An efficient algorithm for T-estimation

Nelo Magalhães, Yves Rozenholc

An efficient algorithm for T-estimation

Nelo Magalhães*[1,3] and Yves Rozenholc†[2,3]

[1]Équipe Probabilités et Statistiques, Université Paris-Sud 11
[2]MAP5 - UMR CNRS 8145, Université Paris Descartes
[3]INRIA team Select

Avril 2014

Abstract

We introduce an efficient and exact algorithm, together with a faster but approximate version, which implements with a sub-quadratic complexity the hold-out derived from T-estimation. We study empirically the performance of this hold-out in the context of density estimation considering well-known competitors (hold-out derived from least-squares or Kullback-Leibler divergence, model selection procedures, etc.) and classical problems including histogram or bandwidth selection. Our algorithms are integrated in a companion R-package called *Density.T.HoldOut* available on the CRAN: `http://cran.r-project.org/web/packages/Density.T.HoldOut/index.html`.

Index terms— T-estimation, density estimation, hold-out, Density.T.HoldOut R-package

1 Introduction

Suppose we have at hand a sample of independent and identically distributed (i.i.d.) random variables from some unknown density s with respect to some dominating measure μ and that we want to estimate s from the sample.

Hundreds of papers have been published about the solution of this estimation problem with as little prior information on s as possible. A widely used strategy consists in starting from a family of preliminary estimators (for instance kernel or histogram estimators) with some varying smoothing parameter (the bandwidth or the partition) and to select one candidate using the sample. Nevertheless, since the 30's (Larson, 1931) it is known that building estimators and evaluating their quality with the same data yields an overoptimistic result. Many solutions exist to overcome this problem. One natural procedure - called *hold-out* - consists in splitting the sample into two subsamples, building a family of estimators using the first subsample (which we shall call the training sample) and proceeding the selection using the second subsample (which we shall call the validation sample).

Concerning this selection part, Birgé (2006, Section 9) proposed a procedure - called *T-hold-out* in what follows - based on robust tests between the preliminary estimators. It can be derived from his construction of T-estimators[1] which are oriented to model selection. The definition of these estimators is introduced in the same paper but relies on old ideas arising from Le Cam (1973); Birgé (1983, 1984b,a). Indeed, conditionally to the training sample, all the estimators are deterministic so that the models are reduced to points and the problem amounts to select one point from the

*nelo.moltermagalhaes@gmail.com
†yves.rozenholc@parisdescartes.fr
[1]"T" refers to test.

validation sample.

The purpose of this paper is to provide an efficient algorithm that implements the T-hold-out, made available in our R-package called *Density.T.HoldOut*. Our motivations are twofold. First, when we started this research in the summer of 2012 there was no practical application of T-estimation[2] and we were very surprised to observe that most of the authors - including Birgé himself - considered this theory only as a theoretical tool, because of its supposed "too high computational complexity" as pointed out in Birgé (2006), Birgé (2007, p.45) and Baraud and Birgé (2009, p.241). Second, we thought it would be of interest to compare empirically T-estimation with classical resampling and penalization procedures since they are motivated by risk estimation, whereas T-estimators are based on robust tests and enjoy therefore some robustness properties. For this purpose we considered several finite collections of preliminary estimators. Histogram or kernel collections - leading to some well-known estimation problems: number of bin selection, partition selection, bandwidth selection, but also more complex collections mixing histograms and kernel estimators potentially completed with some parametric ones[3].

Hold-out is not specific to the density framework. Indeed, in all cases where we have at hand two independent random samples $\mathbb{X}_t$ and $\mathbb{X}_v$, one can build a collection of estimators using the training sample $\mathbb{X}_t$ and proceed to the selection with the validation sample $\mathbb{X}_v$. In density estimation, hold-out has been investigated theoretically for projection estimators (Arlot and Lerasle, 2012, Section 7.1) and kernel density estimates (Devroye and Lugosi, 2001) among other examples. Searching for the best linear (or convex) combination of the preliminary estimators in the validation step leads to the linear (or convex) aggregation problem (see Rigollet and Tsybakov (2007)). Moreover, theoretical properties of the hold-out have also been studied in classification (Bartlett et al., 2002; Blanchard and Massart, 2006) and in regression (Lugosi and Nobel, 1999; Juditsky and Nemirovski, 2000; Nemirovski, 2000; Wegkamp, 2003), among other references.

1.1 Framework

Let us consider a sample $\mathbb{X} = \{X_1, \dots, X_n\}$ of i.i.d. random variables X_i with values in the measured space $(\mathcal{X}, \mathcal{W}, \mu)$. We suppose that the distribution P_s of X_i admits a density s with respect to μ and aim at estimating s. We turn the set $\mathcal{S}$ of all probability densities with respect to μ into a metric space using the Hellinger distance $h(t,u) = H(P_t, P_u)$ where

$$H^2(P,Q) = \frac{1}{2}\int \left(\sqrt{dP} - \sqrt{dQ}\right)^2,$$

dP and dQ being the densities of P and Q with respect to any dominating measure. Although Birgé's procedure relies on this distance, we shall also consider L_q-distances - derived from L_q-norms denoted $\|.\|_q$ - for $q = 1, 2$.

The quality of an approximation $t \in \mathcal{S}$ of the function s is measured by $\ell(t,s)$, where ℓ is a loss function (typically some power of a distance). The risk of an estimator $\tilde{s} = \tilde{s}(\mathbb{X})$ of the function s is defined through this loss function by $R_s(\tilde{s}, \ell) := \mathbb{E}_s[\ell(\tilde{s}, s)]$, where $\mathbb{E}_s$ denotes the expectation when s obtains. The Hellinger risk $R_s(\tilde{s}, h^2)$ comes from the loss $\ell = h^2$. The loss can also be defined as $\ell(t,s) = \gamma(t,X) - \gamma(s,X)$, where $\gamma : \mathcal{S} \times \mathcal{X} \mapsto [0,\infty)$ is a *contrast function* for which s appears as a minimizer of $\mathbb{E}_s[\gamma(t,X)]$ when $t \in \mathcal{S}$ (Birgé and Massart, 1993, Definition 1). In this context, the L_2-loss (resp. the Kullback-Leibler loss) is defined via the contrast function $\gamma(t,x) = \|t\|_2^2 - 2t(x)$ (resp. $\gamma(t,x) = -\log(t(x))$) for any $t \in \mathcal{S}$, $x \in \mathcal{X}$.

[2] recently, Sart has applied robust tests in the special cases of dyadic partition selection (Sart, 2012) and parameter selection (Sart, 2013)

[3] The scripts, developed for this paper using our R-package, are available on the RunMyCode website (`http://www.runmycode.org`) to increase transparency and reproducibility

1.2 About the Hold-Out

Formally, the *hold-out* (HO) is a two-steps estimation procedure which relies on a split of $\mathbb{X}$ into two non-empty complementary subsamples, $\mathbb{X}_t$ and $\mathbb{X}_v$.

- **Step one**: Using the *training* sample $\mathbb{X}_t$, we build a finite set $S = \{\hat{s}_m[\mathbb{X}_t], m \in \mathcal{M}\}$ of preliminary estimators.
- **Step two**: The *validation* sample $\mathbb{X}_v$ is dedicated to the selection of one point $\hat{m}$ in $\mathcal{M}$.

The final estimator might be either $\hat{s}_{\hat{m}}[\mathbb{X}_t]$ or $\hat{s}_{\hat{m}}[\mathbb{X}]$ depending on the authors. The goal is generally to select $\hat{m} \in \mathcal{M}$ such that

$$R_s(\hat{s}_{\hat{m}}[\mathbb{X}_t], \ell) \sim \inf_{m \in \mathcal{M}} R_s(\hat{s}_m[\mathbb{X}_t], \ell) \quad \text{or} \quad R_s(\hat{s}_{\hat{m}}[\mathbb{X}], \ell) \sim \inf_{m \in \mathcal{M}} R_s(\hat{s}_m[\mathbb{X}], \ell),$$

where ℓ is the relevant loss function and the symbol $\sim$ means that quantities on both sides are of the same order.

Usually, after performing *Step one*, one defines some random criterion $\mathrm{crit}(m)$ for each m and selects $\hat{m} \in \mathcal{M}$ that minimizes $\mathrm{crit}(m)$. In the *classical* hold-out, this criterion is an estimation of the risk, made using the empirical contrast based on the validation sample:

$$\mathrm{crit}_{\mathrm{HO}}(m, \mathbb{X}_t, \mathbb{X}_v) = \frac{1}{|\mathbb{X}_v|} \sum_{X_i \in \mathbb{X}_v} \gamma(\hat{s}_m[\mathbb{X}_t], X_i),$$

where $|A|$ denotes the cardinality of the set A. In this context one naturally selects the estimator with the smallest estimated risk,

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} \mathrm{crit}_{\mathrm{HO}}(m, \mathbb{X}_t, \mathbb{X}_v).$$

We shall note in what follows $\hat{m}_{LS}$ and $\hat{m}_{KL}$ for the estimators selected by the classical procedure using the contrast functions $\gamma(t, x) = \|t\|_2^2 - 2t(x)$ and $\gamma(t, x) = -\log(t(x))$ respectively. We call least-squares hold-out (LSHO) and Kullback-Leibler hold-out (KLHO) the corresponding HO procedures. Few theoretical results exist concerning this classical HO in the density framework. Nevertheless, considering projection estimators together with the least-squares contrast, Arlot and Lerasle (2012) have shown that the LSHO criterion can be written as a penalization criterion with some resampling-based penalty. They also proved an oracle inequality and provided variances computations for this criterion (see Theorem 3 and Appendix I in the supplementary material in Arlot and Lerasle (2012)).

1.3 Overview of the paper

In practice the selection problem of *Step two* amounts to select one estimator in a given collection of $|\mathcal{M}|$ initial candidates. While the classical HO relies on the optimization of an empirical contrast function and thus requires at most $|\mathcal{M}|$ computations, T-estimation involves pairwise comparisons based on robust tests between probability balls leading to a quadratic number $O(|\mathcal{M}|^2)$ of tests.

The first goal of this paper is to provide an algorithm in a general framework of T-estimation which allows an efficient and exact implementation of T-estimation in the HO context. This algorithm breaks this quadratic bound. The second goal is to compare the risk performance of this T-hold-out for several losses and a large set of densities. We shall proceed a comparison against two types of procedures: those which select one point in a given family using the validation sample and those which estimate the density from the full sample. Moreover, we provide a faster, but approximate, version of this exact algorithm. We shall study both algorithms from a computational complexity point-of-view as well as the risk performances of the resulting estimators.

The paper is organized as follows. In Section 2 we recall the definition of the T-hold-out in a general framework. We introduce in Section 3 our exact and efficient algorithm which implements exact T-estimation and one approximate version derived from it. Section 4 presents the simulation protocol of our empirical study together with a short description of the R package *Density.T.HoldOut*. Section 5 is dedicated to the study of the quality of the exact T-hold-out in terms of risk. We also provide in this section comparisons with other selection methods. Section 6 is devoted to the empirical study of the complexity of the exact algorithm. Section 7 provides a comparison of exact and approximate algorithms both in terms of quality of estimation and complexity.

2 T-Hold-Out

Let us recall the T-hold-out procedure in a general framework where robust tests exist. We have at hand two independent samples, $\mathbb{X}_t$ and $\mathbb{X}_v$, and we want to estimate some target s belonging to the metric space $(\mathcal{S}, d)$. Suppose that a family $S = \{\hat{s}_m[\mathbb{X}_t], m \in \mathcal{M}\}$ of estimators of s has been built from $\mathbb{X}_t$, and we want to proceed to the validation step with $\mathbb{X}_v$. For $m_1, m_2 \in \mathcal{M}$, we note $d(m_1, m_2)$ instead of $d(\hat{s}_{m_1}[\mathbb{X}_t], \hat{s}_{m_2}[\mathbb{X}_t])$. The key assumption in the construction is the existence of some test having the following robustness property.

Assumption A There exists two constants $a > 0$, $\theta \in (0, 1/2)$, such that, for any m_1 and $m_2 \in \mathcal{M}$, there exists a test $\psi_{m_1,m_2} = \psi_{m_2,m_1}$ which chooses between m_1 and m_2, which satisfies:

$$\sup_{\{s \in \mathcal{S} | d(s,m_1) \leq \theta d(m_1,m_2)\}} \mathbb{P}_s[\psi_{m_1,m_2} = m_2] \leq \exp\Big[- a d^2(m_1, m_2)\Big], \tag{1}$$

$$\sup_{\{s \in \mathcal{S} | d(s,m_2) \leq \theta d(m_1,m_2)\}} \mathbb{P}_s[\psi_{m_1,m_2} = m_1] \leq \exp\Big[- a d^2(m_1, m_2)\Big]. \tag{2}$$

Conditionally to the knowledge of S, that is conditionally to $\mathbb{X}_t$, the test ψ_{m_1,m_2} only depends on the validation sample $\mathbb{X}_v$. Under Assumption A, the T-hold-out (THO) criterion is given by

$$\mathrm{crit}_{\mathrm{THO}}(m, \mathbb{X}_t, \mathbb{X}_v) := \sup_{j \in \mathcal{R}_m} d(\hat{s}_j[\mathbb{X}_t], \hat{s}_m[\mathbb{X}_t]),$$

with $\mathcal{R}_m = \{j \in \mathcal{M},\ j \neq m \mid \psi_{m,j} = j\}$. One finally chooses

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} \mathrm{crit}_{\mathrm{THO}}(m, \mathbb{X}_t, \mathbb{X}_v).$$

There are several theoretical differences with classical HO methods. The $\mathrm{crit}_{\mathrm{THO}}(m, \mathbb{X}_t, \mathbb{X}_v)$ criterion does not estimate the risk but appears instead as a *plausibility index*. Its value is computed through robust tests between estimators, while the classical HO criterion is computed independently for any estimator and thus does not take the geometrical structure of S into account. Theoretical results about this HO procedure can be found in Birgé (2006, Corollary 9) for the Hellinger risk, and in Birgé (2013b, Corollary 1) for the L_2-risk.

In the density framework Assumption A is fulfilled with $d = h$, $a = (1 - 2\theta)^2 |\mathbb{X}_v|$ (see Birgé (2013a)). To the best of our knowledge it is the first HO based on the Hellinger distance. Considering two densities $\hat{s}_i[\mathbb{X}_t]$ and $\hat{s}_j[\mathbb{X}_t]$, the test is defined by

$$\psi_{i,j} = \begin{cases} i & \text{if} \quad T_{i,j} \leq 1 \\ \\ j & \text{otherwise} \end{cases} \tag{3}$$

where

$$T_{i,j} = \prod_{X_k \in \mathbb{X}_v} \frac{\sin(\theta\omega)\sqrt{\hat{s}_i[\mathbb{X}_t]} + \sin(\omega(1-\theta))\sqrt{\hat{s}_j[\mathbb{X}_t]}}{\sin(\theta\omega)\sqrt{\hat{s}_j[\mathbb{X}_t]} + \sin(\omega(1-\theta))\sqrt{\hat{s}_i[\mathbb{X}_t]}}(X_k), \tag{4}$$

with $\omega = \arccos(1 - h^2(\hat{s}_i[\mathbb{X}_t], \hat{s}_j[\mathbb{X}_t]))$.

3 Efficient algorithms for T-estimation

In this section, we detail our algorithms which are at the core of the *Density.T.HoldOut* package to implement THO. Both algorithms may be useful in a general framework of T-estimation as they allow to drop the combinatorial complexity. While our first algorithm computes the true T-estimator, the second implements a lossy approach to reduce the complexity further when the family S is very large, while maintaining good performances in terms of Hellinger risk. In both cases, we assume that *Step one* has already been performed, hence our aim is only to select $\hat{m}$ among the finite S collection of preliminary estimators using $\mathbb{X}_v$, as described in Section 2. Since $\mathcal{M}$ is finite, we assume without loss of generality that $\mathcal{M} = \{1, \dots, M\}$. Since the estimators $\hat{s}_m[\mathbb{X}_t]$ are built from a sample independent of $\mathbb{X}_v$, they are, conditionally to $\mathbb{X}_t$, deterministic points in $\mathcal{S}$. From now on we note them s_m - or m when no confusion is possible - and the THO criterion $\mathrm{crit}_{\mathrm{THO}}(m, \mathbb{X}_t, \mathbb{X}_v)$ is denoted $\mathcal{D}(m) = \max_{i \in \mathcal{R}_m} d(i, m)$, where we recall that $\mathcal{R}_m$ consists of the $j \in \{1, \dots, M\} \setminus \{m\}$ which are chosen against m by the robust tests. Finally let us denote $\bar{\mathcal{B}}(m, r) = \{l \in \{1, \dots, M\} : d(m, l) \leq r\}$ the intersection of $\mathcal{M}$ with the closed ball with center m and radius $r > 0$. From a purely combinatorial point-of-view, the computation of $\hat{m}$ minimizing the plausibility radius $\mathcal{D}(m)$ requires the computation of $O(M^2)$ tests with a "naive" algorithm, which is prohibitive as compared to the $O(M)$ operations needed to compute the classical HO estimator.

3.1 Exact T-Hold-Out

The T-estimator search can be realized with a non-quadratic number of test, thanks to a simple argument which is summarized by the following lemma and its corollary.

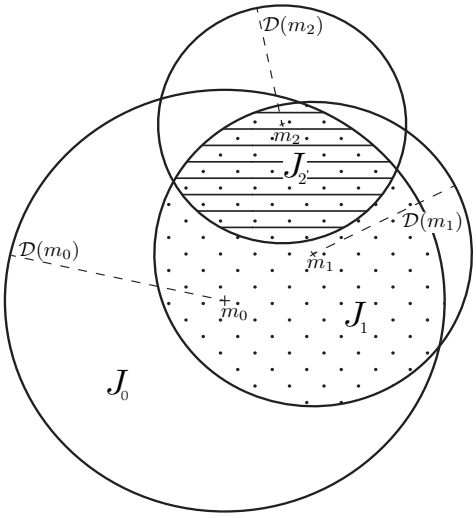

Figure 1: Illustration of our exact search for T-estimation. Along the three first iterations, the estimator m_i, $i = 0, 1, 2$ are considered with associated radius $\mathcal{D}(m_i)$ and the T-estimator belongs successively to J_i where J_0 is $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0))$, J_1 is the dotted and J_2 the hatched area.

Lemma 1. *For any point $m_0 \in \{1, \dots, M\}$, the T-estimator $\hat{m}$ belongs to $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0))$.*

Proof. Suppose that there exists one point $m_0 \in \{1, \dots, M\}$ such that $\hat{m}$ does not belong to the closed ball of radius $\mathcal{D}(m_0)$ centered at m_0. Then it does not belong to $\mathcal{R}_{m_0}$, and it follows that $\psi_{m_0, \hat{m}} = m_0$. Hence m_0 belongs to $\mathcal{R}_{\hat{m}}$ leading to $\mathcal{D}(\hat{m}) \geq d(\hat{m}, m_0) > \mathcal{D}(m_0)$ which provides a contradiction with $\mathcal{D}(\hat{m}) = \min_{m \in \{1, \dots, M\}} \mathcal{D}(m)$. □

Corollary 1. *For any subset $J \subset \{1, \dots, M\}$, the T-estimator $\hat{m}$ belongs to*

$$\bigcap_{m \in J} \bar{\mathcal{B}}(m, \mathcal{D}(m)).$$

Proof. The proof, illustrated by Figure 1, is straightforward using similar arguments as in Lemma 1. □

It follows that, starting from m_0, only a point inside $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0))$ may be the T-estimator. If any point m_1 in this first ball satisfies $\mathcal{D}(m_1) < \mathcal{D}(m_0)$, by Lemma 1, the T-estimator will belong to $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0)) \bigcap \bar{\mathcal{B}}(m_1, \mathcal{D}(m_1))$. Again, criterion $\mathcal{D}$ need to be computed only for points inside this intersection. We keep intersecting balls $\bar{\mathcal{B}}(m, \mathcal{D}(m))$ until there is no more point with a value of $\mathcal{D}$ smaller than its running value. This approach provides an exact computation of the T-estimator.

At each step of the recursion, the current best point is denoted m with associated value $\mathcal{D}(m)$ denoted by $\mathcal{D}$. The running intersection which contains the potentially better points than m is denoted J (this set does not contain m). The recursion stops when J is empty. At a given step of the recursion, a point j in J is better than m - and thus replaces it - if $\mathcal{D}(j) < \mathcal{D}$. In all cases, j is removed from the set J. During the iteration, $|J|$ and $\mathcal{D}$ decrease ensuring that the algorithm stops. The last running m is the T-estimator. The pseudo-code implementing the efficient and exact search of the T-estimator is provided by Algorithm 1.

Algorithm 1: Efficient and exact T-Hold-Out

Input: $m \in J = \{1, \ldots, M\}$
for $(j \neq m)$ **do** compute $\psi_{m,j}(\mathbb{X}_v)$
Compute $\mathcal{D} = \mathcal{D}(m)$ and set $J = \bar{\mathcal{B}}(m, \mathcal{D}) \setminus \{m\}$
while $(|J| > 0)$ **do**
  Set $\mathcal{D}_{tmp} = 0$, select $j \in J$ and set $J = J \setminus \{j\}$
  for $(k \neq j)$ **do**
    Compute $\psi_{k,j}(\mathbb{X}_v)$ `// if it has not been done yet`
    if $(\psi_{k,j}(\mathbb{X}_v) == k)$ **then** `//` $k \in \mathcal{R}_j$
      Set $\mathcal{D}_{tmp} = \max(\mathcal{D}_{tmp}, d(j,k))$
      if $(\mathcal{D}_{tmp} > \mathcal{D})$ **then** break `// break the for loop`
  Set $m = j$, $\mathcal{D} = \mathcal{D}_{tmp}$ and $J = J \bigcap \bar{\mathcal{B}}(m, \mathcal{D})$
Return: m `// the T-estimator`

Comments: This algorithm works for all the statistical frameworks of T-estimation. The "for" loop is realized on all $k \neq j$, as $\mathcal{D}(k)$ depends on all points and not only on those in J. If there are N points in the first ball, the number of computed tests is at most $O(N * M)$. Moreover, if the first ball is empty, i.e. if $\mathcal{D}(m) = 0$, the algorithm stops immediately, returning m for $\hat{m}$. In this case, the complexity of our algorithm is $O(M)$. Any preliminary estimator (maximum likelihood, least-squares, L_1-minimizer, etc.) may be a starting point of our algorithm. We hope that by starting from a good preliminary estimator, there will be only few points in the first ball, resulting in less computations. The computation requires $O(M^2)$ operations if J decreases by only one point at each step of the recursion which happens only if the selected j satisfies

$$\max_{k \in J} d(j,k) = \max_{k,l \in J} d(k,l)$$

at each iteration.

3.2 Fast algorithm for approximate T-Hold-Out

Assumption A ensures that as soon as the Hellinger distance between two estimators of S is large enough, the probability that the robust test does not choose the best estimator is small. However, as shown in Lemma 1 of Le Cam (1973), when this distance is smaller than $c\, n^{-1/2}$, where c is a small positive constant, the two corresponding probabilities cannot be separated by a test built on n observations anymore. From this remark, we derive a lossy version from our efficient and exact algorithm. The main difference consists in ignoring points in S as soon as their Hellinger distance to a previously considered one is smaller than a given threshold $\delta_n > 0$.

We introduce this distance control at two steps of our efficient and exact algorithm. As the interior points of $\bar{\mathcal{B}}(m, \delta_n)$ cannot be properly distinguished from m by any test, the set J becomes, at lines 2 and 10 of Algorithm 1, the intersection of rings instead of balls, obtained by removing from the original ball $\bar{\mathcal{B}}(m, \mathcal{D}(m))$ the ball $\bar{\mathcal{B}}(m, \delta_n)$. In the same spirit, at line 5 of Algorithm 1, the current k, in the *for* loop, is considered if and only if its distance to $\mathcal{T}_j$ is larger than δ_n, where $\mathcal{T}_j$ is made of the running j and the further points which have been tested against j. The pseudo-code of this lossy version is provided by Algorithm 2 and illustrated by Figure 2.

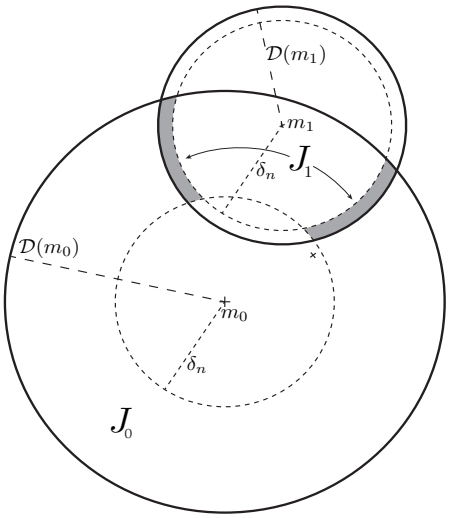

Figure 2: Illustration of the approximate T-estimation search: J_0 is a ring around m_0. The point following m_0 has changed with respect to Figure 1 as the previously selected m_1 is now inside $\bar{\mathcal{B}}(m_0, \delta_n)$. J_1 (in grey) appears as the intersection of two rings.

Algorithm 2: Approximate T-Hold-Out

Input: $m \in J = \{1, \ldots, M\}$; $\delta_n > 0$
1 **for** $(j \neq m)$ **do** compute $\psi_{m,j}(\mathbb{X}_v)$
2 Compute $\mathcal{D} = \mathcal{D}(m)$ and set $J = \bar{\mathcal{B}}(m, \mathcal{D}) \setminus \bar{\mathcal{B}}(m, \delta_n)$
3 **while** $(|J| > 0)$ **do**
4 Set $\mathcal{D}_{tmp} = 0$, select $j \in J$ and set $J = J \setminus \{j\}$
5 Define $\mathcal{T}_j = \{j\}$
6 **for** $(k \neq j)$ **do**
7 **if** $(d(j, \mathcal{T}_j) \leq \delta_n)$ **then** next k `// next k if distance is too small`
8 Set $\mathcal{T}_j = \mathcal{T}_j \cup \{k\}$
9 Compute $\psi_{k,j}(\mathbb{X}_v)$ `// if it has not been done yet`
10 **if** $(\psi_{k,j}(\mathbb{X}_v) == k)$ **then** `// k ∈ R_j`
11 Set $\mathcal{D}_{tmp} = \max(\mathcal{D}_{tmp}, d(j, k))$
12 **if** $(\mathcal{D}_{tmp} > \mathcal{D})$ **then** break `// break the for loop`
13 Set $m = j$, $\mathcal{D} = \mathcal{D}_{tmp}$ and $J = J \bigcap \left[\bar{\mathcal{B}}(m, \mathcal{D}) \setminus \bar{\mathcal{B}}(m, \delta_n)\right]$
Return: m `// the approximate T-estimator`

4 Simulation protocol

We considered $X_1, \ldots, X_n$ i.i.d. random variables from an unknown density s with respect to the Lebesgue measure on $\mathcal{X} = \mathbb{R}$. This framework is motivated by the fact that likelihood ratio tests are not robust in this context and we hoped to observe some differences in terms of risk.

Simulations were carried out with four sample sizes $n = 100, 250, 500, 1000$. Our test functions s vary in a subset $\mathcal{L}$ made of the densities $s_1, \ldots, s_{28}$ of the R-package *benchden*[4] which are in

[4] *Benchden* (see Mildenberger and Weinert (2012)) implements the benchmark distributions of Berlinet and Devroye (1994). Available on the CRAN `http://cran.r-project.org/web/packages/benchden/index.html`.

$L_1 \cap L_2$ - to ensure that risks are computable. This set $\mathcal{L}$ is made of the densities s_i for

$$i \in \{1, \ldots, 5, 7, 11, 12, 13, 16, 17, 21, \ldots, 27\}.$$

We considered several estimator collections:

- S_R made of regular histograms with bin number varying from 1 to $\lceil n/\log(n) \rceil$ as described in Birgé and Rozenholc (2006);
- S_I made of the maximum likelihood irregular histograms when the bin number only varies from 1 to 100 as described in Rozenholc et al. (2010);
- S_K made of Gaussian kernel estimators with varying bandwidths to be specified later;
- S_P made of parametric estimates obtained by moment's method for the Gaussian, exponential, log-normal, chi-square, gamma and beta distributions together with a maximum likelihood estimate of the uniform distribution;
- $S_C = S_R \cup S_I$ as used in Rozenholc et al. (2010);
- $S_1 = S_R \cup S_I \cup S_K$;
- $S_2 = S_R \cup S_I \cup S_K \cup S_P$.

The estimation accuracy of a given procedure $\tilde{s}$ has been evaluated using an empirical version of the risk $R_s(\tilde{s}, \ell) = \mathbb{E}_s[\ell(\tilde{s}, s)]$, obtained by generating 100 samples $\mathbb{X}^{(j)} = (X_1^j, \ldots, X_n^j)$, $1 \leq j \leq 100$, of size n and density s:

$$\bar{R}_s(\tilde{s}, \ell) = \frac{1}{100} \sum_{j=1}^{100} \ell(\tilde{s}[\mathbb{X}^{(j)}], s),$$

where $\ell(t, u) = h^2(t, u)$ for the Hellinger distance, and $\ell(t, u) = \|t - u\|_q^q$ for $q = 1, 2$.

In order to compare two procedures $\tilde{t}_1$ and $\tilde{t}_2$, we introduce the normalized $\log_2$-ratio of their empirical risks, namely:

$$\bar{W}_s(\tilde{t}_1, \tilde{t}_2) = \frac{1}{r} \log_2 \frac{\bar{R}_s(\tilde{t}_1, \ell)}{\bar{R}_s(\tilde{t}_2, \ell)} = \log_2 \bar{R}_s^{1/r}(\tilde{t}_1, \ell) - \log_2 \bar{R}_s^{1/r}(\tilde{t}_2, \ell),$$

where r is equal to q for L_q losses and 2 for the Hellinger loss. The aim of the normalization by r is to provide an easier comparison of $\bar{W}_s$ when the loss changes. In our empirical study, procedure $\tilde{t}_2$ is thus considered better in terms of risk than $\tilde{t}_1$ for a given loss function if the values of $\bar{W}_s(\tilde{t}_1, \tilde{t}_2)$ are positive when the density s varies.

We compared three hold-out methods described above: T-estimation, LS and KL. We first divided $\mathbb{X}$ in $\mathbb{X}_t = (X_1, \ldots, X_{\lfloor pn \rfloor})$ and $\mathbb{X}_v = (X_{\lfloor pn \rfloor + 1}, \ldots, X_n)$ using proportions p equal to 1/2, 2/3 and 3/4. Second, we computed $\hat{s}_m[\mathbb{X}_t]$ for all $m \in \mathcal{M}$. Finally we selected $\hat{m}$ which minimizes the respective HO criterion resulting in $\hat{m}_T$, $\hat{m}_{LS}$ and $\hat{m}_{KL}$ providing $\tilde{s}$ as either $\hat{s}_{\hat{m}}[\mathbb{X}_t]$ or $\hat{s}_{\hat{m}}[\mathbb{X}]$. As $\hat{m}$ depends on the chosen proportion p, in order to explicitly specify the dependency of $\hat{m}$ with respect to this parameter, we will use the following notations $\hat{s}_{\hat{m}[p]}[\mathbb{X}_t]$ or $\hat{s}_{\hat{m}[p]}[\mathbb{X}]$ when needed.

For the family S_K, the bandwidths were chosen as

$$(\max[\mathbb{X}_t] - \min[\mathbb{X}_t])/2j \quad \text{for} \quad j = 1, \ldots, \lceil n/\log(n) \rceil.$$

In Algorithms 1 and 2, the input m has been set to $\hat{m}_{LS}$ and $j = \arg\max_{k \in J} d(k, m)$, at line 4. In Algorithm 2, we fixed $\delta_n = 1/\sqrt{|\mathbb{X}_v|}$ as a lower bound for the Hellinger distance between distinguishable probabilities, following Le Cam (1973).

Moreover, we also considered some calibrated estimation procedures which choose m in some particular families. These are not direct competitors with the T-estimation as they cannot deal with general families S but provide a good benchmark in terms of risk:

- for S_R, S_I, S_C, the penalized maximum likelihood estimators, denoted $\tilde{s}_{\text{pen}}$ introduced in Birgé and Rozenholc (2006); Rozenholc et al. (2010) and implemented in the R package[5] *histogram*,
- for S_K, the L_1-version of the procedure introduced in Goldenshluger and Lepski (2011), denoted $\tilde{s}_{GL}$.

Finally, for the family S_K, we considered some bandwidth selectors (namely nrd, ucv, bcv, SJ) implemented in the *density* generic function available in R , providing some well-known estimators $\tilde{s}_{nrd}$, $\tilde{s}_{bcv}$, $\tilde{s}_{ucv}$, $\tilde{s}_{SJ}$ of the density (Silverman, 1986; Sheather and Jones, 1991; Scott, 1992).

The R package[5] *Density.T.HoldOut* is a ready-for-use software that implements our algorithms in the density framework. The main function - called `densityTestim` - receives as input a sample $\mathbb{X}$ and a family of estimators and returns the selected estimator. The previously described families are available and can be extended or adapted by the user (default family is S_2). Other important input arguments are parameters p, θ and the starting point (default values are $p = 1/2$, $\theta = 1/4$ and $\hat{m}_{LS}$). This function implements the exact and lossy algorithms, through the numeric `csqrt` (default value 0) which controls $\delta_n = \texttt{csqrt}/\sqrt{|\mathbb{X}_v|}$ in Algorithm 2. The default value corresponds to the Algorithm 1 implementing the exact search. The estimator $\hat{s}_{\hat{m}_T}$ is either built with $\mathbb{X}_t$ (`last`='training') or $\mathbb{X}$ (`last`='full', default).

5 Simulation results

This section, which was done thanks to Algorithm 1, is devoted to the study of the quality of the T-hold-out. We illustrate our results showing boxplots of $\bar{W}_s(\tilde{t}_1, \tilde{t}_2)$ for all 18 densities $s \in \mathcal{L}$, various choices of the estimators $\tilde{t}_1$ and $\tilde{t}_2$, various collections of estimators S when n equals to 100, 250, 500 and 1000. We begin by investigating how parameters θ and p influence the THO procedure. Then we provide two main comparison types. First we look at HO methods which select among a family of points using the validation sample. Then we compare the THO against some density estimation methods, which are not necessarily selection procedures anymore. In this subsection, we divide the presentation between calibrated selection procedures and some selectors obtained using asymptotic derivation of the risk for some specific loss.

5.1 Influence of θ

The robustness of the test is controlled through the parameter $\theta < 1/2$ (see Eq. 4), the KLHO corresponding to $\theta = 0$ (no robustness). We computed the empirical risk using the THO procedure with $\theta = 1/16, 1/8, 1/4, 3/8, 7/16$, and $n = 100, 250, 500, 1000$. We observed that θ has little influence in terms of risk ($\theta = 1/16$ being slightly worse) and decided to pursue the empirical study with $\theta = 1/4$.

5.2 Influence of p

We examine the dependence of the THO with respect to p, the proportion of the initial sample dedicated to build the estimators, using the Hellinger risk.

[5]available on the CRAN `http://cran.r-project.org/web/packages/histogram/index.html`.

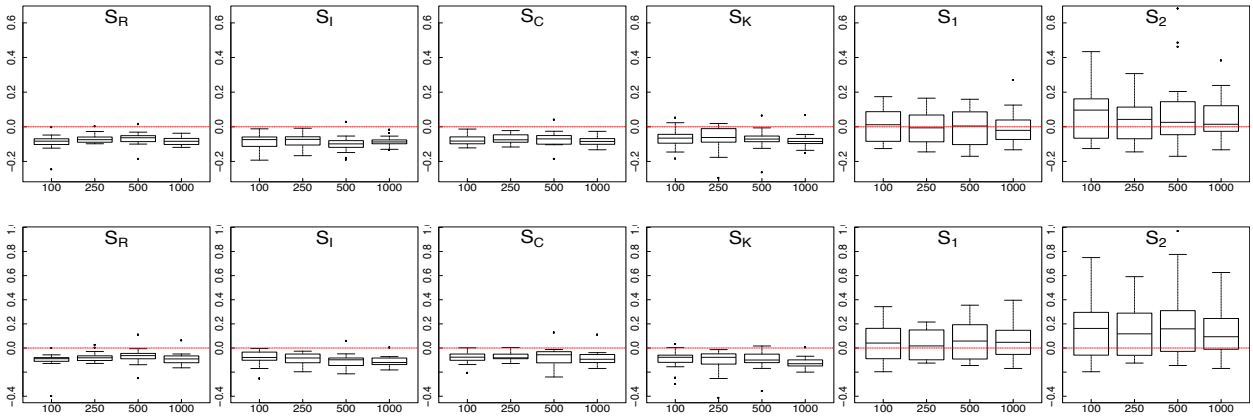

Figure 3: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\hat{s}_{\hat{m}[2/3]}[\mathbb{X}_t], \hat{s}_{\hat{m}[1/2]}[\mathbb{X}_t])$ (upper line) and $\bar{W}_s(\hat{s}_{\hat{m}[3/4]}[\mathbb{X}_t], \hat{s}_{\hat{m}[1/2]}[\mathbb{X}_t])$ (bottom line) for the Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. Each subfigure shows the boxplot for n equals 100, 250, 500 and 1000. The horizontal red dotted line provides the reference value 0.

Figure 3 is built using $\tilde{t}_1 = \hat{s}_{\hat{m}[p]}[\mathbb{X}_t]$ for p equals 2/3 (upper line), 3/4 (bottom line) and $\tilde{t}_2 = \hat{s}_{\hat{m}[1/2]}[\mathbb{X}_t]$. We observe two different behaviors for families S_R, S_I, S_C and S_K on the one hand and for S_1 and S_2 on the other hand. For the first families $p = 2/3$ or $3/4$ is better than $p = 1/2$. For the second ones $p = 2/3$ seems equivalent to $p = 1/2$ but $p = 3/4$ is worst than $p = 1/2$. Hence we consider preferable to use $p = 2/3$, which makes the best compromise for all families.

5.3 Comparing Hold-Out methods

Hold-out procedures are universal since they do not depend on the family S. They can be seen as methods that choose among some family of fixed points. Setting $p = 2/3$, we compare the THO to the KLHO and LSHO introduced in Section 1.2 on the 6 estimator collections described in the previous section.

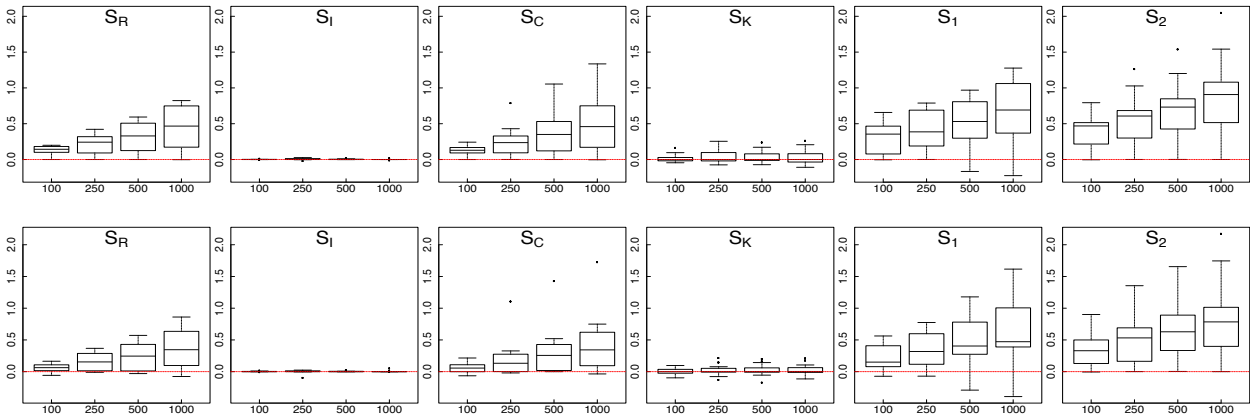

Figure 4: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\hat{s}_{\hat{m}_{KL}}[\mathbb{X}_t], \hat{s}_{\hat{m}_T}[\mathbb{X}_t])$ for $p = 2/3$, using collections S_R, S_I, S_C, S_K, S_1 and S_2. Upper line, using Hellinger loss, bottom line using L_1 loss. See Figure 3 for more details.

Figure 4 is built using $\tilde{t}_1 = \hat{s}_{\hat{m}_{KL}}[\mathbb{X}_t]$ and $\tilde{t}_2 = \hat{s}_{\hat{m}_T}[\mathbb{X}_t]$ considering Hellinger (upper line) and L_1 (bottom line) losses. In all cases, the median and most of the distribution are positive, meaning that the THO outperforms the KLHO estimator. For collections S_I and S_K, empirical risks for both losses are similar, with $\bar{W}_s(\hat{s}_{\hat{m}_{KL}}[\mathbb{X}_t], \hat{s}_{\hat{m}_T}[\mathbb{X}_t])$ being respectively larger than -0.01 (except for the uniform density) for S_I, and -0.2 for S_K. When n grows, while for S_I and S_K the ratio remains stable, it increases for all other families in favor of the THO. Moreover when going from

collection S_1 to S_2, that is adding the parametric collection S_P, we observe that the already good performances of the THO improve. We therefore suspect that the THO chooses the parametric estimator more often than KLHO when facing the corresponding densities.

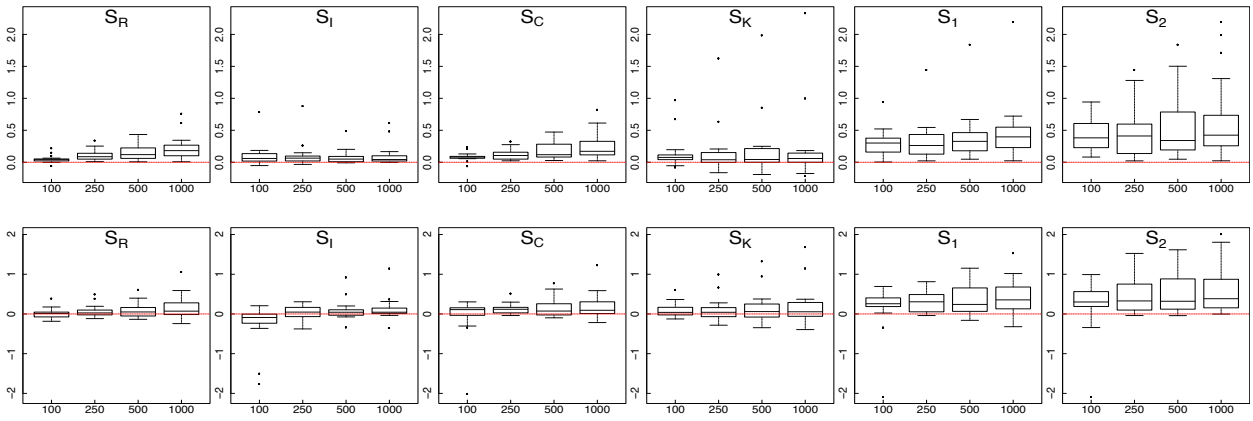

Figure 5: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\hat{s}_{\hat{m}_{LS}}[\mathbb{X}_t], \hat{s}_{\hat{m}_T}[\mathbb{X}_t])$ for $p = 2/3$, using collections S_R, S_I, S_C, S_K, S_1 and S_2. Upper line, using Hellinger loss, bottom line using L_2 loss. See Figure 3 for more details.

Figure 5 is built using $\tilde{t}_1 = \hat{s}_{\hat{m}_{LS}}[\mathbb{X}_t]$ and $\tilde{t}_2 = \hat{s}_{\hat{m}_T}[\mathbb{X}_t]$ considering Hellinger (upper line) and L_2 (bottom line) losses. The THO performs better than the LSHO estimator for all collections except for the collection S_I when $n = 100$. For the larger collections S_1 and S_2, the THO outperforms the LSHO. However, as n grows, we observe that the relative quality of the two procedures remain stable.

5.4 Comparing final strategies for T-Hold-Out

As we now aim at comparing T-estimation versus calibrated methods, we first investigate whether $\hat{s}_{\hat{m}_T}[\mathbb{X}_t]$ or $\hat{s}_{\hat{m}_T}[\mathbb{X}]$ performs better. For this purpose, we study the Hellinger risk of $\hat{s}_{\hat{m}_T}[\mathbb{X}]$ when p varies. Figure 6 is built using $\tilde{t}_1 = \hat{s}_{\hat{m}[p]}[\mathbb{X}]$ for p equals 2/3 (upper line), 3/4 (bottom line) and $\tilde{t}_2 = \hat{s}_{\hat{m}[1/2]}[\mathbb{X}]$. We observe that against $p = 2/3$ or $p = 3/4$, the value $p = 1/2$ provides better results for the large families S_1 and S_2 while for the small families the results are more balanced. Hence we consider preferable to make use of this strategy with $p = 1/2$.

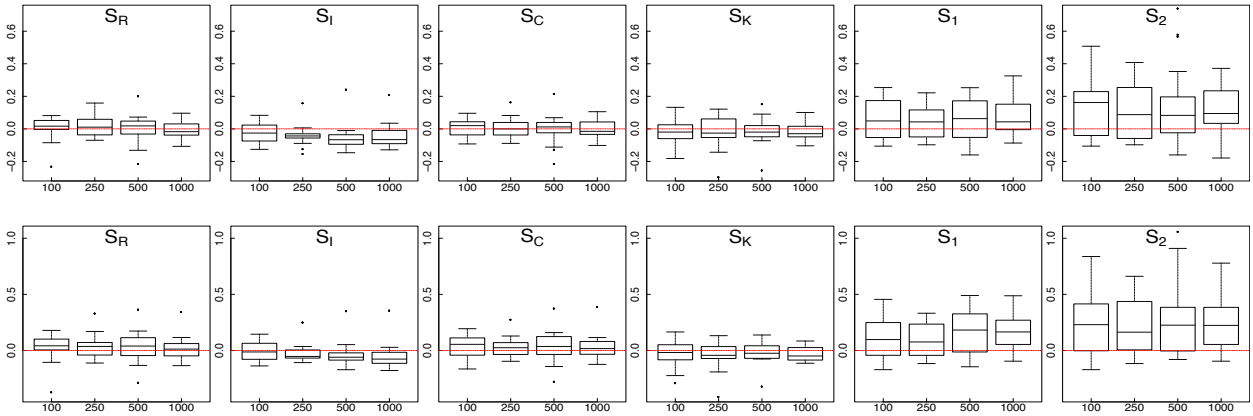

Figure 6: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\hat{s}_{\hat{m}[2/3]}[\mathbb{X}], \hat{s}_{\hat{m}[1/2]}[\mathbb{X}])$ (upper line) and $\bar{W}_s(\hat{s}_{\hat{m}[3/4]}[\mathbb{X}], \hat{s}_{\hat{m}[1/2]}[\mathbb{X}])$ (bottom line) for the Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. See Figure 3 for more details.

We now compare the Hellinger risks of $\hat{s}_{\hat{m}[2/3]}[\mathbb{X}_t]$ - which appeared as the best competitor in

Section 5.2 - and $\hat{s}_{\hat{m}[1/2]}[\mathbb{X}]$. Figure 7 is built using $\tilde{t}_1 = \hat{s}_{\hat{m}[1/2]}[\mathbb{X}]$ and $\tilde{t}_2 = \hat{s}_{\hat{m}[2/3]}[\mathbb{X}_t]$. We observe that the strategy $\hat{s}_{\hat{m}[1/2]}[\mathbb{X}]$ is preferable, since its median (and even most of its distribution) is negative in all considered settings. It should be noticed that our simulations show that, more than the value of p, it is the use of $\mathbb{X}$ instead of $\mathbb{X}_t$ which has the larger influence on the final risk.

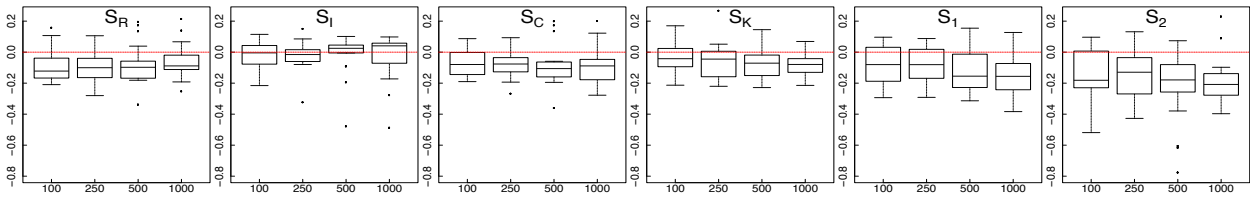

Figure 7: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\hat{s}_{\hat{m}[1/2]}[\mathbb{X}], \hat{s}_{\hat{m}[2/3]}[\mathbb{X}_t])$ for Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. See Figure 3 for more details.

5.5 T-Hold-Out against dedicated estimation procedures

We now compare the THO competitor $\hat{s}_{\hat{m}[1/2]}[\mathbb{X}]$ against the so-called dedicated methods. Figure 8 is built using $\tilde{t}_1 = \tilde{s}[\mathbb{X}]$ ($\tilde{s}$ being either $\tilde{s}_{\text{pen}}$ or $\tilde{s}_{GL}$) and $\tilde{t}_2 = \hat{s}_{\hat{m}[1/2]}[\mathbb{X}]$ considering Hellinger (upper line) and L_1 (bottom line) losses. We observe that the THO is slightly worse than a well-calibrated procedure for histograms but outperforms the L_1-version of the Goldenshluger-Lepski procedure.

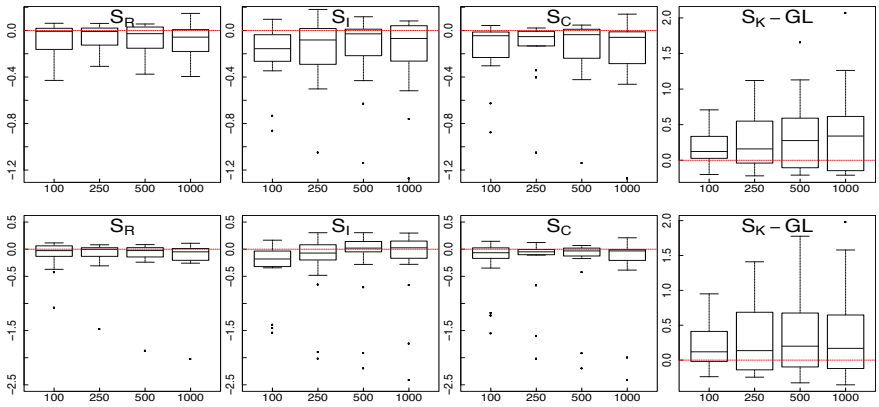

Figure 8: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\tilde{s}[\mathbb{X}], \hat{s}_{\hat{m}[1/2]}[\mathbb{X}])$ using collections S_R, S_I, S_C and S_K with Hellinger (upper line) and L_1 (bottom line) losses. For the 3 first collections $\tilde{s}$ is $\tilde{s}_{\text{pen}}$ and $\tilde{s}_{GL}$ for S_K. Each subfigure shows the boxplot for n equals 100, 250, 500 and 1000. The horizontal red dotted line provides the reference value 0.

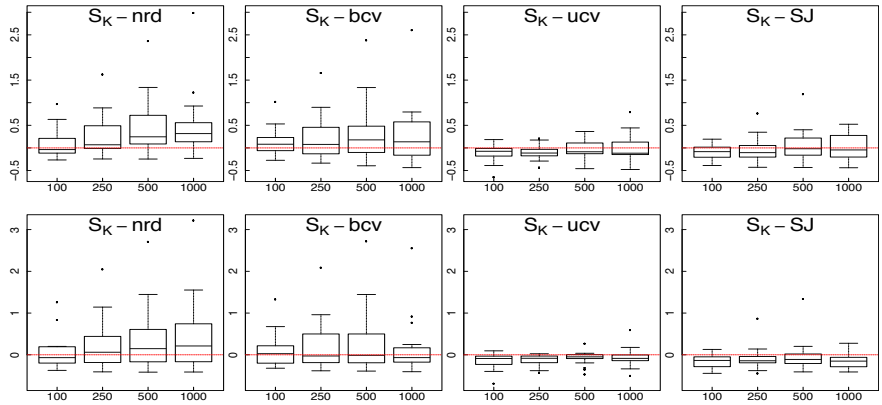

Figure 9: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\tilde{s}[\mathbb{X}], \hat{s}_{\hat{m}[1/2]}[\mathbb{X}])$ for collection S_K. The 3 first competitors $\tilde{s}$ are the kernel estimators with respective bandwidth provided by the bandwidth selectors *nrd*, *bcv*, *ucv* and *SJ* as defined in the function *density* of the *stats* package of *R*. Upper line, using Hellinger loss, bottom line using L_1 loss. See Figure 3 for more details.

For the sake of completeness, we also provide in Figure 9 the comparison between the THO and well-known estimators of the density derived from bandwidth selectors available in the *density* generic function of R. We observe that $\tilde{s}_{ucv}$ and $\tilde{s}_{SJ}$ perform well (particularly for the L_1-loss), whereas the THO outperforms $\tilde{s}_{nrd}$ and $\tilde{s}_{bcv}$.

6 Empirical complexity of the exact algorithm

To evaluate the complexity of our algorithms let us denote by N the number of tests needed in the computation of the THO for each generated sample of our simulations. As N is between $M-1$ and $M(M-1)/2$, we define the so-called "THO complexity" as the ratio of $N-M+1$ over its maximal value, that is

$$\frac{2(N-M+1)}{(M-1)(M-2)}. \tag{5}$$

For any run, this ratio belongs to $[0,1]$ by construction. For each fixed n, we get a global sample of size 10800 corresponding to "18 densities" times "6 families" times "100 simulations". Figure 10 shows the empirical cumulative distribution function (CDF) of the latter sample with the quantiles 0.75, 0.9 and 0.95. We observe from this figure that the complexity our algorithm tends to improve with n. Moreover, 75% of the THO complexities are smaller than 0.1 for n equals 250, 500 and 1000 and 95% are smaller than 0.4 for all values of n.

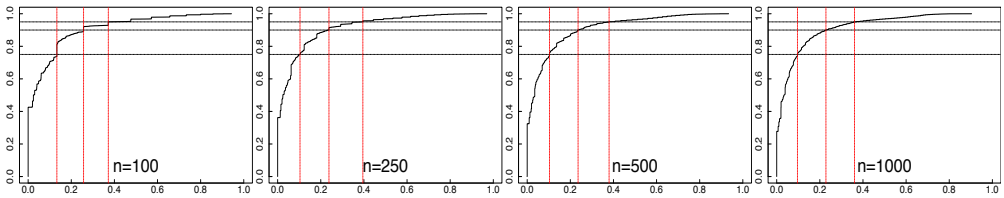

Figure 10: From left to right, the CDF for $n = 100$, 250, 500 and 1000 of the THO complexity in plain line using Algorithm 1. The horizontal black dotted lines provide the values 0.75, 0.9 and 0.95 and the vertical red dotted lines their respective quantiles.

In order to complete this study of the complexity, we focused on the two collections S_R and S_K for which the number of estimators depends on n as $M = \lceil n/\log(n) \rceil$. Having in mind that N is not smaller than $M-1$ and not larger than $M(M-1)/2$, we assumed N to be of order $(M-1)^\beta$ with β in $[1,2]$. For each density and each value of n, we compute the average of $\log(N)$ over the 100 runs. In Figure 11 these average values are drawn versus $\log(M-1)$ for the two collections and for each density.

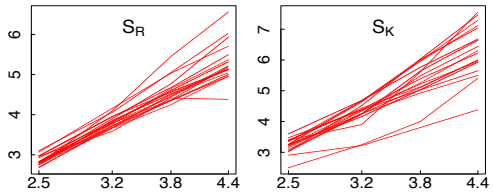

Figure 11: Graphs of $\log(N)$ versus $\log(M-1)$ for each density when using the collections S_R (left) and S_K (right).

As Figure 11 exhibits mostly linear behaviors, we computed the slope in the linear model of $\log(N)$ versus $\log(M-1)$ when n varies as an estimator of β. We observe that this estimator concentrates around respectively 1.2 and 1.4 for the collections S_R and S_K providing a good indicator that our algorithm is typically sub-quadratic. The larger value of β for the collection S_K may be explained by the fact that, for our set of bandwidths, the kernel estimators may be very similar, inducing a slow decrease of the running intersection J in Algorithm 1.

7 Study of the approximate T-Hold-Out

We provide a comparison of the estimators selected respectively using Algorithm 1 and 2, that is the exact T-estimator and its approximate version (denoted here by $\hat{m}_T^g$) computed with $\delta_n = c/\sqrt{|\mathbb{X}_v|}$ for different values of c. We compare these estimators using the two strategies based on $\mathbb{X}_t$ and $\mathbb{X}$.

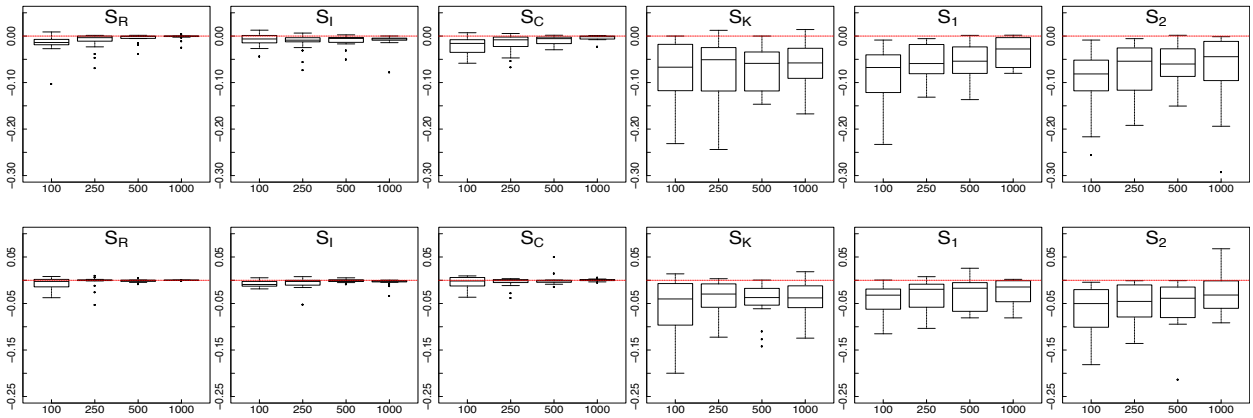

Figure 12: From left to right, normalized $\log_2$-ratio of the empirical risks $\bar{W}_s(\hat{s}_{\hat{m}_T}[\mathbb{X}_t], \hat{s}_{\hat{m}_T^g}[\mathbb{X}_t])$ (upper line) and $\bar{W}_s(\hat{s}_{\hat{m}_T}[\mathbb{X}], \hat{s}_{\hat{m}_T^g}[\mathbb{X}])$ using $c = 1$ (bottom line) for the Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. See Figure 3 for more details.

Figure 12 is built using $\tilde{t}_1 = \hat{s}_{\hat{m}_T}[\mathbb{X}_t]$ and $\tilde{t}_2 = \hat{s}_{\hat{m}_T^g}[\mathbb{X}_t]$ with $p = 2/3$ on the upper line and using $\tilde{t}_1 = \hat{s}_{\hat{m}_T}[\mathbb{X}]$ and $\tilde{t}_2 = \hat{s}_{\hat{m}_T^g}[\mathbb{X}]$ with $p = 1/2$ on the bottom line. As expected, the exact THO is better in terms of risk. For histogram families, the degradation of the Hellinger risk is negligible. For families S_K, S_1 and S_2, we observe that the risk increases not more than 20% in most of the cases (y-axis reference value equals to -0.13). The empirical cumulative distribution function (CDF) of the complexity ratio defined in (5) is shown in Figure 13 for comparison with Figure 10. Clearly the CDFs of the lossy version are more concentrated around 0, showing a significant gain in terms of complexity when using Algorithm 2 (quantiles are divided by more than 2.5).

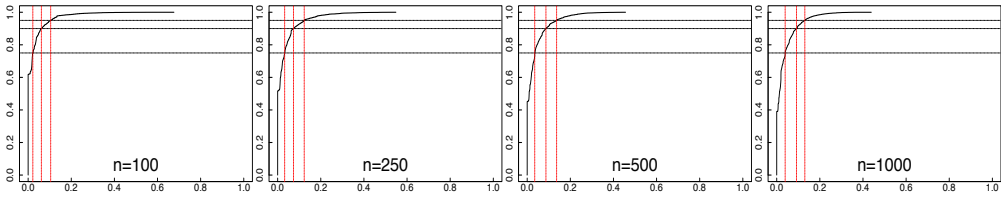

Figure 13: From left to right, the CDF for $n = 100$, 250, 500 and 1000 of the THO complexity in plain line using Algorithm 2 with $c = 1$. See Figure 10 for more details.

A further study, using $c = 2$ in the approximate algorithm, shows that the risk increases up to 75% in most of the cases and does not offer a good trade-off between complexity and accuracy.

8 Conclusion

We introduce an efficient and exact algorithm, together with an approximate version, for T-estimation in the context of hold-out. We study the performances of this T-hold-out in the density framework. Calibration study shows that, when building the final estimate only with the training sample, a good choice of the ratio between training and validation sample sizes is $p = 2/3$. However, risks can be improved using the full sample to build the final estimate when using $p = 1/2$. Our procedure is competitive compared to classical hold-out derived from Kullback-Leibler or least-squares contrasts. It still behaves well against model selection procedures derived from a calibrated penalized contrast for histogram selection, and against most of the bandwidth selectors for kernel estimators. Empirically, we observe that this algorithm improves clearly the combinatorial complexity. Moreover, it can be speeded up thanks to our proposed lossy version, which offers the expected trade-off between complexity and estimation quality.

References

S. Arlot and M. Lerasle. V-fold cross-validation and V-fold penalization in least-squares density estimation. arXiv:1210.5830v1, 2012.

Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143:239–284, 2009.

P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Universite de Paris*, 38(3):3–59, 1994.

L. Birgé. Approximation dans les espaces métriques et theorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 65:181–237, 1983.

L. Birgé. Sur un theorème de minimax et son application aux tests. *Probab. Math. Statist.*, 3: 259–282, 1984a.

L. Birgé. Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Sect. B*, 20:201–223, 1984b.

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Institut Henri Poincare, Probab. et Statist.*, 42:273–325, 2006.

L. Birgé. Model selection for Poisson Processes. *Asymptotic: Particles, processes and inverse problems, Festschrift for Piet Groeneboom (E. Cator, G. Jongbloed, C. Kraaikamp, R. Lopuhaä and J. Wellner, eds)*, IMS Lecture Notes – Monograph Series 55:32–64, 2007.

L. Birgé. Robust tests for Model Selection. *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner (M.Banerjee, F. Bunea, J. Huang, V. Koltchinskii and M. Mathuis,eds)*, IMS Collections – Volume 9:47–64, 2013a.

L. Birgé. Model Selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields*, pages 1–42, 2013b.

L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97:113–150, 1993.

L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Statist.*, 10:24–45, 2006.

G. Blanchard and P. Massart. Discussion: Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2664–2671, 2006.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York, 2001.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Ann. Statist.*, 28:681–712, 2000.

S. C. Larson. The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.*, 22:45–55, 1931.

L. M. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–55, 1973.

G. Lugosi and A.B. Nobel. Adaptive model selection using empirical complexities. *Ann. Statist.*, 27(6):1830–1864, 1999.

T. Mildenberger and H. Weinert. The benchden package: Benchmark densities for nonparametric density estimation. *Journal of Statistical Software*, 46(14):1–14, 2012.

A. Nemirovski. *Topics in Non-Parametric Statistics.* Lecture on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998 (P. Bernard, ed.) Lecture Notes in Math. Springer, Berlin, 2000.

P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

Y. Rozenholc, T. Mildenberger, and U. Gather. Combining regular and irregular histograms by penalized likelihood. *Computational Statistics and Data Analysis*, 54(12):3313–3323, 2010.

M. Sart. Estimation of the transition density of a Markov chain. *Ann. Inst. Henri Poincaré Probab. et Statis. (to appear)*, 2012.

M. Sart. Robust estimation on a parametric model with tests. `http://arxiv.org/abs/1308.2927v2`, 2013.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, 1992.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B.*, 53:683–690, 1991.

B. W. Silverman. *Density Estimation.* London: Chapman and Hall, 1986.

M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.