A note about the critical bandwidth for a kernel density estimator with the uniform kernel

Raphaël Coudret, Gilles Durrieu, Jerome Saracco

A NOTE ABOUT THE CRITICAL BANDWIDTH FOR A KERNEL DENSITY ESTIMATOR WITH THE UNIFORM KERNEL

RAPHAËL COUDRET, GILLES DURRIEU, AND JÉRÔME SARACCO

Abstract. Among available bandwidths for kernel density estimators, the critical bandwidth is a data-driven one, which satisfies a constraint on the number of modes of the estimated density. When using a random bandwidth, it is of particular interest to show that it goes toward 0 in probability when the sample size goes to infinity. Such a property is important to prove satisfying asymptotic results about the corresponding kernel density estimator. It is shown here that this property is not true for the uniform kernel.

1. Introduction

Let consider a sample $\mathbf{X} = (X_1, \dots X_n)$ made of independent and identically distributed random variables generated from the density f. To estimate f from this sample of size n, Parzen (1962) and Rosenblatt (1956) introduced the kernel density estimator $\hat{f}_{K,h}$, defined for every real t by

$$\hat{f}_{K,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right),$$

where K is called the kernel and is most of time a density function while h is a positive real parameter that controls the smoothness of $\hat{f}_{K,h}$. We also refer the interested reader to Scott (1992) and Silverman (1986).

Let $N(\hat{f}_{K,h})$ be the number of modes of $\hat{f}_{K,h}$. To decide how smooth $\hat{f}_{K,h}$ should be, an approach is to set $N(\hat{f}_{K,h})$. To do so, one can use the critical bandwidth h_{crit} introduced by Silverman (1981) for h. It can be defined by

$$h_{crit,k} = \min_{N(\hat{f}_{K,h}) \leq k} h, \tag{1}$$

for any $k \in \mathbb{N}^*$.

When K is the Gaussian kernel, that is $\forall t \in \mathbb{R}, K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$, the bandwidth $h_{crit,k}$ has interesting properties. It can easily be computed and, provided that $k \geq N(f)$, allows $\hat{f}_{K,h}$ to exhibit, in probability, a pointwise convergence, an L_1-convergence and a uniform convergence toward f. Details can be found in Futschik and Isogai (2006), Mammen et al. (1991), Devroye and Wagner (1980), Devroye (1987) and Coudret et al. (2012).

These properties are proven using the key point that $h_{crit,k}$ converges in probability toward 0 when K is the Gaussian kernel. To extend them to other kernels, the asymptotic behavior of $h_{crit,k}$ should thus be studied. In this Note, we focus on the case when K is the uniform kernel defined as $\forall t \in \mathbb{R}, K(t) = \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(t)$, where $\mathbf{1}$ is the indicator function. We will prove in the following section that for this kernel, $h_{crit,k}$ does not converge toward 0, after giving some properties about $\hat{f}_{K,h}$ and $N(\hat{f}_{K,h})$.

2. Properties for the uniform kernel

Let us first introduce some additional notations. Let $\mathbf{A}_h = \cup_{i=1}^{n} \{X_i - \frac{h}{2}\} = \left\{a_{h,(i)}\right\}_{i \in \{1,\dots,\text{card}(\mathbf{A}_h)\}}$ and $\mathbf{B}_h = \cup_{i=1}^{n} \{X_i + \frac{h}{2}\} = \left\{b_{h,(i)}\right\}_{i \in \{1,\dots,\text{card}(\mathbf{B}_h)\}}$. In order to deduce the value of $N(\hat{f}_{K,h})$, we only need to investigate how the points in $\mathbf{A}_h \cup \mathbf{B}_h$ are ordered because of Proposition 1 below. We write

$w = \text{card}(\mathbf{A}_h \cup \mathbf{B}_h) \leq 2n$. We set $\{c_{h,(i)}\}_{i\in\{1,\ldots,w\}}$ as the ordered points in $\mathbf{A}_h \cup \mathbf{B}_h$. Let us also write $b_{h,(0)} = c_{h,(0)} = -\infty$ and $a_{h,(\text{card}(\mathbf{A}_h)+1)} = c_{h,(w+1)} = +\infty$.

Proposition 1. *Let $(X_1, \ldots X_n)$ be a vector of independent random variables generated from f. Let $\hat{f}_{K,h}$ be the kernel estimator of f for the uniform kernel K. Then, $\forall h > 0$, $\forall i \in \{0, 1, \ldots, w\}$, the function $\hat{f}_{K,h}$ is constant on $]c_{h,(i)}, c_{h,(i+1)}[$.*

The proof is given in Appendix A.1.

Remark 1. The reasoning in the proof of Proposition 1 can also be used to obtain the following results:

- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \in \mathbf{A}_h \Leftrightarrow \exists \zeta \in \mathbb{N}^*,\ \forall u \in]c_{h,(i-1)}, c_{h,(i)}[,\ f(u) = f(c_{h,(i)}) - \frac{\zeta}{nh}$,
- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \in \mathbf{B}_h \Leftrightarrow \exists \zeta \in \mathbb{N}^*,\ \forall u \in]c_{h,(i)}, c_{h,(i+1)}[,\ f(u) = f(c_{h,(i)}) - \frac{\zeta}{nh}$,
- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \notin \mathbf{A}_h \Leftrightarrow \forall u \in]c_{h,(i-1)}, c_{h,(i)}[,\ f(u) = f(c_{h,(i)})$,
- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \notin \mathbf{B}_h \Leftrightarrow \forall u \in]c_{h,(i)}, c_{h,(i+1)}[,\ f(u) = f(c_{h,(i)})$.

Proposition 1 and Remark 1 are illustrated in Figure 1 where the sample $(3, 6.5, 6, 1.5, 5)$ of size $n = 5$ is used to compute the function $\hat{f}_{K,2}$. We have here $w = 9$. Note that $N(\hat{f}_{K,2}) = 3$ and that a mode of $\hat{f}_{K,2}$ is actually a single point. This mode vanishes if we use a bandwidth slightly less than 2 so that there is a jump in the estimator $h \mapsto N(\hat{f}_{K,h})$ located at $h = 2$. Figure 1 gives a lead to find $N(\hat{f}_{K,h})$ from $\mathbf{A}_h$ and $\mathbf{B}_h$, as explained in the following property.

Proposition 2. *Let $(X_1, \ldots X_n)$ be a vector of independent random variables generated from f. Let $\hat{f}_{K,h}$ be the kernel estimator of f for the uniform kernel K. The number of modes $N(\hat{f}_{K,h})$ of $\hat{f}_{K,h}$ is such that*

$$N(\hat{f}_{K,h}) = \text{card}\left(\left\{(i,j)\ :\ a_{h,(i)} \in]b_{h,(j-1)}, b_{h,(j)}] \text{ and } b_{h,(j)} \in [a_{h,(i)}, a_{h,(i+1)}[\right\}\right),$$

where $i \in \{1, \ldots, \text{card}(\mathbf{A}_h)\}$ and $j \in \{1, \ldots, \text{card}(\mathbf{B}_h)\}$.

The proof is given in Appendix A.2.

This characterization of $N(\hat{f}_{K,h})$ based on the sets $\mathbf{A}_h$ and $\mathbf{B}_h$, together with an argument about totally positive matrices, allows us to show the following theorem:

Theorem 1. *For any probability density function f of $\mathbf{X}$, let $\hat{f}_{K,h_{crit,k}}$ be the estimator of f when K is the uniform kernel with $h_{crit,k}$ given in (1). Then we have $h_{crit,k}$ increasing with n, for all $k \in \mathbb{N}$.*

The proof is given in Appendix A.3.

To illustrate this theorem, we consider again the sample $(3, 6.5, 6, 1.5, 5)$ and every subsamples made of the first n elements of $(3, 6.5, 6, 1.5, 5)$ for $n \in \{1, \ldots, 5\}$. For each subsample, we compute the function $h \mapsto N(\hat{f}_{K,h})$, where K is the uniform kernel. In Figure 2, we display ranges of values of h and n for which $N(\hat{f}_{K,h})$ is equal to a given number. The increase of $N(\hat{f}_{K,h})$ with n can be observed for small values of h. In the general case, this feature implies that for any $k \in \mathbb{N}$, $h_{crit,k}$ also increases with n.

Theorem 1 means that we can not have that $h_{crit,k}$ goes toward 0 in probability for the uniform kernel. Thus, the proof of the pointwise convergence of $\hat{f}_{K,h_{crit,k}}$ toward f, as given in Futschik and Isogai (2006), does not hold for this kernel. It is also impossible to use the work of Devroye and Wagner (1980) and Devroye (1987) to show the corresponding L_1-convergence and uniform convergence. In addition, the simulation study from Coudret et al. (2012) shows that $\hat{f}_{K,h_{crit,k}}$ is not an accurate estimator of f, when K is the uniform kernel. For these reasons, we recommend not to use $\hat{f}_{K,h_{crit,k}}$ with this kernel in practice.

Appendix A. Proofs

A.1. **Proof of Proposition 1.** Let $(u, v) \in]c_{h,(i)}, c_{h,(i+1)}[\times]u, c_{h,(i+1)}[$ and $\{X_{(i)}\}_{i\in\{1,\ldots n\}}$ be the ordered sequence of the elements of $\mathbf{X}$. We will show that $\hat{f}_{K,h}(u)$ is neither greater nor lesser

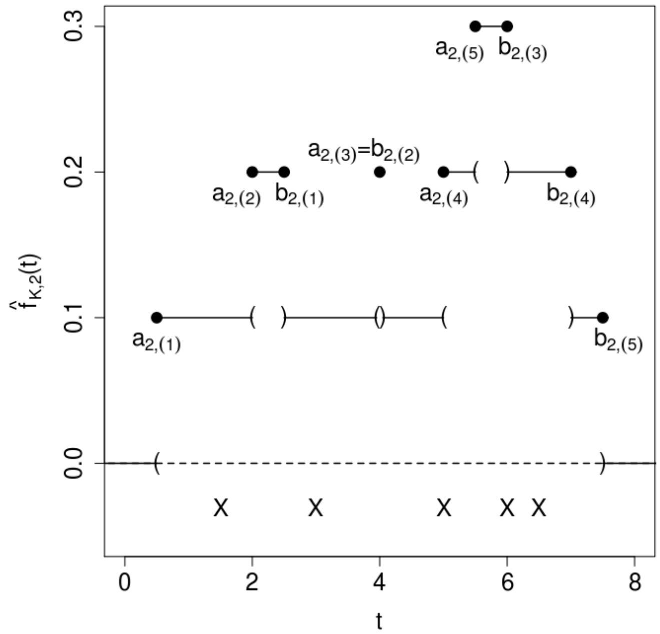

FIGURE 1. Estimation of a density with the estimator $\hat{f}_{K,2}$ and the uniform kernel K (solid line). Solid circles are special points of this estimated density while brackets indicate a difference between the limit and the value of $\hat{f}_{K,2}$. Crosses represent the underlying sample.

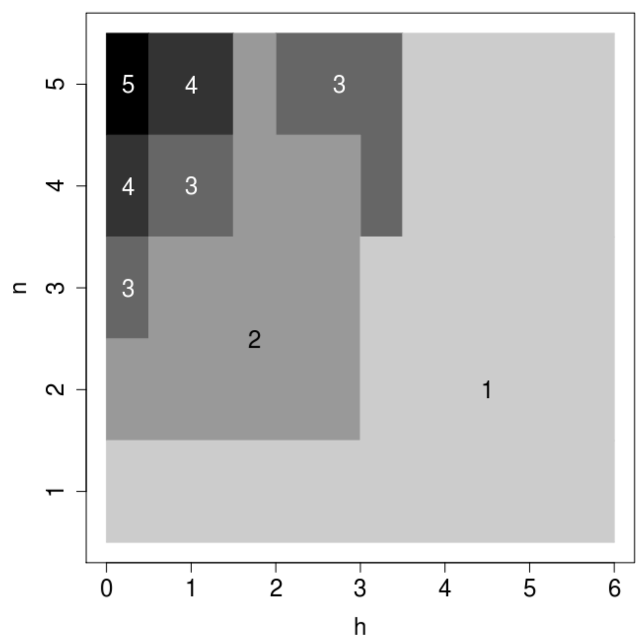

FIGURE 2. Evolution of $N(\hat{f}_{K,h})$ with respect to the bandwidth h and to the sample size n, for the uniform kernel K. The considered samples are the first n values of $(3, 6.5, 6, 1.5, 5)$.

than $\hat{f}_{K,h}(v)$ with a proof by contradiction. Note that for the uniform kernel we have $\hat{f}_{K,h}(u) = \frac{1}{nh}\mathrm{card}\left(\left\{X_k \in [u-\frac{h}{2}, u+\frac{h}{2}]\right\}\right)$.

If $\hat{f}_{K,h}(u) > \hat{f}_{K,h}(v)$, this implies that there exists at least one $k \in \{1,\dots,n\}$, for which we have $X_{(k)} \in [u-\frac{h}{2}, v-\frac{h}{2}[$, which means that there exists $k' \in \{1,\dots,w\}$ which satisfies $c_{h,(k')} = b_{h,(k)} \in [u,v[$. Because $[u,v[\subset]c_{h,(i)}, c_{h,(i+1)}[$, $c_{h,(k')} \in]c_{h,(i)}, c_{h,(i+1)}[$, which is impossible.

Conversely, $\hat{f}_{K,h}(v) > \hat{f}_{K,h}(u)$ implies that there exists $X_{(k)} \in]u+\frac{h}{2}, v+\frac{h}{2}]$. Then there exists $k' \in \{1,\dots,w\}$ such that $c_{h,(k')} = a_{h,(k)} \in]u,v] \subset]c_{h,(i)}, c_{h,(i+1)}[$ and it is impossible.

A.2. **Proof of Proposition 2.** We will show the equivalence between the presence of a mode between $a_{h,(i)}$ and $b_{h,(j)}$ and the inequality $b_{h,(j-1)} < a_{h,(i)} \le b_{h,(j)} < a_{h,(i+1)}$.

At first, we notice that ordered like this, there is no element of $\mathbf{A}_h$ or $\mathbf{B}_h$ that can be between $a_{h,(i)}$ and $b_{h,(j)}$. This is why the last inequality is equivalent to $\exists k \in \{1,\dots,w-1\}$, $a_{h,(i)} = c_{h,(k)}$ and $b_{h,(j)} = c_{h,(k+1)}$, provided that $a_{h,(i)} \neq b_{h,(j)}$.

From Proposition 1, $\hat{f}_{K,h}$ is constant on $]a_{h,(i)}, b_{h,(j)}[$, and thanks to Remark 1, it is equivalent to: $\hat{f}_{K,h}$ is constant on $[a_{h,(i)}, b_{h,(j)}] = [c_{h,(k)}, c_{h,(k+1)}]$. In order for this interval to be a mode, we must prove that there exists $\varepsilon > 0$ for which $\hat{f}_{K,h}$ is increasing on $[c_{h,(k)} - \varepsilon, c_{h,(k)}[$ and decreasing on $]c_{h,(k+1)}, c_{h,(k+1)} + \varepsilon]$, which is also made in Remark 1.

When $a_{h,(i)} = b_{h,(j)} = c_{h,(k)}$, $\hat{f}_{K,h}$ is increasing on $[c_{h,(k)} - \varepsilon, c_{h,(k)}[$ too and decreasing on $]c_{h,(k)}, c_{h,(k)} + \varepsilon]$. The mode is reduced to a single point.

A.3. **Proof of Theorem 1.** First, note that when K is the uniform kernel, for some $h > 0$, we can find $N(\hat{f}_{K,h})$ by counting the number of variations of sign of the following function

$$g'_{h,\varepsilon}(x) = \left\{ \begin{array}{rl} 1 & \text{for } x \in [a_{h,(i)} - \varepsilon, a_{h,(i)}[,\ \forall i \in \{1,\dots,\mathrm{card}(\mathbf{A}_h)\}, \\ -1 & \text{for } x \in [b_{h,(i)}, b_{h,(i)} + \varepsilon[,\ \forall i \in \{1,\dots,\mathrm{card}(\mathbf{B}_h)\}, \\ 0 & \text{elsewhere,} \end{array} \right.$$

where ε is chosen in a way that ensures that $\forall (i,j) \in \{1,\dots,\mathrm{card}(\mathbf{A}_h)\} \times \{1,\dots,\mathrm{card}(\mathbf{B}_h)\}, (a_{h,(i)} - b_{h,(j)}) \in]-\infty,0] \cup]\varepsilon,\infty[$, in order to obtain a unique value of $g'_{h,\varepsilon}(x)$ for each x. The aim of $g'_{h,\varepsilon}$ is to mimic the derivative of $\hat{f}_{K,h}$. It seems easier to use than dirac functions involved in $\hat{f}'_{K,h}$. Besides, one can see that $N(g_{h,\varepsilon}) = N(\hat{f}_{K,h})$, using the fact that Proposition 2 is valid for $g_{h,\varepsilon}$. That is why the number of variations of sign of $g'_{h,\varepsilon}$ is equal to $2N(\hat{f}_{K,h}) - 1$.

Let $\mathbf{C}_{\varepsilon,n} = \left\{c_{h,\varepsilon,(i)}\right\}_{i\in\{1,\dots,w\}}$ be the ordered sequence made of the sets $\left\{a_{h,(i)} - \frac{\varepsilon}{2}\right\}_{i\in\{1,\dots,\mathrm{card}(\mathbf{A}_h)\}}$ and $\left\{b_{h,(i)} + \frac{\varepsilon}{2}\right\}_{i\in\{1,\dots,\mathrm{card}(\mathbf{B}_h)\}}$. Let

$$\begin{array}{rcl} d_{h,\varepsilon,(i)} & = & \mathbf{1}\left(c_{h,\varepsilon,(i)} \in \left\{a_{h,(i)} - \frac{\varepsilon}{2}\right\}_{i\in\{1,\dots,\mathrm{card}(\mathbf{A}_h)\}}\right) \\ & - & \mathbf{1}\left(c_{h,\varepsilon,(i)} \in \left\{b_{h,(i)} + \frac{\varepsilon}{2}\right\}_{i\in\{1,\dots,\mathrm{card}(\mathbf{B}_h)\}}\right), \end{array}$$

and $\mathbf{D}_{\varepsilon,n} = \left\{d_{h,\varepsilon,(i)}\right\}_{i\in\{1,\dots,w\}}$. Every interval where $g'_{h,\varepsilon}(x) \neq 0$ is represented by a $c_{h,\varepsilon,(i)}$, then the number of variations of sign is the same for $g'_{h,\varepsilon}$ and for $\mathbf{D}_{\varepsilon,n}$. We write $v(\mathbf{D}_{\varepsilon,n})$ for the number of variations of sign of $\mathbf{D}_{\varepsilon,n}$ like Schoenberg (1950) did in his article.

Now, we prove that $v(\mathbf{D}_{\varepsilon,n}) \ge v(\mathbf{D}_{\varepsilon,n-1})$, for $n > 1$. This property is satisfied if $\mathbf{D}_{\varepsilon,n-1} = \mathbf{J}\mathbf{D}_{\varepsilon,n}$ where $\mathbf{J}$ is a totally positive matrix, following Schoenberg (1950). To define $\mathbf{J}$, we first focus on the case where the last point in the sample is different from the others. This means that if $\mathbf{\Omega}$ is our sample space, we define $\mathbf{\Omega}_1$ as:

$$\mathbf{\Omega}_1 = \{\omega\ :\ \forall i \in \{1,\dots,n-1\},\ X_i(\omega) \neq X_n(\omega)\}.$$

We remark that, when our sample comes from $\omega \in \mathbf{\Omega}_1$, $\mathbf{D}_{\varepsilon,n-1}$ is constructed by removing two points in $\mathbf{D}_{\varepsilon,n}$. These points correspond to $c_{h,\varepsilon,(\gamma_1)} = X_n - h - \frac{\varepsilon}{2}$ and $c_{h,\varepsilon,(\gamma_2)} = X_n + h + \frac{\varepsilon}{2}$. This is why we have :

$$\mathbf{J} = \begin{pmatrix} & & 0 & \cdots & 0 & \cdots & 0 \\ & \mathbf{I}_{\gamma_1-1} & \vdots & & \vdots & & \vdots \\ & & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & & 0 & \cdots & 0 \\ \vdots & & \vdots & \mathbf{I}_{\gamma_2-\gamma_1-1} & \vdots & & \vdots \\ 0 & \cdots & 0 & & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & & \\ \vdots & & \vdots & & \vdots & \mathbf{I}_{w-\gamma_2} & \\ 0 & \cdots & 0 & \cdots & 0 & & \end{pmatrix}$$

where $\mathbf{I}_\gamma$ is the $\gamma \times \gamma$ identity matrix. It is straightforward to show that $\mathbf{J}$ is a totally positive matrix since every minor of $\mathbf{J}$ is not negative.

If $\omega \notin \mathbf{\Omega}_1$, then $\mathbf{D}_{\varepsilon,n} = \mathbf{D}_{\varepsilon,n-1}$, because $\mathbf{A}_h$ and $\mathbf{B}_h$ stay the same if we build them with $(X_1, \ldots, X_n)$ or with $(X_1, \ldots, X_{n-1})$. Then $\mathbf{J} = \mathbf{I}_w$ and is totally positive.

To conclude, we write $\tilde{N}_{K,h} \ : \ n \mapsto \tilde{N}_{K,h}(n) = N(\hat{f}_{K,k})$. Recall that $\tilde{N}_{K,h}(n) = \frac{v(g'_{h,\varepsilon})+1}{2} = \frac{v(\mathbf{D}_{\varepsilon,n})+1}{2}$. Because $n \mapsto v(\mathbf{D}_{\varepsilon,n})$ is increasing, $\tilde{N}_{K,h}$ is also an increasing function. Let $h_{crit,k,n}$ be the critical bandwidth defined in (1) for a sample of size n, then we have:

$$\forall h < h_{crit,k,n}, \quad \tilde{N}_{K,h}(n) > k.$$

Because $\tilde{N}_{K,h}$ increases with n, it follows that,

$$\forall \eta \in \mathbb{N},\ \forall h < h_{crit,k,n}, \quad \tilde{N}_{K,h}(n+\eta) > k.$$

Thus, with the definition of $h_{crit,k}$,

$$\forall \eta \in \mathbb{N}, \quad h_{crit,k,n+\eta} \geq h_{crit,k,n},$$

and the proof is complete.

References

Coudret, R., Durrieu, G., and Saracco, J. (2012). Comparison of kernel density estimators with assumptions on number of modes. Submitted paper.

Devroye, L. P. (1987). *A Course in Density Estimation*. Birkäuser, Boston.

Devroye, L. P. and Wagner, P. J. (1980). The strong uniform consistency of kernel density estimates. In P. R. Krishnaiah (ed.), *Multivariate Analysis V: Proceedings of the fifth International Symposium on Multivariate Analysis, Volume 5*, North-Holland Pub. Co., 59–77.

Futschik, A. and Isogai, E. (2006). On the consistency of kernel density estimates under modality constraints. *Statist. Probab. Lett.* **76** 431–437.

Mammen, E., Marron, J. S., and Fisher, N. J. (1991). Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Related Fields* **91** 115–132.

Parzen, E. (1962). On estimation of probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **42** 43–47.

Schoenberg, I. J. (1950). On pólya frequency functions II: Variation diminishing integral operators of the convolution type. *Acta. Sci. Math. (Szeged)* **12** 97–106.

Scott, D. W. (1992). *Multivariate Density Estimation : Theory, Practice and Visualization*. John Wiley & Sons.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99.

— (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Raphaël Coudret : Inria cqfd & Imb, umr cnrs 5584, University of Bordeaux, 351, cours de la Libération, 33405 Talence, France,
E-mail address: rcoudret@math.u-bordeaux1.fr

Gilles Durrieu : Laboratoire de Mathématiques de Bretagne Atlantique, umr cnrs 6205, University of South Brittany, Campus de Tohannic, 56017 Vannes, France,
E-mail address: gilles.durrieu@univ-ubs.fr

Jérôme Saracco : Inria cqfd & Imb, umr cnrs 5584, University of Bordeaux, 351, cours de la Libération, 33405 Talence, France,
E-mail address: jerome.saracco@math.u-bordeaux1.fr