\overline{\mathbf{s}}\in\mathbf{S}}[R(\overline{\mathbf{s}}) + \mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}})]$ when $\mathbf{s}$ does not belong to $\mathbf{S}$.

The weak sensibility of this risk bound to small deviations with respect to the Hellinger-type distance $\mathbf{h}$ between $\mathbf{s}$ and an element $\overline{\mathbf{s}}$ of $\mathbf{S}$ covers some classical notions of robustness among which robustness to a possible contamination of the data and robustness to outliers, as we shall see in Section 5.

There are nevertheless some limitations to the properties of $\rho$-estimators as defined in Baraud, Birgé and Sart (2017).

(a) The study of random design regression required that either the distribution of the design be known or that the errors have a symmetric distribution. We want to relax these assumptions and consider the random design regression framework with greater generality.

(b) We always worked with some reference measure $\boldsymbol{\mu}$ and assumed that all the probabilities we considered, including the true distribution $\mathbf{P}$ of $\boldsymbol{X}$, were absolutely continuous with respect to $\boldsymbol{\mu}$. This is quite natural for the probabilities that belong to our models since the models are, by assumption, dominated and, typically, defined via a reference measure $\boldsymbol{\mu}$ and a family of densities with respect to $\boldsymbol{\mu}$. Nevertheless, the assumption that the true distribution $\mathbf{P}$ of the observations be also dominated by $\boldsymbol{\mu}$ is questionable. We therefore would like to get rid of it and let the true distribution be completely arbitrary, relaxing thus the assumption that the density $\mathbf{s}$ exists. Unexpectedly, such an extension leads to subtle complications as we shall see below and this generalization is actually far from being straightforward.

(c) Our construction was necessarily restricted to countable models rather than the uncountable ones currently used in statistics.

We want here to design a method based on "probability models" rather than "density models," which means working with dominated models $\mathcal{P}$ consisting of probabilities rather than of densities as for $\mathbf{S}$. Of course, the choice of a dominating measure $\boldsymbol{\mu}$ and a specific set $\mathbf{S}$ of densities leads to a probability model $\mathcal{P}$. This is by the way what is actually done in statistics, but the converse is definitely not true and there exist many ways of representing a dominated probability model by a reference measure and a set of densities. It turns out (see Section 2.3) that the performance of a very familiar estimator, namely the Maximum Likelihood Estimator (MLE), can be strongly affected by the choice of a specific version of the densities. Our purpose here is to design an estimator the performance of which only depends on the probability model $\mathcal{P}$ and not on the choice of the reference measure and the densities that are used to represent it.

In order to get rid of the above mentioned restrictions, we have to modify our original construction which leads to the new version that we present here. This new version retains all the nice properties that we proved in Baraud, Birgé and Sart (2017) and the numerous illustrations we considered there remain valid for the new version. It additionally provides a general treatment of conditional density estimation and regression, allowing the statistician to estimate both the regression function and the error distribution even when the distribution of the design is totally unknown and the errors admit no finite moments. From this point of view, our approach contrasts very much with that based on the classical least squares. An alternative point of view on the particular problem of estimating a conditional density can be found in Sart (2017).

A thorough study of the performance of the least squares estimator (or truncated versions of it) can be found in Györfi et al. (2002) and we refer the reader to the references therein. A nice feature of these results lies in the fact that they hold without any assumption on the distribution of the design. While few moment conditions on the errors are necessary to bound the $\mathbb{L}_2$-integrated risk of their estimator, much stronger ones, typically boundedness of the errors, are necessary to obtain exponential deviation bounds. In contrast, in linear regression, Audibert and Catoni (2011) established exponential deviation bounds for the risk of some robust versions of the ordinary least squares estimator. Their idea is to replace the sum of squares by the sum of their truncated version in view of designing a new criterion which is less sensitive to possible outliers than the original least squares. Their way of modifying the least squares criterion shares some similarity with our way of modifying the log-likelihood criterion, as we shall see below. However, their results require some conditions on the distribution of the design as well as some (weak) moment condition on the errors while ours do not.

It is known, and we shall give an additional example below, that the MLE, which is often considered as a "universal" estimator, does not possess, in general, the properties that we require and more specifically robustness. An illustration of the lack of robustness of the MLE with respect to Hellinger deviations is provided in Baraud and Birgé (2016). Some other weaknesses of the MLE have been described

in Le Cam (1990) and Birgé (2006), among other authors, and various alternatives aimed at designing some sorts of "universal" estimators (for the problem we consider here) which would not suffer from the same weaknesses have been proposed in the past by Le Cam (1973) and (1975) followed by Birgé (1983) and (2006). The construction of $\rho$-estimators, as described in Baraud, Birgé and Sart (2017) was in this line. In that paper, we actually introduced $\rho$-estimators via a testing argument as was the case for Le Cam and Birgé for their methods. This argument remains valid for the generalized version we consider here (see Lemma D.3 of the Supplementary Material) but $\rho$-estimators can also be viewed as a generalization, and in fact a robustified version, of the MLE. We shall even show, in Section 6, that in favorable situations (i.i.d. observations and a convex separable set of densities as a model for the true density) the MLE is actually a $\rho$-estimator and, therefore, shares their properties.

To explain the idea underlying the construction of $\rho$-estimators, let us assume that we observe an $n$-sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ with an unknown density $q$ belonging to a set $\overline{\mathcal{Q}}$ of densities with respect to some reference measure $\mu$. We may write the log-likelihood of $q$ as $\sum_{i=1}^n \log(q(X_i))$ and the log-likelihood ratios as

$$\mathbf{L}(\boldsymbol{X}, q, q') = \sum_{i=1}^n \log\biggl(\frac{q'(X_i)}{q(X_i)}\biggr) = \sum_{i=1}^n \log\bigl(q'(X_i)\bigr) - \sum_{i=1}^n \log\bigl(q(X_i)\bigr),$$

so that maximizing the likelihood is equivalent to minimizing with respect to $q$

$$\mathbf{L}(\boldsymbol{X}, q) = \sup_{q' \in \overline{\mathcal{Q}}} \sum_{i=1}^n \log\biggl(\frac{q'(X_i)}{q(X_i)}\biggr) = \sup_{q' \in \overline{\mathcal{Q}}} \mathbf{L}(\boldsymbol{X}, q, q').$$

This happens simply because of the magic property of the logarithm which says that $\log(a/b) = \log a - \log b$. However, the use of the unbounded log function in the definition of $\mathbf{L}(\boldsymbol{X}, q)$ leads to various problems that are responsible for some weaknesses of the MLE. Replacing the log function by another function $\varphi$ amounts to replace $\mathbf{L}(\boldsymbol{X}, q, q')$ by

$$\mathbf{T}(\boldsymbol{X}, q, q') = \sum_{i=1}^n \varphi\biggl(\frac{q'(X_i)}{q(X_i)}\biggr) \tag{1}$$

which is different from $\sum_{i=1}^n \varphi(q'(X_i)) - \sum_{i=1}^n \varphi(q(X_i))$ since $\varphi$ is not the log function. We may nevertheless define the analogue of $\mathbf{L}(\boldsymbol{X}, q)$, namely

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, q) = \sup_{q' \in \overline{\mathcal{Q}}} \mathbf{T}(\boldsymbol{X}, q, q') = \sup_{q' \in \overline{\mathcal{Q}}} \sum_{i=1}^n \varphi\biggl(\frac{q'(X_i)}{q(X_i)}\biggr) \tag{2}$$

and define our estimator $\widehat{q}(\boldsymbol{X})$ as a minimizer with respect to $q \in \overline{\mathcal{Q}}$ of the quantity $\boldsymbol{\Upsilon}(\boldsymbol{X}, q)$. The resulting estimator is an alternative to the MLE and we shall show that, for a suitable choice of a bounded function $\varphi$, it enjoys various properties, among which robustness, that are often not shared by the MLE.

To analyse the performance of this new estimator, we have to study the behaviour of the process $\mathbf{T}(\boldsymbol{X}, q, q')$ when $q$ is fixed, $q \cdot \mu$ is close to the true distribution of the $X_i$ and $q'$ varies in $\overline{\mathcal{Q}}$. Since the function $\varphi$ is bounded, the process is similar to those considered in learning theory for the purpose of studying empirical risk minimization. As a consequence, the tools we use are also similar to those described in great detail in Koltchinskii (2006).

It is well known that working with a single model for estimating an unknown distribution is not very efficient unless one has very precise pieces of information about the true distribution, which is rarely the case. Working with many models simultaneously and performing model selection improves the situation drastically. Refining the previous construction of $\rho$-estimators by adding suitable penalty terms to the statistic $\mathbf{T}(\boldsymbol{X}, q, q')$ allows one to work with a finite or countable family of probability models $\{\mathcal{P}_m, m \in \mathcal{M}\}$ instead of a single one, each model $\mathcal{P}_m$ leading to a risk bound of the form $C'[D_n(\mathcal{P}_m) + \mathbf{h}^2(\mathcal{P}_m, \mathbf{P})]$, and to choose from the observations a model with approximately the best possible bound which results in a final estimator $\widehat{\mathbf{P}}$ and a bound for $\mathbf{h}^2(\widehat{\mathbf{P}}, \mathbf{P})$ of the form

$$C'' \inf_{m \in \mathcal{M}} \big[ D_n(\mathcal{P}_m) + \mathbf{h}^2(\mathcal{P}_m, \mathbf{P}) + \Delta_m \big],$$

where the additional term $\Delta_m$ is connected to the complexity of the family of models we use.

The paper is organised as follows. We shall first make our framework, which is based on dominated families of probabilities rather than families of densities with respect to a given dominating measure, precise in Section 2. This section is devoted to the definition of models and of our new version of $\rho$-estimators, then to the assumptions that the function $\varphi$ we use to define the statistic $\mathbf{T}$ in (1) should satisfy. In Section 3, we define the $\rho$-dimension function of a model, a quantity which measures the difficulty of estimation within the model using a $\rho$-estimator, and present the main results, namely the performance of these new $\rho$-estimators. Section 4 is devoted to the extension of the construction from countable to uncountable statistical models (which are the ones currently used in statistics) under suitable assumptions. We describe the robustness properties of $\rho$-estimators in Section 5. In Section 6, we investigate the relationship between $\rho$-estimators and the MLE when the model is a convex set of densities. Section 7 provides various methods that allow one to bound the $\rho$-dimension functions of different types of models and indicates how these bounds are to be used to bound the risk of $\rho$-estimators in typical situations with applications to the minimax risk over classical statistical models. We also provide a few examples of computations of bounds for the $\rho$-dimension function. Many applications of our results about $\rho$-estimators have already been given in Baraud, Birgé and Sart (2017) and we deal here with a new one: estimation of conditional distributions in Section 8. In Section 9, we apply this additional result to the special case of random design regression when the distribution of the design is completely unknown, a situation for which not

many results are known. We provide here a complete treatment of this regression framework with simultaneous estimation of both the regression function and the density of the errors. Section 10 is devoted to estimator selection and aggregation: we show there how our procedure can be used either to select an element from a family of preliminary estimators or to aggregate them in a convex way. The supplemental article [Baraud and Birgé (2018)] contains most of the proofs as well as some additional facts and comments.

**2. Our new framework and estimation strategy.** As already mentioned, our method is based on statistical models which are sets of probability distributions, in opposition with more classical models which are sets of densities with respect to a given dominating measure.

2.1. *A probabilistic framework.* We observe a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ defined on some probability space $(\Omega, \Xi, \mathbb{P})$ with independent components $X_i$ and values in the measurable product space $(\mathscr{X}, \mathscr{B}) = (\prod_{i=1}^n \mathscr{X}_i, \bigotimes_{i=1}^n \mathscr{B}_i)$. We denote by $\mathscr{P}$ the set of all product probabilities on $(\mathscr{X}, \mathscr{B})$ and by $\mathbf{P} = \bigotimes_{i=1}^n P_i \in \mathscr{P}$ the true distribution of $\boldsymbol{X}$. We identify an element $\mathbf{Q} = \bigotimes_{i=1}^n Q_i$ of $\mathscr{P}$ with the $n$-tuple $(Q_1, \ldots, Q_n)$ and extend this identification to the elements $\boldsymbol{\mu} = \bigotimes_{i=1}^n \mu_i$ of the set $\mathscr{M}$ of all $\sigma$-finite product measures on $(\mathscr{X}, \mathscr{B})$.

When $\mathbf{Q}$ is absolutely continuous with respect to $\boldsymbol{\mu} \in \mathscr{M}$ ($\mathbf{Q} \ll \boldsymbol{\mu}$) or, equivalently, $\boldsymbol{\mu}$ dominates $\mathbf{Q}$, each $Q_i$, for $i = 1, \ldots, n$, is absolutely continuous with respect to $\mu_i$ with density $q_i$ so that $Q_i = q_i \cdot \mu_i$. We denote by $\mathcal{L}(\mu_i)$ the set of all densities with respect to $\mu_i$, that is, the set of measurable functions $q$ from $\mathscr{X}_i$ to $\mathbb{R}_+$ such that $\int_{\mathscr{X}_i} q(x)\, d\mu_i(x) = 1$. We then write $\mathbf{Q} = \mathbf{q} \cdot \boldsymbol{\mu}$ where $\mathbf{q}$ is the $n$-tuple $(q_1, \ldots, q_n)$ and we say that $\mathbf{q}$ is a density for $\mathbf{Q}$ with respect to $\boldsymbol{\mu}$. We denote by $\boldsymbol{\mathcal{L}}(\boldsymbol{\mu}) = \prod_{i=1}^n \mathcal{L}(\mu_i)$ the set of such densities $\mathbf{q}$ and by $\mathscr{P}^{\boldsymbol{\mu}}$ the set of all those $\mathbf{P}' \in \mathscr{P}$ which are absolutely continuous with respect to $\boldsymbol{\mu}$.

Our aim is to estimate the unknown distribution $\mathbf{P} = (P_1, \ldots, P_n)$ from the observation of $\boldsymbol{X}$. In order to evaluate the performance of an estimator $\widehat{\mathbf{P}}(\boldsymbol{X}) \in \mathscr{P}$ of $\mathbf{P}$, we shall introduce, following Le Cam (1975), an Hellinger-type distance $\mathbf{h}$ on $\mathscr{P}$. We recall that, given two probabilities $Q$ and $Q'$ on a measurable space $(\mathscr{X}, \mathscr{B})$, the Hellinger distance and the Hellinger affinity between $Q$ and $Q'$ are respectively given by

$$\begin{cases} h^2(Q, Q') = \dfrac{1}{2} \displaystyle\int_{\mathscr{X}} \left( \sqrt{\frac{dQ}{d\mu}} - \sqrt{\frac{dQ'}{d\mu}} \right)^2 d\mu, \\ \rho(Q, Q') = \displaystyle\int_{\mathscr{X}} \sqrt{\frac{dQ}{d\mu} \frac{dQ'}{d\mu}}\, d\mu = 1 - h^2(Q, Q'), \end{cases} \tag{3}$$

where $\mu$ denotes any measure that dominates both $Q$ and $Q'$, the result being independent of the choice of $\mu$. The Hellinger-type distance $\mathbf{h}(\mathbf{Q}, \mathbf{Q}')$ and affinity

$\boldsymbol{\rho}(\mathbf{Q}, \mathbf{Q}')$ between two elements $\mathbf{Q} = (Q_1, \ldots, Q_n)$ and $\mathbf{Q}' = (Q'_1, \ldots, Q'_n)$ of $\mathscr{P}$ are then given by the formulas

$$\mathbf{h}^2(\mathbf{Q}, \mathbf{Q}') = \sum_{i=1}^{n} h^2(Q_i, Q'_i) = \sum_{i=1}^{n} [1 - \rho(Q_i, Q'_i)] = n - \boldsymbol{\rho}(\mathbf{Q}, \mathbf{Q}').$$

We shall denote by $\mathscr{V}$ the topology of the metric space $(\mathscr{P}, \mathbf{h})$.

2.2. *Models and their representations*. Let us start with this definition.

DEFINITION 1. We call *model* any dominated subset $\overline{\mathscr{Q}}$ of $\mathscr{P}$ and we call *representation* of (the model) $\overline{\mathscr{Q}}$ a pair $\mathcal{R}(\overline{\mathscr{Q}}) = (\boldsymbol{\mu}, \overline{\mathbf{Q}})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is a $\sigma$-finite measure which dominates $\overline{\mathscr{Q}}$ and $\overline{\mathbf{Q}}$ is a subset of $\boldsymbol{\mathcal{L}}(\boldsymbol{\mu})$ such that for any $\mathbf{Q}$ in $\overline{\mathscr{Q}}$ there exists a unique density $\mathbf{q} \in \overline{\mathbf{Q}}$ with $\mathbf{Q} = \mathbf{q} \cdot \boldsymbol{\mu}$.

This means that, given a representation $(\boldsymbol{\mu}, \overline{\mathbf{Q}})$ of the model $\overline{\mathscr{Q}}$, we can associate to each probability $\mathbf{Q} \in \overline{\mathscr{Q}}$ a density $\mathbf{q} \in \overline{\mathbf{Q}}$ and vice versa. Clearly, a dominated subset $\overline{\mathscr{Q}}$ has different representations depending on the choice of the dominating measure $\boldsymbol{\mu}$ and the versions of the densities $q_i = dQ_i/d\mu_i$.

Our estimation strategy is based on specific dominated subsets of $\mathscr{P}$ that we call *$\rho$-models*.

DEFINITION 2. A $\rho$-model is a *countable* (which in this paper always means either finite or infinite and countable) subset $\mathscr{Q}$ of $\mathscr{P}$.

A $\rho$-model $\mathscr{Q}$ being countable, it is necessarily dominated. One should think of it as a probability set to which the true distribution is believed to be close (with respect to the Hellinger-type distance $\mathbf{h}$).

2.3. *Construction of a $\rho$-estimator on a model $\mathscr{Q}$*. Given the model $\mathscr{Q}$, our estimator is defined as a random element of $\mathrm{Cl}(\mathscr{Q})$, where $\mathrm{Cl}(\mathscr{R})$ denotes the closure of the subset $\mathscr{R}$ of $\mathscr{P}$ in the metric space $(\mathscr{P}, \mathbf{h})$, and its construction relies on a particular representation $\mathcal{R}(\mathscr{Q})$ of the model $\mathscr{Q}$. It actually depends on three elements with specific properties to be made precise below:

(i) A function $\psi$ (which will serve as a substitute for the logarithm to derive an alternative to the MLE) with the following properties.

ASSUMPTION 1. The function $\psi$ is nondecreasing from $[0, +\infty]$ to $[-1, 1]$, Lipschitz and satisfies

$$\psi(x) = -\psi(1/x) \quad \text{for all } x \in [0, +\infty), \text{ hence } \psi(1) = 0. \tag{4}$$

Throughout this paper, we shall only consider, without further notice, functions $\psi$ satisfying Assumption 1.

(ii) A model $\mathscr{Q} \subset \mathscr{P}$ (in most cases a $\rho$-model) with a representation $\mathcal{R}(\mathscr{Q}) = (\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$.

(iii) A *penalty function* "**pen**" mapping $\mathscr{Q}$ to $\mathbb{R}$, the role of which will be explained later in Section 3. We may, at first reading, assume that this penalty function is identically 0.

It is essential to note that the dominating measure $\boldsymbol{\mu}$ is chosen by the statistician and that there is no reason that the true distribution $\mathbf{P}$ of $\boldsymbol{X}$ be absolutely continuous with respect to $\boldsymbol{\mu}$. On the contrary, all probabilities $\mathbf{P}'$ on $\mathscr{X}$ belonging to $\mathrm{Cl}(\mathscr{Q})$ are absolutely continuous with respect to $\boldsymbol{\mu}$.

Given the function $\psi$ and the representation $\mathcal{R}(\mathscr{Q})$, we define the real-valued function $\mathbf{T}$ on $\mathscr{X} \times \boldsymbol{\mathcal{Q}} \times \boldsymbol{\mathcal{Q}}$ by

$$\mathbf{T}(\mathbf{x}, \mathbf{q}, \mathbf{q}') = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q_i'(x_i)}{q_i(x_i)}}\right) \quad \text{for } \mathbf{x} = (x_1, \ldots, x_n) \in \mathscr{X} \text{ and } \mathbf{q}, \mathbf{q}' \in \boldsymbol{\mathcal{Q}}, \tag{5}$$

with the conventions $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$. We then set (with $\mathbf{Q} = \mathbf{q} \cdot \boldsymbol{\mu}$ and $\mathbf{Q}' = \mathbf{q}' \cdot \boldsymbol{\mu}$)

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, \mathbf{q}) = \sup_{\mathbf{q}' \in \boldsymbol{\mathcal{Q}}} \left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') - \mathbf{pen}(\mathbf{Q}')\right] + \mathbf{pen}(\mathbf{Q}) \quad \text{for all } \mathbf{q} \in \boldsymbol{\mathcal{Q}}. \tag{6}$$

DEFINITION 3 ($\rho$-estimators). Let $\mathscr{E}(\psi, \boldsymbol{X})$ be the (nonvoid) set

$$\mathscr{E}(\psi, \boldsymbol{X}) = \left\{ \mathbf{Q} = \mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \boldsymbol{\mathcal{Q}} \,\middle|\, \boldsymbol{\Upsilon}(\boldsymbol{X}, \mathbf{q}) < \inf_{\mathbf{q}' \in \boldsymbol{\mathcal{Q}}} \boldsymbol{\Upsilon}(\boldsymbol{X}, \mathbf{q}') + \frac{\kappa}{25} \right\}, \tag{7}$$

where the positive constant $\kappa$ is given by (19) below. A *$\rho$-estimator* $\widehat{\mathbf{P}} = \widehat{\mathbf{P}}(\boldsymbol{X})$ relative to $(\mathcal{R}(\mathscr{Q}), \mathbf{pen})$ is any (measurable) element of $\mathrm{Cl}(\mathscr{E}(\psi, \boldsymbol{X}))$.

Since $\widehat{\mathbf{P}}$ belongs to $\mathrm{Cl}(\mathscr{E}(\psi, \boldsymbol{X}))$, the elements of which are dominated by $\boldsymbol{\mu}$, there exists a random density $\widehat{\mathbf{p}} = (\widehat{p}_1, \ldots, \widehat{p}_n)$ with $\widehat{p}_i \in \mathcal{L}(\mu_i)$ for $i = 1, \ldots, n$ such that $\widehat{\mathbf{P}} = \widehat{\mathbf{p}} \cdot \boldsymbol{\mu}$. Note that $\widehat{\mathbf{P}}$ might not belong to $\mathscr{Q}$.

As an immediate consequence of Assumption 1 and the convention $1/0 = +\infty$, $\psi(+\infty) = -\psi(0)$ and

$$\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}') = -\mathbf{T}(\boldsymbol{X}, \mathbf{q}', \mathbf{q}) \quad \text{for all } \mathbf{q}, \mathbf{q}' \in \boldsymbol{\mathcal{Q}}. \tag{8}$$

Moreover,

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, \mathbf{q}) \geq \left[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}) - \mathbf{pen}(\mathbf{Q})\right] + \mathbf{pen}(\mathbf{Q}) = \mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}) = n\psi(1) = 0$$

for all $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$, which implies that any element $\widehat{\mathbf{P}} = \widehat{\mathbf{p}} \cdot \boldsymbol{\mu}$ in $\mathscr{Q}$ such that $\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{p}}) < \kappa/25$ is a $\rho$-estimator. In particular, when $\mathbf{pen}(\mathbf{Q}) = 0$ for all $\mathbf{Q} \in \mathscr{Q}$ (which we shall write in the sequel $\mathbf{pen} = \mathbf{0}$) and $\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{p}}) = 0$, it follows from (6) that

$$\mathbf{T}(\boldsymbol{X}, \widehat{\mathbf{p}}, \mathbf{q}) \leq \boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{\mathbf{p}}) = 0 = \mathbf{T}(\boldsymbol{X}, \widehat{\mathbf{p}}, \widehat{\mathbf{p}}) \leq -\mathbf{T}(\boldsymbol{X}, \widehat{\mathbf{p}}, \mathbf{q}) = \mathbf{T}(\boldsymbol{X}, \mathbf{q}, \widehat{\mathbf{p}})$$

for all $\mathbf{q} \in \mathcal{Q}$. This means that, in this case, $(\widehat{\mathbf{p}}, \widehat{\mathbf{p}})$ is a saddle point of the map $(\mathbf{q}, \mathbf{q}') \mapsto \mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')$.

A $\rho$-estimator $\widehat{\mathbf{P}}$ depends on the chosen representation $\mathcal{R}(\mathscr{Q})$ of $\mathscr{Q}$ and there are different versions of the $\rho$-estimators associated to $\mathscr{Q}$, even though most of the time $\mathscr{Q}$ will directly be given by a specific representation, that is a family $\mathcal{Q}$ of densities with respect to some reference measure $\boldsymbol{\mu}$. Here is the important point, to be proven in Section 3: when $\mathscr{Q}$ is *a $\rho$-model*, the risk bounds we shall derive only depend on $\mathscr{Q}$ and the penalty function but not on the chosen representation of $\mathscr{Q}$, which allows us to choose the more convenient one for the construction. In contrast, the performances of many classical estimators are sensitive to the representation of the model $\mathscr{Q}$ and this is in particular the case of the MLE as shown by the following example.

PROPOSITION 1. *Let us consider a sequence of i.i.d. random variables $(X_k)_{k\geq 1}$ defined on a measurable space $(\Omega, \mathscr{A}, \mathbb{P})$ with normal distribution $P_\theta = \mathcal{N}(\theta, 1)$ for some unknown $\theta \in \mathbb{R}$. We choose for reference measure $\mu = \mathcal{N}(0,1)$ and for the version of $dP_\theta/d\mu$, $\theta \in \mathbb{R}$, the function*

$$p_\theta(x) = \exp\Big[\theta x - (\theta^2/2) + (\theta^2/2)\exp(x^2)\mathbb{1}_\theta(x)\mathbb{1}_{(0,+\infty)}(\theta)\Big]. \tag{9}$$

*Whatever the value of the true parameter $\theta$, on a set of probability tending to 1 when $n$ goes to infinity, the MLE is given by $X_{(n)} = \max\{X_1, \ldots, X_n\}$ and is therefore inconsistent.*

The proof of Proposition 1 is given in Section D.1 of the Supplementary Material [Baraud and Birgé (2018)]. Note that the usual choice for $p_\theta\colon x \mapsto \exp[-x\theta + (\theta^2/2)]$ for $dP_\theta/d\mu$ is purely conventional. Mathematically speaking, our choice (9) is perfectly correct but leads to an inconsistent MLE. Also note that the usual tools that are used to prove consistency of the MLE, like bracketing entropy [see, for instance, Theorem 7.4 of van de Geer (2000)] are not stable with respect to changes of versions of the densities in the family. The same is true for arguments based on VC-classes that we used in Baraud, Birgé and Sart (2017). Choosing a convenient set of densities to work with is well grounded as long as the reference measure $\boldsymbol{\mu}$ not only dominates the model but also the true distribution $\mathbf{P}$. If not, sets of null measure with respect to $\boldsymbol{\mu}$ might have a positive probability under $\mathbf{P}$ and it becomes unclear how the choice of this set of densities influences the performance of the estimator.

2.4. *Notation and conventions.* Throughout this paper, given a representation $\mathcal{R}(\overline{\mathscr{Q}}) = (\boldsymbol{\mu}, \overline{\mathcal{Q}})$ of a model $\overline{\mathscr{Q}}$, we shall use lower case letters $\mathbf{q}, \mathbf{q}', \ldots$ and $q_i, q_i', \ldots$ for denoting the chosen densities of $\mathbf{Q}, \mathbf{Q}', \ldots$ and $Q_i, Q_i', \ldots$ with respect to the reference measures $\boldsymbol{\mu}$ and $\mu_i$, respectively, for all $i = 1, \ldots, n$. We set $\log_+(x) = \max\{\log x, 0\}$ for all $x > 0$; $|A|$ denotes the cardinality of the

set $A$; $\mathscr{B}(\mathbf{P}, r) = \{\mathbf{Q} \in \mathscr{P} \,|\, \mathbf{h}(\mathbf{P}, \mathbf{Q}) \leq r\}$ is the closed Hellinger-type ball in $\mathscr{P}$ with center $\mathbf{Q}$ and radius $r$. Given a set $E$, a nonnegative function $\ell$ on $E \times E$, $x \in E$ and $A \subset E$, we set $\ell(x, A) = \inf_{y \in A} \ell(x, y)$. In particular, for $\mathscr{R} \subset \mathscr{P}$, $\mathbf{h}(\mathbf{P}, \mathscr{R}) = \inf_{\mathbf{R} \in \mathscr{R}} \mathbf{h}(\mathbf{P}, \mathbf{R})$. We set $x \vee y$ and $x \wedge y$ for $\max\{x, y\}$ and $\min\{x, y\}$, respectively. By convention $\sup_{\varnothing} = 0$, the ratio $u/0$ equals $+\infty$ for $u > 0$, $-\infty$ for $u < 0$ and 1 for $u = 0$.

2.5. *Our assumptions.* Given the $\rho$-model $\mathscr{Q}$, let us now indicate what properties the function $\psi$ (satisfying Assumption 1) are required in view of controlling the risk of the resulting $\rho$-estimators.

ASSUMPTION 2. Let $\mathscr{Q}$ be the $\rho$-model to be used for the construction of $\rho$-estimators. There exist three positive constants $a_0$, $a_1$, $a_2$ with $a_0 \geq 1 \geq a_1$ and $a_2^2 \geq 1 \vee (6a_1)$ such that, whatever the representation $\mathcal{R}(\mathscr{Q}) = (\boldsymbol{\mu}, \mathcal{Q})$ of $\mathscr{Q}$, the densities $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$, the probability $\mathbf{R} \in \mathscr{P}$ and $i \in \{1, \ldots, n\}$,

$$\int_{\mathscr{X}_i} \psi\left(\sqrt{\frac{q_i'}{q_i}}\right) dR_i \leq a_0 h^2(R_i, Q_i) - a_1 h^2\big(R_i, Q_i'\big) \tag{10}$$

and

$$\int_{\mathscr{X}_i} \psi^2\left(\sqrt{\frac{q_i'}{q_i}}\right) dR_i \leq a_2^2\big[h^2(R_i, Q_i) + h^2\big(R_i, Q_i'\big)\big]. \tag{11}$$

Note that the left-hand sides of (10) and (11) depend on the choices of the reference measures $\mu_i$ and versions of the densities $q_i = dQ_i/d\mu_i$ and $q_i' = dQ_i'/d\mu_i$ while the corresponding right-hand sides do not.

Given $\psi$ that satisfies Assumption 2, the values of $a_0$, $a_1$ and $a_2$ are clearly not uniquely defined but, in the sequel, when we shall say that Assumption 2 holds, this will mean that the function $\psi$ satisfies (10) and (11) with given values of these constants which will therefore be considered as fixed once $\psi$ has been chosen. When we shall say that some quantity depends on $\psi$, it will implicitly mean that it depends on these chosen values of $a_0$, $a_1$ and $a_2$.

An important consequence of (8), (10) and (11) is the fact that, for all $\mathbf{Q}$, $\mathbf{Q}'$ in $\mathscr{Q}$ and $\mathbf{P} \in \mathscr{P}$,

$$a_1\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) - a_0\mathbf{h}^2\big(\mathbf{P}, \mathbf{Q}'\big) \leq \mathbb{E}\big[\mathbf{T}\big(\boldsymbol{X}, \mathbf{q}, \mathbf{q}'\big)\big] \leq a_0\mathbf{h}^2(\mathbf{P}, \mathbf{Q}) - a_1\mathbf{h}^2\big(\mathbf{P}, \mathbf{Q}'\big). \tag{12}$$

These inequalities follow by summing the inequalities (10) with respect to $i$ with $\mathbf{R} = \mathbf{P}$, then exchanging the roles of $\mathbf{Q}$ and $\mathbf{Q}'$ and applying (8). They imply that the sign of $\mathbb{E}[\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')]$ tells us which of the two distributions $\mathbf{Q}$ and $\mathbf{Q}'$ is closer to the true one when the ratio between the distances $\mathbf{h}(\mathbf{P}, \mathbf{Q})$ and $\mathbf{h}(\mathbf{P}, \mathbf{Q}')$ is far enough from one.

In view of checking that a given function $\psi$ satisfies Assumption 2, the next result to be proved in Section D.3 of the Supplementary Material [Baraud and Birgé (2018)] is useful.

PROPOSITION 2. *If, for a particular representation* $\mathcal{R}(\mathcal{Q}) = (\boldsymbol{\mu}, \mathcal{Q})$ *of the* $\rho$*-model* $\mathcal{Q}$ *and any probability* $\mathbf{R} \in \mathscr{P}^{\boldsymbol{\mu}}$, *the function* $\psi$ *satisfies* (10) *and* (11) *for positive constants* $a_0 > 2$, $a_1 \le [(a_0 - 2)/2] \wedge 1$ *and* $a_2^2 \ge 1 \vee (6a_1)$, *then it satisfies Assumption* 2 *with the same constants* $a_0$, $a_1$ *and* $a_2$.

This proposition means that, up to a possible adjustment of the constants $a_0$ and $a_1$, it is actually enough to check that (10) and (11) hold true for a given representation $(\boldsymbol{\mu}, \mathcal{Q})$ of $\mathcal{Q}$ and all probabilities $\mathbf{R} \ll \boldsymbol{\mu}$.

Let us now introduce two functions $\psi$ which do satisfy Assumption 2.

PROPOSITION 3. *Let* $\psi_1$ *and* $\psi_2$ *be the functions taking the value* 1 *at* $+\infty$ *and defined for* $x \in \mathbb{R}_+$ *by*

$$\psi_1(x) = \frac{x-1}{\sqrt{x^2+1}} \quad \text{and} \quad \psi_2(x) = \frac{x-1}{x+1}.$$

*These two functions are continuously increasing from* $[0, +\infty]$ *to* $[-1, 1]$, *Lipschitz* (*with respective Lipschitz constants* 1.143 *and* 2) *and satisfy Assumption* 2 *for all* $\rho$*-models* $\mathcal{Q}$ *with* $a_0 = 4.97$, $a_1 = 0.083$, $a_2^2 = 3 + 2\sqrt{2}$ *for* $\psi_1$ *and* $a_0 = 4$, $a_1 = 3/8$, $a_2^2 = 3\sqrt{2}$ *for* $\psi_2$.

Both functions can therefore be used everywhere in the applications of the present paper. Nevertheless, we prefer $\psi_2$ because it leads to better constants in the risk bounds of the estimator. Proposition 3 is proved in Section D.4 of the Supplementary Material [Baraud and Birgé (2018)]. Some comments on Assumption 2 can be found in Section D.2 of the Supplementary Material [Baraud and Birgé (2018)]. When the $\rho$-model reduces to two elements, our selection procedure can be interpreted as a robust test between two simple hypotheses. Upper bounds on the errors of the first and second kinds are established in Section D.10 of the Supplementary Material [Baraud and Birgé (2018)].

## 3. The performance of $\rho$-estimators on $\rho$-models.

3.1. *The* $\rho$*-dimension function.* The deviation $\mathbf{h}(\mathbf{P}, \widehat{\mathbf{P}})$ between the true distribution $\mathbf{P}$ and a $\rho$-estimator $\widehat{\mathbf{P}}$ built on the $\rho$-model $\mathcal{Q}$ is controlled by two terms which are the analogue of the classical bias and variance terms and we shall first introduce a function that replaces here the variance.

Let $y > 0$, $\mathbf{P}, \overline{\mathbf{P}} \in \mathscr{P}$ and $\mathscr{P}_0$ be an arbitrary subset of $\mathscr{P}$; we define

$$\mathscr{B}^{\mathscr{P}_0}(\mathbf{P}, \overline{\mathbf{P}}, y) = \left\{ \mathbf{Q} \in \mathscr{P}_0 \,\middle|\, \mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{Q}) < y^2 \right\}$$

and for measurable nonnegative functions $\mathbf{q}$, $\mathbf{q}'$ on $(\mathscr{X}, \mathscr{B})$, we set

$$\mathbf{Z}(X, \mathbf{q}, \mathbf{q}') = \mathbf{T}(X, \mathbf{q}, \mathbf{q}') - \mathbb{E}[\mathbf{T}(X, \mathbf{q}, \mathbf{q}')]. \tag{13}$$

Given a representation $\mathcal{R} = (\boldsymbol{\mu}, \mathcal{Q})$ of $\mathcal{Q} \cup \{\overline{\mathbf{P}}\}$, we define

$$w(\mathcal{R}, \mathcal{Q}, \mathbf{P}, \overline{\mathbf{P}}, y) = \mathbb{E}\left[\sup_{\mathbf{Q} \in \mathcal{B}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y)} |\mathbf{Z}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})|\right], \tag{14}$$

where, for $\mathbf{Q} \in \mathcal{B}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) \subset \mathcal{Q}$, $\mathbf{q}$ denotes the (unique) element of $\mathcal{Q}$ such as $\mathbf{Q} = \mathbf{q} \cdot \boldsymbol{\mu}$ and $\overline{\mathbf{p}}$ denotes the element of $\mathcal{Q}$ such that $\overline{\mathbf{P}} = \overline{\mathbf{p}} \cdot \boldsymbol{\mu}$. We recall that we use the convention $\sup_{\varnothing} = 0$. Since $\mathcal{Q}$ is countable, so is $\mathcal{B}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) \subset \mathcal{Q}$. Therefore, the supremum of $|\mathbf{Z}(\boldsymbol{X}, \overline{\mathbf{p}}, \cdot)|$ over $\mathcal{B}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y)$ is measurable and the right-hand side of (14) is well defined. Also note that, since $\mathbf{T}(\boldsymbol{X}, \overline{\mathbf{p}}, \overline{\mathbf{p}}) = n\psi(1) = 0$,

$$\mathbb{E}\left[\sup_{\mathbf{Q} \in \mathcal{B}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y)} |\mathbf{Z}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})|\right] = \mathbb{E}\left[\sup_{\mathbf{Q} \in \mathcal{B}^{\mathcal{Q} \cup \{\overline{\mathbf{P}}\}}(\mathbf{P}, \overline{\mathbf{P}}, y)} |\mathbf{Z}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})|\right].$$

Hence $w(\mathcal{R}, \mathcal{Q}, \mathbf{P}, \overline{\mathbf{P}}, y) = w(\mathcal{R}, \mathcal{Q} \cup \{\overline{\mathbf{P}}\}, \mathbf{P}, \overline{\mathbf{P}}, y)$.

DEFINITION 4 ($\rho$-dimension function). Let $\mathcal{Q}$ be a $\rho$-model and $\psi$ some function satisfying Assumption 2 with constants $a_0$, $a_1$ and $a_2$. The *$\rho$-dimension function* $D^{\mathcal{Q}}$ of $\mathcal{Q}$ is the mapping from $\mathcal{P} \times \mathcal{P}$ to $[1, +\infty)$ given by

$$D^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}) = \left[\beta^2 \sup\left\{y^2 \,\middle|\, \mathbf{w}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) > \frac{a_1 y^2}{8}\right\}\right] \vee 1 \tag{15}$$

with $\beta = a_1/(4a_2)$ and

$$\mathbf{w}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) = \inf_{\mathcal{R}} w(\mathcal{R}, \mathcal{Q}, \mathbf{P}, \overline{\mathbf{P}}, y) \quad \text{for all } y > 0,$$

where the infimum runs over all the representations $\mathcal{R} = (\boldsymbol{\mu}, \mathcal{Q})$ of $\mathcal{Q} \cup \{\overline{\mathbf{P}}\}$.

Note that the $\rho$-dimension function of $\mathcal{Q}$ depends on the choice of the function $\psi$ and not on the choice of the representations of $\mathcal{Q} \cup \{\overline{\mathbf{P}}\}$. Since it measures the local fluctuations of the centred empirical process $\mathbf{Z}(\boldsymbol{X}, \overline{\mathbf{p}}, \mathbf{q})$ indexed by $\mathbf{q} \in \mathcal{Q}$, it is quite similar to the local Rademacher complexity introduced in Koltchinskii (2006) for the purpose of studying empirical risk minimization. Its importance comes from the following property.

PROPOSITION 4. *Let $\mathcal{Q}$ be a $\rho$-model, $\overline{\mathbf{P}} \in \mathcal{P}$ and $\mathcal{R} = (\boldsymbol{\mu}, \mathcal{Q})$, an arbitrary representation of $\mathcal{Q} \cup \{\overline{\mathbf{P}}\}$. Whatever $\mathbf{P} \in \mathcal{P}$,*

$$w(\mathcal{R}, \mathcal{Q}, \mathbf{P}, \overline{\mathbf{P}}, y) \le \mathbf{w}^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}, y) + 8\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) \quad \textit{for all } y > 0, \tag{16}$$

*hence, for all $y > \beta^{-1}\sqrt{D^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}})}$*

$$w(\mathcal{R}, \mathcal{Q}, \mathbf{P}, \overline{\mathbf{P}}, y) \le (a_1 y^2/8) + 8\mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}). \tag{17}$$

The proof is provided in Section D.5 of the Supplementary Material [Baraud and Birgé (2018)].

3.2. *Exponential deviation bounds.* Our first theorem, to be proven in Section A.2 of the Supplementary Material [Baraud and Birgé (2018)], deals with the situation of a null penalty function $\mathbf{pen} = \mathbf{0}$.

THEOREM 1. *Let* $\mathbf{P}$ *be an arbitrary distribution in* $\mathscr{P}$, $\mathscr{Q}$ *a* $\rho$*-model and* $\psi$ *a function satisfying Assumption 2. Whatever the representation* $\mathcal{R}$ *of* $\mathscr{Q}$, *a* $\rho$*-estimator* $\widehat{\mathbf{P}}$ *relative to* $(\mathcal{R}, \mathbf{0})$ *as defined in Section 2.3 satisfies, for all* $\overline{\mathbf{P}} \in \mathscr{Q}$ *and* $\xi > 0$,

$$\mathbb{P}\Big[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \gamma \mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \frac{4\kappa}{a_1}\Big(\frac{D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}})}{4.7} + 1.49 + \xi\Big)\Big] \ge 1 - e^{-\xi}, \tag{18}$$

*with*

$$\gamma = \frac{4(a_0 + 8)}{a_1} + 2 + \frac{84}{a_2^2} \quad \text{and} \quad \kappa = \frac{35a_2^2}{a_1} + 74, \ \text{hence} \ \frac{\kappa}{25} \ge 11.36. \tag{19}$$

*In particular, if the* $\rho$*-dimension function* $D^{\mathscr{Q}}$ *is bounded on* $\mathscr{P} \times \mathscr{Q}$ *by* $D_n \ge 1$, *then*

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P}, \mathscr{Q}) + D_n + \xi\Big] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0 \tag{20}$$

*and for some constant* $C > 0$ *which only depends on the choice of* $\psi$.

None of the quantities involved in (18) depends on the chosen representation $\mathcal{R}$ of $\mathscr{Q}$, which means that the performance of $\widehat{\mathbf{P}}$ does not depend on $\mathcal{R}$ although its construction depends on it. We shall therefore (abusively) refer to $\widehat{\mathbf{P}}$ as *a* $\rho$*-estimator on* $\mathscr{Q}$ omitting to mention what representation is used for its construction.

Introducing a nontrivial penalty function allows one to favour some probabilities as compared to others in $\mathscr{Q}$ and gives thus a Bayesian flavour to our estimation procedure. We shall mainly use it when we have at our disposal not only one single $\rho$-model for $\mathbf{P}$ but rather a countable collection $\{\mathscr{Q}_m, m \in \mathcal{M}\}$ of candidate ones, in which case $\mathscr{Q} = \bigcup_{m\in\mathcal{M}} \mathscr{Q}_m$ is still a $\rho$-model that we call the *reference* $\rho$*-model*. The penalty function may not only be used for estimating $\mathbf{P}$ but also for performing model selection among the family $\{\mathscr{Q}_m, m \in \mathcal{M}\}$ by deciding that the procedure selects the $\rho$-model $\mathscr{Q}_{\widehat{m}}$ if the resulting estimator $\widehat{\mathbf{P}}$ belongs to $\mathscr{Q}_{\widehat{m}}$. Since $\widehat{\mathbf{P}}$ may belong to several $\rho$-models, this selection procedure may result in a (random) set of possible $\rho$-models for $\mathbf{P}$ and a common way of selecting one is to choose that with the smallest *complexity* in a suitable sense. In the present paper, the complexity of a $\rho$-model $\mathscr{Q}_m$ will be measured by means of a nonnegative weight function $\Delta$ mapping $\mathcal{M}$ into $\mathbb{R}_+$ and which satisfies

$$\sum_{m\in\mathcal{M}} e^{-\Delta(m)} \le 1, \tag{21}$$

where the number "1" is chosen for convenience. When equality holds in (21), $e^{-\Delta(\cdot)}$ can be viewed as a prior distribution on the family of $\rho$-models $\{\mathcal{Q}_m, m \in \mathcal{M}\}$.

In such a context, we shall describe how our penalty term should depend on this weight function $\Delta$ in view of selecting a suitable $\rho$-model for $\mathbf{P}$. The next theorem is proved in Section A.3 of the Supplementary Material [Baraud and Birgé (2018)].

THEOREM 2. *Let $\mathbf{P}$ be an arbitrary distribution in $\mathcal{P}$, $\{\mathcal{Q}_m, m \in \mathcal{M}\}$ be a countable collection of $\rho$-models, $\Delta$ a weight function satisfying (21), $\mathcal{R}(\mathcal{Q})$ a representation of $\mathcal{Q} = \bigcup_{m\in\mathcal{M}} \mathcal{Q}_m$, $\psi$ a function satisfying Assumption 2 and $\kappa$ be given by (19). Assume that there exists a mapping $D_n : \mathcal{M} \to \mathbb{R}_+$ and a number $K \geq 0$ such that, whatever $m \in \mathcal{M}$,*

$$D^{\mathcal{Q}_m}(\mathbf{P}, \overline{\mathbf{P}}) \leq D_n(m) + K D_n(m') \quad \text{for all } (\mathbf{P}, \overline{\mathbf{P}}) \in \mathcal{P} \times \mathcal{Q}_{m'}. \tag{22}$$

*Let the penalty function satisfy, for some constant $\kappa_1 \in \mathbb{R}$,*

$$\mathbf{pen}(\mathbf{Q}) = \kappa_1 + \kappa \inf_{\{m\in\mathcal{M} | \mathcal{Q}_m \ni \mathbf{Q}\}} \left[ \frac{D_n(m)}{4.7} + \Delta(m) \right] \quad \text{for all } \mathbf{Q} \in \mathcal{Q}. \tag{23}$$

*Then any $\rho$-estimator $\widehat{\mathbf{P}}$ relative to $(\mathcal{R}(\mathcal{Q}), \mathbf{pen})$ satisfies, for all $\xi > 0$ with probability at least $1 - e^{-\xi}$ and with $\gamma$ given by (19),*

$$\begin{aligned} \mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \leq \inf_{m\in\mathcal{M}} \left[ \gamma \mathbf{h}^2(\mathbf{P}, \mathcal{Q}_m) + \frac{4\kappa}{a_1} \left( \frac{K+1}{4.7} D_n(m) + \Delta(m) \right) \right] \\ + \frac{4\kappa}{a_1}(1.49 + \xi). \end{aligned} \tag{24}$$

3.3. *The case of density estimation.* Of special interest is the situation where the $X_i$ are *assumed* to be i.i.d. with values in a measurable set $(\mathcal{X}, \mathcal{B})$ in which case $\mathcal{X} = \mathcal{X}^n$, $\mathcal{B} = \mathcal{B}^{\otimes n}$, $\mathcal{P}$ and $\mathcal{M}$ denote respectively the set of all probability distributions and all positive $\sigma$-finite measures on $(\mathcal{X}, \mathcal{B})$ and $\mathbf{P}$ is *expected* (although this is not necessarily true) to belong to $\mathcal{P}^n = \{P^{\otimes n}, P \in \mathcal{P}\}$. Note that the Hellinger distance $h(\cdot, \cdot)$ on $\mathcal{P}$ is related to the Hellinger-type distance $\mathbf{h}(\cdot, \cdot)$ on $\mathcal{P}^n$ in the following way:

$$\mathbf{h}^2(\mathbf{Q}, \mathbf{Q}') = nh^2(Q, Q') \quad \text{for all } Q, Q' \in \mathcal{P} \text{ with } \mathbf{Q} = Q^{\otimes n}, \mathbf{Q}' = (Q')^{\otimes n}.$$

If $\mathbf{P} = P^{\otimes n} \in \mathcal{P}^n$, estimating $\mathbf{P}$ then amounts to estimating the marginal distribution $P$ and we model the probability $P$ rather than $\mathbf{P}$.

DEFINITION 5. We call *density $\rho$-model* any countable subset $\mathcal{Q}$ of $\mathcal{P}$.

Given such a density $\rho$-model $\mathcal{Q}$ for $P$ with representation $(\mu, \mathcal{Q})$ (which implies that the mapping $q \mapsto Q = q \cdot \mu$ is one to one), the corresponding $\rho$-model for $\mathbf{P}$ is simply $\mathcal{Q} = \{\mathbf{Q} = Q^{\otimes n}, Q \in \mathcal{Q}\}$ with representation $(\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$, $\boldsymbol{\mu} = \mu^{\otimes n}$

and $\mathcal{Q} = \{\mathbf{q} : (x_1, \dots, x_n) \mapsto (q(x_1) \dots q(x_n)), q \in \mathcal{Q}\}$. In this case, for simplicity, we write $\mathbf{T}(\boldsymbol{X}, q, q')$ and $\boldsymbol{\Upsilon}(\boldsymbol{X}, q)$ for $\mathbf{T}(\boldsymbol{X}, \mathbf{q}, \mathbf{q}')$ and $\boldsymbol{\Upsilon}(\boldsymbol{X}, \mathbf{q})$, respectively. Examples involving density estimation will be considered in Sections 5, 6, 8 and 9 below.

We may also work with several density $\rho$-models $\{\mathscr{Q}_m, m \in \mathcal{M}\}$ for $P$ simultaneously, in which case $\mathscr{Q} = \bigcup_{m \in \mathcal{M}} \mathscr{Q}_m$ is also a density $\rho$-model. A penalty function pen on $\mathscr{Q}$ leads to a penalty function **pen** on $\boldsymbol{\mathscr{Q}} = \bigcup_{m \in \mathcal{M}} \boldsymbol{\mathscr{Q}}_m$ defined by $\mathbf{pen}(\mathbf{Q}) = \mathbf{pen}(Q^{\otimes n}) = \mathrm{pen}(Q)$ for all $Q \in \mathscr{Q}$. Any $\rho$-estimator $\widehat{\mathbf{P}}$ relative to $((\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}}), \mathbf{pen})$ is of the form $\widehat{\mathbf{P}} = \widehat{P}^{\otimes n}$ with $\widehat{P} \in \mathrm{Cl}(\mathscr{Q})$ and $\widehat{P}$ will be called a *density $\rho$-estimator* for $P$ relative to $((\mu, \mathcal{Q}), \mathrm{pen})$.

We deduce that, under the assumptions of Theorem 1, if $\mathbf{P}$ is truly of the form $\mathbf{P} = P^{\otimes n}$, for all $\overline{P} \in \mathscr{Q}$,

$$\mathbb{P}\left[ Ch^2(P, \widehat{P}) \le h^2\big(P, \overline{P}\big) + \frac{D^{\boldsymbol{\mathscr{Q}}}\Big(\mathbf{P}, \overline{P}^{\otimes n}\Big)}{n} + \frac{\xi}{n} \right] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

Under the assumptions of Theorem 2, for all $\xi > 0$ and a positive constant $C$ depending only on $\psi$,

$$\mathbb{P}\left[ Ch^2\big(P, \widehat{P}\big) \le \inf_{m \in \mathcal{M}} \left[ h^2(P, \mathscr{Q}_m) + \frac{D_n(m) + \Delta(m)}{n} \right] + \frac{\xi}{n} \right] \ge 1 - e^{-\xi}.$$

**4. From $\rho$-models to uncountable statistical models.** The previous results apply to statistical models $\boldsymbol{\mathscr{Q}}$ that are countable, which is not the common case in statistics. The aim of this section is to explain how our general theory on $\rho$-models can be used to solve estimation problems on models that are possibly uncountable. Hereafter, we shall denote by $\overline{\boldsymbol{\mathscr{Q}}}$ a *general statistical model*, that is, an arbitrary subset of $\boldsymbol{\mathscr{P}}$.

4.1. *Working with nets.* Let us first recall this classical definition.

DEFINITION 6. Given $\eta \ge 0$, a subset $\boldsymbol{\mathscr{Q}}$ of $\overline{\boldsymbol{\mathscr{Q}}}$ such that $\mathbf{h}(\mathbf{Q}, \boldsymbol{\mathscr{Q}}) \le \eta$ for all $\mathbf{Q} \in \overline{\boldsymbol{\mathscr{Q}}}$ is called an $\eta$-net of $\overline{\boldsymbol{\mathscr{Q}}}$. The case $\eta = 0$ corresponds to the situation where $\boldsymbol{\mathscr{Q}}$ is $\mathscr{V}$-dense in $\overline{\boldsymbol{\mathscr{Q}}}$.

If there exists a countable $\eta$-net $\boldsymbol{\mathscr{Q}}$ for $\overline{\boldsymbol{\mathscr{Q}}}$, it is a $\rho$-model. If its $\rho$-dimension $D^{\boldsymbol{\mathscr{Q}}}$ is bounded by $D_n = D_n(\eta) \ge 1$ on $\boldsymbol{\mathscr{P}} \times \boldsymbol{\mathscr{Q}}$, we deduce from Theorem 1 and the inequality $\mathbf{h}(\mathbf{P}, \boldsymbol{\mathscr{Q}}) \le \mathbf{h}(\mathbf{P}, \overline{\boldsymbol{\mathscr{Q}}}) + \eta$ that any $\rho$-estimator on $\boldsymbol{\mathscr{Q}}$ satisfies

$$\mathbb{P}\Big[ C\mathbf{h}^2\big(\mathbf{P}, \widehat{\mathbf{P}}\big) \le \mathbf{h}^2\big(\mathbf{P}, \overline{\boldsymbol{\mathscr{Q}}}\big) + D_n(\eta) + \eta^2 + \xi \Big] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0, \tag{25}$$

hence

$$\mathbb{E}\Big[ \mathbf{h}^2\big(\mathbf{P}, \widehat{\mathbf{P}}\big) \Big] \le C' \Big[ \mathbf{h}^2\big(\mathbf{P}, \overline{\boldsymbol{\mathscr{Q}}}\big) + D_n(\eta) + \eta^2 \Big] \tag{26}$$

for some constants $C, C' > 0$ depending on $\psi$ only. Most of the statistical models $\overline{\mathcal{Q}}$ that are used in statistics possess $\eta$-nets for all values of $\eta \geq 0$. Since the $\rho$-dimension function $D^{\mathcal{Q}}$ can only increase with inclusion, choosing for each $\eta \geq 0$ an $\eta$-net with the smallest possible cardinality and then the value $\eta^\star$ of $\eta$ that minimizes $D_n(\eta) + \eta^2$ leads to a $\rho$-estimator $\widehat{\mathbf{P}}$ with the smallest possible risk bound in (26). This risk bound turns out to be minimax (up to possible extra logarithmic factors) in all cases we know; see Section 7.1.

4.2. *Models that are universally separable.* Following Pollard (1984), we shall say that a class of densities $\overline{\mathcal{Q}} \subset \mathcal{L}(\boldsymbol{\mu})$ is *universally separable* if one can find a countable subset $\mathcal{Q} \subset \overline{\mathcal{Q}}$ such that, for each $\mathbf{q} \in \overline{\mathcal{Q}}$, there exists a sequence $\left(\mathbf{q}^{(j)}\right)_{j \geq 1}$ in $\mathcal{Q}$ which converges towards $\mathbf{q}$ pointwise, that is,

$$q_i^{(j)}(x) \underset{j \to +\infty}{\longrightarrow} q_i(x) \quad \text{for } 1 \leq i \leq n \text{ and all } x \in \mathcal{X}_i. \tag{27}$$

We shall then say that $\mathcal{Q}$ is $\mathcal{T}$*-dense* in $\overline{\mathcal{Q}}$. Note that if $\left(\mathbf{q}^{(j)}\right)_{j \geq 1}$ converges towards $\mathbf{q}$ pointwise, by Scheffé's lemma, the sequence of probabilities $\mathbf{Q}_j = \mathbf{q}^{(j)} \cdot \boldsymbol{\mu}$ converges in total variation, hence in Hellinger distance, towards $\mathbf{Q} = \mathbf{q} \cdot \boldsymbol{\mu}$. This implies that if $\mathcal{Q}$ is $\mathcal{T}$-dense in $\overline{\mathcal{Q}}$, the set of probabilities $\mathcal{Q} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \mathcal{Q}\}$ is $\mathcal{V}$-dense in $\overline{\mathcal{Q}} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\mathcal{Q}}\}$.

We shall work here within the following framework. For some $\boldsymbol{\mu} \in \mathcal{M}$, let $\{\overline{\mathcal{Q}}_m, m \in \mathcal{M}\}$ be a countable family of universally separable subsets of $\mathcal{L}(\boldsymbol{\mu})$ with $\mathcal{Q}_m \subset \overline{\mathcal{Q}}_m$ a countable and $\mathcal{T}$-dense subset of $\overline{\mathcal{Q}}_m$. We set $\overline{\mathcal{Q}} = \bigcup_{m \in \mathcal{M}} \overline{\mathcal{Q}}_m$, $\mathcal{Q} = \bigcup_{m \in \mathcal{M}} \mathcal{Q}_m$, $\overline{\mathcal{Q}}_m = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\mathcal{Q}}_m\}$ for all $m \in \mathcal{M}$, $\overline{\mathcal{Q}} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\mathcal{Q}}\}$ and $\mathcal{Q} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \mathcal{Q}\}$. Note that $\mathcal{Q}$ is a $\rho$-model since $\mathcal{Q}$ is countable and that $\mathcal{Q}$ is $\mathcal{V}$-dense in $\overline{\mathcal{Q}}$ since $\mathcal{Q}$ is $\mathcal{T}$-dense in $\overline{\mathcal{Q}}$. Let now $\overline{\mathbf{pen}}$ be some penalty function on $\overline{\mathcal{Q}}$ with the following property.

ASSUMPTION 3. There exists a function $p : \mathcal{M} \to \mathbb{R}$ such that

$$\overline{\mathbf{pen}}(\mathbf{Q}) = \inf_{m \in \mathcal{M}, \overline{\mathcal{Q}}_m \ni \mathbf{Q}} p(m) \quad \text{for all } \mathbf{Q} \in \overline{\mathcal{Q}} \tag{28}$$

and, for any $\mathbf{Q} \in \overline{\mathcal{Q}}$, there exists some $m_{\mathbf{Q}} \in \mathcal{M}$ such that $\mathbf{Q} \in \overline{\mathcal{Q}}_{m_{\mathbf{Q}}}$ and $\overline{\mathbf{pen}}(\mathbf{Q}) = p(m_{\mathbf{Q}})$.

Note that this assumption holds in particular in the case of a single model with $\overline{\mathbf{pen}} = \mathbf{0}$. Within this framework, we can prove the following result.

THEOREM 3. *Let $\{\overline{\mathcal{Q}}_m, ; m \in \mathcal{M}\}$ be a countable family of universally separable subsets of $\mathcal{L}(\boldsymbol{\mu})$ and $\overline{\mathbf{pen}}$ a penalty function on $\overline{\mathcal{Q}}$ that satisfies Assumption 3. Any $\rho$-estimator $\widehat{\mathbf{P}}$ on $\overline{\mathcal{Q}}$ relative to $((\boldsymbol{\mu}, \overline{\mathcal{Q}}), \overline{\mathbf{pen}})$ is also a $\rho$-estimator on the $\rho$-model $\mathcal{Q}$ relative to $((\boldsymbol{\mu}, \mathcal{Q}), \mathbf{pen})$ where $\mathbf{pen}$ is the restriction of $\overline{\mathbf{pen}}$ to $\mathcal{Q}$.*

The proof is postponed to Section A.5 of the Supplementary Material [Baraud and Birgé (2018)].

This result says that, provided that the penalty function satisfies (28), which is consistent with (23), the construction of a $\rho$-estimator on the possibly uncountable set $\overline{\mathscr{Q}}$ with representation $(\boldsymbol{\mu}, \overline{\mathcal{Q}})$ actually results in a $\rho$-estimator based on the $\rho$-model $\mathscr{Q}$.

As soon as we can control the $\rho$-dimension function of $\mathscr{Q}$ by some features of $\overline{\mathscr{Q}}$, in the case of a single model, or the $\rho$-dimension functions of the $\rho$-models $\mathscr{Q}_m = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \mathcal{Q}_m\}$ by the features of the models $\overline{\mathscr{Q}}_m$, in the general case, we are able to bound the risk of the $\rho$-estimator relative to $((\boldsymbol{\mu}, \overline{\mathcal{Q}}), \overline{\mathbf{pen}})$ using the results of Theorems 1 and 2.

For illustration, let us mention a few examples of density sets that are universally separable:

(a) the set $\overline{\mathcal{H}}_D$ of right-continuous histograms on $\mathbb{R}$ with at most $D \geq 1$ pieces;

(b) for $L > 0$ and $\alpha = r + \beta$ with $r \in \mathbb{N}$, $\beta \in (0, 1]$, the set $\overline{\mathcal{H}}_\alpha(L)$ of functions $f$ on $[0, 1]$ that are $r$-times differentiable and satisfy

$$\left| f^{(r)}(x) - f^{(r)}(y) \right| \leq L|x - y|^\beta \quad \text{for all } x, y \in [0, 1];$$

(c) the set $\overline{\mathcal{H}}_\downarrow$ of nonincreasing and right-continuous densities on $(0, +\infty)$.

The set $\overline{\mathcal{H}}_\alpha(L)$ is universally separable because the larger set consisting of continuous functions on $[0, 1]$ is separable for the topology induced by the norm of the uniform convergence; hence all its subsets are separable with respect to this topology which implies pointwise convergence. We prove that the sets $\overline{\mathcal{H}}_D$ and $\overline{\mathcal{H}}_\downarrow$ are universally separable in Section B of the Supplementary Material [Baraud and Birgé (2018)]. We shall see in Section 6 that the MLE on the convex density sets $\overline{\mathcal{H}}_\alpha(L)$ and $\overline{\mathcal{H}}_\downarrow$ is actually a $\rho$-estimator.

**5. Why is a $\rho$-estimator robust?** The aim of this section is to analyse the robustness properties of $\rho$-estimators. For the sake of simplicity, we shall restrict ourselves to the particular case of density estimation as described in Section 3.3.

5.1. *Misspecification and contamination.* We assume here that we work with a single $\rho$-model $\mathscr{Q}$ (so that Theorem 1 applies) for which $D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}})$ is bounded from above independently of $\mathbf{P} \in \mathscr{P}$ and $\overline{\mathbf{P}} \in \mathscr{Q}$ by some some number $D_n(\mathcal{Q}) \geq 1$ depending on the marginal model $\mathcal{Q}$ and the number $n$ of marginals. Examples of such situations will be provided in Section 7.

When $\mathbf{P} = P^{\otimes n}$, that is when the data are truely i.i.d. with marginal distribution $P$, (18) becomes

$$\mathbb{P}\Big[Ch^2\big(P, \widehat{P}\big) \leq h^2(P, \mathcal{Q}) + n^{-1}\big[D_n(\mathcal{Q}) + \xi\big]\Big] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0, \tag{29}$$

where $C$ is a positive constant only depending on $\psi$.

The bias term in (29), namely $h^2(P, \mathscr{Q})$, accounts for the robustness property of the $\rho$-estimator with respect to the Hellinger distance and measures the additional loss we get as compared to the case when $P$ belongs to $\mathscr{Q}$. If this quantity is small, the performance of the $\rho$-estimator will not deteriorate too much as compared to the ideal situation where $P$ does belong to $\mathscr{Q}$. In fact, if there exists some probability $\overline{P} \in \mathscr{Q}$ such that $h^2(P, \mathscr{Q}) = h^2(P, \overline{P})$ is small as compared to $D_n(\mathscr{Q})/n$, everything is almost as if the $\rho$-estimator $\widehat{P}$ were built from an i.i.d. sample with distribution $\overline{P}$. The $\rho$-estimators under $P$ and $\overline{P}$ would therefore look the same. This includes the following situations.

*Misspecification.* The true distribution $P$ of the observations does not belong to $\mathscr{Q}$ but is close to $\mathscr{Q}$. For example, let $\mathscr{Q}$ be countable and $\mathscr{V}$-dense in the set of all Gaussian distributions on $\mathbb{R}^k$ with identity covariance matrix and mean vector belonging to a linear subspace $\overline{S} \subset \mathbb{R}^k$. Assume that the true distribution $P$ has the same form except for the fact that its mean does not belong to $\overline{S}$ but is at Euclidean distance $\varepsilon > 0$ from $\overline{S}$. Then it follows from classical formulas that

$$h^2(P, \mathscr{Q}) = 1 - e^{-\varepsilon^2/8} \leq \varepsilon^2/8.$$

*Contamination.* The true distribution $P$ is of the form $(1 - \varepsilon)\overline{P} + \varepsilon R$ with $\overline{P} \in \mathscr{Q}$ and $R \neq \overline{P}$ but otherwise arbitrary. This situation arises when a proportion $\varepsilon \in (0, 1)$ of the sample $X_1, \ldots, X_n$ is contaminated by another sample. It follows from the convexity property of the Hellinger distance that

$$h^2(P, \mathscr{Q}) \leq h^2\bigl(P, \overline{P}\bigr) \leq \varepsilon h^2\bigl(R, \overline{P}\bigr) \leq \varepsilon,$$

and this bound holds whatever the contaminating distribution $R$. From a more practical point of view, one can see the contaminated case as follows: for each $i$, one decides between no contamination with a probability $1 - \varepsilon$ and contamination with a probability $\varepsilon$ and draws $X_i$ accordingly with distribution either $\overline{P}$ or $R$. If it were possible to extract from the sample $X_1, \ldots, X_n$ these $N$ data, with $N \sim \mathcal{B}(n, 1 - \varepsilon)$, which are really distributed according to the distribution $\overline{P} \in \mathscr{Q}$, we would build a $\rho$-estimator $\widetilde{P}$ on these data. The robustness property ensures that the $\rho$-estimator $\widehat{P}$ based on the whole data set remains close to $\widetilde{P}$. Everything works almost as if the $\rho$-estimator $\widehat{P}$ only considered the noncontaminated subsample and ignored the other data, at least when $\varepsilon$ is small enough.

5.2. *More robustness.* There is an additional aspect of robustness that is not apparent in (29). Our general result about the performance of $\rho$-estimators, as stated in (18), actually allows that our observations be independent but not necessarily i.i.d., in which case the joint distribution $\mathbf{P}$ of $(X_1, \ldots, X_n)$ is actually of the form $\bigotimes_{i=1}^n P_i$ but not necessarily of the form $P^{\otimes n}$. Of course we do not know whether $\mathbf{P}$ is of the first form or the second and, proceeding as if $X_1, \ldots, X_n$ were

i.i.d., we build a $\rho$-estimator $\widehat{P} \in \mathrm{Cl}(\mathscr{Q})$ of the presumed common density $P$ and make a mistake which is no longer $h^2(P, \widehat{P})$ but

$$\frac{1}{n}\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \quad \text{with } \widehat{\mathbf{P}} = \widehat{P}^{\otimes n} \qquad \text{and} \qquad \mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) = \sum_{i=1}^{n} h^2(P_i, \widehat{P}),$$

which is consistent with the i.i.d. case $P_i = P$ for all $i$. In this context, we actually get the following analogue of (29): for all $\xi > 0$,

$$\mathbb{P}\left[\frac{C}{n}\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \inf_{Q \in \mathscr{Q}}\left(\frac{1}{n}\sum_{i=1}^{n} h^2(P_i, Q)\right) + \frac{D_n(\mathscr{Q}) + \xi}{n}\right] \ge 1 - e^{-\xi}. \tag{30}$$

This allows many more possibilities of deviations between $\mathbf{P}$ and the statistical model $\{Q^{\otimes n}, Q \in \mathscr{Q}\}$. For instance, we may have $h(P_i, \overline{P}) \le \varepsilon$ for some $\overline{P} \in \mathscr{Q}$ and all $i$, $P_i \ne P_{i'}$ for all $i \ne i'$, and nevertheless

$$\inf_{Q \in \mathscr{Q}}\left(\frac{1}{n}\sum_{i=1}^{n} h^2(P_i, Q)\right) \le \varepsilon^2.$$

An alternative situation corresponds to a small number of "outliers", namely, $P_i = P$ except on a subset $J \subset \{1, \dots, n\}$ of indices of small cardinality and, for $i \in J$, $P_i$ is completely arbitrary, for instance a Dirac measure. In such a case, for any probability $Q$,

$$\left(1 - \frac{|J|}{n}\right)h^2(P, Q) \le \frac{1}{n}\sum_{i=1}^{n} h^2(P_i, Q) \le \left(1 - \frac{|J|}{n}\right)h^2(P, Q) + \frac{|J|}{n},$$

and we deduce from (30) that, on a set of probability at least $1 - e^{-\xi}$,

$$\frac{C(n - |J|)}{n}h^2(P, \widehat{P}) \le C\frac{\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}})}{n} \le \left[\left(\frac{n - |J|}{n}\right)h^2(P, \mathscr{Q}) + \frac{|J|}{n}\right] + \frac{D_n(\mathscr{Q}) + \xi}{n}.$$

Finally,

$$\mathbb{P}\left[Ch^2(P, \widehat{P}) \le h^2(P, \mathscr{Q}) + \frac{|J| + D_n(\mathscr{Q}) + \xi}{n - |J|}\right] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

When $|J|/n$ is small enough, this bound appears to be a slight modification of what we would get from (29) if $\mathbf{P}$ were of the form $P^{\otimes n}$. This means that the $\rho$-estimator $\widehat{P}$ is also robust with respect to a possible departure from the assumption that the $X_i$ are i.i.d.

**6. The $\rho$-estimators and the MLE.** As mentioned in the Introduction, there are some deep connexions between the MLE and $\rho$-estimators which are mostly due to the similarities in the neighbourhood of 1 between the logarithm and the functions $\psi$ of Proposition 3. A nice result in this direction was communicated to the authors by Weijie Su in October 2016. It concerns the case of density estimation, as described in Section 3.3 with a single density model $\overline{\mathscr{D}} = \{q \cdot \mu, q \in \overline{\mathcal{Q}}\}$ where $\overline{\mathcal{Q}}$ is universally separable as defined in Section 4.2.

ASSUMPTION 4. The function $x \mapsto \varphi(x) = \psi(\sqrt{x})$, where $\psi$ is the function used to define the statistic $\mathbf{T}$ in (5), satisfies $\varphi(1) = 0$, is concave and admits a positive derivative at 1.

PROPOSITION 5 [Weijie Su, private communication (2016)]. *Let Assumption 4 hold, $\overline{\mathcal{Q}}$ be a convex set of densities on the measured space $(\mathscr{X}, \mathscr{B}, \mu)$ and the likelihood be not identically equal to 0 on $\overline{\mathcal{Q}}$. The maximum likelihood estimator $\widehat{Q} = \widehat{q} \cdot \mu$ on the density model $\overline{\mathscr{D}} = \{q \cdot \mu, q \in \overline{\mathcal{Q}}\}$, when it exists, satisfies*

$$\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{q}) = \sup_{q' \in \overline{\mathcal{Q}}} \mathbf{T}(\boldsymbol{X}, \widehat{q}, q') = 0 = \inf_{q \in \overline{\mathcal{Q}}} \sup_{q' \in \overline{\mathcal{Q}}} \mathbf{T}(\boldsymbol{X}, q, q') = \inf_{q \in \overline{\mathcal{Q}}} \boldsymbol{\Upsilon}(\boldsymbol{X}, q)$$

*and is therefore a $\rho$-estimator relative to $(\overline{\mathcal{Q}}, 0)$.*

PROOF. Given the data $X_1, \ldots, X_n$, if the maximum likelihood $\widehat{q}$ exists, it is unique since the logarithm is strictly concave. Moreover, $\widehat{q}(X_i) > 0$ for all $i \in \{1, \ldots, n\}$. Since $\boldsymbol{\Upsilon}(\boldsymbol{X}, \widehat{q}) \geq \mathbf{T}(\boldsymbol{X}, \widehat{q}, \widehat{q}) = 0$, it suffices to prove that

$$L(q) = \mathbf{T}(\mathbf{x}, \widehat{q}, q) = \sum_{i=1}^{n} \varphi\left(\frac{q(X_i)}{\widehat{q}(X_i)}\right) \leq 0 \quad \text{for all } q \in \overline{\mathcal{Q}}.$$

For $q \in \overline{\mathcal{Q}}$ and $\varepsilon \in [0, 1]$, $(1 - \varepsilon)\widehat{q} + \varepsilon q \in \overline{\mathcal{Q}}$ and, when $\varepsilon \to 0$,

$$\begin{aligned} L\big((1 - \varepsilon)\widehat{q} + \varepsilon q\big) &= n\varphi(1) + \varepsilon\left[\varphi'(1) \sum_{i=1}^{n} \frac{q(X_i)}{\widehat{q}(X_i)} + o(1)\right] \\ &= \varepsilon\left[\varphi'(1) \sum_{i=1}^{n} \frac{q(X_i)}{\widehat{q}(X_i)} + o(1)\right] \end{aligned} \tag{31}$$

since $\varphi(1) = 0$. When $\varphi$ is the logarithm and $\varepsilon > 0$, the right-hand side of (31) is negative since $\widehat{q}$ is the unique MLE. Letting $\varepsilon$ go to 0 we derive that

$$\sum_{i=1}^{n} \frac{q(X_i)}{\widehat{q}(X_i)} \leq 0 \quad \text{for all } q \in \overline{\mathcal{Q}}. \tag{32}$$

Moreover, the concavity of $\varphi$ implies that for all $\varepsilon \in [0,1]$

$$\varphi\biggl(\frac{(1-\varepsilon)\widehat{q}(X_i)+\varepsilon q(X_i)}{\widehat{q}(X_i)}\biggr) \le \varphi(1)+\varepsilon\frac{q(X_i)}{\widehat{q}(X_i)}\varphi'(1) = \varepsilon\varphi'(1)\frac{q(X_i)}{\widehat{q}(X_i)}$$

so that, for all $q \in \overline{\mathcal{Q}}$, $L((1-\varepsilon)\widehat{q}+\varepsilon q) \le \varepsilon\varphi'(1)\sum_{i=1}^n q(X_i)/\widehat{q}(X_i)$ and

$$\mathbf{T}\bigl(\mathbf{x},\widehat{q},q\bigr) = L(q) \le \varphi'(1)\sum_{i=1}^n \frac{q(X_i)}{\widehat{q}(X_i)} \le 0 \quad \text{for all } q \in \overline{\mathcal{Q}}$$

by (32), which completes the proof. □

Note that both functions $\psi_1$ and $\psi_2$ of Proposition 3 satisfy Assumption 4.

We may now derive the following relationship between the MLE and $\rho$-estimators, the proof of which immediately follows from Theorem 3 and Su's proposition.

COROLLARY 1. *Let $\overline{\mathcal{Q}}$ be a convex set of densities on the measured space $(\mathscr{X},\mathscr{B},\mu)$ which is universally separable on $\mathscr{X}$ with countable and $\mathscr{T}$-dense subset $\mathcal{Q}$ and $\psi$ satisfy Assumptions 1 and 4. The maximum likelihood estimator $\widehat{Q}=\widehat{q}\cdot\mu$ on the density model $\overline{\mathscr{Q}}=\{q\cdot\mu, q\in\overline{\mathcal{Q}}\}$, when it exists, is a $\rho$-estimator on the $\rho$-density model $\mathscr{Q}=\{q\cdot\mu, q\in\mathcal{Q}\}$.*

For illustration, the set $\overline{\mathcal{Q}}=\overline{\mathcal{H}}_{\mathcal{I}}$ of right-continuous histograms based on a fixed partition $\mathcal{I}$ of $[0,1)$ into $D\ge 1$ intervals is convex and obviously universally separable. The usual histogram $\widehat{p}$ based on $\mathcal{I}$, which corresponds to the MLE on $\overline{\mathcal{H}}_{\mathcal{I}}$, can be viewed as a $\rho$-estimator on a countable subset of $\overline{\mathcal{H}}_{\mathcal{I}}$. Taking back some of the examples of convex and universally separable density sets given in Section 4.2, we deduce that the MLE on $\overline{\mathcal{H}}_{\downarrow}$, that is, the Grenander estimator, or on the set $\overline{\mathcal{H}}_{\alpha}(L)$ are also $\rho$-estimators.

**7. Bounding the $\rho$-dimension function of a $\rho$-model with applications to the risk of $\rho$-estimators.** It clearly follows from the results of Section 3 that bounding the risk of $\rho$-estimators amounts to bounding the $\rho$-dimension of $\rho$-models which we shall now do under various assumptions. Throughout this section, we fix the function $\psi$ satisfying Assumption 2 (typically $\psi_1$ or $\psi_2$) and when we shall say that some quantity depends on $\psi$, this will mean that it actually depends on $a_1$ and $a_2$.

In view of (20), of special interest is the situation where the $\rho$-dimension function $(\mathbf{P},\overline{\mathbf{P}})\mapsto D^{\mathscr{Q}}(\mathbf{P},\overline{\mathbf{P}})$ of the $\rho$-model $\mathscr{Q}$ is uniformly bounded from above on $\mathscr{P}\times\mathscr{Q}$ by some constant $D_n\ge 1$. Let us begin by a few elementary considerations. If one can find a representation $\mathcal{R}=(\boldsymbol{\mu},\boldsymbol{\mathcal{Q}})$ of $\mathscr{Q}\cup\{\overline{\mathbf{P}}\}$ such that $w(\mathcal{R},\mathscr{Q},\mathbf{P},\overline{\mathbf{P}},y)\le a_1y^2/8$ for all $y\ge\beta^{-1}\sqrt{D}$, we immediately derive from the definition of $D^{\mathscr{Q}}$ that

$$D^{\mathscr{Q}}\bigl(\mathbf{P},\overline{\mathbf{P}}\bigr)\le D\vee 1 \quad \text{for } \bigl(\mathbf{P},\overline{\mathbf{P}}\bigr)\in\mathscr{P}^2. \tag{33}$$

In particular, since $|\psi| \le 1$, the expectation in (14) is never larger than $2n$ so that $w(\mathcal{R}, \mathscr{Q}, \mathbf{P}, \overline{\mathbf{P}}, y) \le a_1 y^2/8$ for $y \ge 4\sqrt{(n/a_1)}$ and (33) always holds with

$$\sqrt{D} = 4\beta\sqrt{(n/a_1)} = \sqrt{na_1}/a_2 \quad \text{or equivalently} \quad D = na_1/a_2^2 \le n/6.$$

Finally, whatever the choices of $\mathscr{Q}$ and $\psi$,

$$D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \le n/6 \quad \text{for all } (\mathbf{P}, \overline{\mathbf{P}}) \in \mathscr{P}^2. \tag{34}$$

More precise bounds will now be given that depend on some specific features of $\mathscr{Q}$.

7.1. *The finite case.* Given a finite subset $\mathscr{Q} \subset \mathscr{P}$, let us set

$$\mathscr{H}(\mathscr{Q}, y) = \sup_{\mathbf{P} \in \mathscr{P}} \log_+\bigl(2|\mathscr{Q} \cap \mathscr{B}(\mathbf{P}, y)|\bigr) \quad \text{for all } y > 0 \tag{35}$$

and, for $x_0 = \sqrt{2}[\sqrt{1 + (\beta/a_2)} + 1]$,

$$\overline{\eta} = \sup\left\{z > 0, \sqrt{\mathscr{H}(\mathscr{Q}, z/\beta)} > z/x_0\right\}. \tag{36}$$

Since $\mathscr{Q}$ is finite, the function $y \mapsto \mathscr{H}(\mathscr{Q}, y)$ is bounded by $\log(2|\mathscr{Q}|)$ and since $\beta/a_2 = a_1/(4a_2^2) \le 1/24$,

$$\overline{\eta} \le x_0\sqrt{\log\bigl(2|\mathscr{Q}|\bigr)} < 3\sqrt{\log\bigl(2|\mathscr{Q}|\bigr)}. \tag{37}$$

PROPOSITION 6. *If $\mathscr{Q}$ is a finite subset of $\mathscr{P}$ and $\overline{\eta}$ is defined by* (36),

$$D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \le D_n(\mathscr{Q}) = \overline{\eta}^2 \vee 1 < 9\log\bigl(2|\mathscr{Q}|\bigr) \quad \textit{for all } (\mathbf{P}, \overline{\mathbf{P}}) \in \mathscr{P}^2.$$

The proof of this result is given in Section D.6 of the of the Supplementary Material [Baraud and Birgé (2018)]. The first upper bound $\overline{\eta}^2 \vee 1$ for $D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}})$ neither depends on $\mathbf{P}$ nor on $\overline{\mathbf{P}}$ but might depend on $\beta$. The second bound only depends on the cardinality of $\mathscr{Q}$ and, therefore, holds whatever $\psi$.

If a model $\overline{\mathscr{Q}}$ is a totally bounded subset of the metric space $(\mathscr{P}, \mathbf{h})$ and $\eta > 0$, one can cover $\overline{\mathscr{Q}}$ by a finite number of closed balls of radius $\eta$ and the set $\mathscr{Q}[\eta]$ of their centers is an $\eta$-net for $\overline{\mathscr{Q}}$ (see Definition 6), which means that $\mathbf{h}(\mathbf{Q}, \mathscr{Q}[\eta]) \le \eta$ for all $\mathbf{Q} \in \overline{\mathscr{Q}}$. The function $y \mapsto \mathscr{H}(\mathscr{Q}[\eta], y)$ measures in a sense the massiveness of $\mathscr{Q}[\eta]$ and turns out to be a useful tool to measure that of $\overline{\mathscr{Q}}$. We shall in particular use the following classical notions of dimension based on the metric structure of $\overline{\mathscr{Q}}$.

DEFINITION 7. A set $\overline{\mathscr{Q}} \subset \mathscr{P}$ is said to have a metric dimension bounded by $\widetilde{D}$, where $\widetilde{D}$ is a right-continuous function from $(0, +\infty)$ to $[1/2, +\infty]$, if, for any positive $\eta$, there exists an $\eta$-net $\mathscr{Q}[\eta]$ for $\overline{\mathscr{Q}}$ which satisfies

$$\mathscr{H}\bigl(\mathscr{Q}[\eta], y\bigr) \le (y/\eta)^2 \widetilde{D}_n(\eta) \quad \text{for all } y \ge 2\eta. \tag{38}$$

We shall say that $\overline{\mathscr{Q}}$ has an entropy dimension bounded by $V \geq 0$ if, for any $\eta > 0$, there exists an $\eta$-net $\mathscr{Q}[\eta]$ of $\overline{\mathscr{Q}}$ such that

$$\mathscr{H}\big(\mathscr{Q}[\eta], y\big) \leq V \log(y/\eta) \quad \text{for all } y \geq 2\eta. \tag{39}$$

For the sake of convenience, we have slightly modified the original definition of the metric dimension due to Birgé [(2006), Definition 6, page 293] which is actually obtained by replacing the left-hand side of (38) by $\mathscr{H}(\mathscr{Q}[\eta], y) - \log 2$. Since in both definitions the metric dimension is not smaller than $1/2$, it is easy to check that, if $\overline{\mathscr{Q}}$ has a metric dimension bounded by $D_M$ in Birgé's sense, it has a metric dimension bounded by $\widetilde{D} = (1 + (\log 2)/2)D_M$ in our sense and, conversely, if $\overline{\mathscr{Q}}$ has a metric dimension bounded by $\widetilde{D}$ in our sense, it also has a metric dimension bounded by $\widetilde{D}$ in Birgé's sense. Hence, changing $\widetilde{D}$ into $D_M$ only changes the numerical constants.

The logarithm being a slowly varying function, it is not difficult to see that the notion of metric dimension is more general than the entropy one in the sense that if $\overline{\mathscr{Q}}$ has an entropy dimension bounded by $V$, then it also has a metric dimension bounded by $\widetilde{D}_n(\cdot)$ with

$$\widetilde{D}_n(\eta) \leq (1/2) \vee \big[V(\log 2)/4\big] \quad \text{for all } \eta > 0. \tag{40}$$

If $\overline{\mathscr{Q}}$ has a metric dimension bounded by $\widetilde{D}$ and if $\eta$ is a positive number satisfying

$$\widetilde{D}_n(\eta) \leq (\beta\eta/x_0)^2, \tag{41}$$

with $x_0$ given by (36), we deduce from (38) that there exists an $\eta$-net $\mathscr{Q}[\eta]$ for $\overline{\mathscr{Q}}$ for which

$$\sqrt{\mathscr{H}\big(\mathscr{Q}[\eta], z/\beta\big)} \leq z/x_0 \quad \text{for all } z \geq 2\beta\eta.$$

It then follows that $\overline{\eta}$, as defined in (36), satisfies $\overline{\eta} \leq 2\beta\eta$ and we deduce from Proposition 6 that the $\rho$-dimension function $D^{\mathscr{Q}}$ of $\mathscr{Q} = \mathscr{Q}[\eta] \subset \overline{\mathscr{Q}}$ satisfies

$$D^{\mathscr{Q}}\big(\mathbf{P}, \overline{\mathbf{P}}\big) \leq D_n(\mathscr{Q}) = (2\beta\eta)^2 \vee 1 \quad \text{for all } \big(\mathbf{P}, \overline{\mathbf{P}}\big) \in \mathscr{P}^2. \tag{42}$$

If, in particular, $\overline{\mathscr{Q}}$ has an entropy dimension bounded by $V \geq 0$ we deduce from (40) that (41) holds for

$$\eta^2 = \frac{x_0^2}{2\beta^2}\left(1 \vee \frac{V \log 2}{2}\right) < \frac{9}{2\beta^2}\left(1 \vee \frac{V \log 2}{2}\right) \tag{43}$$

and we derive from (42) that

$$D^{\mathscr{Q}}\big(\mathbf{P}, \overline{\mathbf{P}}\big) \leq D_n(\mathscr{Q}) = 18\left(1 \vee \frac{V \log 2}{2}\right) \quad \text{for all } \big(\mathbf{P}, \overline{\mathbf{P}}\big) \in \mathscr{P}^2. \tag{44}$$

Since in both cases $\mathbf{h}(\mathbf{P}, \mathscr{Q}) \leq \mathbf{h}(\mathbf{P}, \overline{\mathscr{Q}}) + \eta$ for all $\mathbf{P} \in \mathscr{P}$ because $\mathscr{Q}$ is a $\eta$-net for $\overline{\mathscr{Q}}$, we obtain from (42), (44) and (25) the following result.

COROLLARY 2. *Let $\psi$ be a function satisfying Assumption 2.*

(i) *If $\overline{\mathscr{Q}}$ has a metric dimension bounded by $\widetilde{D}$ and $\eta$ satisfies (41), any $\rho$-estimator $\widehat{\mathbf{P}}$ based on a suitable $\eta$-net $\mathscr{Q}$ for $\overline{\mathscr{Q}}$ satisfies for all $\mathbf{P} \in \mathscr{P}$ and $\xi > 0$:*

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P}, \overline{\mathscr{Q}}) + (\eta^2 \vee 1) + \xi\Big] \ge 1 - e^{-\xi}. \tag{45}$$

(ii) *If $\overline{\mathscr{Q}}$ has an entropy dimension bounded by $V$ and $\eta$ satisfies (43), any $\rho$-estimator $\widehat{\mathbf{P}}$ based on a suitable $\eta$-net $\mathscr{Q}$ for $\overline{\mathscr{Q}}$ satisfies for all $\mathbf{P} \in \mathscr{P}$ and $\xi > 0$:*

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P}, \overline{\mathscr{Q}}) + (V \vee 1) + \xi\Big] \ge 1 - e^{-\xi}. \tag{46}$$

*In both cases, $C$ is a constant depending only on the choice of $\psi$.*

7.2. *Bounds based on the VC-index.* In this section, we investigate the case of a model $\overline{\mathscr{Q}}$ given by a specific representation $(\boldsymbol{\mu}, \overline{\mathcal{Q}})$ where the density set $\overline{\mathcal{Q}}$ is possibly uncountable but satisfies some property to be described below.

A density $\mathbf{q} = (q_1, \ldots, q_n) \in \boldsymbol{\mathcal{L}}(\boldsymbol{\mu})$ can be viewed as a function on $\overline{\mathscr{X}} = \bigcup_{i=1}^n (\{i\} \times \mathscr{X}_i)$ defined, for $\overline{x} = (i, x)$ with $x \in \mathscr{X}_i$, by $\mathbf{q}(i, x) = q_i(x)$ so that a subset $\overline{\mathcal{Q}} \subset \boldsymbol{\mathcal{L}}(\boldsymbol{\mu})$ is now viewed as a class of real-valued functions on $\overline{\mathscr{X}}$. A common notion of dimension for the class $\overline{\mathcal{Q}}$ is the following one.

DEFINITION 8. A class $\mathscr{F}$ of functions from a set $\mathscr{X}$ with values in $(-\infty, +\infty]$ is VC-subgraph with index $\overline{V}$ (or equivalently with dimension $\overline{V} - 1 \ge 0$) if the class of subgraphs $\{(x, u) \in \mathscr{X} \times \mathbb{R} \mid f(x) > u\}$ as $f$ varies in $\mathscr{F}$ is a VC-class of sets in $\mathscr{X} \times \mathbb{R}$ with index $\overline{V}$ (or dimension $\overline{V} - 1$).

We recall that, by definition, the index $\overline{V}$ of a VC-class is a positive integer, hence its dimension $\overline{V} - 1 \in \mathbb{N}$. For additional information about VC-classes and related notions, we refer to van der Vaart and Wellner (1996) and Baraud, Birgé and Sart (2017), Section 8.

PROPOSITION 7. *Let $\psi$ satisfy Assumption 2 and $\overline{\mathcal{Q}} \subset \boldsymbol{\mathcal{L}}(\boldsymbol{\mu})$ be a VC-subgraph class of densities on $\overline{\mathscr{X}}$ with index not larger than $\overline{V}$. For any $\rho$-model $\mathscr{Q} \subset \overline{\mathscr{Q}} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\mathcal{Q}}\}$, for all $(\mathbf{P}, \overline{\mathbf{P}}) \in \mathscr{P} \times \mathscr{P}^{\boldsymbol{\mu}}$*

$$D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \le D_n(\overline{\mathscr{Q}}) = C_1(\overline{V} \wedge n)\big[1 + \log_+(n/\overline{V})\big],$$

*where $C_1$ is a universal constant.*

The proof is given in Section D.7 of the Supplementary Material [Baraud and Birgé (2018)]. A nice feature of this bound lies in the fact that it neither depends on the choices of $\psi$ nor on the cardinality of $\mathscr{Q}$ which can therefore be arbitrarily large. In particular, when $\mathscr{Q}$ is $\mathscr{V}$-dense in $\overline{\mathscr{Q}}$ we deduce the following result from Proposition 7, (25) (with $\eta = 0$) and Theorem 3.

COROLLARY 3. *Let $\psi$ be a function satisfying Assumption 2 and $\overline{\mathcal{Q}} \subset \mathcal{L}(\boldsymbol{\mu})$ a VC-subgraph class of densities on $\overline{\mathscr{X}}$ with index $\overline{V}$. Any $\rho$-estimator $\widehat{\mathbf{P}}$ built on a countable and $\mathscr{V}$-dense subset $\mathscr{Q}$ of $\overline{\mathscr{Q}} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\mathcal{Q}}\}$ satisfies, for all $\mathbf{P} \in \mathscr{P}$ and $\xi > 0$,*

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P}, \overline{\mathscr{Q}}) + (\overline{V} \wedge n)\big[1 + \log_+(n/\overline{V})\big] + \xi\Big] \ge 1 - e^{-\xi}, \tag{47}$$

*where the constant $C$ only depends on the choice of $\psi$. If, moreover, $\overline{\mathcal{Q}}$ is universally separable, then (47) holds for any $\rho$-estimator relative to $((\boldsymbol{\mu}, \overline{\mathcal{Q}}), \mathbf{0})$.*

In the particular case of density estimation, the following result is useful in view of applying Proposition 7.

PROPOSITION 8. *If $\overline{\mathcal{Q}} \subset \mathcal{L}(\mu)$ is VC-subgraph on $\mathscr{X}$ with index $\overline{V}$, then the set $\overline{\mathcal{Q}} = \{\mathbf{q} = (q, \dots, q), q \in \overline{\mathcal{Q}}\} \subset \mathcal{L}(\boldsymbol{\mu})$ is VC-subgraph on $\overline{\mathscr{X}}$ with an index not larger than $\overline{V}$.*

PROOF. If the class of subgraphs $\{(\overline{x}, u) \in \overline{\mathscr{X}} \times \mathbb{R} \,|\, \mathbf{q}(\overline{x}) > u\}$, with $\mathbf{q}$ running in $\overline{\mathcal{Q}}$, shatters the subset $\{(\overline{x}_1, u_1), \dots, (\overline{x}_k, u_k)\}$ of $\overline{\mathscr{X}} \times \mathbb{R}$, then whatever $J \subset \{1, \dots, k\}$, there exists $\mathbf{q} \in \overline{\mathcal{Q}}$ such that $j \in J$ is equivalent to $\mathbf{q}(\overline{x}_j) = q(x_j) > u_j$. Hence, the class of subgraphs $\{(x, u) \in \mathscr{X} \times \mathbb{R} \,|\, q(x) > u\}$ with $q$ running in $\overline{\mathcal{Q}}$ shatters the subset $\{(x_1, u_1), \dots, (x_k, u_k)\}$ of $\mathscr{X} \times \mathbb{R}$ and, therefore, $k + 1 \le \overline{V}$. □

7.3. *More bounds.* As we have observed in the previous sections [see (45), (46) and (47)], there are various situations for which, given a model $\overline{\mathscr{Q}} \subset \mathscr{P}$, it is possible to build a $\rho$-estimator $\widehat{\mathbf{P}}$ with values in $\overline{\mathscr{Q}}$ satisfying for all $\mathbf{P} \in \mathscr{P}$ and $\xi > 0$,

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P}, \overline{\mathscr{Q}}) + D_n(\overline{\mathscr{Q}}) + \xi\Big] \ge 1 - e^{-\xi}$$

for some quantity $D_n(\overline{\mathscr{Q}}) \ge 1$ only depending on the specific features of $\overline{\mathscr{Q}}$ and some constant $C > 0$ depending on the choice of $\psi$. Such an inequality leads to a risk bound of the following form (with $C' > 0$):

$$\mathbb{E}\Big[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}})\Big] \le C'\Big[\mathbf{h}^2(\mathbf{P}, \overline{\mathscr{Q}}) + D_n(\overline{\mathscr{Q}})\Big] \quad \text{for all } \mathbf{P} \in \mathscr{P} \tag{48}$$

and allows us to bound from above the minimax risk over $\overline{\mathscr{Q}}$ by $C' D_n(\overline{\mathscr{Q}})$.

However, not all statistical models admit a finite minimax risk and for such models there is consequently no hope to bound from above the $\rho$-dimension function uniformly as we did in the previous sections. One such example is the set of probabilities on $(0, +\infty)$ with nonincreasing densities with respect to the Lebesgue measure. More examples can also be found in Baraud and Birgé (2016). For some of these models, it is possible to build a $\rho$-estimator the risk of which does not

degenerate, a typical example being the Grenander estimator which is, as already seen, a $\rho$-estimator.

Following Baraud (2016), we introduce this definition.

DEFINITION 9. A class of functions $\mathscr{F}$ defined on a set $\mathscr{X}$ and with values in $[-\infty, +\infty]$ is said to be weak VC-major with dimension not larger than $k \in \mathbb{N}$ if, for all $u \in \mathbb{R}$, the class of subsets

$$\mathscr{C}_u(\mathscr{F}) = \big\{\{x \in \mathscr{X} \,|\, f(x) > u\},\ f \in \mathscr{F}\big\}$$

is a VC-class with dimension not larger than $k$ (index not larger than $k+1$). The weak dimension of $\mathscr{F}$ is the smallest of such integers $k$.

DEFINITION 10. Let $\mathscr{F}$ be a class of real-valued functions on $\mathscr{X}$. We shall say that an element $\overline{f} \in \mathscr{F}$ is extremal in $\mathscr{F}$ with degree $d \in \mathbb{N}$ if the class of functions

$$\big(\mathscr{F}/\overline{f}\big) = \big\{f/\overline{f},\ f \in \mathscr{F}\big\}$$

is weak VC-major with dimension $d$.

For $\boldsymbol{\mu} \in \boldsymbol{\mathscr{M}}$, we consider a density set $\overline{\boldsymbol{\mathcal{Q}}} \subset \boldsymbol{\mathcal{L}}(\boldsymbol{\mu})$ which is viewed as a class of real-valued functions on $\overline{\mathscr{X}} = \bigcup_{i=1}^{n}(\{i\} \times \mathscr{X}_i)$ as we did in Section 7.2. The corresponding model $\overline{\boldsymbol{\mathcal{Q}}}$ is $\{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\boldsymbol{\mathcal{Q}}}\}$ and, for $d \in \{1, \ldots, n\}$, we denote by $\overline{\boldsymbol{\mathcal{Q}}}_d$ the subset of $\overline{\boldsymbol{\mathcal{Q}}}$ of those densities $\mathbf{q}$ which are extremal in $\overline{\boldsymbol{\mathcal{Q}}}$ with degree $d$. We set $\overline{\boldsymbol{\mathcal{Q}}}_d = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\boldsymbol{\mathcal{Q}}}_d\}$ and let $\mathcal{D}$ be the subset of $\{1, \ldots, n\}$ consisting of those $d$ such that $\overline{\boldsymbol{\mathcal{Q}}}_d \neq \varnothing$.

PROPOSITION 9. *Let $\psi$ satisfy Assumption 2. For all $\rho$-models $\boldsymbol{\mathcal{Q}} \subset \overline{\boldsymbol{\mathcal{Q}}}$ and all $d \in \mathcal{D}$,*

$$D^{\boldsymbol{\mathcal{Q}}}(\mathbf{P}, \overline{\mathbf{P}}) \leq 33d\big[\log\big(e^2 n/d\big)\big]^3 \quad \textit{for all } (\mathbf{P}, \overline{\mathbf{P}}) \in \boldsymbol{\mathscr{P}} \times \overline{\boldsymbol{\mathcal{Q}}}_d.$$

The proof is given in Section D.8 of the Supplementary Material [Baraud and Birgé (2018)]. This upper bound, although depending on the specific features of $\overline{\boldsymbol{\mathcal{Q}}}$, is free from the choices of $\psi$. We immediately derive from Proposition 9 and Theorem 1 with a suitable choice of $\overline{\mathbf{P}}$ the following result.

COROLLARY 4. *Let $\psi$ satisfy Assumption 2 and assume that $\mathcal{D} \neq \varnothing$ and that $\overline{\boldsymbol{\mathcal{Q}}}_d$ contains a countable and $\mathbf{V}$-dense subset $\boldsymbol{\mathcal{Q}}_d$ for all $d \in \mathcal{D}$. Any $\rho$-estimator $\widehat{\mathbf{P}}$ on a $\rho$-model $\boldsymbol{\mathcal{Q}} \subset \overline{\boldsymbol{\mathcal{Q}}}$ containing $\bigcup_{d\in\mathcal{D}} \boldsymbol{\mathcal{Q}}_d$ satisfies, for all $\mathbf{P} \in \boldsymbol{\mathscr{P}}$ and $\xi > 0$,*

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \leq \inf_{d\in\mathcal{D}}\Big[\mathbf{h}^2(\mathbf{P}, \overline{\boldsymbol{\mathcal{Q}}}_d) + d\big[\log\big(e^2 n/d\big)\big]^3\Big] + \xi\Big] \geq 1 - e^{-\xi}, \tag{49}$$

*for some constant $C$ depending on $\psi$ only. If, moreover, $\overline{\boldsymbol{\mathcal{Q}}}$ is universally separable, any $\rho$-estimator relative to $((\boldsymbol{\mu}, \overline{\boldsymbol{\mathcal{Q}}}), \mathbf{0})$ also satisfies (49).*

PROOF. Proposition 9 and Theorem 1 lead to (49). When $\overline{\mathcal{Q}}$ is universally separable there exists a countable and $\mathcal{T}$-dense subset $\mathcal{Q}' \subset \overline{\mathcal{Q}}$. The countable set $\mathcal{Q} = \mathcal{Q}' \cup (\bigcup_{d\in\mathcal{D}} \mathcal{Q}_d)$ is still countable and $\mathcal{T}$-dense in $\overline{\mathcal{Q}}$ and the corresponding $\rho$-model $\mathcal{Q}$ is $\mathcal{V}$-dense in $\overline{\mathcal{Q}}$ and contains $\bigcup_{d\in\mathcal{D}} \mathcal{Q}_d$. By Theorem 3, any $\rho$-estimator relative to $((\boldsymbol{\mu}, \overline{\mathcal{Q}}), \mathbf{0})$ is also a $\rho$-estimator relative to $((\boldsymbol{\mu}, \mathcal{Q}), \mathbf{0})$ and, therefore, satisfies (49). □

Note that the bound depends on the initial representation $(\boldsymbol{\mu}, \overline{\mathcal{Q}})$ because the sets $\overline{\mathcal{Q}}_d$, hence the sets $\overline{\mathcal{Q}}_d$, depend on $\overline{\mathcal{Q}}$. This result looks like a model selection result among the sets $\{\overline{\mathcal{Q}}_d, d \in \mathcal{D}\}$ although the $\rho$-estimator $\widehat{\mathbf{P}}$ is built on a single $\rho$-model $\mathcal{Q} \subset \overline{\mathcal{Q}}$ with a null penalty function. It implies that the minimax risk over each set $\overline{\mathcal{Q}}_d$ is necessarily finite, while that on $\overline{\mathcal{Q}}$ might not be.

In the particular case of density estimation, the following result turns to be useful in view of applying Proposition 9.

PROPOSITION 10. *Let $\overline{\mathcal{Q}}$ be a subset of $\mathcal{L}(\mu)$ viewed as a class of functions on $\mathcal{X}$. If $\overline{p}$ is extremal in $\overline{\mathcal{Q}}$ with degree $d$, $\overline{\mathbf{p}} = (\overline{p}, \ldots, \overline{p})$ is extremal in $\overline{\mathcal{Q}} = \{\mathbf{q} = (q, \ldots, q), q \in \overline{\mathcal{Q}}\}$, viewed as a class of functions on $\overline{\mathcal{X}}$, with degree not larger than $d$.*

PROOF. Let $u \in \mathbb{R}$. If $\mathscr{C}_u((\overline{\mathcal{Q}}/\overline{\mathbf{p}}))$ shatters $\{\overline{x}_1, \ldots, \overline{x}_k\} \subset \overline{\mathcal{X}}$, for all $J \subset \{1, \ldots, k\}$ there exists $\mathbf{q} \in \overline{\mathcal{Q}}$ such that $j \in J$ if and only if $(\mathbf{q}/\overline{\mathbf{p}})(\overline{x}_j) = (q/\overline{p})(x_j) > u$. Hence, $\mathscr{C}_u((\overline{\mathcal{Q}}/\overline{p}))$ shatters $\{x_1, \ldots, x_k\} \subset \mathcal{X}$ which is only possible for $k \le d$. □

7.4. *Some examples of statistical models.* Let us restrict ourselves here to the density framework where $X_1, \ldots, X_n$ are assumed to be i.i.d. with common distribution $P$ on $\mathcal{X}$ and we have at hand a set of candidate probabilities $Q = q \cdot \mu$ with $q \in \overline{\mathcal{Q}} \subset \mathcal{L}(\mu)$ for $P$. We shall provide here some examples of density sets $\overline{\mathcal{Q}}$ to which Proposition 7 or 9 applies.

*Piecewise constant functions.* Let $k$ be a positive integer and $\mathcal{X}$ an arbitrary interval of $\mathbb{R}$ (possibly $\mathcal{X} = \mathbb{R}$). We define $\mathcal{F}_k$ as the class of functions $f$ on $\mathcal{X}$ such that there exists a partition $\mathcal{I}(f)$ of $\mathcal{X}$ into at most $k$ intervals (of positive lengths) with $f$ constant on each of these intervals. Note that $\mathcal{I}(f)$ depends on $f$. The following result is to be proved in Section C.1 of the Supplementary Material [Baraud and Birgé (2018)].

PROPOSITION 11. *The set $\mathcal{F}_k$ is VC-subgraph with dimension bounded by $2k$.*

Let us apply this to histogram estimation on $\mathcal{X} = \mathbb{R}$. For a positive integer $D$, we denote by $\overline{\mathcal{Q}}_D$ the subset of $\mathcal{F}_{D+2}$ of right-continuous densities with respect to the Lebesgue measure $\mu$, that is, the set of right-continuous and piecewise

constant densities on $\mathbb{R}$ with at most $D$ pieces and by $\overline{\mathcal{Q}}_D = \{q \cdot \mu, q \in \overline{\mathcal{Q}}_D\}$ the corresponding model for $P$. We derive from Propositions 7 and 11 that, for some universal constant $C > 0$ and all $\rho$-models $\mathcal{Q} \subset \overline{\mathcal{Q}}_D$,

$$D^{\mathcal{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \leq C(D \wedge n)\big[1 + \log_+(n/D)\big] \quad \text{for all } (\mathbf{P}, \overline{\mathbf{P}}) \in \mathcal{P} \times \mathcal{P}^{\boldsymbol{\mu}}.$$

Hence, by Corollary 3, for all $\rho$-estimators $\widehat{\mathbf{P}}$ on some countable and $\mathcal{V}$-dense subset $\mathcal{Q}_D$ of $\overline{\mathcal{Q}}_D$,

$$C\mathbb{E}\Big[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}})\Big] \leq \mathbf{h}^2(\mathbf{P}, \overline{\mathcal{Q}}_D) + (D \wedge n)\big[1 + \log_+(n/D)\big]. \tag{50}$$

Since $\overline{\mathcal{Q}}_D$ is universally separable (see Section B of the Supplementary Material [Baraud and Birgé (2018)]), we deduce from Theorem 3 that (50) remains true for any $\rho$-estimator $\widehat{\mathbf{P}}$ on the noncountable model $\overline{\mathcal{Q}}_D$ relative to the representation $(\boldsymbol{\mu}, \overline{\mathcal{Q}}_D)$ (with a null penalty function).

The logarithmic factor in this bound turns out to be necessary. The argument is as follows. When $\mathbf{P} \in \overline{\mathcal{Q}}_D$, it follows from (50) that $\mathbb{E}[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}})] \leq C'(D \wedge n)[1 + \log_+(n/D)]$ for some universal constant $C' > 0$. This inequality appears to be optimal (up to the numerical constant $C'$) in view of the lower bound established in Proposition 2 of Birgé and Massart (1998). This also shows that the logarithmic factor involved in the bound of the $\rho$-dimension function established in Proposition 7 is necessary, at least for some VC-subgraph classes.

*Piecewise exponential families*. Using similar arguments based on Corollaries 3 and 4 as we did above, we may establish risk bounds of the same flavour as (50) with the following density sets.

DEFINITION 11. Let $g_1, \ldots, g_J$ be $J \geq 1$ real-valued functions on a set $\mathcal{X}$. We shall say that a class $\mathcal{F}$ of positive functions on $\mathcal{X}$ is an exponential family based on $g_1, \ldots, g_J$ if the elements $f$ of $\mathcal{F}$ are of the form

$$f = \exp\left[\sum_{j=1}^{J} \beta_j g_j\right] \quad \text{for } \beta_1, \ldots, \beta_J \in \mathbb{R}. \tag{51}$$

If $\mathcal{X}$ is a nontrivial interval of $\mathbb{R}$ and $k$ a positive integer, we shall say that $\mathcal{F}$ is a $k$-piecewise exponential family based on $g_1, \ldots, g_J$ if for any $f \in \mathcal{F}$ there exists a partition $\mathcal{I}(f)$ of $\mathcal{X}$ into at most $k$ intervals such that for all $I \in \mathcal{I}(f)$, the restriction $f_I$ of $f$ to $I$ is of the form (51) with coefficients $\beta_j$ depending on $I$.

The properties of exponential and piecewise exponential families are described by the following proposition to be proven in Section C of the Supplementary Material [Baraud and Birgé (2018)].

PROPOSITION 12. *Let $\overline{\mathcal{Q}}$ be a class of functions on $\mathcal{X}$.*

(i) *If $\overline{\mathcal{Q}}$ is an exponential family based on $J \geq 1$ functions, $\overline{\mathcal{Q}}$ is VC-subgraph with index not larger than $J+2$.*

(ii) *Let $\mathcal{I}$ be a partition of $\mathcal{X}$ with cardinality not larger than $k \geq 1$. If for all $I \in \mathcal{I}$, the family $\overline{\mathcal{Q}}_I$ consisting of the restrictions of the functions $q$ in $\overline{\mathcal{Q}}$ to the set $I$ is an exponential family on $I$ based on $J \geq 1$ functions, $\overline{\mathcal{Q}}$ is VC-subgraph with index not larger than $k(J+2)$.*

(iii) *If $\mathcal{X}$ is a nontrivial interval of $\mathbb{R}$ and $\overline{\mathcal{Q}}$ is a $k$-piecewise exponential family based on $J$ functions, all densities $\overline{p} \in \overline{\mathcal{Q}}$ are extremal in $\overline{\mathcal{Q}}$ with degree $d$ not larger than $\lceil 9.4k(J+2) \rceil = \inf\{j \in \mathbb{N},\ j \geq 9.4k(J+2)\}$.*

## 8. Estimating a conditional distribution.

8.1. *Description of the framework.* Let us now apply our result to the estimation of a conditional distribution. We consider i.i.d. pairs $X_i = (W_i, Y_i)$, $i = 1, \ldots, n$ of random variables with values in the product space $(\mathcal{W} \times \mathcal{Y}, \mathcal{B}(\mathcal{W}) \otimes \mathcal{B}(\mathcal{Y}))$ and common distribution $P$, assuming that truly $\mathbf{P} = P^{\otimes n}$. We denote by $P_W$ the marginal distribution of $W$ and assume the existence of a conditional distribution $P_w$ of $Y$ when $W = w$, which means that for all bounded measurable functions $f$ on $\mathcal{Y}$,

$$\mathbb{E}\big[f(Y)|W = w\big] = \int_{\mathcal{Y}} f(y)\, dP_w(y) \qquad P_W\text{-a.s.}$$

and for all bounded measurable functions $g$ on $\mathcal{W} \times \mathcal{Y}$,

$$\mathbb{E}\big[g(W, Y)\big] = \int_{\mathcal{W}} \left[ \int_{\mathcal{Y}} g(w, y)\, dP_w(y) \right] dP_W(w).$$

Our purpose is to estimate the conditional distribution $P_w$ without the knowledge of $P_W$ which may therefore be completely arbitrary. To do so, we consider a reference measure $\lambda$ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ and the set $\mathcal{L}_c(\mathcal{W}, \lambda)$ of conditional densities with respect to $\lambda$, that is, the set of measurable functions $t$ from $(\mathcal{W} \times \mathcal{Y}, \mathcal{B}(\mathcal{W}) \otimes \mathcal{B}(\mathcal{Y}))$ to $\mathbb{R}_+$ such that for all $w \in \mathcal{W}$, the function $t_w : y \mapsto t(w, y) \in \mathcal{L}(\lambda)$. Then to each element $t \in \mathcal{L}_c(\mathcal{W}, \lambda)$ is associated a conditional distribution $t_w \cdot \lambda$ for $Y$. In order to build our estimators, we first introduce a countable family $\{S_m, m \in \mathcal{M}\}$ of countable subsets $S_m$ of $\mathcal{L}_c(\mathcal{W}, \lambda)$, and a nonnegative weight function $\Delta$ on $\mathcal{M}$ satisfying (21). To each $S_m$, we associate the $\rho$-density model $\mathcal{Q}_m = \{Q_t, t \in S_m\}$ for $P$, where the probability $Q_t$ on $\mathcal{W} \times \mathcal{Y}$ is given by

$$Q_t(A \times B) = \int_A \left[ \int_B t_w(y)\, d\lambda(y) \right] dP_W(w),$$

or equivalently

$$\frac{dQ_t}{dP_W \otimes d\lambda}(w, y) = t(w, y).$$

This means that $Q_t$ has a marginal distribution $P_W$ on $\mathscr{W}$ and a conditional distribution given $W = w$ with density $t_w$ with respect to $\lambda$. Note that the $\rho$-models $\mathscr{Q}_m$ depend on the unknown distribution $P_W$ but the densities with respect to the dominating measure $P_W \otimes \lambda$ do not. This leads to a family of $\rho$-models $\boldsymbol{\mathscr{Q}}_m$ for $\mathbf{P}$ and a reference $\rho$-model $\boldsymbol{\mathscr{Q}} = \bigcup_{m\in\mathcal{M}} \boldsymbol{\mathscr{Q}}_m$. If we introduce a suitable penalty **pen** on $\boldsymbol{\mathscr{Q}}$, we may build a $\rho$-estimator of $\mathbf{P}$ from our sample $X_1, \dots, X_n$ according to the recipe of Section 2.3 since its values only depend on the family of densities in $\mathcal{Q} = \bigcup_{m\in\mathcal{M}} S_m$. As a consequence, our estimation strategy neither needs to know $P_W$ nor to estimate it. Such a $\rho$-estimator will be of the form $Q_{\widehat{s}}^{\otimes n}$ with $Q_{\widehat{s}} = \widehat{s} \cdot (P_W \otimes \lambda)$ and will provide an estimator $\widehat{s}_w \cdot \lambda$ of the conditional probability $P_w$.

Within this framework, the Hellinger distance between the probabilities at hand writes, for any measure $\nu$ that dominates both $P$ and $P_W \otimes \lambda$,

$$\frac{1}{2}\int_{\mathscr{W}\times\mathscr{Y}} \left(\sqrt{\frac{dP_W(w)\,dP_w(y)}{d\nu}} - \sqrt{t(w,y)\frac{dP_W(w)\,d\lambda(y)}{d\nu}}\right)^2 d\nu(w,y)$$
$$= h^2(P, Q_t) = \int_{\mathscr{W}} h^2(P_w, t_w \cdot \lambda)\, dP_W(w).$$

Therefore,

$$h^2(P, \mathscr{Q}_m) = \inf_{t\in S_m} \int_{\mathscr{W}} h^2(P_w, t_w \cdot \lambda)\, dP_W(w). \tag{52}$$

Note that $h^2(P, Q_t)$ can actually be viewed as a loss function for the conditional distributions, of the form $\ell(P_w, t_w \cdot \lambda)$ since it actually only depends on $P_w$ and $t_w$.

8.2. *Assumptions and results.* Let us assume the following.

ASSUMPTION 5. For all $m \in \mathcal{M}$, $S_m$ is VC-subgraph with index not larger $\overline{V}_m$.

We may then deduce from Theorem A.1 of the Supplementary Material [Baraud and Birgé (2018)] the following result.

COROLLARY 5. *Let $\{S_m, m \in \mathcal{M}\}$ be a family of countable subsets of $\mathcal{L}_c(\mathscr{W}, \lambda)$ satisfying Assumption 5, $\Delta$ be a weight function on $\mathcal{M}$ which satisfies (21), $\psi$ a function satisfying Assumption 2, $\mathscr{Q} = \bigcup_{m\in\mathcal{M}}\{Q_t, t \in S_m\}$ and* $\mathrm{pen} : \mathscr{Q} \to \mathbb{R}_+$ *given, for all $Q \in \mathscr{Q}$, by*

$$\mathrm{pen}(Q) = \kappa \inf_{\{m\in\mathcal{M}\,|\,Q=Q_t \text{ with } t\in S_m\}} \left[\frac{C_1}{4.7}(\overline{V}_m \wedge n)\left[1 + \log_+\left(\frac{n}{\overline{V}_m}\right)\right] + \Delta(m)\right],$$

*where $\kappa$ is given by (19) and $C_1$ is the constant appearing in Proposition 7. Then any density $\rho$-estimator $Q_{\widehat{s}}$ relative to $[(\mu, \mathcal{Q}), \mathrm{pen}]$ satisfies, for some constant $C' > 0$ depending on the choice of $\psi$ only,*

$$\mathbb{E}\big[h^2(P, Q_{\widehat{s}})\big] \le C' \inf_{m\in\mathcal{M}} \left[h^2(P, \mathscr{Q}_m) + \frac{\overline{V}_m \wedge n}{n}\left(1 + \log_+\left(\frac{n}{\overline{V}_m}\right)\right) + \frac{\Delta(m)}{n}\right].$$

Note that this result does not require any information or assumption on the distribution of $W$. If, in particular, the conditional probability $P_w$ is absolutely continuous with respect to $\lambda$ for almost all $w$ with density $dP_w/d\lambda = s_w$, $P_W$-a.s., one can write

$$\begin{cases} h^2(P, Q_{\widehat{s}}) = \displaystyle\int_{\mathscr{W}} h^2(s_w \cdot \lambda, \widehat{s}_w \cdot \lambda)\, dP_W(w), \\ h^2(P, \mathscr{Q}_m) = \displaystyle\inf_{t \in S_m} \int_{\mathscr{W}} h^2(s_w \cdot \lambda, t_w \cdot \lambda)\, dP_W(w). \end{cases}$$

PROOF OF COROLLARY 5. If we apply Propositions 7 and 8 to each $\rho$-model $\mathscr{Q}_m$ with $m \in \mathcal{M}$, we obtain under Assumption 5 the existence of a universal constant $C_1 > 0$ such that for all $(\mathbf{P}, \overline{\mathbf{P}}) \in \mathscr{P} \times \mathscr{P}^{\boldsymbol{\mu}}$:

$$D^{\mathscr{Q}_m}(\mathbf{P}, \overline{\mathbf{P}}) \le D_n(m) = C_1(\overline{V}_m \wedge n)\big[1 + \log_+(n/\overline{V}_m)\big].$$

Inequality (22) is fulfilled with $K = 0$ and the penalty function therefore satisfies (23) with $\kappa_1 = 0$ for all $m \in \mathcal{M}$. The result follows from Theorem 2, then an integration of (24), with respect to $\xi > 0$. $\square$

**9. Regression with a random design.** In this section, we assume that the observations $X_i = (W_i, Y_i)$, $1 \le i \le n$ are i.i.d. copies of a random pair

$$X = (W, Y) \quad \text{with } Y = f(W) + \epsilon, \tag{53}$$

where $W$ is a random variable with distribution $P_W$ on a measurable space $(\mathscr{W}, \mathscr{B}(\mathscr{W}))$, $f$ is an unknown regression function mapping $\mathscr{W}$ into $\mathbb{R}$ and $\epsilon$ is a real-valued random variable with distribution $P_\epsilon$, which is independent of $W$. Both distributions $P_W$ and $P_\epsilon$ are assumed to be unknown. We shall use the specific notation introduced in Section 3.3 when the data are i.i.d. and denote by $\mu$ the product measure $P_W \otimes \lambda$ where $\lambda$ is the Lebesgue measure on $\mathbb{R}$. Note that $\mu$ is unknown since it depends on the distribution $P_W$ of the design $W$.

If $\epsilon$ had a density $s$ with respect to $\lambda$, the distribution $P$ of $X = (W, Y)$ would be absolutely continuous with respect to $\mu$ with density $p$ given by

$$p(w, y) = s\big(y - f(w)\big) \quad \text{for } (w, y) \in \mathscr{X}, \tag{54}$$

depending thus on two parameters: the density $s$ of the errors and the regression function $f$.

Denoting by $\mathscr{D}$ the set of all densities on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ and $\mathscr{F}$ the set of all measurable functions mapping $\mathscr{W}$ into $\mathbb{R}$, our aim is to estimate $P$ assuming that it is close to some distribution of the form $p \cdot \mu$ with $p$ given by (54) for some $s \in \mathscr{D}$ and $f \in \mathscr{F}$. Besides, when $P$ is truly of this form we shall also derive estimators for both $s$ and $f$.

9.1. *The main result.* For $r \in \mathscr{D}$ and $g \in \mathscr{F}$, we set

$$Q_{r,g} = q_{r,g} \cdot \mu \quad \text{with } q_{r,g}(w, y) = r\big(y - g(w)\big),$$

which means that $Q_{r,g}$ is the distribution of $X$ in (53) when $f = g$ and $\epsilon$ is distributed according to $R = r \cdot \lambda$. Given a density $r \in \mathscr{D}$ and a countable subset $F$ of $\mathscr{F}$, we define the $\rho$-model:

$$\mathscr{Q}_m = \{Q_{r,g}, g \in F\} \quad \text{for } m = (r, F).$$

Given a countable subset $\mathcal{D}$ of $\mathscr{D}$ and a countable family $\mathbb{F}$ of countable subsets $F$ of $\mathscr{F}$, we estimate $P$ on the basis of the collection of $\rho$-models $\{\mathscr{Q}_m, m \in \mathcal{M}\}$ with $\mathcal{M} \subset \mathcal{D} \times \mathbb{F}$. We endow the family $\{\boldsymbol{\mathscr{Q}}_m, m \in \mathcal{M}\}$ with a weight function $\Delta$ satisfying (21) and assume the following.

ASSUMPTION 6.
(i) The densities $r \in \mathcal{D}$ are unimodal.
(ii) Each $F$ in $\mathbb{F}$ is VC-subgraph with index $\overline{V}(F)$.
(iii) The function $\psi$ satisfies Assumption 2 with $\boldsymbol{\mathscr{Q}} = \{Q^{\otimes n}, Q \in \mathscr{Q} = \bigcup_{m \in \mathcal{M}} \mathscr{Q}_m\}$.

Under Assumptions 6-(i) and 6-(ii), the family of densities $\mathcal{Q}_m$ is a VC-subgraph on $\mathscr{X}$ with index not larger than

$$\overline{V}_m = 9.41\overline{V}(F) \quad \text{for all } m = (r, F) \in \mathcal{M}. \tag{55}$$

This result derives from Baraud, Birgé and Sart (2017), Proposition 42. Besides, under Assumption 6-(iii), Proposition 7 applies and implies that, for some universal constant $C_1 > 0$, all $m \in \mathcal{M}$, $\mathbf{P} \in \boldsymbol{\mathscr{P}}$ and $\overline{\mathbf{P}} \in \boldsymbol{\mathscr{Q}}$, $D^{\boldsymbol{\mathscr{Q}}_m}(\mathbf{P}, \overline{\mathbf{P}}) \le D_n(m)$ with

$$D_n(m) = C_1\big(\overline{V}_m \wedge n\big)\big[1 + \log_+\big(n/\overline{V}_m\big)\big] \qquad \text{for all } m = (r, F) \in \mathcal{M}, \tag{56}$$

so that (22) holds with $K = 0$. Setting

$$\mathbf{pen}(\mathbf{Q}) = \kappa \inf_{\{m \in \mathcal{M} \mid \mathscr{Q}_m \ni Q\}} \left\{ \frac{D_n(m)}{4.7} + \Delta(m) \right\},$$

we may apply Theorem 2 with $\kappa_1 = 0$, which leads, in this particular case, to the following analogue of (24).

THEOREM 4. *Assume that Assumption 6 holds. For any distribution $P \in \boldsymbol{\mathscr{P}}$ and $\mathbf{P} = P^{\otimes n}$, any $\rho$-estimator $\widehat{\mathbf{P}} = (\widehat{P}, \dots, \widehat{P})$ satisfies, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$,*

$$\begin{aligned} &Ch^2(P, \widehat{P}) \\ &\quad \le \inf_{m \in \mathcal{M}} \left[ h^2(P, \mathscr{Q}_m) + \frac{\overline{V}_m \wedge n}{n} \left[ 1 + \log_+\left( \frac{n}{\overline{V}_m} \right) \right] + \frac{\Delta(m)}{n} \right] + \frac{\xi}{n}, \end{aligned} \tag{57}$$

*for some constant $C > 0$ only depending on the choice of $\psi$.*

At this stage, some comments are in order:

(a) This result holds without any assumption on the distribution $P_W$ of the design.

(b) The result is true even if the regression framework (53) is not exact as long as the $X_i$ are i.i.d. In particular, the distribution $P$ needs not have a density with respect to $\mu = P_W \otimes \lambda$.

(c) If $r$ admits $k$ modes with $k > 1$ and $F$ is a VC-subgraph with index not larger than $\overline{V}$, $\mathscr{Q}_{(r,F)}$ remains a VC-subgraph and its index is still bounded by $C(k)\overline{V}$ for some constant $C(k)$ that now depends on $k$. Consequently, the above result generalizes to families $\mathcal{D}$ of densities admitting more than a single mode in which case $\overline{V}(F)$ should be replaced by $c(r)\overline{V}(F)$ where $c(r)$ is a positive number depending on the number of modes of the density $r$.

(d) With Theorem 4 at hand, we could obtain in the present random design context an analogue of Corollary 39 in Baraud, Birgé and Sart (2017) which was established when the $W_i$ were deterministic (fixed design regression).

9.2. *Estimation of $s$ and $f$.* Let us now consider the situation where the regression framework (53) is exact and $\epsilon$ has an unknown density $s$ with respect to the Lebesgue measure $\lambda$. Then $P = Q_{s,f}$ admits a density $q_{s,f}$ with respect to $\mu$ which is given by (54) with $s$ belonging to $\mathscr{D}$ and $f$ to $\mathscr{F}$ but not necessarily to our $\rho$-models $\mathcal{D}$ and $\mathcal{F} = \bigcup_{F\in\mathbb{F}} F$, respectively. Since we may choose our $\rho$-estimator of the form

$$\widehat{P} = q_{\widehat{s},\widehat{f}} \cdot \mu \quad \text{with } (\widehat{s}, \widehat{f}) \in \mathcal{D} \times \mathcal{F},$$

our procedure results in estimators $\widehat{s}$ and $\widehat{f}$ for $s$ and $f$, respectively, and our aim in this section is to establish risk bounds for these two estimators.

Since the map $(r, g) \mapsto Q_{r,g}$ is not necessarily one-to-one from $\mathscr{D} \times \mathscr{F}$ to $\mathscr{P}$, an identifiability condition is required on our $\rho$-model $\mathscr{Q}$ so that the equality $Q_{r,g} = Q_{r',g'}$ with $r, r' \in \mathcal{D}$ and $g, g' \in \mathcal{F}$ implies that $r = r'$ $\lambda$-a.e. and $g = g'$ $P_W$-a.s. In order to state this identifiability condition, let us introduce the following notation. For $r \in \mathscr{D}$ and $a \in \mathbb{R}$, we shall denote by $R_a$ the probability on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ with density $r_a(\cdot) = r(\cdot - a)$. When $\epsilon$ has density $r$ and $a = g(w)$ for some $w \in \mathscr{W}$, $R_a$ can be viewed as the conditional distribution of $Y = g(W) + \epsilon$ given $W = w$. Given $r, r' \in \mathscr{D}$, $g, g' \in \mathscr{F}$ and $w \in \mathscr{W}$, the Hellinger distance between the probabilities $R_{g(w)}$, and $R'_{g'(w)}$ is given by

$$h^2\big(R_{g(w)}, R'_{g'(w)}\big) = \frac{1}{2}\int_{\mathbb{R}}\Big[\sqrt{r\big(y - g(w)\big)} - \sqrt{r'\big(y - g'(w)\big)}\Big]^2 d\lambda(y)$$

and the Hellinger distance between the corresponding probabilities $Q_{r,g}$ and $Q_{r',g'}$ on $(\mathscr{X}, \mathscr{B})$ writes

$$h^2(Q_{r,g}, Q_{r',g'}) = \int_{\mathscr{W}} h^2\big(R_{g(w)}, R'_{g'(w)}\big)\, dP_W(w). \tag{58}$$

We recall that the Hellinger distance is translation invariant which means that for all densities $r, r' \in \mathscr{D}$, $a, a' \in \mathbb{R}$,

$$h^2(R_a, R'_{a'}) = h^2(R_{a-a'}, R'). \tag{59}$$

In particular, taking $a = g(w)$ and $a' = g'(w)$ for $g, g' \in \mathscr{F}$ and $w \in \mathscr{W}$ and integrating (59) with respect to $P_W$ we get for all $(g, g') \in \mathscr{F}^2$ and $(r, r') \in \mathscr{D}^2$

$$h^2(Q_{r,g}, Q_{r',g'}) = h^2(Q_{r,g-g'}, Q_{r',0}). \tag{60}$$

In order to warrant identifiability, we assume the following.

ASSUMPTION 7. There exists a positive constant $A$ such that, for all $r, r' \in \mathcal{D}$, $R = r \cdot \lambda$ and $R' = r' \cdot \lambda$,

$$h(R, R') \le A \inf_{a \in \mathbb{R}} h(R_a, R').$$

When $Q_{r,g} = Q_{r',g'}$, (58) asserts that $h(R_{g(w)}, R'_{g'(w)}) = 0$ for $P_W$-almost all $w \in \mathscr{W}$ and (59) implies that $h(R_{g(w)-g'(w)}, R') = 0$ for all such $w \in \mathscr{W}$. Applying Assumption 7 with $a = g(w) - g'(w)$ leads to $R = R'$ and $g(w) = g'(w)$ which solves our identifiability problem.

In order to evaluate the risk of our estimator $\widehat{f}$ of $f$, we endow $\mathscr{F}$ with the loss function $d_s$ defined on $\mathscr{F} \times \mathscr{F}$ by

$$d_s^2(g, g') = \frac{1}{2} \int_{\mathscr{W} \times \mathbb{R}} \left( \sqrt{s_{g(w)}(y)} - \sqrt{s_{g'(w)}(y)} \right)^2 dP_W(w)\, dy \quad \text{for } g, g' \in \mathscr{F}.$$

This loss function depends on the true density $s$ of the errors $\epsilon$ and on the distribution $P_W$ of the design, hence on $P$. We have seen in Section 6.3 of Baraud, Birgé and Sart (2017) that, if the density $s$ is of order $\alpha$ with $\alpha \in (-1, 1]$ (see Definition 26 of that paper for the order of a function), the restriction of $d_s$ to the $\mathbb{L}_\infty(P_W)$-ball $\mathscr{B}_\infty(b)$ centred at 0 with radius $b$ is equivalent (up to factors depending on $b$ and $s$) to

$$\|g - g'\|_{1+\alpha, P_W}^{(1+\alpha)/2} \quad \text{with } \|g - g'\|_{1+\alpha, P_W} = \left[ \int_{\mathscr{W}} |g - g'|^{1+\alpha}\, dP_W \right]^{1/(1+\alpha)}.$$

In particular, if $\mathcal{F} \subset \mathscr{B}_\infty(b)$ and the true regression function $f$ also belongs to $\mathscr{B}_\infty(b)$,

$$c(s, b) \|f - g\|_{1+\alpha, P_W}^{(1+\alpha)/2} \le d_s(f, g) \le C(s, b) \|f - g\|_{1+\alpha, P_W}^{(1+\alpha)/2} \quad \text{for all } g \in \mathcal{F}$$

and suitable positive numbers $c(s, b)$ and $C(s, b)$. Of special interest is the case of $\alpha = 1$ for which $d_s(f, g)$ is of the order of the $\mathbb{L}_2(P_W)$-distance between $f$ and $g$ for all $g \in \mathcal{F}$. This situation is met when the translation $\rho$-model associated to $s$ is regular which is the case when $s$ is Cauchy, Gaussian, Laplace, etc. When $s$ is uniform or exponential, $d_s^2(\cdot, \cdot)$ is then equivalent to the $\mathbb{L}_1(P_W)$-norm.

Furthermore, when $\mathcal{F}$ is a subset of a $k$-dimensional linear space $\overline{\mathcal{F}}$ generated by $\phi_1, \ldots, \phi_k$, these norms can in turn be translated into a norm on $\mathbb{R}^k$ between the coefficients relative to this basis. More precisely, if $f$ belongs to $\overline{\mathcal{F}}$ and writes as $\sum_{j=1}^k \beta_j \phi_j$ and $\overline{f} = \sum_{j=1}^k \overline{\beta}_j \phi_j$ is an element of $\overline{\mathcal{F}}$, there exist two positive constants $c'(s,b,k)$ and $C'(s,b,k)$ such that

$$c'(s,b,k)\Big[\max_{j=1,\ldots,k} |\beta_j - \overline{\beta}_j|\Big]^{(1+\alpha)/2} \le d_s(f,\overline{f}) \le C'(s,b,k)\Big[\max_{j=1,\ldots,k} |\beta_j - \overline{\beta}_j|\Big]^{(1+\alpha)/2}.$$

In particular, if $\overline{f} = \overline{f}(n)$ converges towards $f$ at a rate $v_n$ with respect to the distance $d_s(\cdot,\cdot)$, the coefficients of $\overline{f}$ converge in sup-norm towards those of $f$ at rate $v_n^{2/(1+\alpha)}$; this latter rate being faster than $v_n$ when $\alpha < 1$.

In view of evaluating the risk of our estimator $\widehat{s}$ of the density $s$, we shall consider the loss between two densities $r, r' \in \mathscr{D}$ induced by the Hellinger distance between the two corresponding measures $r \cdot \lambda$ and $r' \cdot \lambda$ and we shall write this loss $h(r,r')$ so that

$$h(r,r') = h(r \cdot \lambda, r' \cdot \lambda) = h(Q_{r,0}, Q_{r',0}) \quad \text{for all } r, r' \in \mathscr{D}.$$

We deduce from Theorem 4 the following result.

COROLLARY 6. *Assume that the $X_i$ are i.i.d. with density $p$ given by* (54) *and that Assumptions 6 and 7 are satisfied. For all $\xi > 0$ and all $\rho$-estimators $Q_{\widehat{s},\widehat{f}}$, with $\widehat{s} \in \mathcal{D}$ and $\widehat{f} \in \mathcal{F}$, based on the family of $\rho$-models $\mathscr{Q}_m$ defined in Section 9.1, with probability at least $1 - e^{-\xi}$,*

$$C \max\Big\{d_s^2(f,\widehat{f}), h^2(s,\widehat{s})\Big\} \le \inf_m \left\{ d_s^2(f,F) + h^2(s,r) + \frac{\overline{V}_m \wedge n}{n}\left[1 + \log_+\left(\frac{n}{\overline{V}_m}\right)\right] + \frac{\Delta(m)}{n} \right\} + \frac{\xi}{n},$$

*where the infimum runs among all $m = (r,F) \in \mathcal{M}$, $C$ is a positive constant depending on $A$ and the choice of $\psi$ and $\overline{V}_m$ is given by* (55).

The risk bound is the same for the two estimators and depends on the approximation properties of $\mathcal{F}$ and $\mathcal{D}$ with respect to $f$ and $s$, respectively. The proof of this corollary is given in Section D.9 of the Supplementary Material [Baraud and Birgé (2018)].

**10. Estimator selection and aggregation.** In the case of density estimation, $\rho$-estimators can also be used to perform selection or aggregation of preliminary estimators. In this case, we assume that we have at hand a set $\boldsymbol{X}_1 = (X_1, \ldots, X_n)$

of $n$ independent random variables with an unknown joint distribution $\mathbf{P}$ to be estimated. We also have at hand a finite family $\mathscr{Q} = \{\mathbf{P}_j, j \in \mathcal{J}\}$ of probabilities that can be considered as candidate estimators for $\mathbf{P}$. These are completely arbitrary, but in a typical situation, it is assumed (although this may not be true) that the observations $X_i$ are i.i.d. and the $\mathbf{P}_j$ are preliminary estimators of the form $\mathbf{P}_j = P_j^{\otimes n}(\boldsymbol{X}_2)$, where $\boldsymbol{X}_2$ is a second sample independent from $\boldsymbol{X}_1$, and the $P_j = P_j(\boldsymbol{X}_2)$ are estimators that derive from various procedures applied to the sample $\boldsymbol{X}_2$.

10.1. *Estimator selection.* Taking $\mathcal{M} = \mathcal{J}$, we view each probability $\mathbf{P}_j$ as a $\rho$-model $\mathscr{Q}_j = \{\mathbf{P}_j\}$ with a single element. As a consequence, it follows from Proposition 6 that $D^{\mathscr{Q}_j}(\mathbf{P}, \mathbf{P}_j) \le 9 \log 2 < 6.3$ so that (22) holds with $D_n(j) = 6.3$ for all $j$ and $K = 0$. Then we choose the weights $\Delta(j)$ satisfying (21). We may choose $\Delta(j) = \log|\mathcal{J}|$ for all $j \in \mathcal{J}$ but other more Bayesian choices are possible, or choices based on the confidence we have in the various procedures used to build the preliminary estimators. To compute the penalized $\rho$-estimator $\widehat{\mathbf{P}}$ of $\mathbf{P}$, we may use the penalty function $\mathbf{pen}(\mathbf{P}_j) = \kappa\Delta(j)$ for all $j \in \mathcal{J}$ which is of the form (23) with $\kappa_1 = -\kappa(6.3/4.7)$. Finally, (24) shows that, for all $\xi > 0$ and $\mathbf{P} \in \mathscr{P}$,

$$\mathbb{P}\Big[\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le C \inf_{j \in \mathcal{J}} \Big(\mathbf{h}^2(\mathbf{P}, \mathbf{P}_j) + \Delta(j) + 1 + \xi\Big)\Big] \ge 1 - e^{-\xi},$$

where $C$ denotes a suitable constant depending on $\psi$ only.

10.2. *Convex estimator aggregation.* In this case, we set $\mathcal{J} = \{1, \dots, N\}$, $N \ge 2$ and $\mathscr{C}$ for the $N$-dimensional simplex:

$$\mathscr{C} = \left\{(\alpha_1, \dots, \alpha_N) \text{ such that } \alpha_j \ge 0 \text{ for } 1 \le j \le N \text{ and } \sum_{j=1}^N \alpha_j = 1\right\}.$$

We select a dominating measure $\mu$, densities $p_j = dP_j/d\mu$ and we then consider a single density model:

$$\overline{\mathcal{Q}} = \left\{\sum_{j=1}^N \alpha_j p_j \text{ for } (\alpha_1, \dots, \alpha_N) \in \mathscr{C}\right\}.$$

The following result then holds.

PROPOSITION 13. *Let $\psi$ satisfy Assumption 2. Any $\rho$-estimator $\widehat{\mathbf{P}}$ on $\overline{\mathscr{Q}} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \overline{\mathcal{Q}}\}$ relative to $((\boldsymbol{\mu}, \overline{\mathcal{Q}}), \mathbf{0})$ satisfies*

$$\mathbb{P}\Big[Ch^2(\mathbf{P}, \widehat{\mathbf{P}}) \le h^2(\mathbf{P}, \overline{\mathscr{Q}}) + N \log n + \xi\Big] \ge 1 - e^{-\xi} \quad \textit{for all } \xi > 0$$

*and some constant $C > 0$ depending on $\psi$ only.*

PROOF. The map $(\alpha_1, \dots, \alpha_n) \mapsto \sum_{j=1}^N \alpha_j p_j$ from $\mathscr{C}$ to $\overline{\mathcal{Q}}$ is continuous if we equip $\mathscr{C}$ with the usual Euclidean distance and $\overline{\mathcal{Q}}$ with the topology of pointwise convergence. Since $\mathscr{C}$ is separable, $\overline{\mathcal{Q}}$ is universally separable and contains thus a countable subset $\mathcal{Q}$ which is $\mathscr{T}$-dense in $\overline{\mathcal{Q}}$. The set $\overline{\mathcal{Q}}$ being furthermore a subset of an $N$-dimensional linear space, it is a VC-subgraph with index $\overline{V}$ not larger than $N+2$ and it follows from Proposition 7 (and Proposition 8) that, for all $\rho$-models $\mathscr{Q} \subset \overline{\mathscr{Q}}$,

$$D^{\mathscr{Q}}(\mathbf{P}, \overline{\mathbf{P}}) \le D_n = CN \log n \quad \text{for all } (\mathbf{P}, \overline{\mathbf{P}}) \in \mathscr{P} \times \mathscr{P}^{\boldsymbol{\mu}}.$$

We may apply Theorem 3 with $p \equiv 0$ to the single model $\overline{\mathscr{Q}}$. A $\rho$-estimator $\widehat{\mathbf{P}}$ on $\overline{\mathscr{Q}}$ relative to $((\boldsymbol{\mu}, \overline{\mathcal{Q}}), \mathbf{0})$ is also a $\rho$-estimator on the $\rho$-model $\mathscr{Q} = \{\mathbf{q} \cdot \boldsymbol{\mu}, \mathbf{q} \in \mathcal{Q}\}$ and we deduce from Theorem 1, more precisely from (20), that

$$\mathbb{P}\Big[C\mathbf{h}^2(\mathbf{P}, \widehat{\mathbf{P}}) \le \mathbf{h}^2(\mathbf{P}, \mathscr{Q}) + N \log n + \xi\Big] \ge 1 - e^{-\xi} \quad \text{for all } \xi > 0,$$

some constant $C$ depending on $\psi$ only and all $\mathbf{P} \in \mathscr{P}$. We conclude using the fact that $\mathscr{Q}$ is $\mathscr{V}$-dense in $\overline{\mathscr{Q}}$ since $\mathcal{Q}$ is $\mathscr{T}$-dense in $\overline{\mathcal{Q}}$. □

It should be noted that there is no $\mathbb{L}_2$-type argument here; the densities $p_j$ can be absolutely anything and the true distribution $\mathbf{P}$ should be a product measure but not necessarily of the form $P^{\otimes n}$.

*Practical implementation.* Since the set $\overline{\mathcal{Q}}$ is convex, Proposition 5 applies with $\psi = \psi_1$ or $\psi = \psi_2$ of Proposition 3 and the MLE on $\overline{\mathcal{Q}}$ is a $\rho$-estimator. It is obtained by maximizing over the convex set $\mathscr{C}$ the concave map

$$(\alpha_1, \dots, \alpha_N) \mapsto \sum_{i=1}^n \log\left(\sum_{j=1}^N \alpha_j p_j(X_i)\right).$$

**Acknowledgements.** The authors are grateful to Weijie Su for letting them know about the nice connection between the MLE and $\rho$-estimators in the case of a convex parameter set and for allowing them to include his result (Proposition 5) in this paper.

## SUPPLEMENTARY MATERIAL

**Supplement to "Rho-estimators revisited: general theory and applications"** (DOI: 10.1214/17-AOS1675SUPP; .pdf). This supplement provides the proofs of most results given in the paper and an additional section (D.10) devoted to robust tests.

## REFERENCES


AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *Ann. Statist.* **39** 2766–2794. MR2906886

BARAUD, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class. *Electron. J. Stat.* **10** 1709–1728. MR3522658

BARAUD, Y. and BIRGÉ, L. (2016). Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.* **126** 3888–3912.

BARAUD, Y. and BIRGÉ, L. (2018). Supplement to "Rho-estimators revisited: General theory and applications." DOI:10.1214/17-AOS1675SUPP.

BARAUD, Y., BIRGÉ, L. and SART, M. (2017). A new method for estimation and model selection: $\rho$-estimation. *Invent. Math.* **207** 425–517.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129

BIRGÉ, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** 273–325.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

GINÉ, E. and KOLTCHINSKII, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143–1216. MR2243881

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

LE CAM, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic Processes and Related Topics* (*Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes*, *Indiana Univ.*, *Bloomington*, *Ind.*, 1974, *Vol.* 1; *Dedicated to Jerzy Neyman*) 13–54. Academic Press, New York.

LE CAM, L. (1990). Maximum likelihood: An introduction. *Int. Stat. Rev.* **58** 153–171.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

SART, M. (2017). Estimating the conditional density by histogram type estimators and model selection. *ESAIM, Probab. Stat.* **21** 34–55.

VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. MR1739079

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York.



UNIVERSITÉ CÔTE D'AZUR
UMR CNRS 7351
LABORATOIRE JEAN ALEXANDRE DIEUDONNÉ
PARC VALROSE
06108 NICE CEDEX 02
FRANCE
E-MAIL: baraud@unice.fr

SORBONNE UNIVERSITÉ ET CNRS
LABORATOIRE DE PROBABILITÉS, STATISTIQUE ET MODÉLISATION (LPSM)
CASE COURRIER 188
F-75252 PARIS CEDEX 05
FRANCE
E-MAIL: lucien.birge@upmc.fr