

Density Estimation with Imprecise Kernels: Application to Classification

Guillaume Dendievel, Sébastien Destercke, Pierre Wachalski

► **To cite this version:**

Guillaume Dendievel, Sébastien Destercke, Pierre Wachalski. Density Estimation with Imprecise Kernels: Application to Classification. *Soft Methods in Probability and Statistics (SMPS 2019)*, Sep 2018, Compiègne, France. pp.59-67, 10.1007/978-3-319-97547-4_9 . hal-02079843

HAL Id: hal-02079843

<https://hal.archives-ouvertes.fr/hal-02079843>

Submitted on 26 Mar 2019

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

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

Density estimation with imprecise kernels: application to classification

Guillaume Dendievel, Sebastien Destercke, and Pierre Wachalski

Technologic University of Compiègne, CNRS, UMR 7253 - Heudiasyc, Centre de Recherche de Royallieu, Compiègne, France

Openvalue, 58 Avenue Charles de Gaulle, 92200 Neuilly-sur-Seine

sebastien.destercke@utc.fr

{guillaume.dendievel, pierre.wachalski}@gmail.com

Abstract. In this paper, we explore the problem of estimating lower and upper densities from imprecisely defined families of parametric kernels. Such estimations allow to rely on a single bandwidth value, and we show that it provides good results on classification tasks when extending the naive Bayesian classifier

Keywords: density estimation, kernel, imprecision, classification.

1 Introduction

Estimating probability densities is a key task in many problems: signal filtering, classification, risk and uncertainty analysis, ... When the densities are known to belong to some parametric family, one can use efficient estimators of the parameters, yet when it is not the case, non-parametric methods such as kernel-based estimation must be used [3].

To perform this estimation, we need a kernel shape and a kernel bandwidth. It is commonly recognized that the resulting estimation will often not be sensitive to the kernel shape, but can be highly sensitive to the choice of the bandwidth [4]. A too low bandwidth will capture very local variations, while a too high bandwidth will provide a too smooth density. This is particularly true when the number of samples is low.

Except for specific cases, finding an optimal bandwidth for a finite sample of values is not doable. It may therefore be interesting to let the bandwidth vary in a pre-determined interval, obtaining upper and lower values of the estimated density. Such bounds can then be used in robustness analysis, ensuring that the inferences do not depend too much on the bandwidth value. For example, Destercke and Strauss [2] consider so-called cloudy kernels (pairs of possibility distributions) to perform signal filtering.

In this paper, we study how lower and upper density bounds can be obtained from imprecise bandwidth defined for a given family of kernels. The approach is described Section 2, and Section 3 deals with the practical problem of computing those bounds for specific kernels. We apply in Section 4 our findings to the naive Bayesian classifier, obtaining a non-parametric credal naive classifier that can deal with continuous data.

2 From precise to imprecise kernel density estimation

A common problem when observing a sample x_1, \dots, x_N of a random variable $X \in \mathbb{R}$ is to estimate its density function $f : \mathbb{R}^+ \rightarrow \mathbb{R}$. When f can be assumed to follow some

parametric model, estimating it comes down to estimate its parameters. There are cases, though, where f cannot be satisfactorily approximated by a simple parametric family.

In such cases, non-parametric kernel density estimation can be used to estimate density values without making a priori assumptions about its shape. Given a scaled kernel K with bandwidth h^1 , the estimated density at point x is

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x-x^{(i)}}{h}\right).$$

It is well known that the exact shape of K has in general a small influence on the end-result, while different choices of bandwidth h may lead to very different results.

This is why it could be interesting to develop tools that allow one to consider sets of bandwidth at once, thus providing a way to perform a global sensitivity analysis. The basic idea is the following: given an interval $H = [\underline{h}, \bar{h}]$ of possible values, how can we determine, for a point x , the upper and lower bounds of the corresponding density, i.e.,

$$\underline{\hat{f}}_H(x) = \inf_{h \in H} \hat{f}_h(x) \text{ and } \bar{\hat{f}}_H(x) = \sup_{h \in H} \hat{f}_h(x). \quad (1)$$

Finding the solutions to these equations is non-trivial in general, as the functions to optimize are usually non-convex in h . In practice, we should try to find kernels for which efficient algorithmic solutions exist. This is what we do next, for the cases of triangular and Epanechnikov kernels, recalled in Table 1.

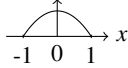
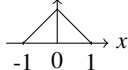
Name	K	Shape
Epanechnikov	$K(x) = \frac{3}{4}(1-x^2)I_{ x \leq 1}$	
Triangular	$K(x) = (1- x)I_{ x \leq 1}$	

Table 1. Triangular and Epanechnikov kernels

3 Particular tractable cases

In this section, we study how solutions for Equations (1) can be found for some specific kernels, namely the Triangular and Epanechnikov ones, and for a specific value x . Since we focus on a particular value x , we will consider the re-indexing $x_{(1)}, \dots, x_{(N)}$ of the sample in an increasing sequence with respect to their distance of x , that is such that

¹ A kernel is here a symmetric, non-negative function with $\int_{\mathbb{R}} K(y)dy = 1$ and mean 0.

$|x_{(i)} - x| \leq |x_{(i+1)} - x|$ for any $i = 1, \dots, N-1$. We will also use the notations $D_{(i)} = |x_{(i)} - x|$ and $\mathbb{E}(D_{(i)}) = \left(\sum_{j=1}^i D_{(j)}\right)/i$ to simplify further proofs.

3.1 Triangular Kernel

To compute the bounds given by Equations (1) over the interval $[\underline{h}, \bar{h}]$, we need to identify points that will reach this global optimum, and to do this we show that local optima are easy to obtain within each interval $[D_{(i)}, D_{(i+1)}]$. This is shown below.

Proposition 1. For a triangular kernel K and values of $h \in [D_{(i)}, D_{(i+1)}]$, we have:

- if $D_{(i+1)} < 2\mathbb{E}(D_{(i)})$ or $D_{(i)} > 2\mathbb{E}(D_{(i)})$,

$$\max_{h \in [D_{(i)}, D_{(i+1)}]} \hat{f}_h(x) = \max(\hat{f}_{D_{(i+1)}}(x), \hat{f}_{D_{(i)}}(x))$$

- if $D_{(i)} \leq 2\mathbb{E}(D_{(i)}) \leq D_{(i+1)}$, $\hat{f}_h(x)$ has one maximal value in h and

$$\max_{h \in [D_{(i)}, D_{(i+1)}]} \hat{f}_h(x) = \hat{f}_{2\mathbb{E}(D_{(i)})}(x)$$

and the minimal value is given by

$$\min_{h \in [D_{(i)}, D_{(i+1)}]} \hat{f}_h(x) = \min(\hat{f}_{D_{(i)}}(x), \hat{f}_{D_{(i+1)}}(x))$$

Proof (sketch). If $h \in [D_{(i)}, D_{(i+1)}]$, we can write

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{j=1}^i K\left(\frac{x-x_{(j)}}{h}\right).$$

as $K(x-x_{(j)}/h)$ will be null for any $j \geq i+1$, and non-negative for any $j < i+1$. The derivative in h is $\partial \hat{f}_h(x)/\partial h = -i/Nh^2 + 2^* \sum_{j=1}^i D_{(j)}/N * h^3$. The following table shows the sign variation of this function, which is sufficient to obtain the proposition

$$\frac{h}{\partial \hat{f}_h(x)/\partial h} \begin{array}{|c|} \hline < 2\mathbb{E}(D_{(i)}) \\ \hline & = 2\mathbb{E}(D_{(i)}) \\ \hline > 2\mathbb{E}(D_{(i)}) \\ \hline \end{array} \begin{array}{|c|} \hline < 0 \\ \hline = 0 \\ \hline > 0 \\ \hline \end{array}$$

Next we show that $\hat{f}_h(x)$ is continuous in h , hence that going from $[\underline{h}, \bar{h}]$ to $[\underline{h} - \varepsilon, \bar{h} + \varepsilon]$ will not induce "jumps" in our results.

Proposition 2. $\hat{f}_h(x) = \frac{1}{Nh} \sum_{j=1}^i K\left(\frac{x-x_{(j)}}{h}\right)$ is piecewise continuous

Proof. Let us show the continuity in $D_{(i+1)}$. Consider first $h \in [D_{(i)}, D_{(i+1)}[$, we have

$$\lim_{\substack{h \rightarrow D_{(i+1)} \\ h < D_{(i+1)}}} \hat{f}_{h,i}(x) = \frac{i}{Nh} - \frac{\left(\sum_{j=1}^i D_{(j)}\right)}{N * h^2} = \frac{i}{ND_{(i+1)}} - \frac{\sum_{j=1}^i D_{(j)}}{N * D_{(i+1)}^2}$$

with the last equality being obtained by taking $h = D_{(i+1)}$. Now consider $h \in [D_{(i+1)}; D_{(i+2)}[$

$$\lim_{\substack{h \rightarrow D_{(i+1)} \\ h > D_{(i+1)}}} \hat{f}_{h,i+1}(x) = \frac{i+1}{Nh} - \frac{\sum_{j=1}^{i+1} D_{(j)}}{N * h^2} = \frac{i+1}{ND_{(i+1)}} - \frac{\sum_{j=1}^{i+1} D_{(j)}}{N * D_{(i+1)}^2}$$

and we have that these two values are equal.

Algorithm 1: Find $\underline{\hat{f}}(x), \overline{\hat{f}}(x)$ in $H = [\underline{h}, \overline{h}]$

Input: $x_{(i)}, D_{(i)}$ sorted in ascending order, $H, x, i = 1$

Output: Bounds $\underline{\hat{f}}(x), \overline{\hat{f}}(x)$

while $i \neq N$ **do**

if $[a, b] = H \cap [D_{(i)}, D_{(i+1)}] \neq \emptyset$ **then**

$\underline{\hat{f}}(x) \leftarrow \min(\underline{\hat{f}}(x), \hat{f}(a), \hat{f}(b))$;

$\overline{\hat{f}}(x) \leftarrow \max(\overline{\hat{f}}(x), \hat{f}(a), \hat{f}(b))$

if $2 * \mathbb{E}(D_{(i)}) \in [a, b]$ **then**

$\overline{\hat{f}}(x) \leftarrow \max(\overline{\hat{f}}(x), \hat{f}(2 * \mathbb{E}(D_{(i)})))$

$i \leftarrow i + 1$

3.2 Epanechnikov kernel

Results similar to the previous case can be given for the Epanechnikov case. Due to page limit restriction, we only provide the main results.

Proposition 3. For an Epanechnikov kernel K and values of $h \in [D_{(i)}, D_{(i+1)}]$, we have:

– if $D_{(i+1)} < (3 * \mathbb{E}(D_{(i)}^2))^{\frac{1}{2}}$ or $D_{(i)} > (3 * \mathbb{E}(D_{(i)}^2))^{\frac{1}{2}}$,

$$\max_{h \in [D_{(i)}, D_{(i+1)}]} \hat{f}_h(x) = \max(\hat{f}_{D_{(i+1)}}(x), \hat{f}_{D_{(i)}}(x))$$

– if $D_{(i)} \leq (3 * \mathbb{E}(D_{(i)}^2))^{\frac{1}{2}} \leq D_{(i+1)}$, $\hat{f}_h(x)$ has one maximal value in h and

$$\max_{h \in [D_{(i)}, D_{(i+1)}]} \hat{f}_h(x) = \hat{f}_{(3 * \mathbb{E}(D_{(i)}^2))^{\frac{1}{2}}}(x)$$

and the minimal value is given by

$$\min_{h \in [D_{(i)}, D_{(i+1)}]} \hat{f}_h(x) = \min(\hat{f}_{D_{(i)}}(x), \hat{f}_{D_{(i+1)}}(x))$$

Proposition 4. $\hat{f}_h(x) = \frac{1}{Nh} \sum_{j=1}^i K\left(\frac{x-x_{(j)}}{h}\right)$ is piecewise continuous

3.3 Illustrative experiments

To illustrate the results provided by imprecise kernels, we perform experiments where we modify the number of observations and the size of the interval H . To do that, we generated points from a bimodal mixture of Gaussian $X \sim 0.6\mathcal{N}(-1, 1) + 0.4\mathcal{N}(5, 1)$. To perform our experiments, we started from a reference bandwidth that corresponds to the optimal one for the normal case $h^* = (1.06 \cdot \hat{\sigma} \cdot N)^{-\frac{1}{5}}$

Increasing sample Figure 1 shows the results of the following experiments: we generate 1500 points and pick a fixed $H = [h^* - 0.2 * h^*, h^* + 0.2 * h^*]$. We then randomly shuffle the samples, and take each time the first n samples to achieve density estimation. From the picture, we can easily see that the more points we get, the less imprecise we are.

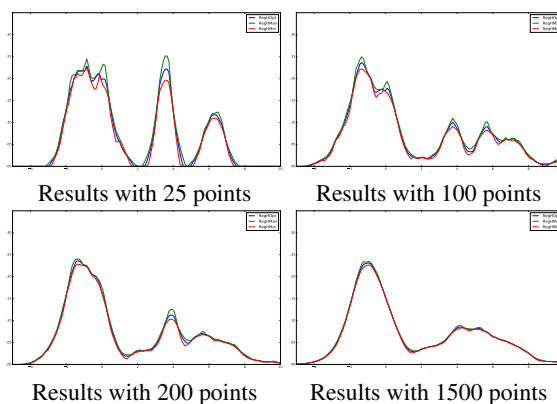


Fig. 1. Imprecise estimation with varying sample sizes

Increasing bandwidth we now fix the number of samples to 75, and make estimations using intervals going from $H = [h^* - 0.05 * h^*, h^* + 0.05 * h^*]$ to $H = [h^* - 0.90h^*, h^* + 0.90h^*]$. The results are shown in Figure 2, and again we can easily see the increase of imprecision, as well as the increasing noise as the lower bound of H gets close to 0.

4 Application to Naive credal classification

As an illustration of our approach, we will apply it to the popular naive Bayes classifier, and will turn it into a credal naive classifier [5] on continuous variables, whereas most of the previous versions only accept discrete variables [1]

4.1 Naive credal classification: brief reminder

The Naive Bayes model is a very popular classification model that considers inputs from a multivariate space $X = X^1 \times \dots \times X^p$ and outputs in the form of a discrete class

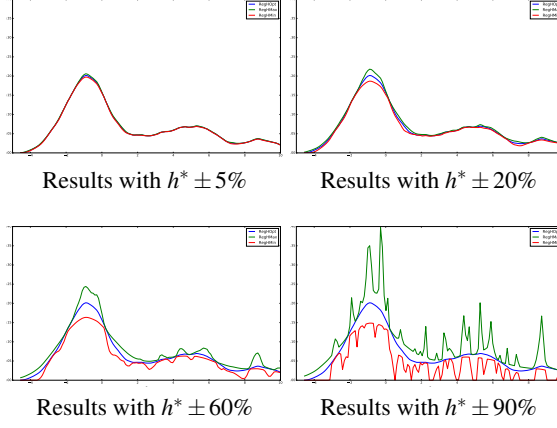


Fig. 2. Imprecise estimation with varying bandwidths

Y . The Naive Bayes model proposes to estimate the posterior probability $p(y|x)$ of class y given observation $x = (x^1, \dots, x^p)$ by assuming that input variables are independent, given the class. That is, $p(y|x)$ can be rewritten

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_{y \in Y} p(x|y)p(y)} = \frac{p(y) \prod_{i=1}^p p(x^i|y)}{\sum_{y \in Y} p(y) \prod_{i=1}^p p(x^i|y)}. \quad (2)$$

Given two classes y and y' , checking that $p(y|x) \geq p(y'|x)$ comes down to check that

$$\frac{p(y|x)}{p(y'|x)} = \frac{p(y) \prod_{i=1}^p p(x^i|y)}{p(y') \prod_{i=1}^p p(x^i|y')} \quad (3)$$

is higher than one. In this case, we say that y is preferred to y' , noted $y \succ y'$. When probabilities become imprecise, this comes down to test whether the infimum value of Eq. (3) is higher than 1. This becomes, when class prior probabilities are assumed precise

$$\inf_{p(x^i|y) \in [\underline{p}(x^i|y), \bar{p}(x^i|y)]} \frac{p(y|x)}{p(y'|x)} = \frac{p(y) \prod_{i=1}^p \underline{p}(x^i|y)}{p(y') \prod_{i=1}^p \bar{p}(x^i|y')}$$

This may result in a partial order, in which case our prediction consists in taking all the maximal elements of the resulting order.

4.2 Experimental protocol

In our training, for any pair feature/class, we consider symmetric intervals around the estimation $h^* = (1.06 \cdot \hat{\sigma} \cdot N)^{-\frac{1}{5}}$. To avoid zero-probability of $\underline{p}(x^i|y)$ or of $p(x^i|y)^2$, we take $\underline{p}(x^i|y) = \max(0, 0.1 * p(x^i|y))$, and we set $p(x^i|y) = 10^{-3}$ if it is null.

² In some sense, to regularize our model

The protocol adopted is the following: for each data set, we decide how imprecise our kernels will be by setting ε and taking $H = [h^* - \varepsilon h^*, h^* + \varepsilon h^*]$, and a split ratio of training/test. We then perform ten repetitions for each couple $(\varepsilon, \text{ratio})$. The selected data sets are summarised in Table 2

#	a	b	c	d	e	f	g	h	i	j
Names	Breasts	Iris	Wine	Automobile	Seed	Glass	Forest	Dermatology	Diabete	Segment
Instances	106	150	178	205	210	214	325	366	769	2310
Features	10	4	13	26	7	9	27	34	8	19
Labels	6	3	3	7	3	7	4	6	2	7

Table 2. Selected data sets

As we deal with imprecise results, we have chosen to use well-motivated utility-discounted accuracies u_{65} and u_{80} [6]

$$u_{65} = \frac{1}{T} \sum_{t=1}^T -0.6a_t^2 + 1.6a_t \text{ and } u_{80} = \frac{1}{T} \sum_{t=1}^T -1.2a_t^2 + 2.2a_t$$

where $a_t = \mathbb{1}_{y_t \in Y_t} / |Y_t|$, Y_t being the predicted set. In practice, u_{80} rewards more imprecise predictions than u_{65} , hence should be more favourable to imprecise methods in comparisons. Table 3 summarizes the obtained results, showing that our method is clearly superior in terms of u_{80} for most configurations, and quite competitive in terms of u_{60} . Figure 3 shows the accuracy on those instances who were imprecisely predicted by our approach, on which we can notice that the precise model accuracy drops (e.g. for data set c , accuracy on those instances is less than 10%, while it is more than 60% in average), while our approach is almost systematically right.

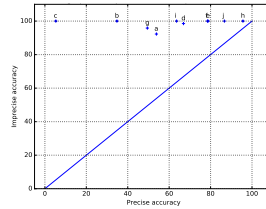


Fig. 3. Accuracy of imprecise predictions for $\varepsilon=0.2$ and split ratio = 0.3

5 Conclusion and perspectives

In this paper, we have introduced the idea of using imprecise parametric kernels in order to estimate density bounds. We have shown that for some kernels, efficient algorithmic

#	stats	SR = 50%		$\varepsilon = 20\%$		#	SR = 50%		$\varepsilon = 20\%$	
		$\varepsilon = 10\%$	$\varepsilon = 40\%$	SR=30%	SR=75%		$\varepsilon = 10\%$	$\varepsilon = 40\%$	SR=30%	SR=75%
a	precise	56.4	56.4	56.9	57.0	f	80.7	80.7	80.1	82.6
	u_{65}	61.4	49.0	55.2	55.9		52.7	39.6	45.9	46.3
	u_{80}	71.1	61.1	67.5	67.6		64.0	50.6	57.1	57.4
b	precise	97.1	97.1	96.3	95.8	g	87.4	87.4	87.2	87.3
	u_{65}	97.3	96.6	96.9	96.1		88.5	88.9	88.2	88.4
	u_{80}	97.5	97.1	97.3	96.3		89.1	90.9	89.4	89.4
c	precise	62.7	62.7	61.3	62.9	h	98.9	98.9	99.1	99.0
	u_{65}	86.2	82.0	84.9	85.9		96.9	78.3	92.2	92.8
	u_{80}	92.0	88.5	91.0	91.6		98.2	84.7	95.4	95.8
d	precise	80.0	80.0	79.6	79.8	i	79.2	79.2	79.7	79.7
	u_{65}	82.8	61.0	74.9	74.0		79.7	79.6	80.0	79.5
	u_{80}	86.2	71.6	81.5	80.9		81.5	85.7	83.5	83.1
e	precise	93.1	93.1	93.6	94.0	j	89.3	89.3	89.2	89.3
	u_{65}	92.4	91.6	92.2	92.2		61.7	50.1	56.7	56.3
	u_{80}	93.1	93.4	93.7	93.8		71.8	60.9	66.9	66.6

Table 3. Experimental results

procedure can be developed, and that good results can be obtained in classification problems (at least for the naive credal classifier). Since density estimation plays an important role in many applications, we expect our approach to be of interest to many people.

Possible extensions to our paper include the study of more generic forms of kernels (e.g., polynomials), as well as the extension of the current study to multi-dimensional kernels and density estimation.

References

1. G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. *Journal of Machine Learning Research*, 9(Apr):581–621, 2008.
2. S. Destercke and O. Strauss. Filtering with clouds. *Soft Computing*, 16(5):821–831, 2012.
3. B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
4. M. P. Wand and M. C. Jones. *Kernel smoothing*. Crc Press, 1994.
5. M. Zaffalon. The naive credal classifier. *J. Probabilistic Planning and Inference*, 105:105–122, 2002.
6. M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.