UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Présenté par

Thomas LALOË

On some problems of statistical learning, level set estimation and neuroscience

Soutenue publiquement le **29 Novembre 2018** devant le jury composé de

Mathilde MOUGEOT	-	ENSIIE Paris-Evry	-	*Rapporteur*
Gérard BIAU	-	Sorbonne Université	-	*Examinateur*
Charles BOUVEYRON	-	Université de Nice Sophia Antipolis	-	*Examinateur*
Delphine BLANKE	-	Université d'Avignon	-	*Examinateur*
Anne-Laure FOUGÈRES	-	Université Claude Bernard Lyon 1	-	*Examinateur*
Patricia REYNAUD-BOURET	-	Université de Nice Sophia Antipolis	-	*Examinateur*

Et au vu des rapports écrits par

Antonio CUEVAS	-	Universidad Autónoma de Madrid	-	*Rapporteur*
Michael KOHLER	-	Universität Darmstadt	-	*Rapporteur*

Remerciements

Je souhaite tout d'abord remercier Mathilde Mougeot, Antonio Cuevas et Michael Kohler qui m'ont fait l'honneur d'accepter de rapporter ce manuscrit d'HDR. Merci pour votre travail minutieux et vos commentaires enthousiastes. Thanks Antonio and Michael, your respectives works on the level set and regression estimation have been really inspiring for me. Je tiens également à remercier chaleureusement Gérard Biau, Charles Bouveyron, Delphine Blanke, Anne-Laure Fougères et Patricia Reynaud-Bouret qui ont gentiment accepté de participer à mon Jury. J'ai une pensée supplémentaire pour Gérard qui (aouch il y a maintenant un peu plus de 12 ans!) a accepté de m'encadrer pour mon stage de M2, puis pour ma thèse. C'est grâce à toi et à Benoît si j'en suis là aujourd'hui.

Ce manuscrit n'existerait pas sans toutes les personnes avec lesquelles j'ai collaboré. Je remercie donc tous mes co-auteurs et en particulier Elena et Rémi qui partagent ma vie professionnelle et personnelle depuis tant d'années.

Je veux également remercier mes collègues du laboratoire Dieudonné, ainsi que le personnnel administratif. Quelques mentions particulières : Merci Yannick de m'avoir emmené au festival de Cannes, Patricia pour avoir cru en moi pour co-encadrer une thèse (et avoir relu mon manuscrit!), Mathieu, Roland, Seb et Stella pour les échanges au labo et en dehors, François pour tes conseils toujours pertinents (et pour m'avoir fait découvrir le Q-Ponk), Christine pour me supporter dans notre bureau commun. Merci également à l'équipe pédagogique de l'IUT.

Enfin, merci à ma femme et mes enfants pour tout le bonheur qu'ils m'apportent au quotidien. Je vous aime.

Contents

General Introduction

These notes present an overview of my research in statistics, with some incursions in industrial collaboration, since my PhD thesis [Laloë, 2009]. In this PhD I focused on the area of statistical learning [Laloë, 2008, 2010], in particular for functional data. I started at this moment to work with R. Servien on the topics of level sets (for the regression function [Laloë and Servien, 2013]) and density estimation [Laloë and Servien, 2016].

During my ATER year at the "Institut de Science Financière et d'Assurance" (ISFA), I started a collaboration with E. Di Bernardino who was interested by the level set estimation in order to define new risk measures. In particular we succeeded to estimate the level sets of a multivariate distribution function and provided a multi-dimensional extension of the Conditional Tailed Expectation [Di Bernardino et al., 2013]. Next we considered a combination of the estimation of the regression function and of the distribution level sets to provide an estimator of a new risk measure : the Covariate Conditional Tailed Expectation (CCTE, [Di Bernardino et al., 2015]). A difficulty of this work was to deal with the non compactness of the setting for the estimation of the regression function. This leaded me to work on a new estimator of the regression function [Chagny et al., 2017] (with G. Chagny and R. Servien) to handle this non compact setting.

Another axis of my research concerns clustering and classification. In particular I started to study quantization of functional data (and applications to clustering) during my PhD. I have since continued to work on this subject, and in particular proposed in [Laloë and Servien, 2013] a practical way to use the estimators proposed in [Laloë, 2010].

I also had at heart to have interactions with other disciplines. That leaded me to collaborations in halieutic [Brehmer et al., 2011] and neuroscience [Chevallier and Laloë, 2015] with J. Chevallier. I am also engaged with F. Mathy and P. Reynaud-Bouret in the co-direction of a PhD on cognitive science.

Finally I had the chance to start an industrial collaboration with the start-up Option Way, on the complex area of airfare prediction.

Without giving all the details (all the proofs and simulation studies are available in the corresponding articles, mentioned at the beginning of each chapter), these notes provide some overview of my main topics of research. To keep the presentation as clear as possible, many mathematical details have been left apart and most results are presented in an informal way on simple examples rather than in their full generality. Rigorous results with many discussions and examples can be found in the articles and precise references are given in the notes. I also changed some notation to keep a coherent presentation in the manuscript.

Personal Bibliography

P. Brehmer, P. Fernandes, and T. Laloë. Three-dimensional internal spatial structure of young-of-the-year pelagic freshwater fish provides evidence for the identification of fish school species. *Limnology and Oceanography, Methods*, 9:322–328, 2011. URL https://doi.org/10.4319/lom.2011.9.322.

G. Chagny, T. Laloë, and R. Servien. Multivariate adaptive warped kernel estimation. preprint, 2017. URL https://hal.archives-ouvertes.fr/hal-01616373.

J. Chevallier and T. Laloë. Detection of dependence patterns with delay. *Biometrical Journal*, 57(6):1110–1130, 2015. URL https://doi.org/10.1002/bimj.201400235.

E. Di Bernardino, T. Laloë, V. Maume-Deschamps, and C. Prieur. Plug-in estimation of level sets in a non-compact setting with applications in multivariate risk theory. *ESAIM. Probability and Statistics*, 17:236–256, 2013. URL https://doi.org/10.1051/ps/2011161.

E. Di Bernardino, T. Laloë, and R. Servien. Estimating covariate functions associated to multivariate risks: a level set approach. *Metrika*, 78(5):497–526, 2015. URL https://hal.archives-ouvertes.fr/hal-00800461.

T. Laloë. A k-nearest neighbor approach for functional regression. *Statistics & Probability Letters*, 78(10): 1189–1193, 2008. URL https://doi.org/10.1016/j.spl.2007.11.014.

T. Laloë. *Sur Quelques Problèmes d'Apprentissage Supervisé et Non Supervisé*. PhD thesis, University Montpellier II, 2009. URL https://tel.archives-ouvertes.fr/tel-00455528.

T. Laloë. L_1 quantizationand clustering in banach spaces. *Mathematical Methods of Statistics*, 19(2):136–150, 2010. URL https://hal.archives-ouvertes.fr/hal-01292694.

T. Laloë and R. Servien. Nonparametric estimation of regression level sets using kernel plug-in estimator. *Journal of the Korean Statistical Society*, 42(3):301–311, 2013. URL https://doi.org/10.1016/j.jkss.2012.10.001.

T. Laloë and R. Servien. The X-Alter algorithm : a parameter-free method to perform unsupervised clustering. *Journal of Modern Applied Statistical Methods*, 12(1):90–102, 2013. URL https://hal.archives-ouvertes.fr/hal-00674407.

T. Laloë and R. Servien. A note on the asymptotic law of the histogram without continuity assumptions. *Brazilian Journal of Probability and Statistics*, 30(4):562–569, 2016. URL https://doi.org/10.1214/15-BJPS294.

Part I

A journey in the estimation of level sets

Introduction

The estimation of level sets is widely studied in the literature. In particular the estimation of density level sets has been studied in [Baíllo, 2003, Baíllo et al., 2001, Biau et al., 2007, Butucea et al., 2007, Cadre, 2006, Polonik, 1995, Rigollet and Vert, 2009, Tsybakov, 1997]. This profusion of articles on the subject is due to the various practical uses of level sets. As an example, in the context of environmental monitoring, [Rahimi et al., 2004] consider the estimation of level sets of quantities such as solar radiation, humidity, etc. In a different category, [Bryan et al., 2006] want to determine the subset of a parameter space that represents "acceptable" hypotheses. We also will see in this part that level set estimation can be a suitable tool to generalize univariate risk measures (Value at Risk, Conditional Tailed Expectation) in multivariate settings. So far, the level set estimation has mostly be done under compactness assumption on the level sets. In particular, the problem of estimating general level sets under compactness assumptions has been discussed by [Cuevas et al., 2006].

I worked at first on the level set estimation for regression functions [Laloë and Servien, 2013]. More recently, I considered with E. Di Bernardino, V. Maume-Deschamps, C. Prieur [Di Bernardino et al., 2013] the level set estimation of a cumulative distribution function (c.d.f.), to propose new risk measures. The main difference in our setting is that the level set cannot supposed to be compact any more. Thus we had to find a way to deal with this non compactness.

Then, with E. Di Bernardino and R. Servien, I studied the behaviour of the regression function on a level set of the c.d.f. of the covariates [Di Bernardino et al., 2015]. One problem is to construct an estimator of the regression function, without assuming the compactness of the support of the covariates. This leads to a work with G. Chagny and R. Servien proposing a warped kernel estimator of the regression function [Chagny et al., 2017].

This part is organised as follows. Chapter 1 is devoted to the estimation of the level sets of the c.d.f. In Chapter 2, I define and estimate our two new risk measures. Chapter 3 details our warped estimator of the regression function. Finally I end this part by several prospects and ongoing works.

Chapter 1

Estimating the level sets

This chapter, related to the article [Di Bernardino et al., 2013] I wrote in collaboration with E. Di Bernardino, V. Maume-Deschamps and C. Prieur , is dedicated to the cumulative distribution function (c.d.f.) level set estimation. I skip a previous work on the regression function level set estimation [Laloë and Servien, 2013]. However I will discuss in the prospects a new development I recently started on the subject. Moreover, in order to be consistent with Chapters 2 and 3, I consider here random variables defined on $\mathbb{R}_+^d$ (instead of $\mathbb{R}_+^2$ as in [Di Bernardino et al., 2013]). This generalization is straightforward.

1.1 Definitions, Notations and Assumptions

Note $\mathcal{F}$ the set of continuous c.d.f. $\mathbb{R}_+^d \to [0,1]$ and Consider $(\mathbf{X}, Y)$ a random pair on $\mathbb{R}_+^d \times \mathbb{R}$, with $F_{\mathbf{X}} \in \mathcal{F}$ the c.d.f. of $\mathbf{X}$. We are interested in estimating the level sets of this c.d.f., defined by

$$\mathcal{L}(\alpha) := \{\mathbf{x} \in \mathbb{R}_+^d : F_{\mathbf{X}}(\mathbf{x}) > \alpha\}, \quad \alpha \in (0,1).$$

We set

$$\{F_{\mathbf{X}} = \alpha\} = \{\mathbf{x} \in \mathbb{R}_+^d : F_{\mathbf{X}}(\mathbf{x}) = \alpha\},$$

and given $T > 0$, truncated versions

$$\mathcal{L}(\alpha)^T = \{\mathbf{x} \in [0,T]^d : F_{\mathbf{X}}(\mathbf{x}) \geq \alpha\}, \text{ and } \{F_{\mathbf{X}} = \alpha\}^T = \{\mathbf{x} \in [0,T]^d : F_{\mathbf{X}}(\mathbf{x}) = \alpha\}.$$

As discussed in [Di Bernardino et al., 2013] and in the following, this truncation is necessary in order to deal with the non compactness of the considered level sets. Furthermore, for any $A \subset \mathbb{R}_+^d$ we denote by ∂A its boundary. Note that we restrict ourselves to $\mathbb{R}_+^d$ for convenience but the following results are completely adaptable in $\mathbb{R}^d$.

In the metric space $(\mathbb{R}_+^d, d)$, where d stands for the Euclidean distance, we denote by $B(\mathbf{x}, \rho)$ the closed ball centred on x and with positive radius ρ. Let $B(S, \rho) = \bigcup_{\mathbf{x} \in S} B(\mathbf{x}, \rho)$, with S a closed set of $\mathbb{R}_+^d$. For $t > 0$, $\zeta > 0$ and $\alpha \in (0,1)$, define

$$E = B(\{\mathbf{x} \in \mathbb{R}_+^d : \mid F_{\mathbf{X}}(\mathbf{x}) - \alpha \mid \leq t\}, \zeta),$$

and, for a twice differentiable function $F_{\mathbf{X}}$,

$$m^{\nabla} = \inf_{\mathbf{x} \in E} \|(\nabla F_{\mathbf{X}})_x\|, \qquad M_H = \sup_{\mathbf{x} \in E} \|(HF_{\mathbf{X}})_{\mathbf{x}}\|,$$

where $(\nabla F_{\mathbf{X}})_x$ is the gradient vector of $F_{\mathbf{X}}$ evaluated at x and $\|(\nabla F_{\mathbf{X}})_x\|$ its Euclidean norm, $(HF_{\mathbf{X}})_{\mathbf{x}}$ the Hessian matrix evaluated in x and $\|(HF_{\mathbf{X}})_{\mathbf{x}}\|$ its matrix norm induced by the Euclidean norm.

We chose to study the consistency properties of an estimator $\mathcal{L}_n(\alpha)^T$ of $\mathcal{L}(\alpha)^T$ with respect to two different distances.

The Hausdorff distance:

The Hausdorff distance corresponds to an intuitive notion of "physical proximity" between sets. However the metric d_H (see (1.1) and (1.2) below) is not always completely successful in capturing the shape properties: two sets can be very close in d_H and still have quite different shapes. In order to avoid these situations, following [Cuevas and Rodríguez-Casal, 2004] and [Cuevas et al., 2006], a way to reinforce the notion of visual proximity between two sets provided by d_H is to impose the proximity of the respective boundaries. Thus we will consider $d_H(\partial\mathcal{L}_n(\alpha)^T, \partial\mathcal{L}(\alpha)^T)$.

For sake of completeness, we recall that if A_1 and A_2 are compact sets in $(\mathbb{R}^d_+, d)$, the Hausdorff distance between A_1 and A_2 is defined by

$$d_H(A_1, A_2) = \inf\{\rho > 0 : A_1 \subset B(A_2, \rho), A_2 \subset B(A_1, \rho)\}, \tag{1.1}$$

or equivalently by

$$d_H(A_1, A_2) = \max\left\{ \sup_{\mathbf{x}\in A_1} \inf_{\mathbf{y}\in A_2} \| \mathbf{x}-\mathbf{y} \|, \sup_{\mathbf{x}\in A_2} \inf_{\mathbf{y}\in A_1} \| \mathbf{x}-\mathbf{y} \| \right\}. \tag{1.2}$$

The above expression is well-defined even when A_1 and A_2 are just closed (not necessarily compact) sets but in this case the value $d_H(A_1, A_2)$ could be infinite. Then in our setting, in order to avoid these situations we need a truncated version of Assumption (T) in [Cuevas et al., 2006]. More precisely we introduce the following assumption:

H: There exist $\gamma > 0$ and $A > 0$ such that, if $|\, t - c\,| \le \gamma$ then $\forall\ T > 0$ such that $\{F_{\mathbf{X}} = c\}^T \neq \emptyset$ and $\{F_{\mathbf{X}} = t\}^T \neq \emptyset$,

$$d_H(\{F_{\mathbf{X}} = c\}^T, \{F_{\mathbf{X}} = t\}^T) \le A\,|\,t - c\,|\,.$$

Assumption **H** is satisfied under mild conditions. Proposition 1.1.1 below is a slight modification of Proposition 3.1 in the PhD Thesis of Rodríguez-Casal [Rodríguez-Casal, 2003] in order to deal with non-compact sets, and is proved in [Di Bernardino et al., 2013].

Proposition 1.1.1 *Let $\alpha \in (0,1)$. Let $F_{\mathbf{X}} \in \mathcal{F}$ be twice differentiable on $\mathbb{R}^{d*}_+$. Assume there exist $r > 0$, $\zeta > 0$ such that $m^{\nabla} > 0$ and $M_H < \infty$. Then $F_{\mathbf{X}}$ satisfies Assumption* **H**, *with $A = \frac{2}{m^{\nabla}}$.*

The L_1 distance:

Another possibility is to consider the volume (in the Lebesgue measure sense) of the symmetric difference between $\mathcal{L}(\alpha)^{T_n}$ and $\mathcal{L}_n(\alpha)^{T_n}$:

$$d_\lambda(\mathcal{L}_n(\alpha)^T, \mathcal{L}(\alpha)^T) = \lambda\left(\mathcal{L}^T_n(\alpha)\Delta\mathcal{L}^T(\alpha)\right) = \lambda\left[\mathcal{L}_n(\alpha)^T \cap \overline{\mathcal{L}^T(\alpha)}) \cup (\overline{\mathcal{L}^T_n(\alpha)} \cap \mathcal{L}(\alpha)\right],$$

where λ stands for the Lebesgue measure on $\mathbb{R}^d$ and $\triangle$ for the symmetric difference.

1.2 Plug-in estimation of the Level Sets

Given $\{\mathbf{X}_i\}_{i=1}^n$ an *i.i.d* sample in $\mathbb{R}_+^d$, drawn from a c.d.f. $F_{\mathbf{X}}$, we denote by $F_n(\cdot) = F_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \cdot)$ an estimator of $F_{\mathbf{X}}$ based on $\{\mathbf{X}_i\}_{i=1}^n$.

1.2.1 Hausdorff Distance

From now on we note, for $n \in \mathbb{N}^*$,

$$\|F_{\mathbf{X}} - F_n\|_\infty = \sup_{\mathbf{x} \in \mathbb{R}_+^d} |\, F_{\mathbf{X}}(\mathbf{x}) - F_n(\mathbf{x}) \,|,$$

and for $T > 0$

$$\|F_{\mathbf{X}} - F_n\|_\infty^T = \sup_{\mathbf{x} \in [0,T]^d} |\, F_{\mathbf{X}}(\mathbf{x}) - F_n(\mathbf{x}) \,| \,.$$

The following result, proved in [Di Bernardino et al., 2013] can be considered as an adapted version of Theorem 2 in [Cuevas et al., 2006].

Theorem 1.2.1 *Let $\alpha \in (0,1)$. Let $F_{\mathbf{X}} \in \mathcal{F}$ be twice differentiable on $\mathbb{R}_+^{d*}$. Assume that there exist $t > 0$, $\zeta > 0$ such that $m^\nabla > 0$ and $M_H < \infty$. Let $T_1 > 0$ such that for all $\alpha' : |\, \alpha' - \alpha \,| \leq t$, $\partial\mathcal{L}(\alpha')^{T_1} \neq \emptyset$. Let $(T_n)_{n \in \mathbb{N}^*}$ be an increasing sequence of positive values. Assume that, for each n and for almost all samples of size n, F_n is an upper semi-continuous function and that*

$$\|F_{\mathbf{X}} - F_n\|_\infty \to 0, \quad a.s. \tag{1.3}$$

Then

$$d_H(\partial\mathcal{L}(\alpha)^{T_n}, \partial\mathcal{L}_n(\alpha)^{T_n}) = O(\|F_{\mathbf{X}} - F_n\|_\infty), \qquad a.s.$$

Remark 1.2.1 *1. Theorem 1.2.1 is slightly different from its equivalent in [Di Bernardino et al., 2013] where F_n was supposed continuous. The weakening of this assumption requires a trivial modification of the proof and allows the use of the empirical c.d.f. for F_n.*

2. Theorem 1.2.1 provides an asymptotic result for a fixed level α. In particular following the proof in [Di Bernardino et al., 2013] we remark that, for n large enough,

$$d_H(\partial\mathcal{L}(\alpha)^{T_n}, \partial\mathcal{L}_n(\alpha)^{T_n}) \leq 6\, A\, \|F_{\mathbf{X}} - F_n\|_\infty^{T_n}, \quad a.s.,$$

where $A = \frac{2}{m^\nabla}$. Note that if α is close to one, then m^∇ can be very close to 0. This leads to a large value of the constant A, degrading the performances of the estimation.

1.2.2 L_1 Consistency

The previous section was devoted to the consistency of $\mathcal{L}_n$ in terms of the Hausdorff distance. We consider now another consistency criterion introduced in the previous section: the consistency of the volume (in the Lebesgue measure sense) of the symmetric difference between $\mathcal{L}(\alpha)^{T_n}$ and $\mathcal{L}_n(\alpha)^{T_n}$. Let us introduce the following assumption:

A1 There exist positive increasing sequences $(v_n)_{n \in \mathbb{N}^*}$ and $(T_n)_{n \in \mathbb{N}^*}$ such that

$$v_n \int_{[0,T_n]^d} |\, F_{\mathbf{X}} - F_n \,|^p \; \lambda(\mathrm{d}x) \xrightarrow[n \to \infty]{\mathbb{P}} 0,$$

for some $1 \leq p < \infty$.

We now establish our consistency result in terms of the volume of the symmetric difference. We can interpret the following theorem as a generalization of Theorem 3 in [Cuevas et al., 2006] in the case of non-compact level sets. In addition, we provide also a convergence rate for the symmetric difference between $\mathcal{L}(\alpha)^{T_n}$ and $\mathcal{L}_n(\alpha)^{T_n}$.

Theorem 1.2.2 *Let $\alpha \in (0,1)$. Let $F_{\mathbf{X}} \in \mathcal{F}$ be twice differentiable on $\mathbb{R}_+^{d*}$. Assume that there exist $t > 0$, $\zeta > 0$ such that $m^{\nabla} > 0$ and $M_H < \infty$. Assume that for each n, with probability one, F_n is measurable. Let $(v_n)_{n \in \mathbb{N}^*}$ and $(T_n)_{n \in \mathbb{N}^*}$ positive increasing sequences such that Assumption* **A1** *is satisfied and that for all $\alpha' : \mid \alpha' - \alpha \mid \leq t$, $\partial \mathcal{L}(\alpha')^{T_1} \neq \emptyset$. Then it holds that*

$$p_n \, d_{\lambda}(\mathcal{L}(\alpha)^{T_n}, \mathcal{L}_n(\alpha)^{T_n}) \underset{n \to \infty}{\overset{\mathbb{P}}{\to}} 0,$$

with p_n an increasing positive sequence such that $p_n = o\left(v_n^{\frac{1}{p+1}} / T_n^{\frac{(d-1)*p}{p+1}} \right)$.

The proof and a discussion around an assumption concerning $\|F_{\mathbf{X}} - F_n\|_{\infty}$ are given in [Di Bernardino et al., 2013]. Theorem 1.2.2 provides a convergence rate, which is closely related to the choice of the sequence T_n. A sequence T_n whose divergence rate is large implies a convergence rate p_n quite slow. An example using the empirical c.d.f. (in the bivariate case) for F_n is also provided in [Di Bernardino et al., 2013].

Conclusion

At this point, we dispose of an estimator for the level sets of a multivariate c.d.f. The empirical study we provided in [Di Bernardino et al., 2013] illustrates perfectly the rates of consistency of Theorem 1.2.2 and emphasizes the key role of the choice of T_n. This choice remains a difficult and open problem. I will now present how we use these level sets to define new risk measures.

Chapter 2

Introduction of new risk measures

In this chapter, related to two papers I wrote with E. Di Bernardino, V. Maume-Deschamps, C. Prieur and R. Servien [Di Bernardino et al., 2013, 2015], I introduce and estimate two new risk measures: the multivariate Conditional Tailed Expectation (Section 2.1) and the Covariate Conditional Tailed expectation (Section 2.2). As in the previous chapter I consider random variables defined on $\mathbb{R}_+^d$ (instead of $\mathbb{R}_+^2$ in [Di Bernardino et al., 2013]).

2.1 Multivariate Conditional Tailed Expectation (CTE)

From the usual definition in the univariate setting we know that the quantile function Q_X provides a point which accumulates a probability α to the left tail and $1-\alpha$ to the right tail. More precisely, given an univariate continuous and strictly monotonic cumulative distribution function (c.d.f.) F_X,

$$Q_X(\alpha) = F_X^{-1}(\alpha), \quad \forall\, \alpha \in (0,1). \tag{2.1}$$

The notion of univariate quantile function Q_X is used in risk theory to define an univariate measure of risk: the Value-at-Risk (VaR). This measure is defined as

$$\mathrm{VaR}_\alpha(X) = Q_X(\alpha), \quad \forall\, \alpha \in (0,1). \tag{2.2}$$

Following the general ideas of [Embrechts and Puccetti, 2006] and [Nappo and Spizzichino, 2009] an intuitive generalization of the VaR measure in the case of a multidimensional c.d.f. $F_{\mathbf{X}}$ is given by its α-quantile curves. More precisely:

Definition 2.1.1 *For $\alpha \in (0,1)$ and $F_{\mathbf{X}} \in \mathcal{F}$, the multidimensional Value-at-Risk at probability level α is the boundary of its α-level set, i.e. $VaR_\alpha(F_{\mathbf{X}}) = \partial\mathcal{L}(\alpha)$.*

For details about a parametric formulation of the quantile curve $\partial\mathcal{L}(\alpha)$ see [Belzunce et al., 2007]. For details about its properties see for instance [Fernández-Ponce and Suárez-Lloréns, 2002] (and references therein) and [Nappo and Spizzichino, 2009].

Then, using an estimator F_n as in Section 1.2, we can define our plug-in estimator of the multivariate Value-at-Risk by

$$\mathrm{VaR}_\alpha(F_n) = \partial\mathcal{L}_n(\alpha), \qquad \text{for } \alpha \in (0,1).$$

Moreover, under assumptions of Theorem 1.2.1, we obtain a consistency result, with respect to the Hausdorff distance, for the $\mathrm{VaR}_\alpha(F_n)$ on the quadrant $\mathbb{R}_+^d$ i.e.

$$d_H(\mathrm{VaR}_\alpha(F_{\mathbf{X}})^{T_n}, \mathrm{VaR}_\alpha(F_n)^{T_n}) = O(\|F_{\mathbf{X}} - F_n\|_\infty), \quad a.s.$$

As in the univariate case, the multi-dimensional VaR at a predetermined level α does not give any information about the thickness of the upper tail of the c.d.f. This is a considerable shortcoming of VaR measure because in practice we are not only concerned with frequency of the default but also with the severity of loss in case of default. In other words we are interested to analyse the behaviour of $\mathbf{X}$ not only on the boundary but also in the whole α-level set.

In dimension one, in order to overcome this problem, another risk measure has recently received growing attention in insurance and finance literature: the Conditional Tail Expectation (CTE). Following [Artzner et al., 1999] and [Dedu and Ciumara, 2010], for a continuous c.d.f. F_X the CTE at level α is defined by

$$\mathrm{CTE}_\alpha(X) = \mathbf{E}[\, X \,|\, X \geq \mathrm{VaR}_\alpha(X)\,],$$

where $\mathrm{VaR}_\alpha(X)$ is the univariate VaR in (2.2). For a comprehensive treatment and for references to the extensive literature on $\mathrm{VaR}_\alpha(X)$ and $\mathrm{CTE}_\alpha(X)$ one may refer to [Denuit et al., 2005].

In [Di Bernardino et al., 2013] we discussed the several bivariate generalizations of the classical univariate CTE and we proposed a new one. I propose here the same generalization but in the d-dimensional case. This generalization (from bivariate to multivariate CTE) is straightforward and I refer to [Di Bernardino et al., 2013] for the proofs of the theoretical results. Let us first introduce the following assumption:

A2: $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ is a random vector on $\mathbb{R}^d_+$ fulfilling $\mathbf{E}((X^{(i)})^2) < \infty$ $\forall i = 1, \ldots, d$. The random vector $\mathbf{X}$ has a λ-density function f such that $\int f^{1+\epsilon} \mathrm{d}\lambda < \infty$, for some $\epsilon > 0$.

Definition 2.1.2 *Consider a random vector* $\mathbf{X}$ *satisfying Assumption* **A2**, *with associate distribution function* $F_{\mathbf{X}} \in \mathcal{F}$. *For* $\alpha \in (0, 1)$, *we define*

1. *the multivariate* α*-Conditional Tail Expectation*

$$\mathrm{CTE}_\alpha(\mathbf{X}) = \mathbf{E}[\mathbf{X} | \mathbf{X} \in \mathcal{L}(\alpha)] = \Big(\mathbf{E}[X^{(i)} | \mathbf{X} \in \mathcal{L}(\alpha)]\Big)_{1 \leq i \leq d}$$

2. *the estimated multivariate* α*-Conditional Tail Expectation*

$$\widehat{\mathrm{CTE}}_\alpha(\mathbf{X}) = \left(\frac{\sum_{j=1}^n X_j^{(i)} 1_{\{(\mathbf{X}_j) \in \mathcal{L}_n(\alpha)\}}}{\sum_{j=1}^n 1_{\{\mathbf{X}_j^{(i)} \in \mathcal{L}_n(\alpha)\}}} \right)_{1 \leq i \leq d} \tag{2.3}$$

An advantage of our $\mathrm{CTE}_\alpha(\mathbf{X})$ is that it does not use an aggregate (over the d elements of the vector) variable (sum, min, max ...) in order to analyse the multivariate risk's issue. By contrast, using a geometric approach, $\mathrm{CTE}_\alpha(\mathbf{X})$ rather deals with the simultaneous joint damages considering the multivariate dependence structure of data in a specific risk's area $\mathcal{L}(\alpha)$. We also introduce truncated versions of the theoretical and estimated CTE_α:

$$\mathrm{CTE}^T_\alpha(\mathbf{X}) = \mathbf{E}[\mathbf{X} | \mathbf{X} \in \mathcal{L}(\alpha)^T],$$

and

$$\widehat{\mathrm{CTE}}^T_\alpha(\mathbf{X}) = \left(\frac{\sum_{j=1}^n X_j^{(i)} 1_{\{(\mathbf{X}_j) \in \mathcal{L}_n(\alpha)^T\}}}{\sum_{j=1}^n 1_{\{\mathbf{X}_j^{(i)} \in \mathcal{L}_n(\alpha)^T\}}} \right)_{1 \leq i \leq d},$$

where $\mathcal{L}(\alpha)^T$ and $\mathcal{L}_n(\alpha)^T$ are the truncated versions of theoretical and estimated upper α-level set defined in Chapter 1.

Theorem 2.1.1 *Under Assumption* **A2***, Assumptions of Theorem 1.2.2 and with the notation of Theorem 1.2.2, it holds that*

$$\beta_n \big| \, \mathrm{CTE}_\alpha^{T_n}(\mathbf{X}) - \widehat{\mathrm{CTE}}_\alpha^{T_n}(\mathbf{X}) \, \big| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{2.4}$$

where $\beta_n = \min\{ p_n^{\frac{\epsilon}{2(1+\epsilon)}}, a_n \}$*, with* $\epsilon > 0$ *satisfying Assumption* **A2** *and* $a_n = o(\sqrt{n})$*.*

The convergence in (2.4) must be interpreted as a componentwise convergence. In the case of a bounded density function $f_{(X,Y)}$ we obtain $\beta_n = \min\{\sqrt{p_n}, a_n\}$. The special case where F_n is the empirical c.d.f. is discussed in [Di Bernardino et al., 2013]

Remark 2.1.1 *It could be interesting to consider the convergence* $\big| \mathrm{CTE}_\alpha(\mathbf{X}) - \widehat{\mathrm{CTE}}_\alpha^{T_n}(\mathbf{X}) \big|$*. We remark that in this case the speed of convergence will also depend on the convergence rate to zero of* $\big| \, \mathrm{CTE}_\alpha(\mathbf{X}) - \mathrm{CTE}_\alpha^{T_n}(\mathbf{X}) \big|$*, then, in particular of* $\mathbb{P}[(\mathbf{X}) \in \mathcal{L}(\alpha) \setminus \mathcal{L}(\alpha)^{T_n}]$ *for* $n \to \infty$*. More precisely* $\big| \mathrm{CTE}_\alpha(\mathbf{X}) - \mathrm{CTE}_\alpha^{T_n}(\mathbf{X}) \big|$ *decays to zero at least with a convergence rate* $(\mathbb{P}[\bigcup_i (\mathbf{X}_i \geq T_n)])^{-1}$*.*

2.2 Covariate Conditional Tailed Expectation (CCTE)

In this section we are interested in another notion of risk studied in [Di Bernardino et al., 2015]: the Covariate Conditional Tailed Expectation (CCTE). The goal is to provide an indicator of the behaviour of a covariate Y on the level sets of a d-dimensional vector of risk factors $\mathbf{X}$. More precisely, adapting the multivariate CTE defined in the previous section we define the Covariate Conditional Tailed Expectation by

$$\mathrm{CCTE}_\alpha(\mathbf{X}, Y) := \mathbf{E}[Y \,|\, \mathbf{X} \in \mathcal{L}(\alpha)], \quad \text{with } \alpha \in (0,1). \tag{2.5}$$

In [Di Bernardino et al., 2015] we first studied the L_p-consistency of the estimation of the regression function $r(\mathbf{x}) = r_{Y|\mathbf{X}}(\mathbf{x}) = \mathbf{E}[Y|\mathbf{X} = \mathbf{x}]$ on the level sets of $F_{\mathbf{X}}$. We will focus only here on the estimation of the CCTE and its “truncated version”

$$\mathrm{CCTE}_\alpha^T(\mathbf{X}, Y) := \mathbf{E}[Y \,|\, \mathbf{X} \in \mathcal{L}(\alpha)^T], \text{ for T} > 0.$$

2.2.1 Covariable Y is measured

Let us first assume that the covariable is measured for all the data. Assume that we have an i.i.d. sample $\{(\mathbf{X}_i, Y_i)\}_{i=1,\dots,n}$. We introduce the following assumption:

A3 There exist three positive increasing sequences $(v_{1,n})_{n\in\mathbb{N}^*}$ and $(T_n)_{n\in\mathbb{N}^*}$ such that $v_{1,n} \to \infty$, $T_n \to \infty$, and

$$v_{1,n} \int_{[0,T_n]^d} | \, F_{\mathbf{X}}(\mathbf{x}) - F_n(\mathbf{x}) \, |^p \; \lambda(\mathrm{d}x) \xrightarrow{\mathbb{P}} 0,$$

for some $1 \leq p < \infty$.

We can now define our estimator for the CCTE and establish the consistency of this estimator with a convergence rate.

Definition 2.2.1 *Consider a random vector* $\mathbf{X}$ *with c.d.f.* $F_{\mathbf{X}}$ *and a random variable* Y*. For* $\alpha \in (0,1)$*, we estimate* $\mathrm{CCTE}_\alpha^T(\mathbf{X}, Y)$ *by*

$$\widehat{\mathrm{CCTE}}_{\alpha,n}^{T_n}(\mathbf{X}, Y) = \mathbf{E}_n \left[Y | \mathbf{X} \in \mathcal{L}_n(\alpha)^{T_n} \right],$$

where $\mathbf{E}_n$ *denotes the empirical version of the expected value.*

Theorem 2.2.1 *Let* α *be in* $(0,1)$. *Let* $F_{\mathbf{X}} \in \mathcal{F}$ *be a twice differentiable c.d.f. on* $\mathbb{R}_+^{d*}$ *with an associated density* f *such that Assumption* **A2** *is satisfied. Assume that there exist* $t > 0$, $\zeta > 0$ *such that* $m^{\nabla} > 0$ *and* $M_H < \infty$. *Let* $T_1 > 0$ *such that for all* $\alpha' : \mid \alpha' - \alpha \mid \leq t$, $\partial \mathcal{L}(\alpha')^{T_1} \neq \emptyset$. *Assume that for each* n, F_n *is measurable. Let* $(v_{1,n})_{n \in \mathbb{N}^*}$ *and* $(T_n)_{n \in \mathbb{N}^*}$ *positive increasing sequences such that Assumption* **A3** *is satisfied. We have*

$$\beta_n \left| \widehat{\mathrm{CCTE}}_{\alpha,n}^{T_n}(\mathbf{X}, Y) - \mathrm{CCTE}_{\alpha}^{T_n}(\mathbf{X}, Y) \right| \xrightarrow[n \to \infty]{\mathbb{P}} 0,$$

with $\beta_n = \min\left(p_n^{\frac{\epsilon}{2(1+\epsilon)}}, d_n \right)$, *where* $p_n = o\left(v_{1,n}^{\frac{p}{p+1}} / T_n^{\frac{d+(d-1)p}{p+1}} \right)$ *and* $d_n = o(\sqrt{n})$.

Remark 2.2.1 *Similarly to Remark 2.1.1, it could also be very interesting to consider the convergence* $\left| \widehat{\mathrm{CCTE}}_{\alpha,n}^{T_n}(\mathbf{X}, Y) - \mathrm{CCTE}_{\alpha}(\mathbf{X}, Y) \right|$. *We remark that in this case the speed of convergence will also depend on the convergence rate to zero of* $\left| \mathrm{CCTE}_{\alpha}^{T_n}(\mathbf{X}, Y) - \mathrm{CCTE}_{\alpha}(\mathbf{X}, Y) \right|$, *then, in particular of* $\mathbb{P}[\mathbf{X} \in \mathcal{L}(\alpha) \setminus \mathcal{L}(\alpha)^{T_n}]$, *for* $n \to \infty$. *This could not be done without adding strong assumptions and careful developments and has not been considered yet.*

The particular case of the empirical c.d.f. has been studied in [Di Bernardino et al., 2015].

2.2.2 Covariable Y is partially unknown

We deal now with a more difficult case. We suppose that the covariable Y cannot be measured for all the individuals. It could happen if a measure of Y is very expensive or invasive (in some medical treatment, for example). So we have two different i.i.d. samples: $S_{N_1}^1 = \{(\mathbf{X}_i, Y_i)\}_{i=1}^{N_1}$ and $S_{N_2}^2 = \{\mathbf{X}_j\}_{j=1}^{N_2}$, with N_2 potentially much bigger than N_1.

In this case we use $S_{N_1}^1$ to get an estimator r_{N_1} of the regression function r. Then, we apply this estimator on the sample $S_{N_2}^2$ in order to estimate the CCTE measure. To this aim we define:

Definition 2.2.2 *Consider a random vector* $\mathbf{X}$ *with c.d.f.* $F_{\mathbf{X}}$, *a random variable* Y *and two i.i.d. samples* $S_{N_1}^1$ *and* $S_{N_2}^2$. *For* $\alpha \in (0,1)$, *we define our estimator of the CCTE by*

$$\widehat{\mathrm{CCTE}^{\star}}_{\alpha,N_1,N_2}^{T_{N_2}}(\mathbf{X}, Y) = \mathbf{E}_{N_2}\left[r_{N_1}(\mathbf{X}) | \mathbf{X} \in \mathcal{L}_{N_2}(\alpha)^{T_{N_2}} \right].$$

Before establishing a rate of consistency we need to introduce a new assumption:

A3′ There exist three positive increasing sequences $(v_{1,n})_{n \in \mathbb{N}^*}$, $(v_{2,n})_{n \in \mathbb{N}^*}$ and $(T_n)_{n \in \mathbb{N}^*}$ such that $v_{1,n} \to \infty$, $v_{2,n} \to \infty$, $T_n \to \infty$, and

$$v_{1,N_1} \, \|r_{N_1} - r\|_p \xrightarrow{\mathbb{P}} 0, \quad v_{2,N_2} \int_{[0,T_{N_2}]^d} \mid F_{\mathbf{X}}(\mathbf{x}) - F_{N_2}(\mathbf{x}) \mid^p \, \lambda(\mathrm{d}x) \xrightarrow{\mathbb{P}} 0, \text{ for some } 1 \leq p < \infty.$$

The following result proves the consistency of the estimator introduced in Definition 2.2.2.

Theorem 2.2.2 *Let* α *be in* $(0,1)$. *Let* $F_{\mathbf{X}} \in \mathcal{F}$ *be a twice differentiable c.d.f. on* $\mathbb{R}_+^{d*}$ *with an associated density* f *such that Assumption* **A2** *is satisfied. Assume that there exist* $t > 0$, $\zeta > 0$ *such that* $m^{\nabla} > 0$ *and* $M_H < \infty$. *Let* $T_1 > 0$ *such that for all* $\alpha' : \mid \alpha' - \alpha \mid \leq t$, $\partial \mathcal{L}(\alpha')^{T_1} \neq \emptyset$. *Let* $(v_{1,n})_{n \in \mathbb{N}^*}$, $(v_{2,n})_{n \in \mathbb{N}^*}$ *and* $(T_n)_{n \in \mathbb{N}^*}$ *positive increasing sequences such that Assumption* **A3′** *is satisfied. Assume that* r *is a continuous and positive regression function such that* $E\left[r(\mathbf{X})^2 \right] < \infty$. *We have*

$$\beta_{N_1,N_2} \left| \widehat{\mathrm{CCTE}^{\star}}_{\alpha,N_1,N_2}^{T_{N_2}}(\mathbf{X}, Y) - \mathrm{CCTE}_{\alpha}^{T_{N_2}}(\mathbf{X}, Y) \right| \xrightarrow{\mathbb{P}} 0, \quad \textit{for } N_1, N_2 \to \infty,$$

with $\beta_{N_1,N_2} = \min\left(p_{N_2}^{\frac{\epsilon}{2(1+\epsilon)}}, c_{N_1}, d_{N_2}\right)$, $p_{N_2} = o\left(v_{2,N_2}^{\frac{p}{p+1}} / T_n^{\frac{d+(d-1)p}{p+1}}\right)$, $c_{N_1} = o\left(\mathbf{E}|r_{N_1}(\mathbf{X}) - r(\mathbf{X})|\right)$ *and* $d_{N_2} = o\left(\sqrt{N_2}\right)$.

Note that, compared to Theorem 2.2.1, we have a supplementary term c_{N_1}. This term controls the rate of convergence of r_{N_1} toward r. Again, the particular case of the empirical c.d.f. has been studied in [Di Bernardino et al., 2015]. Note that assumption **A3′** implies to consider an estimator of the regression function that do not assume a bounded support. The next chapter is devoted to the introduction of such an estimator.

Conclusion

Starting from the estimation of the level sets of a multivariate cumulative distribution function (see Chapter 1), I presented two new risk measures: the CTE and CCTE. In both cases a plug-in estimator was provided and studied. We provide an empirical study of the performance of our estimator of the CTE in [Di Bernardino et al., 2013]. It confirms the key role of T_n, specially for high values of α. We also provide an illustration on real data. We present an equivalent empirical study for our estimator of the CCTE in [Di Bernardino et al., 2015], confirming the theoretical rates of consistency. In order to deal with the non compactness of the setting to estimate the regression function, we used a kernel estimator provided by [Kohler et al., 2009]. Moreover we perform an comparison with parametric and semi-parametric approaches, and establish the relevance of our estimator in general cases. Finally we provide an application on a real data-set. The next chapter provides a new and estimator of the regression function, suitable to our estimator of the CCTE.

Chapter 3

Estimating the regression function

In this chapter, dedicated to the article I wrote with G. Chagny and R. Servien [Chagny et al., 2017], I use a warping approach to propose a new estimator of the regression function in a non compact support design. The goal was to propose an efficient estimator satisfying Assumption **A3'** of the previous chapter.

3.1 Preliminary notions

Let $(\mathbf{X}, Y)$ be a couple of random variables taking values on $\mathbb{R}^d \times \mathbb{R}$ such that

$$Y = r(\mathbf{X}) + \varepsilon, \tag{3.1}$$

with ε a centred real random variable with finite variance independent of $\mathbf{X} = (X_1, \ldots, X_d)$. Assume that we have an independent identically distributed (*i.i.d.* in the sequel) sample $(\mathbf{X}_i, Y_i)_{i=1\ldots n}$ distributed as $(\mathbf{X}, Y)$. The subject of the paper is the estimation of the multivariate regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ on a subset $A \subset \mathbb{R}^d$, with a warping device described below, that also requires the estimation of the dependence structure between the coordinates of $\mathbf{X}$.

Regression estimation is a classical problem in statistics, addressed in a significant number of research works frequently based on non-parametric methods such as kernel estimators [Nadaraya, 1964, Watson, 1964], local polynomial estimators [Fan and Gijbels, 1996], orthogonal series or spline estimators [Antoniadis et al., 1997, Baraud, 2002, Efromovich, 1999, Golubev and Nussbaum, 1992] and nearest neighbour-type estimators [Guyader and Hengartner, 2013, Stute, 1984]. Among kernel methods, the most popular estimator is the well-known Nadaraya-Watson estimator, defined for model (3.1) by

$$\widehat{r}^{NW}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i)}{\sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i)}, \tag{3.2}$$

where $\mathbf{h} = {}^t(h_1, \ldots, h_d)$ is the bandwidth of the kernel K, $K_{\mathbf{h}}(\mathbf{x}) = K_{1,h_1}(x_1)K_{2,h_2}(x_2)\ldots K_{d,h_d}(x_d)$, with $K_{l,h_l}(x) = K_l(x/h_l)/h_l$ for $h_l > 0$, and $K_l : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} K_l(x)dx = 1$, $l = 1, \ldots, d$.

A commonly shared assumption for regression analysis is that the support of $\mathbf{X}$ is a compact subset of $\mathbb{R}^d$ [Furer and Kohler, 2015, Guyader and Hengartner, 2013, Györfi et al., 2002]. It could be very restrictive in some situations such as for example the estimation of the regression function on the level sets of the cumulative distribution function (c.d.f.) [Di Bernardino et al., 2015]. To weaken this assumption, [Kohler et al., 2009] assume some smoothness properties on the regression function. It requires that the partial derivatives of the regression function r are k-Hölderian with a constant C (for further details see Definition 1 in [Kohler et al., 2009]). So far, this estimator was the only one we found adapted to estimate our CCTE. With the "warping trick" we intend to further weaken the assumptions on the regression function.

Warped estimators [Kerkyacharian and Picard, 2004, Yang, 1981] require very few assumptions on the support of $\mathbf{X}$. If we assume, for a sake of clarity, that $d = 1$, the warped method is based on the introduction of the auxiliary function $g = r \circ F_{\mathbf{X}}^{-1}$, where $F_{\mathbf{X}} : x \in \mathbb{R} \mapsto \mathbb{P}(\mathbf{X} \leq x)$ is the c.d.f. of the design $\mathbf{X}$. First, an estimator $\hat{g}$ is proposed for g, and then, the regression r is estimated using $\hat{g} \circ \hat{F}$, where $\hat{F}$ is the empirical c.d.f. of $\mathbf{X}$. This strategy has already been applied in the regression setting using projection methods [Chagny, 2013, Kerkyacharian and Picard, 2004, Pham Ngoc, 2009] but also for other estimation problems (conditional density estimation, hazard rate estimation based on randomly right-censored data, and c.d.f. estimation from current-status data, see *e.g.* [Chagny, 2015, Chesneau and Willer, 2015]). If the warping device permits to weaken the assumptions on the design support, the warped estimators also depend on a unique bandwidth (for $d = 1$), whereas the ratio form of the kernel estimator (3.2) requires the selection of two smoothing parameters (one for the numerator, one for the denominator). In return, the c.d.f. $F_{\mathbf{X}}$ of $\mathbf{X}$ has to be estimated, but this can simply be done using its empirical counterpart. This does not deteriorate the optimal convergence rate, since this estimator converges at a parametric rate. A data-driven selection of the unique bandwidth involved in the resulting warped kernel estimator, in the spirit of [Goldenshluger and Lepski, 2011] leads to non-asymptotic risk bounds when $d = 1$ [Chagny, 2015]. To our knowledge, this adaptive estimation has never been carried out for a ratio regression estimator, the only reference on this subject being [Ngoc Bien, 2014] who assumes that the design $\mathbf{X}$ has a known uniform distribution.

Nevertheless, the extension of the warped strategy to the multivariate framework is not trivial, and we propose to deal with this problem here. The key question is to take into account the dependence between the multivariate components of each $\mathbf{X}_i$. We propose to tackle this problem by using copulas, that permit to describe the dependence structure between random variables [Jaworski et al., 2010, Sklar, 1959]. The price to pay is the additional estimation of the copula density of the design: the complete strategy requires the plug-in of such estimator in the final warped regression estimator. The sequel is thus organized as follows. I explain in Section 3.2 how we extend the [Yang, 1981] estimator to the multivariate design framework. For sake of clarity, I first concentrate on the simple toy case of known design distribution (Section 3.3): under mild assumptions, we derive (i) a non-asymptotic oracle type inequality for an integrated criterion for a warped kernel estimator with a data-driven bandwidth selected with a Lepski-type method, and (ii) an optimal convergence rate over possibly anisotropic functional classes. Then, a kernel copula estimator that also adapts automatically to the unknown smoothness of the design is exhibited and studied in Section 3.4. An oracle type inequality is then proved. Finally, warped regression estimation with unknown copula density is the subject of Section 3.5: as expected, the risk of the final estimator depends on the risks of both the copula estimator and the regression estimator with known design density.

3.2 Multivariate warping strategy

If $d = 1$, the warping device is based on the transformation $F(X_i)$ of the data X_i, $i = 1, \ldots, n$. For $d > 1$, a natural extension is to use $F_l(X_{l,i})$, for $l = 1, \ldots, d$ and $i = 1, \ldots, n$, where F_l is the marginal c.d.f. of X_l. Let us introduce $\widetilde{F}_{\mathbf{X}} : \mathbf{x} = (x_l)_{l=1,\ldots,d} \in \mathbb{R}^d \mapsto (F_1(x_1), \ldots, F_d(x_d))$. Assume that $\widetilde{F}_{\mathbf{X}}^{-1} : \mathbf{u} \in [0,1]^d \mapsto (F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))$ exists, and let

$$g = r \circ \widetilde{F}_{\mathbf{X}}^{-1},$$

in such a way that $r = g \circ \widetilde{F}_{\mathbf{X}}$. If we consider that the marginal variables X_l of $\mathbf{X}$ are independent, the estimator of [Yang, 1981] can immediately be adapted to the multivariate setting. We set

$$\hat{g}_{\perp\!\!\!\perp} : \mathbf{u} \mapsto \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)) \tag{3.3}$$

to estimate g, and it remains to compose by the empirical counterpart of $\widetilde{F}_{\mathbf{X}}$ to estimate r. However, a dependence between the coordinates $X_{l,i}$ of $\mathbf{X}_i$ generally appears. The usual model for this dependence use a copula C [Jaworski et al., 2010, Sklar, 1959], and the c.d.f. $F_{\mathbf{X}}$ of $\mathbf{X}$ can be written

$$F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d)) = C(\widetilde{F}_{\mathbf{X}}(\mathbf{x})).$$

Denoting the copula density by c, we have

$$c(\mathbf{u}) = \frac{\partial^d C}{\partial u_1 \dots \partial u_d}(\mathbf{u}), \quad \mathbf{u} \in [0;1]^d,$$

and the density $f_{\mathbf{X}}$ of $\mathbf{X}$ can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}) = c(\widetilde{F}_{\mathbf{X}}(\mathbf{x})) \prod_{l=1}^{d} f_l(x_l), \quad \mathbf{x} = (x_l)_{l=1,\dots,d} \in \mathbb{R}^d,$$

where $(f_l)_{l=1,\dots,d}$ are the marginal densities of $\mathbf{X} = (X_1, \dots, X_d)$. It can then be proved that the previous estimator given by (3.3) estimates cg and not g (see the computation (3.7) below). As a consequence, we propose to set, as an estimator for g,

$$\widehat{g}_{\mathbf{h}}(\mathbf{u}) = \frac{1}{n\widehat{c}(\mathbf{u})} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in [0,1]^d,$$

where $\widehat{c}$ is an estimator of the copula density. We denote by $\widehat{\widetilde{F}}_{\mathbf{X}} : \mathbb{R}^d \to [0;1]^d$ the empirical multivariate marginal c.d.f.:

$$\widehat{\widetilde{F}}_{\mathbf{X}} = (\widehat{\widetilde{F}}_{\mathbf{X},1}, \dots, \widehat{\widetilde{F}}_{\mathbf{X},d}), \; \widehat{\widetilde{F}}_{\mathbf{X},l}(x_l) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_{l,i} \le x_l}, \; x_l \in \mathbb{R}, l \in \{1, \dots, d\}, \tag{3.4}$$

and finally set

$$\widehat{r}_{\mathbf{h}}(\mathbf{x}) = \widehat{g}_{\mathbf{h}} \circ \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\widehat{c}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}))} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)) \tag{3.5}$$

to rebuild our target function r from the data. In the sequel, we denote by $\| \cdot \|$ the (unweighted) L^2-norm on $L^2(\mathbb{R}^d)$ and, more generally, by $\| \cdot \|_{L^p(\Theta)}$ the classical L^p-norm on a set Θ.

3.3 The simple case of known design distribution

3.3.1 Collection of kernel estimators

For the sake of clarity, we first consider the regression estimation problem with a known design distribution. In this section, the copula density c and the marginal c.d.f. $\widetilde{F}_{\mathbf{X}}$ are consequently considered to be known. Thus, (3.5) becomes

$$\widehat{r}_{\mathbf{h}}(\mathbf{x}) = \widehat{g}_{\mathbf{h}} \circ \widetilde{F}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{nc(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}) - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)), \tag{3.6}$$

where we denote $\widehat{g}_{\mathbf{h}}(\mathbf{u}) = \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i))/(nc(\mathbf{u}))$, $\mathbf{u} \in [0,1]^d$. The following computation enlights the definitions (3.5) and (3.6) above. For any $\mathbf{u} \in [0,1]^d$,

$$
\begin{aligned}
\mathbb{E}[\widehat{g}_{\mathbf{h}}(\mathbf{u})] &= \mathbb{E}\left[\frac{YK_{\mathbf{h}}(\mathbf{u}-\widetilde{F}_{\mathbf{X}}(\mathbf{X}))}{c(\mathbf{u})}\right], \\
&= \mathbb{E}\left[\frac{r(\mathbf{X})K_{\mathbf{h}}(\mathbf{u}-\widetilde{F}_{\mathbf{X}}(\mathbf{X}))}{c(\mathbf{u})}\right], \\
&= \frac{1}{c(\mathbf{u})}\int_{\mathbb{R}^d} r(\mathbf{x})K_{\mathbf{h}}(\mathbf{u}-\widetilde{F}_{\mathbf{X}}(\mathbf{x}))c(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))\prod_{l=1}^{d} f_l(x_l)d\mathbf{x}, \\
&= \frac{1}{c(\mathbf{u})}\int_{[0,1]^d} g(\mathbf{u}')K_{\mathbf{h}}(\mathbf{u}-\mathbf{u}')c(\mathbf{u}')d\mathbf{u}', \\
&= \frac{K_{\mathbf{h}}\star(cg\mathbf{1}_{[0,1]^d})}{c}(\mathbf{u}). \qquad (3.7)
\end{aligned}
$$

where $\star$ is the convolution product. For small $\mathbf{h}$, the convolution product $K_{\mathbf{h}}\star(cg)\mathbf{1}_{[0,1]^d}$ is supposed to be closed to cg:this justifies that $\widehat{g}_{\mathbf{h}}$ is suitable to estimate g, and $\widehat{r}_{\mathbf{h}}$ suits well to recover the target r.

3.3.2 Risk of the estimator with fixed bandwidth

A global integrated criterion is considered to study the properties of our estimators. Let $\|.\|_{f_{\mathbf{X}}}$ be the classical L^2-norm on the space of squared integrable functions with respect to the Lebesgue measure weighted by $f_{\mathbf{X}}$ on A: for any function t in this space,

$$
\|t\|^2_{f_{\mathbf{X}}} = \int_A t^2(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \int_{\widetilde{F}_{\mathbf{X}}(A)} t^2\circ\widetilde{F}_{\mathbf{X}}^{-1}(\mathbf{u})c(\mathbf{u})d\mathbf{u}.
$$

The mean integrated squared risk of the estimator $\widehat{r}_{\mathbf{h}}$ can thus be written

$$
\mathbf{E}\|\widehat{r}_{\mathbf{h}}-r\|^2_{f_{\mathbf{X}}} = \mathbf{E}\int_A\left(\widehat{r}_{\mathbf{h}}(\mathbf{x})-r(\mathbf{x})\right)^2 f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \mathbf{E}\int_{\widetilde{F}_{\mathbf{X}}(A)}\left(\widehat{g}_{\mathbf{h}}(\mathbf{u})-g(\mathbf{u})\right)^2 c(\mathbf{u})d\mathbf{u}
$$

and, using a classical bias-variance decomposition, we have $\mathbf{E}[\|\widehat{r}_{\mathbf{h}}-r\|^2_{f_{\mathbf{X}}}] = B(\mathbf{h})+V(\mathbf{h})$, where

$$
\begin{aligned}
B(\mathbf{h}) &= \int_{\widetilde{F}_{\mathbf{X}}(A)} c(\mathbf{u})\left(\frac{K_{\mathbf{h}}\star(cg\mathbf{1}_{[0,1]^d})}{c}(\mathbf{u})-g(\mathbf{u})\right)^2 d\mathbf{u}, \\
V(\mathbf{h}) &= \int_{\widetilde{F}_{\mathbf{X}}(A)} c(\mathbf{u})\left(\widehat{g}_{\mathbf{h}}(\mathbf{u})-\frac{K_{\mathbf{h}}\star(cg\mathbf{1}_{[0,1]^d})}{c}(\mathbf{u})\right)^2 d\mathbf{u}.
\end{aligned}
$$

To obtain upper-bounds for these two terms, we introduce the following assumptions.

($H_{c,low}$) The copula density is lower bounded: $\exists m_C>0$, $\forall\mathbf{u}\in[0,1]^d$, $c(\mathbf{u})\geq m_C$.

($H_{cg,\beta}$) The function $cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)}$ belongs to an anisotropic Nikol'skiĭ ball $\mathcal{N}_2(\beta,L)$, with $L>0$ and $\beta=(\beta_1,\dots,\beta_d)\in(\mathbb{R}_+^*)^d$ [Nikol'skiĭ, 1975]. This is the set of functions $f:\mathbb{R}^d\mapsto\mathbb{R}$ such that f admits derivatives with respect to x_l up to the order $\lfloor\beta_l\rfloor$ (where $\lfloor\beta_l\rfloor$ denotes the largest integer less than β_l), and

(i) for all $l\in\{1,\dots,d\}$, $\|\partial^{\lfloor\beta_l\rfloor}f/(\partial x_l)^{\lfloor\beta_l\rfloor}\|_{L^2(\mathbb{R}^d)}\leq L$,

(ii) for all $l\in\{1,\dots,d\}$ and $t\in\mathbb{R}$,

$$
\int_{\mathbb{R}^d}\left|\frac{\partial^{\lfloor\beta_l\rfloor}f}{(\partial x_l)^{\lfloor\beta_l\rfloor}}(x_1,\dots,x_{l-1},x_l+t,x_{l+1},\dots,x_d)-\frac{\partial^{\lfloor\beta_l\rfloor}f}{(\partial x_l)^{\lfloor\beta_l\rfloor}}(\mathbf{x})\right|^2 d\mathbf{x}\leq L^2|t|^{2(\beta_l-\lfloor\beta_l\rfloor)}.
$$

($H_{K,\ell}$) The kernel K is of order $\ell\in(R_+)^d$, *i.e.*

(i) $\forall l \in \{1,\dots,d\}$, $\forall k \in \{1,\dots,\ell_l\}$, $\int_{\mathbb{R}^d} x_l^k K(\mathbf{x})d\mathbf{x} = 0$.

(ii) $\forall l \in \{1,\dots,d\}$, $\int_{\mathbb{R}^d} (1+x_l)^{\ell_l} K(\mathbf{x})d\mathbf{x} < \infty$.

Note that Assumption $(H_{c,low})$ is verified for the Frank copula. In the the bivariate case, the Frank copula (with parameter $\theta \neq 0$) is defined by

$$C(u,v) = -\frac{1}{\theta}\log\left(1 + \frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)-1}{\exp(-\theta)-1}\right).$$

For other copulas we always can restrict the estimation to a subset $A \subset \mathbb{R}^d$ to exclude problematic points of $\widetilde{F}_{\mathbf{X}}(A)$ (see [Balakrishnan and Lai, 2009] for more details on other specific copula densities). Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ are classical for non-parametric multivariate kernel estimation [Comte and Lacour, 2013, Goldenshluger and Lepski, 2011] to control the bias term $B(\mathbf{h})$. Assumption $(H_{K,\ell})$ is not restrictive since a wide range of kernels could be chosen. In $(H_{cg,\beta})$, the index β measures the smoothness of the function cg. The difficulty is that this smoothness assumption is made directly on cg, and not on the targeted function r. It is for example satisfied if the two functions c and g separately belong to $\mathcal{N}_2(\beta, L')$ $(L' > 0)$, for β such that each $\beta_l \leq 1$, $l \in \{1,\dots,d\}$. Furthermore, the fact that the assumption is carried by the auxiliary function g and not r is classical in warped methods [Chagny, 2015, Pham Ngoc, 2009]. Another solution is to consider weighted spaces: lots of details can be found in [Kerkyacharian and Picard, 2004]. Using these assumptions we state the following proposition.

Proposition 3.3.1 *Assume $(H_{c,low})$, $(H_{cg,\beta})$ and $(H_{K,\ell})$ for an index $\ell \in \mathbb{R}_+^d$ such that $\ell_j \geq \lfloor \beta_j \rfloor$. Then,*

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|_{f_{\mathbf{X}}}^2] \leq \frac{1}{m_c}\left(L\sum_{l=1}^{d} h_l^{2\beta_l} + \|K\|^2 \mathbb{E}[Y_1^2]\frac{1}{nh_1\dots h_d}\right).$$

This is a non-asymptotic bias-variance upper bound for the quadratic risk. The first term of the right-hand-side of the inequality of Proposition 3.3.1 is an upper-bound for the bias term $B(\mathbf{h})$. The second one bounds the variance term. Another choice would have been to kept $B(\mathbf{h})$ in the inequality (in this case, Assumptions $(H_{cg,\beta})$ and $(H_{K,l})$ are not required). But now we could immediately deduce the following convergence rate.

Corollary 3.3.1 *Under the same assumptions as Proposition 3.3.1, there exists a bandwidth $\mathbf{h}(\beta)$ such that*

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h}(\beta)} - r\|_{f_{\mathbf{X}}}^2] = O\left(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+d}}\right),$$

where $\bar{\beta}$ is the harmonic mean of $\beta_1,\dots,\beta_d$: $d\bar{\beta}^{-1} = \beta_1^{-1} + \dots + \beta_d^{-1}$.

Thus the usual convergence rate in multivariate non-parametric estimation can be achieved by our estimator, provided that its bandwidth is carefully chosen. Here, the bandwidth $\mathbf{h}(\beta)$ that minimizes the upper-bound of the inequality of Proposition 3.3.1 depends on the smoothness index of the function cg. The challenge of adaptive estimation is to propose a data-driven choice that also leads to an estimator with the same optimal convergence rate.

Remark 3.3.1 *Note that Corollary 3.3.1 implies that our estimator satisfies Assumption* **A3'** *of the previous chapter.*

3.3.3 Estimator selection

Let $\mathcal{H}_n \subset (\mathbb{R}_+^*)^d$ a finite bandwidth collection. We set

$$\widehat{B}(\mathbf{h}) = \max_{\mathbf{h}' \in \mathcal{H}_n}\left\{\left\|\frac{K_{\mathbf{h}} \star (c\widehat{g}_{\mathbf{h}'})}{c} \circ \widetilde{F}_{\mathbf{X}} - \widehat{r}_{\mathbf{h}'}\right\|_{f_{\mathbf{X}}}^2 - \widehat{V}(\mathbf{h}')\right\}_+$$

with

$$\widehat{V}(\mathbf{h}) = \kappa \frac{\sum_{i=1}^{n} Y_i^2}{\widehat{m}_c} \frac{1}{nh_1 \dots h_d}, \tag{3.8}$$

where $\kappa > 0$ is a tuning constant and $\widehat{m}_c$ an estimator for m_c. We define

$$\widehat{\mathbf{h}} = \arg \min_{\mathbf{h} \in \mathcal{H}_n} \{\widehat{B}(\mathbf{h}) + \widehat{V}(\mathbf{h})\}, \tag{3.9}$$

and the final estimator $\widehat{r}_{\widehat{\mathbf{h}}}$. The criterion (3.9), inspired from [Goldenshluger and Lepski, 2011], is known to mimic the optimal "bias-variance" trade-off that has to be realized in a data-driven way. We also introduce $\widetilde{\mathbf{h}} = \arg\min_{\mathbf{h}\in\mathcal{H}_n}\{\widetilde{B}(\mathbf{h}) + \widetilde{V}(\mathbf{h})\}$ with

$$\widetilde{B}(\mathbf{h}) = \max_{\mathbf{h}' \in \mathcal{H}_n} \left\{ \left\| \frac{K_{\mathbf{h}} \star (c\widehat{g}_{\mathbf{h}'} \mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})}{c} \circ \widetilde{F}_{\mathbf{X}} - \widehat{r}_{\mathbf{h}'} \right\|_{f_{\mathbf{X}}}^2 - \widetilde{V}(\mathbf{h}') \right\}_+$$

and

$$\widetilde{V}(\mathbf{h}) = \kappa_0 \frac{\mathbb{E}[Y_1^2]}{m_c} \frac{1}{nh_1 \dots h_d}, \quad \kappa_0 > 0.$$

We start with the study of the estimator $\widehat{r}_{\widetilde{h}}$. The collection $\mathcal{H}_n$ is chosen such that

$$\exists \alpha_0 > 0, \kappa_1 > 0, \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \cdots h_d} \leq \kappa_1 n^{\alpha_0} \tag{3.10}$$

$$\text{and } \forall \kappa_1 > 0, \exists C_0 > 0, \sum_{\mathbf{h} \in \mathcal{H}_n} \exp\left(-\frac{\kappa_1}{h_1 \cdots h_d}\right) \leq C_0.$$

These assumptions are very common to derive such estimators [Chagny, 2015, Comte and Lacour, 2013]. For example, $\mathcal{H}_n = \{k_1^{-1} \cdots k_d^{-1}, k_l \in \{1, \dots, \lfloor n^{1/r} \rfloor\}, l = 1, \dots, d\}$ satisfies them with $\alpha_0 = 2d/r$.

We also introduce additional assumptions:

(H_ε) The noise ε is $p+2$ integrable, for some $p > 2\alpha_0$: $\mathbb{E}[|\varepsilon|^{2+p}] < \infty$.

$(H_{c,high})$ The copula density is upper-bounded over $[0,1]^d$: $\exists M_C > 0,\ \forall \mathbf{u} \in [0,1]^d,\ c(\mathbf{u}) \leq M_C$.

The assumption $(H_{c,high})$ is quite restrictive. However, it will also be required for copula density estimation (Section 3.4), and is classical for adaptive density estimation purpose. It is satisfied also by the Frank copula density for example. Moreover, the same upper-bound is assumed in [Autin et al., 2010]. It is then possible to set the following upper bound.

Theorem 3.3.1 *Assume that $\mathcal{H}_n$ satisfies (3.10) and assume also (H_ε), $(H_{c,low})$ and $(H_{c,high})$. Then there exist two constants c_1 et c_2 such that*

$$\begin{aligned} \mathbf{E}[\|\widehat{r}_{\widetilde{\mathbf{h}}} - r\|_{f_{\mathbf{X}}}^2] \quad &\leq \quad c_1 \min_{\mathbf{h} \in \mathcal{H}_n} \left\{ \frac{1 + \|K\|_{L_1([0,1]^d)}^2}{m_c} \left\| K_{\mathbf{h}} \star (cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)}) - cg \right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2 \right. \\ &\left. + \|K\|^2 \mathbb{E}[Y_1^2] \frac{1}{nm_c h_1 \dots h_d} \right\} + \frac{c_2}{n}. \end{aligned}$$

This result is an oracle-type inequality which assesses that the selected estimator performs as well as the best estimator of the collection $(\widehat{r}_{\mathbf{h}})_{\mathbf{h}\in\mathcal{H}_n}$, up to multiplicative constants and a remainder term: it achieves the best bias-variance trade-off (see Proposition 3.3.1). If we add Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ (for an index $\ell \in \mathbb{R}_+^d$ such that $\ell_j \geq \lfloor \beta_j \rfloor$, $j = 1, \dots, d$) to the assumptions of Theorem 3.3.1, we obtain the same convergence rate as the one of Corollary 3.3.1 for the estimator $\widehat{r}_{\widetilde{\mathbf{h}}}$. The difference is that the smoothness index β is not required to build $\widehat{r}_{\widetilde{\mathbf{h}}}$: our selected estimator automatically adapts to the unknown smoothness of the function cg, as it is stated in the following corollary.

Corollary 3.3.2 *Under the same assumptions as Theorem 3.3.1, if we also assume that* $(H_{cg,\beta})$ *and* $(H_{K,\ell})$ *are fulfilled for an index* $\ell \in \mathbb{R}_+^d$ *such that* $\ell_j \geq \lfloor \beta_j \rfloor$*, we have*

$$\mathbf{E}[\|\widehat{r}_{\widetilde{\mathbf{h}}} - r\|_{f_X}^2] = O\left(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+d}}\right),$$

where $\bar{\beta}$ *is the harmonic mean of* $\beta_1, \dots, \beta_d$*:* $d\bar{\beta}^{-1} = \beta_1^{-1} + \cdots + \beta_d^{-1}$.

To switch from $\widehat{r}_{\widetilde{\mathbf{h}}}$ to $\widehat{r}_{\widehat{\mathbf{h}}}$, it remains then to first replace the unknown expectation $\mathbb{E}[Y_1^2]$ by its empirical counterpart $\frac{1}{n}\sum_{i=1}^n Y_i^2$ and to change $\widetilde{V}(\mathbf{h})$ in $\widehat{V}(\mathbf{h})$. This is quite classical, and can be done for example like in Theorem 3.4 p.465 of [Brunel and Comte, 2005]. The replacement of $\widetilde{F}_{\mathbf{X}}$ by its empirical counterpart is discussed at the end of Section 3.5. Before that, the next section is devoted to the estimation of the copula density.

3.4 Copula density estimation

The estimator defined by (3.5) involves an estimator $\widehat{c}$ of the copula density c that was assumed to be known in the previous section. This section is devoted to the question of copula density estimation. An adaptive estimator based on wavelets is defined in [Autin et al., 2010] but, to be consistent with the previous kernel regression estimator already chosen, we propose to use the kernel estimator defined by [Fermanian, 2005]. Consider $\mathbf{b} = {}^t(b_1, \dots, b_d) \in (\mathbb{R}_+^*)^d$ a multivariate bandwidth, a kernel $W_{\mathbf{b}}(\mathbf{u}) = W_{1,b_1}(u_1)W_{2,b_2}(u_2)\dots W_{d,b_d}(u_d)$, with $W_{l,b_l}(u) = W_l(u/b_l)/b_l$ for $b_l > 0$, and $W_l : \mathbb{R} \to \mathbb{R}$ such that $\int_0^1 W_l(u)du = 1$, $l \in \{1, \dots, d\}$. Let us introduce

$$\widehat{c}_{\mathbf{b}}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n W_{\mathbf{b}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in [0,1]. \tag{3.11}$$

The estimator is very close to the classical kernel density estimator, up to the warping of the data through the empirical c.d.f. Remark that if we replace the estimator $\widehat{\widetilde{F}}_{\mathbf{X}}$ in (3.11) by its target $\widetilde{F}_{\mathbf{X}}$, like in the previous section, then $\widehat{c}_{\mathbf{b}}(\mathbf{u})$ is the density estimator of the random vector $(F_1(X_1), \dots, F_d(X_d))$, with uniformly distributed marginal distributions. We easily obtain the following upper-bound for the risk of the copula density estimator when the marginal distributions are known:

$$\mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|_{L^2([0,1]^d)}^2\right] \leq \|W_{\mathbf{b}} \star c - c\|_{L^2([0,1]^d)}^2 + \frac{\|W\|_{L^2([0,1]^d)}^2}{nb_1 \cdots b_d}. \tag{3.12}$$

The results of [Fermanian, 2005] are asymptotic. Since our goal is to prove non-asymptotic adaptive upper-bounds, the Goldenshluger-Lepski method allows us to select a bandwidth $\widehat{\mathbf{b}}$ among a finite collection $\mathcal{B}_n \subset (\mathbb{R}_+^*)^d$. The collection $\mathcal{B}_n$ should satisfy

$$\exists \alpha_1 > 0,\ \kappa_2 > 0,\ \sum_{\mathbf{b} \in \mathcal{B}_n} \frac{1}{b_1 \cdots b_d} \leq \kappa_2 n^{\alpha_1},$$

and one of the following constraints

$$|\mathcal{B}_n| \leq \ln(n), \quad \text{or} \quad \forall \kappa_3 > 0,\ \exists C_0 > 0,\ \sum_{\mathbf{b} \in \mathcal{B}_n} \exp\left(-\frac{\kappa_3}{b_1 \cdots b_d}\right) \leq C_0, \tag{3.13}$$

where $|\mathcal{B}_n|$ is the cardinal of the set $\mathcal{B}_n$. These assumptions are similar to (3.10). Let

$$\widehat{B}_c(\mathbf{b}) = \max_{\mathbf{b}' \in \mathcal{B}_n} \left\{\|W_{\mathbf{b}} \star \widehat{c}_{\mathbf{b}'} - \widehat{c}_{\mathbf{b}'}\|_{L^2([0,1]^d)}^2 - V_c(\mathbf{b}')\right\}_+$$

with

$$V_c(\mathbf{b}) = \kappa_c \frac{\|W\|_{L_1([0,1]^d)}^2 \|W\|_{L^2([0,1]^d)}^2}{nb_1 \cdots b_d}, \quad \kappa_c > 0. \tag{3.14}$$

Like above for regression estimation, $\widehat{B}_c$ stands for an empirical counterpart of the bias term of the risk, and V_c has the same order as the variance term (compare to (3.12)). An oracle-type inequality could be derived for the final copula density estimator $\widehat{c}_{\widehat{\mathbf{b}}}$, with $\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}\in\mathcal{B}_n}\{\widehat{B}_c(\mathbf{b}) + V_c(\mathbf{b})\}$.

Proposition 3.4.1 *Assume $(H_{c,high})$, and assume that the marginal c.d.f. of the vector $\mathbf{X}$ are known. Then, there exist some non-negative constants c_1 and c_2 such that*

$$\mathbb{E}\left[\|\widehat{c}_{\widehat{\mathbf{b}}} - c\|^2_{L^2([0,1]^d)}\right] \leq c_1 \min_{\mathbf{b}\in\mathcal{B}_n}\left\{\|W_{\mathbf{b}} \star c - c\|^2_{L^2([0,1]^d)} + \frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}\right\} + \frac{c_2\ln(n)}{n}.$$

Note that the L_1-norm of the kernel does not appear in (3.12), but only in the variance term of the Goldenshluger-Lepski method, namely (3.14), for technical reasons (more details on the proof in [Chagny et al., 2017]). The logarithmic term in the upper-bound of the inequality can be avoided by assuming the second part of (3.13), instead of $|\mathcal{B}_n| \leq \ln(n)$. Like the tuning constant κ in $\widehat{V}$ (see (3.8)), the constant κ_c in (3.14) has to be calibrated. The bound that we obtain in the proof is unfortunately not accurate (this is a consequence of numerous technical upper bound, based on a concentration inequality), and cannot be used for practical purpose. The tuning of this parameter is discussed in the simulation section of [Chagny et al., 2017].

Proposition 3.4.1 also permits to derive an adaptive convergence rate for our copula density estimator: if the copula density c belongs to a Nikol'skiĭ ball $\mathcal{N}_2(\alpha, L')$ for $L' > 0$ and $\alpha = {}^t(\alpha_1, \ldots, \alpha_d) \in (\mathbb{R}_+^*)^d$, and if the kernel W is of order $\ell \in \mathbb{R}_+^d$ such that $\ell_j \geq \lfloor \alpha_j \rfloor$ for $j = 1, \ldots, d$, (see Assumption $(H_{K,\ell})$), then $\widehat{c}_{\widehat{\mathbf{b}}}$ automatically achieves the convergence rate $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+d}}$ where $\bar{\alpha}$ is the harmonic mean of the components of α. Following [Autin et al., 2010], this is also the lower bound for the minimax risk, and thus our estimators is minimax optimal (with no additional logarithm factor, comparing to Corollary 4.1 of [Autin et al., 2010]).

3.5 Plug-in regression estimate

Now we consider the general case of unknown copula density c to estimate the regression function r. The idea is to plug the kernel estimator $\widehat{c}_{\mathbf{b}}$ (defined by (3.11)) of c in (3.6) for a well-chosen bandwidth $\mathbf{b}$. We consider the case of fixed bandwidth, both for the regression and the copula estimators. Let us plug in $\hat{r}_{\mathbf{h}}$ the estimator $\widehat{c}_{\mathbf{b}}$: for any $\mathbf{b}, \mathbf{h} > 0$, under Assumption $(H_{c,low})$,

$$\widehat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = \frac{1}{n\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))}\sum_{i=1}^n Y_i K_{\mathbf{h}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}) - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i))\mathbf{1}_{\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))\geq m_c/2}, \quad \mathbf{x} \in A. \tag{3.15}$$

This means that $\widehat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = ((c \times \widehat{g}_{\mathbf{h}})/\widehat{c}_{\mathbf{b}}) \circ \widetilde{F}_{\mathbf{X}}(\mathbf{x})\mathbf{1}_{\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))\geq m_c/2}$. To make the estimator fully computable, one needs to know the lower bound m_c of the copula: in practice it is possible to replace it by a lower bound of an estimator. To avoid making the proofs more technical and cumbersome, and since the minimum of classical copula density are very small (see the example of the Franck copula used for the simulation study in [Chagny et al., 2017]), we choose to not consider the problem from a theoretical point of view. We obtain the following upper-bound for our ratio estimator. Its risk has the order of magnitude of the worst risk between the risk of $\widehat{r}_{\mathbf{h}}$ and $\widehat{c}_{\mathbf{b}}$.

Proposition 3.5.1 *Assume $(H_{c,low})$ and $(H_{c,high})$. Then,*

$$\begin{aligned}\mathbf{E}[\|\widehat{r}_{\mathbf{h},\mathbf{b}} - r\|^2_{f_{\mathbf{X}}}] \leq \frac{4M_c}{m_c^2} \quad & \{\, 2M_c\mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|^2_{f_{\mathbf{X}}}] \\ & +(2\|g\|^2_{L^\infty(\widetilde{F}_{\mathbf{X}}(A))} + \|g\|^2_{L^2(\widetilde{F}_{\mathbf{X}}(A))})\mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|^2_{L^2([0,1]^d)}\right]\Big\}.\end{aligned}$$

The result is not surprising, and we cannot expect to obtain a sharper bound for the plug-in estimator. We thus have to add smoothness assumptions both on the regression function and on the copula density to derive the convergence rate of the plug-in estimator.

Finally, to obtain a fully computable estimator, one needs to replace the c.d.f. $\widetilde{F}_{\mathbf{X}}$ by its empirical counterpart introduced in (3.4). The switch is not a problem: the idea is that the empirical c.d.f. converges at a parametric rate, that does not deteriorate our slower non-parametric decrease of the risk. The multivariate setting does not change anything for the substitution compare to the univariate case. The scheme of the switching can now be considered as classical, since it has been widely detailed both by [Kerkyacharian and Picard, 2004] and [Chagny, 2015].

Conclusion

In this chapter a new estimator of the regression function is given. This estimator does not require a bounded support for $\mathbf{X}$. In the complete simulation study we provided in [Chagny et al., 2017], we illustrate the fact that a plug-in estimation of the marginals does not seem to to affect the quality of the estimation and establish the relevance of the bandwidth selection for the copula estimator. We also see that our estimator performs better than the classical Nadaraya-Watson one. Finally, we compare our estimator to the one proposed by [Kohler et al., 2009] (that is the one we used to estimate the CCTE in Chapter 2). We obtain better results but for regression functions that not necessarily satisfies the required theoretical assumptions of smoothness. A remaining work is to use this estimator to compute the CCTE estimator presented in Chapter 2.

Conclusion and prospects

In this first part, I presented some of my contributions in the area of level sets estimation and associated risk measures. However many questions remain.

Normality Regression Level Sets

As said in the introduction, I studied during my PhD the problem of estimating regression level sets. In particular we proposed in [Laloë and Servien, 2013] a plug-in estimator and proved its consistency.

In [Mason and Polonik, 2009] a similar estimator of the density level sets is studied. Here is their set-up. Let $X_1, X_2, \ldots$ be i.i.d. with density function f, and consider the kernel density estimator of f based on a sample $X_1, \ldots, X_n$,

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n^{1/d}}\right) \ , \ x \in \mathbb{R}^d,$$

where K is a kernel and $h_n > 0$ is a bandwidth parameter. Define the α level set of f and its plug-in estimator by

$$\mathcal{L}_f(\alpha) = \{x : f(x) \geq \alpha\} \quad \text{and} \quad \mathcal{L}_{f,n}(\alpha) = \{x : f_n(x) \geq \alpha\}.$$

Let G be a positive measure dominated by Lebesgue measure λ. Their interest is to establish the asymptotic normality of $\lambda(\mathcal{L}_{f,n}(\alpha)\triangle\mathcal{L}_f(\alpha))$. Their main result states that under suitable regularity conditions, there exists a constant $0 < \sigma_G^2 < \infty$ such that

$$(n/h_n)^{1/4}\{\lambda(\mathcal{L}_{f,n}(\alpha)\triangle\mathcal{L}_f(\alpha))\} \to \sigma_G Z$$

as $n \to \infty$, where Z denotes a standard normal random variable.

It should be possible to obtain similar results for the level sets of the regression function, and maybe for the level sets of the c.d.f. I am now supervising a Master's internship on this subject.

Ultimately, I'm interested by the approach of [Cuevas et al., 2006]. In this article the authors consider the problem of estimating general level sets (regardless to the function of interest: density, regression, distribution,...). Under mild condition on the interest function g they derive consistency results for the estimator of the level sets. As an example, given an estimator g_n of g, they obtain that for the Hausdorff distance, the rate of consistency is of the same order as $\|g_n - g\|_\infty$. Following this philosophy I would like to get the necessary assumptions on the interest function to obtain asymptotic normality for kernel plug-in estimators of general level sets. It could be a nice start for a PhD subject.

Generalizations of the CCTE

In a recent work, [Di Bernardino et al., 2018] consider the problem of estimating extreme risk region and provide an *out-sample* method to estimate the directional multivariate quantiles recently introduced in [Torres et al., 2015, 2017]. There are here two major interests. First the inclusion of a direction parameter allows to analyse the observations from several interesting perspectives. Second the authors consider the case of very high levels α.

Following this approach, I would like to use the approach proposed in [Di Bernardino et al., 2018] in order to provide an estimation of the CCTE (Section 2.2) for very high levels. I also think it would be interesting to consider an alternative (and complementary) approach to deal with the problem of the direction for the level sets by replacing the c.d.f. by a depth function [Tukey, 1974]. This would provide an estimated cost associated to high level of risks regardless to a specified direction.

Extreme CCTE

The idea here is to characterize the points of $\mathcal{L}(\alpha)$ at very high levels, using the heuristic ideas of the bivariate quantile parametrization given in [Dehaan and Huang, 1995] extended to a general multivariate context by [Di Bernardino et al., 2018]. Therefore, we will have to adopt conditions from Extreme Value Theory. In particular we will have to impose assumptions on the right tail of X:

Given the marginal cumulative distributions F_i $(i = 1, \dots, d)$, consider the tail quantile functions $U_i = \left(\frac{1}{1-F_i}\right)^{\leftarrow}$ (where $\leftarrow$ denotes the left-continuous inverse). We assume as usual in financial or non-life insurance settings that each X_i follows a distribution with heavy right tail: for each $1 \leq i \leq d$ there exists $\gamma_i > 0$ such that for $x > 0$,

$$\lim_{t \to \infty} \frac{U_i(tx)}{U_i(x)} = x^{\gamma_i}.$$

This ensure that the marginal c.d.f. of the X_i belongs to the domain of attraction of a non-degenerate function G_i. Hopefully we could then follow the estimation procedure proposed in [Di Bernardino et al., 2018] to estimate the γ_i and then the G_i. Moreover we will have to add an assumption on the domain of attraction of the joint law of $\mathbf{X}$. From this it would not be difficult to consistently estimate the level sets.

To get an estimation of the CCTE we also will have to impose (in addition to the assumption on $\mathbf{X}$ mentioned above) assumptions on the right-upper tail dependence of $(\mathbf{X}, Y)$. Roughly speaking, considering applications in financial or non-life insurance settings, we expect that high values of Y correspond to high values of the X_i. More precisely we will have to suppose that for all $(\mathbf{x}, y) \in [0, \infty)^{d+1}$ the following limit exists:

$$\lim_{t \to \infty} t\mathbb{P}(1 - F_i(X_i) \leq x_i/t, 1 - F_Y(Y) \leq y/t) := R(\mathbf{x}, y),$$

R determining the stable tail dependence function (see [Beirlant et al., 2004, Drees and Huang, 1998]). We also will have to think about an assumption linking the joint law of $\mathbf{X}$ and Y

From a real-life applications point of view the possibility to extrapolate out-sample is a crucial point. For this reason this extreme CCTE can be applied to several environmental and economic problems. I am working on this subject with E. Di Bernardino, R. Servien and R. Torres.

CCTE using depth level sets

An interesting point in [Di Bernardino et al., 2018] is the inclusion of a direction parameter. Indeed in our definition of the CCTE we only consider the canonical direction. Another way to consider risks in several direction is to replace the c.d.f. by a function who's definition is not based on any direction: a statistical depth.

In 1975, John Tukey proposed a multivariate median which is the “deepest” point in a given data cloud in $\mathbb{R}^d$ [Tukey, 1974]. Since then, this idea has proved extremely fruitful. A rich statistical methodology based on data depth has been developed [Chebana and Ouarda, 2011, Cuevas and Fraiman, 2009, Mosler, 2002]. General notions of data depth have been introduced as well as many special ones. These notions vary regarding their computability and robustness and their sensitivity to reflect asymmetric shapes of the data. According to their different properties they fit to particular applications. The upper level sets of a depth statistic provide a family of set-valued statistics, named *depth-trimmed* or *central regions*. They describe the distribution regarding its location, scale and shape. The most central region serves as a *median*. The notion of depth has been extended from data clouds, that is empirical distributions, to general probability distributions on $\mathbb{R}^d$, thus allowing for laws of large numbers and consistency results.

Formally, for a random variable X with distribution P, a *depth function* is a function D: $(z,\ X) \mapsto D(z|X)$, that satisfies the following restrictions:

D1 *Translation invariant*: $D(z+b|X+b) = D(z|X)$ for all $b \in E$.

D2 *Linear invariant*: $D(Az|AX) = D(z|X)$ for every bijective linear transformation $A : E \to E$.

D3 *Null at infinity*: $\lim_{\|z\|\to\infty} D(z|X) = 0$.

D4 *Monotone on rays*: If a point z^* has maximal depth, that is $D(z^*|X) = \max_{z\in E} D(z|X)$, then for any r in the unit sphere of E the function $\alpha \mapsto D(z^* + \alpha r|X)$ decreases, in the weak sense, with $\alpha > 0$.

D5 *Upper semi-continuous*: The level sets $\mathcal{L}_D(\alpha) = \{z \in E : D(z|X) \geq \alpha\}$ are closed for all α.

It appears that depths function could be perfectly suitable tools to redefine our CCTE. Indeed a depth function D orders data by their degree of centrality. Given a sample, it provides a centre-outward *order statistic*. The depth induces an *outlyingness function* $\mathbb{R}^d \to [0, \infty[$ by

$$Out(z|X) = \frac{1}{D(z|X)} - 1,$$

which is zero at the centre and infinite at infinity. Studying the level-sets $\mathcal{L}_{Out}(\alpha)$ of this *outlyingness function* instead of those of the c.d.f. will naturally leads us to define a CCTE independent to a specified direction.

I just started to work on this idea but so far I am convinced that the plug-in estimation procedure of the level sets presented in our article [Di Bernardino et al., 2013] can be adapted with this notion of Depth. Moreover after discussions with R. Diel we also started to consider the estimation of depth-based level sets on functional spaces. The tricky part is the extension of the definition of differentiability notions for a depth calculated on a functional space.

Warped regression estimation

The aim of Chapter 3 was to extend the “warping” device to a multivariate framework, through the study of regression kernel estimation. When the design distribution is known, the extension of the method can be done and similar results as the ones obtained in the univariate framework (non-asymptotic risk bound and optimal convergence rate) are proved: compare for example Theorem 3.3.1 and Corollary 3.3.2 to Theorem 1 and Corollary 1 of [Chagny, 2015]. When the design distribution is unknown, the challenge is to cope with the possible dependence structure between the coordinates of the design, and the extension can be done only through the additional estimation of the copula density. This can be done separately in an adaptive way, also with kernel estimators.

The copula density kernel estimator satisfies an oracle type inequality. The challenge is thus now to study the warped kernel regression estimate that involve the adaptive copula density estimator. Section 3.5 paves the way of this future study: the risk of the regression estimator with fixed bandwidth and after plug in of the copula estimator (also with fixed bandwidth) is computed. As expected it depends both on the risk of the estimator with known marginal distribution and on the risk of the copula estimator. The convergence rate of the resulting estimator is the worst of these two risks. The next step, which we have not studied yet, is to propose a bandwidth selection rule for the regression estimate computed with the copula density adaptive estimate (that is with selected bandwidth). It requires to replace the copula density c in the GL estimation of the bias term of the risk (see 3.10) by $\hat{c}_{\hat{b}}$. The difficulties are numerous, owing first to the problem of dependence (the regression estimator AND copula estimator AND selected bandwidth depends on the design $\{\mathbf{X}_i\}_{1\leq i\leq n}$): it makes difficult to isolate the risk of the copula adaptive estimator from the risk of the regression estimator with known marginal distribution. A natural idea is to imagine that we have at our disposal an additional sample of the design, independent from the data. We can perhaps then conduct the study in the spirit of [Bertin et al., 2016], who deal with similar questions for conditional density estimators (that involve the plug-in of marginal density estimators). But additional difficulties are probably due to the regression framework: even with the warping device, we are brought back to the initial problem of the Nadaraya kernel estimator. We have to select simultaneously two multiple bandwidths (one for the regression estimate, one for the copula density estimate), and to my knowledge, this have never been done yet in the multivariate regression framework.

Bibliography of the first part

Bibliography

A. Antoniadis, G. Grégoire, and P. Vial. Random design wavelet curve smoothing. *Statistics & Probability Letters*, 35(3):225–232, 1997. URL https://doi.org/10.1016/S0167-7152(97)00017-5.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3): 203–228, 1999. URL https://doi.org/10.1111/1467-9965.00068.

F. Autin, E. Le Pennec, and K. Tribouley. Thresholding methods to estimate copula density. *Journal of Multivariate Analysis*, 101(1):200–222, 2010. URL http://dx.doi.org/10.1016/j.jmva.2009.07.009.

A. Baíllo. Total error in a plug-in estimator of level sets. *Statistics & Probability Letters*, 65(4):411–417, 2003. URL https://doi.org/10.1016/j.spl.2003.08.007.

A. Baíllo, J. Cuesta-Albertos, and A. Cuevas. Convergence rates in nonparametric estimation of level sets. *Statistics & Probability Letters*, 53:27–35, 2001. URL https://doi.org/10.1016/S0167-7152(01)00006-2.

N. Balakrishnan and C. Lai. *Continuous Bivariate Distributions*. Springer Science+Business Media, 2009. URL https://link.springer.com/book/10.1007%2Fb101765.

Y. Baraud. Model selection for regression on a random design. *ESAIM. Probability and Statistics*, 6:127–146, 2002. URL https://doi.org/10.1051/ps:2002007.

J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of extremes*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2004. URL https://doi.org/10.1002/0470012382.

F. Belzunce, A. Castaño, A. Olvera-Cervantes, and A. Suárez-Llorens. Quantile curves and dependence structure for bivariate distributions. *Computational Statistics & Data Analysis*, 51(10):5112–5129, 2007. URL https://doi.org/10.1016/j.csda.2006.08.017.

K. Bertin, C. Lacour, and V. Rivoirard. Adaptive pointwise estimation of conditional density function. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 52(2):939–980, 2016. URL http://dx.doi.org/10.1214/14-AIHP665.

G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM. Probability and Statistics*, 11:272–280, 2007. URL http://dx.doi.org/10.1051/ps:2007019.

E. Brunel and F. Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, 67(3):441–475, 2005. URL https://hal.archives-ouvertes.fr/hal-00138765.

B. Bryan, R. C. Nichol, C. R. Genovese, J. Schneider, C. J. Miller, and L. Wasserman. Active learning for identifying function threshold boundaries. In *Advances in Neural Information Processing Systems 18*, pages 163–170. MIT Press, 2006. URL http://papers.nips.cc/paper/2940-active-learning-for-identifying-function-threshold-boundaries.pdf.

C. Butucea, M. Mougeot, and K. Tribouley. Functional approach for excess mass estimation in the density model. *Electronic Journal of Statistics*, 1:449–472, 2007. doi: 10.1214/07-EJS079. URL https://doi.org/10.1214/07-EJS079.

B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4):999–1023, 2006. URL https://doi.org/10.1016/j.jmva.2005.05.004.

G. Chagny. Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM. Probability and Statistics*, 17:328–358, 2013. URL http://dx.doi.org/10.1051/ps/2011165.

G. Chagny. Adaptive warped kernel estimators. *Scandinavian Journal of Statistics*, 42(2):336–360, 2015. URL https://doi.org/10.1111/sjos.12109.

G. Chagny, T. Laloë, and R. Servien. Multivariate adaptive warped kernel estimation. preprint, 2017. URL https://hal.archives-ouvertes.fr/hal-01616373.

F. Chebana and T. B. Ouarda. Depth based multivariate descriptive statistics with hydrological applications. *Journal of Geophysical Research: Atmospheres*, 116(D10), 2011. URL https://doi.org/10.1029/2010JD015338.

C. Chesneau and T. Willer. Estimation of a cumulative distribution function under interval censoring "case 1" via warped wavelets. *Communications in Statistics. Theory and Methods*, 44(17):3680–3702, 2015. URL http://dx.doi.org/10.1080/03610926.2013.851231.

F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 49(2):569–609, 2013. URL https://doi.org/10.1214/11-aihp470.

A. Cuevas and R. Fraiman. On depth measures and dual statistics. a methodology for dealing with general data. *Journal of Multivariate Analysis*, 100(4):753 – 766, 2009. URL https://doi.org/10.1016/j.jmva.2008.08.002.

A. Cuevas and A. Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2):340–354, 2004. URL https://doi.org/10.1239/aap/1086957575.

A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian & New Zealand Journal of Statistics*, 48(1):7–19, 2006. URL https://doi.org/10.1111/j.1467-842X.2006.00421.x.

S. Dedu and R. Ciumara. Restricted optimal retention in stop-loss reinsurance under VaR and CTE risk measures. *Proceedings of the Romanian Academy*, 11(3):213–217, 2010. URL http://www.academiaromana.ro/sectii2002/proceedings/doc2010-3/03-Dedu.pdf.

L. Dehaan and X. Huang. Large quantile estimation in a multivariate setting. *Journal of Multivariate Analysis*, 53(2):247–263, 1995. URL https://EconPapers.repec.org/RePEc:eee:jmvana:v:53:y:1995:i:2:p:247-263.

M. Denuit, J. Dhaene, M. Goovaerts, and R. Kaas. *Actuarial Theory for Dependent Risks*. Wiley, 2005. URL https://doi.org/10.1002/0470016450.

E. Di Bernardino, T. Laloë, V. Maume-Deschamps, and C. Prieur. Plug-in estimation of level sets in a non-compact setting with applications in multivariate risk theory. *ESAIM. Probability and Statistics*, 17:236–256, 2013. URL https://doi.org/10.1051/ps/2011161.

E. Di Bernardino, T. Laloë, and R. Servien. Estimating covariate functions associated to multivariate risks: a level set approach. *Metrika*, 78(5):497–526, 2015. URL https://hal.archives-ouvertes.fr/hal-00800461.

E. Di Bernardino, H. Laniado, R. E. Lillo, and R. Torres. On the estimation of extreme directional multivariate quantiles. preprint, 2018. URL https://arxiv.org/abs/1610.08386v3.

H. Drees and X. Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–47, 1998. URL https://doi.org/10.1006/jmva.1997.1708.

S. Efromovich. *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York, 1999. URL https://link.springer.com/book/10.1007%2Fb97679.

P. Embrechts and G. Puccetti. Bounds for functions of multivariate risks. *Journal of Multivariate Analysis*, 97 (2):526–547, 2006. URL http://dx.doi.org/10.1016/j.jmva.2005.04.001.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. 1996.

J.-D. Fermanian. Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95(1):119–152, 2005. URL https://doi.org/10.1016/j.jmva.2004.07.004.

J. M. Fernández-Ponce and A. Suárez-Lloréns. Central regions for bivariate distributions. *Austrian Journal of Statistic*, 31(2–3):141–156, 2002. URL https://doi.org/10.17713/ajs.v31i2&3.477.

D. Furer and M. Kohler. Smoothing spline regression estimation based on real and artificial data. *Metrika*, 78 (6):711–746, 2015. URL https://doi.org/10.1007/s00184-014-0524-6.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011. URL https://doi.org/10.1214/11-AOS883.

G. K. Golubev and M. Nussbaum. Adaptive spline estimates in a nonparametric regression model. *Theory of Probability and Its Applications*, 37(3):521–529, 1992. URL https://doi.org/10.1137/1137102.

A. Guyader and N. Hengartner. On the mutual nearest neighbors estimate in regression. *Journal of Machine Learning Research*, 14:2361–2376, 2013. URL http://jmlr.org/papers/v14/guyader13a.html.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer New York, 2002. URL https://doi.org/10.1016/s0165-0270(99)00126-0.

P. Jaworski, F. Durante, W. Härdle, and T. Rychlik. *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009.* Lecture Notes in Statistics. Springer Berlin Heidelberg, 2010. URL https://books.google.fr/books?id=vX233feaA6MC.

G. Kerkyacharian and D. Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 12 2004. URL https://doi.org/10.3150/bj/1106314850/.

M. Kohler, A. Krzyzak, and H. Walk. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4):1286–1296, 2009. URL http://dx.doi.org/10.1016/j.jspi.2008.07.012.

T. Laloë and R. Servien. Nonparametric estimation of regression level sets using kernel plug-in estimator. *Journal of the Korean Statistical Society*, 42(3):301–311, 2013. URL https://doi.org/10.1016/j.jkss.2012.10.001.

M. Mason and W. Polonik. Asymptotic normality of plug-in level set estimates. *Annals of Applied Probability*, 19:1108–1142, 2009. URL https://doi.org/10.1214/08-aap569.

K. Mosler. *Multivariate Dispersion, Central Regions and Depth.* Springer, 2002. URL https://doi.org/10.1007/978-1-4613-0045-8.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964. URL https://doi.org/10.1137/1109020.

G. Nappo and F. Spizzichino. Kendall distributions and level sets in bivariate exchangeable survival models. *Information Sciences*, 179:2878–2890, 2009. URL https://doi.org/10.1016/j.ins.2009.02.007.

N. Ngoc Bien. *Adaptation via des inéqualités d'oracle dans le modèle de régression avec design aléatoire*. PhD thesis, Université d'Aix-Marseille, 2014. URL https://www.theses.fr/2014AIXM4716.

S. M. Nikol'skiĭ. *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, 1975. URL https://link.springer.com/book/10.1007%2F978-3-642-65711-5.

T. M. Pham Ngoc. Regression in random design and Bayesian warped wavelets estimators. *Electronic Journal of Statistics*, 3:1084–1112, 2009. URL https://doi.org/10.1214/09-ejs466.

W. Polonik. Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995. URL http://dx.doi.org/10.1214/aos/1176324626.

M. Rahimi, R. Pon, W. J. Kaiser, G. S. Sukhatme, D. Estrin, and M. Srivastava. Adaptive sampling for environmental robotics. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, volume 4, pages 3537–3544 Vol.4, 2004. URL https://doi.org/10.1109/ROBOT.2004.1308801.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009. URL http://dx.doi.org/10.3150/09-BEJ184.

A. Rodríguez-Casal. *Estimacíon de conjuntos y sus fronteras. Un enfoque geometrico*. PhD thesis, University of Santiago de Compostela, 2003.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

W. Stute. Asymptotic normality of nearest neighbor regression function estimates. *The Annals of Statistics*, 12:917–926, 1984. URL https://doi.org/10.1214/aos/1176346711.

R. Torres, R. E. Lillo, and H. Laniado. A directional multivariate value at risk. *Insurance: Mathematics and Economics*, 65:111 – 123, 2015. URL https://doi.org/10.1016/j.insmatheco.2015.09.002.

R. Torres, C. De Michele, H. Laniado, and R. E. Lillo. Directional multivariate extremes in environmental phenomena. *Environmetrics*, 28(2), 2017. URL http://dx.doi.org/10.1002/env.2428.

A. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25:948–969, 1997. URL https://doi.org/10.1214/aos/1069362732.

J. W. Tukey. Mathematics and the Picturing of Data. In R. D. James, editor, *International Congress of Mathematicians 1974*, volume 2, pages 523–532, 1974.

G. S. Watson. Smooth regression analysis. *Sankhyā, Series A*, 26:359–372, 1964. URL https://www.jstor.org/stable/25049340?seq=1#page_scan_tab_contents.

S.-S. Yang. Linear combination of concomitants of order statistics with application to testing and estimation. *Annals of the Institute of Statistical Mathematics*, 33(1):463–470, 1981. URL http://dx.doi.org/10.1007/BF02480956.

Part II

Statistical Learning

Introduction

In this second part, I focus on the general topic of statistical learning. I started to study this topic during my PhD with two principal contributions. First I studied the problem of the estimation of a regression function in a functional setting [Laloë, 2008]. Second I considered the problem of clustering using quantization in Banach spaces [Laloë, 2010].

After that I continued to be interested by the subject in a more applied way. First I started a collaboration with P. Brehmer, P. Fernandes and J. Guillard and applied my clustering procedure on real data [Brehmer et al., 2011]. The idea was to propose a way to identify fish school species using sonar recordings (a fish school is a set of fishes staying together for social reasons and swimming in the same direction). This work is not detailed in this manuscript

Second I proposed with R. Servien an efficient adaptation of the clustering algorithm presented in [Laloë, 2010], in order to lower its high complexity (Chapter 1 of this second part).

Finally, together with F. Delarue and R. Diel, I started an industrial collaboration with the start-up Option Way on the topic of airfare prediction (Chapter 2 of this second part).

Chapter 1

X Alter

In this chapter, related to the article I wrote with R. Servien [Laloë and Servien, 2013], I present a practical extension of my PhD work (and more specifically to [Laloë, 2010]). The Idea is to combine the X-means method proposed by [Pelleg and Moore, 2000] to the Alter algorithm I proposed in [Laloë, 2010].

Introduction

Clustering consists in partitioning a data set into subsets (or clusters), so that the data in each subset share some common trait. Proximity is determined according to some distance measure. For a thorough introduction to the subject, we refer to the book [Kaufman and Rousseeuw, 1990]. The origin of clustering goes back to several decades ago, when some biologists and sociologists began to search for automatics methods to build different groups with their data. Today, clustering is used in many fields. For example, in medical imaging, it can be used to differentiate between types of tissue and blood in a three dimensional image. Market researchers use it to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. There are also many different applications in artificial intelligence, sociology, medical research, or political sciences.

The K-means clustering is the most popular method [Hartigan and Wong, 1979, MacQueen, 1967]. Its attractiveness lies in its simplicity and its fast execution. It has however two main drawbacks. On the one hand, the number of clusters K has to be supplied by the user. Thus, different ways to determine K have been studied in the literature [Li et al., 2008, Pham et al., 2005]. On the other hand, the algorithm strongly depends on the initialization and can easily converge to a local minimum. [Pelleg and Moore, 2000] offer a solution for the first problem with a building-block algorithm called X-means which quickly estimates K. After each run of 2-means, local decisions are done whether subsets of the current centroid should be splitted or not. The splitting decision is done by computing the Bayesian Information Criterion (BIC). In a different approach, I proposed in [Laloë, 2010] a consistent algorithm to address the second problem, called Alter algorithm, which also needs the specification of K. This algorithm search an optimal set of centres among the data set and does not need an initialization. Moreover it is proved to converge to a global optimum.

Our purpose here is to combine the X-means and the Alter algorithm in order to overcome the drawbacks of both algorithms. The complexity of the Alter algorithm decreases and an automatic selection of the number of clusters is simultaneously performed. Moreover, the convergence properties of the Alter algorithm will overcome the local optimality problem of the X-means algorithm, inherited from the K-means one.

The Chapter is organized as follows: the different algorithms are presented in Section 1.1. Performances of X-Alter, X-means and other methods are compared in Section 1.2.

1.1 Methodology

1.1.1 The Alter algorithm

Let us summary the Alter algorithm. All the theoretical results presented in this section come from a previous work I performed during my PhD [Laloë, 2010]. The method is based on quantization. It is a commonly used technique in signal compression [Graf and Luschgy, 2000, Linder, 2002]. Consider $(\mathcal{H}, \|.\|)$ a normed space. We let X be a $\mathcal{H}$-valued random variable with distribution μ such as $\mathbb{E}\|X\| < \infty$.

Given a set $\mathcal{C}$ of points in $\mathcal{H}^k$, any Borel function $q : \mathcal{H} \to \mathcal{C}$ is called a quantizer. The set $\mathcal{C}$ is called a codebook, and the error made by replacing X by $q(X)$ is measured by the distortion:

$$D(\mu, q) = E\,\|X - q(X)\| = \int_{\mathcal{H}} \|x - q(x)\|\mu(dx).$$

I choosed to consider a distortion based on the L_1 distance for robustness properties. We will discuss it in the simulations. Note that $D(\mu, q) < \infty$ since $\mathbb{E}\|X\| < \infty$. For a given k, the aim is to minimize $D(\mu, .)$ among the set $\mathcal{Q}_k$ of all possible k-quantizers. The optimal distortion is then defined by

$$D_k^*(\mu) = \inf_{q \in \mathcal{Q}_k} D(\mu, q).$$

When it exists, a quantizer q^* satisfying $D(\mu, q^*) = D_k^*(\mu)$ is said to be an optimal quantizer.

As detailed in [Laloë, 2010], a quantizer is characterized by its codebook $\mathcal{C} = \{y_i\}_{i=1}^k$ and a partition of $\mathcal{H}$ in cells $S_i = \{x \in \mathcal{H} : q(x) = y_i\}$, $i = 1, \dots, k$ via the rule

$$q(x) = y_i \iff x \in S_i.$$

Moreover I proved in [Laloë, 2010] that for a given codebook the optimal partition is the nearest neighbour one. So we can consider only nearest neighbour quantizers, which means that a quantizer q will be characterized by its codebook $\mathcal{C} = \{y_i\}_{i=1}^k$ and the rule

$$q(x) = y_i \iff \forall 1 \leq j \leq k, j \neq i, \|x - y_i\| \leq \|x - y_j\|.$$

Thus, a quantizer can be defined by its codebook only. Moreover the aim is to minimize the distortion among all possible nearest neighbour quantizers.

However, in practice, the distribution μ of the observations is unknown, and we only have at hand n independent observations $X_1, \dots, X_n$ with the same distribution than X. The goal is then to minimize the empirical distortion:

$$\frac{1}{n}\sum_{i=1}^{n} \|X_i - q(X_i)\|.$$

As said before, I choosed to consider the L_1-based distortion to obtain more robust estimators (see [Kemperman, 1987]). Then, clustering is done by regrouping the observations that have the same image by q. More precisely, we define a cluster $\mathcal{C}$ by $\mathcal{C} = \{X_i : q(X_i) = \hat{x}_{\mathcal{C}}\}$, $\hat{x}_{\mathcal{C}}$ being the representative of cluster $\mathcal{C}$.

I stated the theoretical results (consistency and rate of consistency) in [Laloë, 2010]. In particular the rate of consistency is closely related to the metric entropy of the data space. However, the minimization of the empirical distortion is not possible in practice and that is why I proposed an alternative: the Alter algorithm. The idea is to select an optimal codebook among the data set. More precisely the outline of the algorithm is:

1. List all possible codebooks , i.e., all possible K-tuples of data;
2. Compute the empirical distortion associated to the first codebook. Each observation X_i is associated with its closed centre;
3. For each successive codebook, compute the associated empirical distortion. Each time a codebook has an associated empirical distortion smaller than the previous smallest one, store the codebook;
4. Return the codebook that has the smallest distortion.

Again, theoretical results of consistency and rate of consistency are available in [Laloë, 2010]. In particular I stated that the convergence rate is of the same order than the theoretical method described above (minimization of the empirical distortion over all possible quantizers). Moreover, this algorithm does not depend on initial conditions (unlike the K-means algorithm) and it converges to the optimal distortion. Unfortunately its complexity is $O(n^{K+1})$ and it is impossible to use it for high values of n or K.

1.1.2 The X-Means algorithm

In a different approach, [Pelleg and Moore, 2000] define the X-means algorithm which is adapted from the K-means one. The idea is to perform successive run of K-means, making local decisions about which cluster of the current partition should be splitted in order to improve the data fit. The splitting decision is done by computing the BIC criterion. This new approach proposes an efficient solution to one major drawbacks of K-means: the choice of the number of clusters K. Moreover, X-means has a low computational cost. But results suffer from the non-convergence property of the K-means algorithm. The outline of this algorithm is:

1. Perform 2-means. This gives us clustering C;
2. Evaluate the relevance of the classification C with a BIC Criterion: if the BIC criterion calculated with one cluster is greater than the one calculated with two clusters, validate the discrimination in two clusters;
3. Iterate step one and two in each cluster of C. Keep going until there is no more relevant discrimination.

1.1.3 The X-Alter Algorithm

Following the idea of X-means, a recursive use of Alter with $K = 2$ can simultaneously allow us to combine both advantages of these two methods: estimation of K / low computational cost for X-means and convergence / parameter-free character for Alter. We also add an aggregation step at the end of our algorithm to prevent the creation of too many clusters.

Note that no parameter is needed by the algorithm. Though, the user can specify a range in which the true K reasonably lies if he wishes to (this is $[2, +\infty[$ if one had no information). More precisely, the outline of the algorithm is the following:

1. Perform Alter with $K = 2$. This gives us clustering C;

2. Evaluate the relevance of the classification C (Figure 1.1) with a BIC Criterion;
3. Iterate step one and two in each cell of C (Figure 1.2). Keep going until there is no more relevant discrimination (Figure 1.3);
4. Final step of aggregation (Figure 1.4). For each pair of clusters we compute the difference $DIFF_{1,2} = BIC(K=1) - BIC(K=2)$. Then, according to the decreasing values of $DIFF_{1,2}$ aggregation can be considered if $DIFF_{1,2} > 0$. After each aggregation we actualise the $DIFF_{1,2}$ values involving the clusters just aggregated.

The algorithm starts by performing Alter with $K = 2$ centres. At this point, a model selection criterion (BIC, detailed below) is performed on all the data set. Using this criterion, we check the suitability of the discrimination by comparing $BIC(K = 1)$ and $BIC(K = 2)$. In another way, the criterion asks if the model with the two clusters is better than the one with only one. If the answer is yes, the iterative procedure occurs in the two subsets.

The structure improvement operation begins by splitting each cluster into two sub-clusters. The procedure is local on that the sub-clusters are fighting each other only for the points in the parent's cluster. Up to there, the only difference with X-means is that we use Alter instead of 2-means because the consistent property of Alter must improve results. Finally, when all regions are asleep and no more clusters are needed, the aggregative step starts to prevent the creation of too many clusters or the presence of splitted clusters (as in Figure 1.2).

The complexity of this algorithm in the worst case scenario (that is when it creates n clusters with one data set) is $O(n^4)$, which is less than the initial Alter algorithm. However, the computational cost is still higher than for X-mean and can be a problem for massive data sets.

The BIC criterion

We use here the same criterion than in [Pelleg and Moore, 2000], that is the formula from [Kass and Wasserman, 1995]. It evaluates the relevance of the classification C with

$$BIC(C) = l - \frac{p}{2} \log n$$

where l is the log-likelihood of the data according to the clustering C and taken at the maximum likelihood point, and p is the number of parameters in C. The number of free parameters p is simply the sum of $K-1$ cluster affectation probabilities, $d * K$ centroids coordinates, and one variance estimate. Note that we suppose here that in each cluster, the data are normally distributed around the centre. We will see in the empirical study that it performs well on real data.

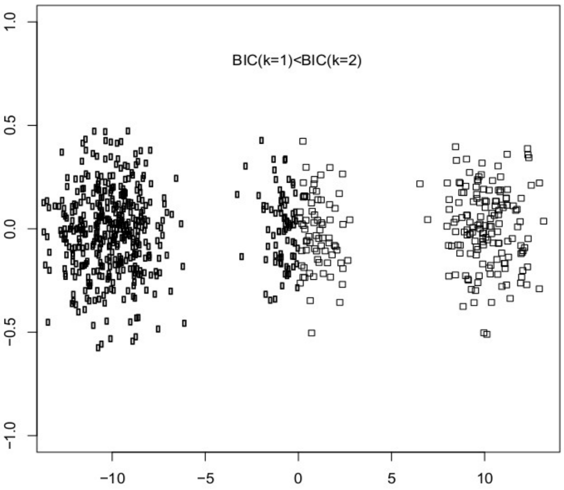


Figure 1.1: First iteration of X-Alter. The discrimination in 2 clusters (Step 1) is validated by BIC criterion (Step 2). In each cluster, observations are represented by a different symbol.

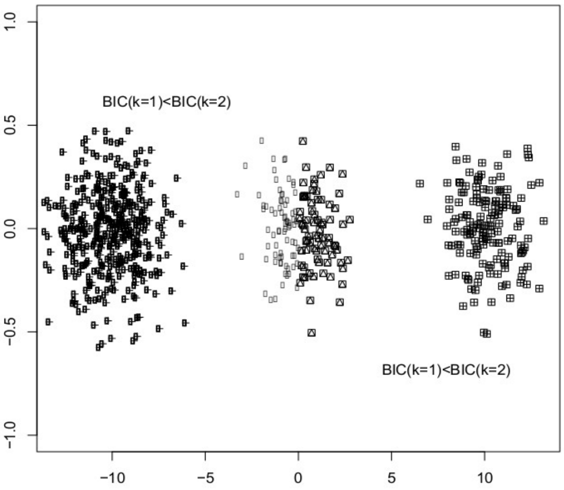


Figure 1.2: Second iteration of X-Alter: the sub-classification is done in the two relevant clusters (Step 1). Sub-classifications are validated by BIC (Step 2) so we obtain four clusters.

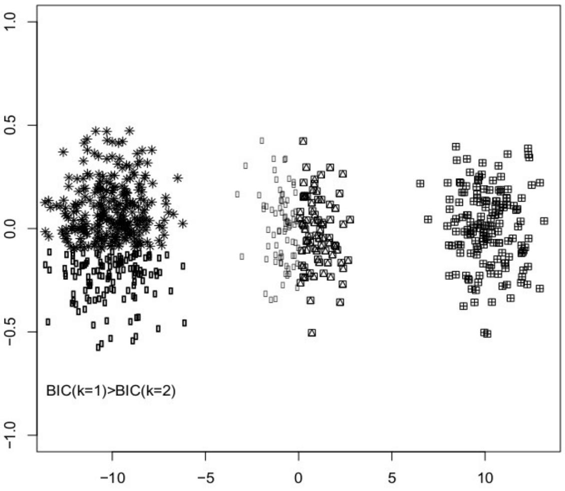


Figure 1.3: No relevant sub-classification in the left cluster according to BIC. In the three other clusters, we obtain the same rejection of sub-classification (Step 3).

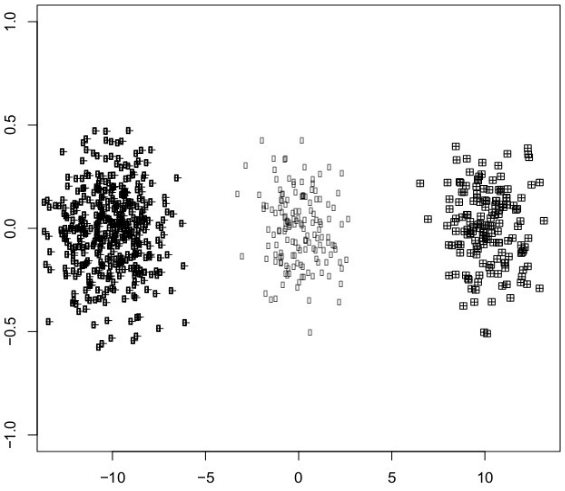


Figure 1.4: Final discrimination. The two middle clusters have been aggregated in Step 4.

1.2 Empirical results

In this section, an empirical study is performed to show the relevance of our method. We confront our method to various simulated data sets, but also on classical real data sets. We consider three criterion: the number of detected clusters, the Adjusted Rand Index (A.R.I.) [Hubert and Arabie, 1985, Rand, 1971] and the Dunn index [Dunn, 1974, Handl et al., 2005]. The Rand Index is a measure of the similarity between two clusters. A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value (say zero). So, [Hubert and Arabie, 1985] defined the A.R.I. which is the corrected-for-chance version of the Rand index. Studies have shown the need and usefulness of the adjusted measures [Nguyen et al., 2009]. The more similar (respectively dissimilars) the clusters are , the closer to 1 (respectively 0) the A.R.I. is. On another way, the Dunn Index measures the "compactness" of the clusters and is a sort of the worst case indicator. The goal is to identify sets of clusters that are compact, with a small variance between individuals in the same cluster, and well separated, where the centres of different clusters are sufficiently far apart, as compared to the within cluster variance. Higher is the Dunn Index, better the clustering is. For more details on this classical cluster validation indexes we refer the reader to [Dunn, 1974, Handl et al., 2005].

Pelleg and Moore show that X-means performs better and faster than repeatedly using accelerated K-means for different values of K. So, we compare our X-Alter algorithm to X-means and to X-means with the aggregation step, called X-means-R. That is we obtain a clustering using X-means and we compute the aggregation procedure (Step 4 on Section 1.1.3) on this clustering. It allows us to assess the usefulness and the computational time of the aggregation step.

1.2.1 Simulated data

We choose to consider here functional data. We consider two configurations:

First configuration:

First, we generate three clusters of functions defined on $[0, 1]$ (and discretized 20 times). A data is a function generated around one of the three generative functions ($\sqrt{x}$, x and x^2), perturbed by a common term $\cos(10x + \pi/2 - 10)/5$. Moreover, each data is noised with a vector composed by twenty Gaussian law $\mathcal{N}(0, \sigma)$ where the value of σ is selected for each data using $\sigma \sim \mathcal{N}(0.1, 0.02)$. Figure 1.5 shows examples of some of the functions that we want to classify. Three clusters of size randomly chosen between 15 and 25 are simulated 300 times. Results are presented in Table 1.1 (time is given in seconds) and show that our method gives better results, mostly on the search of the number of clusters.

Second configuration:

Second, we consider a slightly more difficult case. We construct this configuration on the same model than the first, but based on functions $\sqrt{x}$, $x^{3/4}$ and x which are closer than the previous ones as we can see in Figure 1.6. Results are gathered in Table 1.2. Again, we see that our method retrieves more often the correct number of clusters. Note that if the complexity of our algorithm is larger than the X-means one, it is still much smaller than the Alter one. Moreover Alter does not estimate the number of clusters.

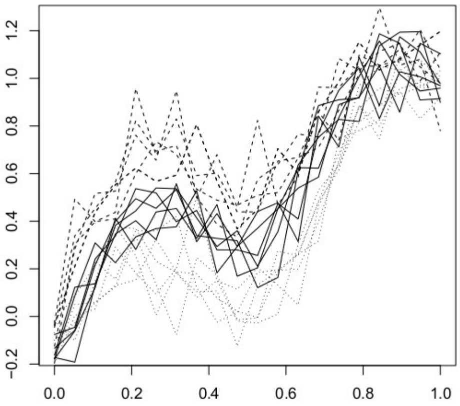


Figure 1.5: Example of functions. Functions based on $\sqrt{x}$ are on dashed lines, ones based on x are on solid lines and ones based on x^2 are on dotted lines.

Table 1.1: Results for the three algorithms on the functional data.

Algorithm	% of correct number of clusters	A.R.I.	Dunn	Time
X-means	81	0.88	0.63	2.0
X-means-R	85	0.88	0.63	3.5
X-Alter	95	0.89	0.63	27.6

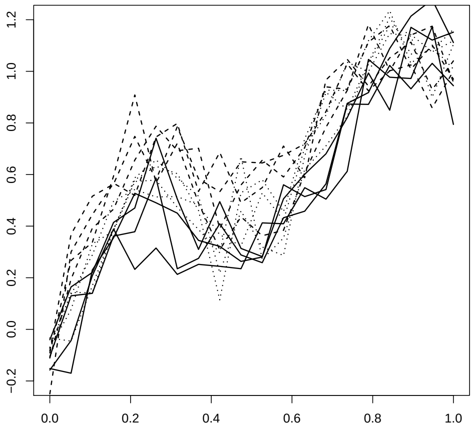


Figure 1.6: Example of functions. Functions based on $\sqrt{x}$ are on dashed lines, ones based on x are on solid lines and ones based on $x^{3/4}$ are on dotted lines.

Table 1.2: Results for the three algorithms on the functional data.

Algorithm	% of correct number of clusters	A.R.I.	Dunn	Time
X-means	26	0.75	0.43	2.4
X-means-R	31	0.75	0.46	3.2
X-Alter	40	0.77	0.46	28.7

Robustness study

In this paragraph, we illustrate the robustness properties of the L_1 distance. We consider as a starting point the first functional configuration above used in Figure 1.5. To obtain noisy data we use the following protocol: we add a value $x \in [-0.30; -0.15] \cup [0.15; 0.30]$ to $a \in [10; 25]$ percent of points (randomly chosen) of $b \in [10; 25]$ percent of data (randomly chosen). An example is given in Figure 1.7. We repeat this 300 times and give averaged results in Table 1.3.

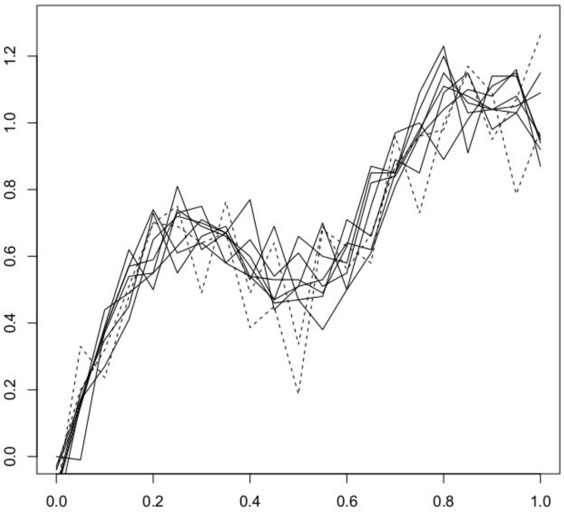


Figure 1.7: Example of the results of the perturbation of $\sqrt{x}+\cos(10x+\pi/2-10)/5$. Affected (by the previously described protocol) functions are on dashed lines.

Table 1.3: Results for the three algorithms on the perturbed functional data sets.

Algorithm	% of correct number of clusters	A.R.I.	Dunn	Time
X-means	77	0.87	0.52	2.6
X-means-R	79	0.87	0.52	3.8
X-Alter	95	0.88	0.53	29.4

The relevance of the L_1-based distance error, which is much more robust to extreme values, is shown here. Indeed, if we compare to the results gathered in Table 1.1 we still find the correct number of clusters 95% of the time while X-means and X-means-R do not (a loss of respectively 4% and 6%).

1.2.2 Real data

In this section, we confront our method to two conventional data sets from the UCI Machine Learning Repository [Frank and Asuncion, 2010]: the wine and iris ones. In this case, we do not know if the spherical Gaussian assumption of the BIC criterion is verified. We compare our method to the X-means algorithm but also to the K-means algorithm with K known to be 3 (the real number of clusters here). So, 3-means have a significant advantage over others methods by knowing the number of clusters. In these two real cases, as suggested in the description of the data sets, we centre and standardize each variable before performing clustering. Since K-means, X-means and X-means-R depends on the initialisation, we give averaged results (over 50 running) for these methods.

Wine data set

We consider first the wine data set. We have 178 instances and 13 variables found in each of the three types of wines. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. In a classification context, this is a well posed problem with "well behaved" class structures. The results for the 4 methods are presented in Table 1.4. We can see that our method retrieves the real number of clusters, and that we get the same adjusted rand index than 3-means and slightly less than the 2 others. On the other hand, we do not have a good Dunn Index because one extreme instance is bad classified. We can also compare X-Alter to other methods used on this data set and listed on the UCI Machine Learning [Frank and Asuncion, 2010]. For example, we better estimate the number of clusters than [Dy and Brodley, 2004] with their different methods.

Table 1.4: Results for the wine data set.

Algorithm	Number of clusters	A.R.I.	Dunn
X-means	8.67 (var=6.92)	0.78 (var=0.03)	0.162 (var=2.10^{-4})
X-means-R	8.54 (var=6.01)	0.78 (var=0.03)	0.165 (var=10^{-4})
3-means	-	0.76 (var=0.03)	0.163 (var=0.0002)
X-Alter	3	0.76	0.142

Iris data set

We consider now the Iris data set. We have 150 instances and 4 variables of 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are not linearly separable from each other which makes it more difficult to classify. The results are gathered in Table 1.5.

Table 1.5: Results for Iris data set.

Algorithm	Number of clusters	A.R.I.	Dunn
X-means	13.7 (var=6.2)	0.46 (var=0.07)	0.0405 (var=6.10^{-5})
X-means-R	8 (var=1.56)	0.57 (var=0.03)	0.0398 (var=0)
3-means	-	0.46 (var=0.0036)	0.04 (var=0)
X-Alter	6	1	0.402

It appears that our method do not find the real number of clusters but gets closer to it than others. While Adjusted Rand Index were previously very close for all methods, X-Alter is here significantly better and can not be improved. Indeed, as we consider here the Adjusted Rand Index (and not the Rand Index), it does not mean that our classification is perfect. However the high value of the A.R.I. informs us that the great majority of iris plant are well-classified, the 3 additional clusters are in fact very small and do not affect the A.R.I and the global quality of the obtained clustering. In [Dy and Brodley, 2004], the estimation of the number of clusters is slightly better but the quality of our clustering seems (as we don't use the same criterion) to be better. Moreover, we observe the interest of the aggregation step in X-means-R and it seems to appear that the spherical Gaussian assumption required for the BIC is acceptable for some real data sets.

Finally, we see that in all cases (simulated or real data sets) our method performs better than others to estimate the number of clusters. This confirms that we avoid the local convergence of X-means, which is inherited from K-means. Furthermore, according to Adjusted Rand Index and to Dunn Index, quality of clustering is either equal or significantly better than other methods.

Conclusion

I presented in this chapter a simple new algorithm to perform clustering. The main advantage of this method is that it is parameter-free. So, it can be easily used without an expertise knowledge of the data. This algorithm combines Alter and X-means algorithm to obtain of qualities of both algorithms (respectively the convergence and the automatic selection of the number of clusters). Moreover, we avoid the main drawbacks of these two methods: the high complexity for Alter and the dependence on initials conditions for X-means. A confrontation on both simulated and real data sets showed the relevance of this method.

Chapter 2

Airfare prediction

I present in this chapter the results of an industrial collaboration with the start-up "Option Way" on Airfare prediction. Since I am subject to a confidentiality agreement I cannot give details of the developed procedure. However I will try to give the general idea of the project and some efficiency results. I would like to thank Ali Sofiane who worked on this project as an engineer.

Introduction

Option Way is a French start-up specialized in Airfare prediction, with the objective to share benefits with customers. The start-up is born in 2012 on the following simple statement: the ticket prices vary constantly according to the Yield Management of the airlines. How can we use this to the travellers benefits? As an example, Figure 2.1 shows the variation of the daily minimal prices (by company), starting four months before the take off, of all direct round trip Paris - New York on fixed dates. We can see that prices can decrease at any time.

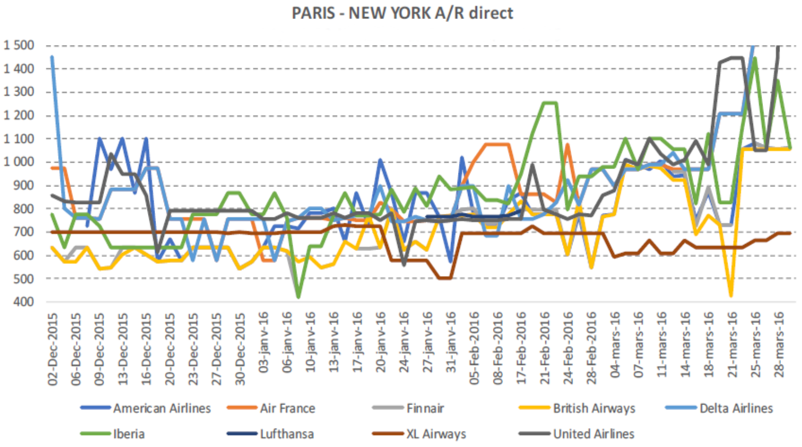


Figure 2.1: Trajectories of the daily minimal prices observed on a direct round trip Paris -New York

In order to allow the traveller to benefit from the price decrease, Option Way first proposed an "option strategy": the traveller agrees to buy his/her ticket at a given price (lower than the current price). Option way tracks the daily prices and buys the ticket as soon as it reaches this price. The main issue is to evaluate the probability of success. However, in this scenario the traveller is taking the risk to never get the ticket. Marketing studies showed that a majority of customers will not accept such a risk. Thus Option Way is interested in a new concept: the guaranteed option. The idea is to sell a plane ticket with a price lower than the actual price, knowing that with a great probability there will be a corresponding price decrease. This time it is the company who takes the risk to pay the ticket at an higher price than the one proposed to the customer.

On this basis, we started a collaboration between Option Way and a group of researchers of the probability and statistics team of J.A. Dieudonné Laboratory (F. Delarue, R. Diel and myself). This collaboration is partly funded by "BPIFrance" via the 2nd edition of the "Concours d'Innovation Numérique". This allowed us to hire a research engineer (Sofiane Ali), who worked with us on the project.

As said in the preamble, a confidentiality agreement prevents me from giving much details on the statistical procedure but I will provide general considerations about the method and the data preprocessing (Section 2.1). I will finish by a presentation of some results of our evaluation process (Section 2.2).

2.1 Presentation of the method

Our final objective is to predict the minimal price of a flight, or a group of flights, on a given period. To do so we dispose of the customers criteria and a data base gathering daily records of prices of previous flights. We have focused on the flights between Paris and New-York with travel from week-end to week-end.

In order to avoid availabilities issues on a precise flight (availability which we are not able to infer), the customer can only precise if he accepts low cost companies or not. Finally, the customer can only specify a wide time slot for his departure. Thus we always consider the minimum price on at least 10 flights. In what follows the term "travel" define the set of flights corresponding to a customer request.

2.1.1 General idea of our algorithm

At first we considered a two step algorithm:

1. Perform an unsupervised clustering on the historical travel lists of prices in order to define a certain number of classes of optimal prices (optimal meaning the minimal price ever reached for a given travel);

2. Given this classification, use a supervised approach in order to map the customers criteria to a class of minimal prices.

The first tests were encouraging but a precise look at the past minimal prices on travels shows that the unsupervised creation of classes of optimal prices, only based on historical travel list of prices, is a little artificial. Thus we decided to switch to a regression approach instead of a classification approach. That is we want to directly predict the minimal price coming for a customer request, using the past observed minimal prices associated to the travel criteria. To do so we choose to use a random forest based algorithm [Breiman, 2001]. The most important part of our work was to define the most efficient inputs to extract or compute from the database and from external variables.

2.1.2 Data preprocessing

The original database contains lists of successive prices of previous flights, with associated flights criteria (time slot, dates, departure and arrival location, company etc.). Figure 2.2 shows an example of some prices trajectories associated to a travel (from the historical database). Since our goal is to predict the future minimal price for a given travel, we perform a preprocessing and create a new database containing lists of daily minimal prices associated to a travel. Each list contains the minimal price of all the flights composing a travel, and is associated to the criteria of the travel (time of departure, allowed companies, ...). Figure 2.3 shows the example of the list of such minimal prices associated to a travel presented in Figure 2.2.

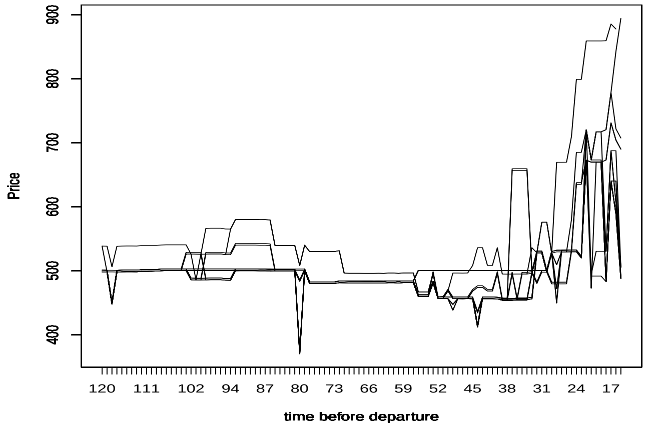


Figure 2.2: Examples of prices trajectories for one travel, up to 120 days before departure.

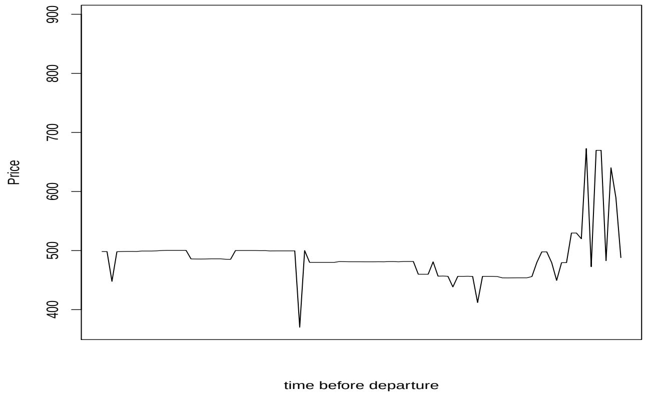


Figure 2.3: Trajectory of the daily minimal prices observed on a travel, from 120 to days before departure.

In addition, we add a certain number of variables to model the market condition and the time before departure. Without going deeper into the details of the preprocessing and the extraction of suitable informations, I provide in the next section a comparison between our final algorithm and a naive version using the basic informations extracted from the request customer and the database (day of departure, current price).

2.2 Evaluation of our algorithm

In this section, I show the results of our algorithm on historical data. We consider direct round trip between Paris and New-York, from week-end to week-end. The period of the evaluation starts on January 2017 and ends on January 2018. On this period we simulate the use of our algorithm. That is for each week end on period and each travel we run our algorithm to predict the minimal incoming price. Based on the prediction, we propose or not a guaranteed option and assume that all options are sold. A travel is given by a time of departure and return (morning, afternoon, evening) and a class of companies (low-cost or not low-cost). The prediction are actualised on a daily basis. The indicators used to evaluate the quality of our algorithm are the following:

- $\#_{AO}$ is the number of available options (that is the travels for which the minimal incoming price is lower than the current price, only known because we test our algorithm on a past period);
- $\#_{GO}$ is the number of guaranteed options we sold in the period;
- α_{GO} represent the proportion of successful guaranteed options (that is we succeed to buy under the proposed price);
- $Loss$ is the total of (financial) losses on the period;
- $Gain$ is the total of (financial) gains on the period;
- $GLRatio$ is the ratio Gain over Loss (the higher the better);
- $Benefits$ is the benefit on the period;
- μ_E is the mean economy (over all the guaranteed options) for the customer (expressed in percentage of reduction).

At this time we apply a very simple rule to exercise the option: we buy the travel as soon as the price goes under the option price. Moreover we apply a stop rule: whatever happens we buy the tickets 16 days before the departure at the latest. Moreover, in order to address the problem of availability, it was decided by the company to propose only one option at a time for a given travel. That is once an option is sold, we wait that it succeed (that is that we actually bought the ticket) before selling a new one (this explains the weak ratio $\frac{\#_{GO}}{\#_{AO}}$). Besides we compare our algorithm to a naive version using the basic informations extracted from the request customer (time of departure, low-cost or not, . . .) and the database. The results are gathered in Table 2.2 below. We see on this test that we have a $GLRratio$ at almost 6, which means that our gains are almost six times superior than our losses. Moreover we see that the mean economy provided to the customer is 8% of the initial price. This is very satisfying since marketing studies have shown that customers begin to be interested with reduction of at least 5%. These results were welcomed by Option Way, who intends to sell this product in the coming months. We also see the importance of our work on the preprocessing of the data, and the definition of efficient (in the sense that they improve α_{GO}, the $GLRario$ and the $Benefits$) indicators.

Table 2.1: Evaluation of our algorithm

Indicator	$\#_{AO}$	$\#_{GO}$	α_{GO}	$Loss$	$Gain$	$GLRatio$	$Benefits$	μ_E
Final algorithm	14580	1838	0.97	8734	146321	16.8	137587	8%
Naive version	14580	3909	0.87	142852	213990	1.5	71138	10%

Conclusion

I presented in this chapter the results of the collaboration with the start-up Option Way. We succeed to propose an efficient algorithm to anticipate price decreasing, which will allow to start a commercial phase. However some problems remains open, such as the anticipation of the availability of a flight or travel.

Conclusion and prospects

X Alter

I have presented in Chapter 1 a practical development to the method proposed in [Laloë, 2010]. However, if the X-Alter algorithm is far less complex than the Alter one, it is still too important for the method to be applied on really big data sets. I proposed in my PhD [Laloë, 2009] an accelerated version of Alter called Alter-Fast. The principle was simple: instead of performing Alter on the whole data set we can use it on several (very) small subset. Thus each run of Alter gives a quantizer and we take finally the best one. This process clearly helps to save computational time but at the price of a significant loss of efficiency. So as a future work, it could be interesting to look for another way to accelerate Alter while preserving (as much as possible) its properties of convergence.

Airfare prediction

As presented in Chapter 2, we have an algorithm able to predict a future minimal price of a travel, which open the possibility to sell option with a substantial reduction comparing to a current price. However a certain number of problems remains unanswered.

Buying strategy

Now that we can propose options with reduced prices, we have to think about the best manner to exercise it. In fact we have two problem to solve:

First, we have to imagine a way to determine the maximal delay we allow ourselves before exercising an option. One idea is to use historical data to evaluate the evolution of the probability that an option will succeed. This idea has to be specified but it seems that it could provide us a good indicator of the moment when the probability that we can successfully exercise an option becomes too weak. Using this kind of indicator we will then have to think about a way to minimize our losses.

Second, we want to propose a method to optimally exercise the option. Indeed we think that our actual strategy (exercising the option as soon as the price goes under the option price) is far from optimal. One idea is to dynamically use our algorithm to actualize our prediction and maximize the benefits. This point is of most importance for Option Way. Indeed their main source of profits will come from the difference between the option price and the exercise price.

Risk evaluation

In order to evaluate the required liquidity of the company we need to provide an indicator of exposure to the risk. Our first idea is to evaluate the worst scenario: given an ongoing option we consider the maximal price that we could have to pay if we exercise the option at the worst moment. The distribution of such individual exposures is illustrated in Figure 2.4.

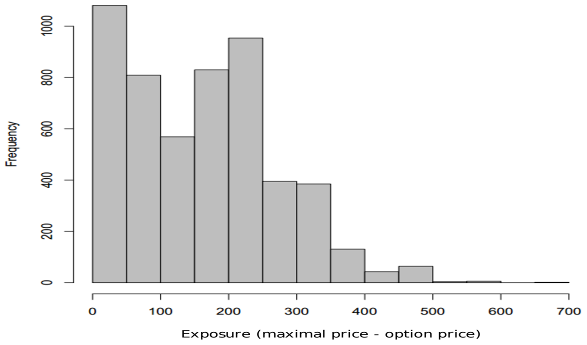


Figure 2.4: Distribution of the maximal individual exposures.

We also are interested in a global exposure at a given time, which can be the sum of the exposure of all ongoing options. Obviously this overestimate our exposure since such a scenario in which we will exercise all our options at the worst moment will never occur in practice. Moreover, in order to predict a future exposure, we also adapted our algorithm to predict the maximal incoming price. Figure 2.5 represents the "real" and "predicted" exposure.

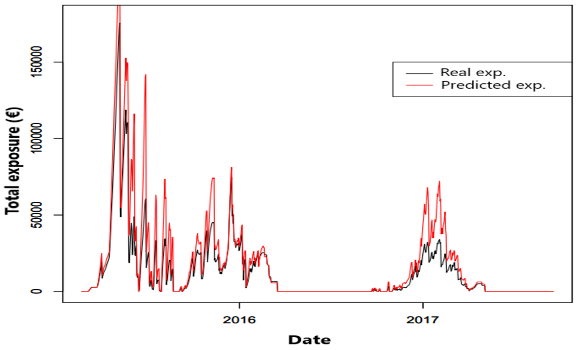


Figure 2.5: Real and predicted exposure

This indicator (which is a simple adaptation of our algorithm) is a good starting point but overestimate too much our exposure. Thus we have to improve it before considering a practical use by the company.

Finally, we have two more prospects in mind. First, if the company wants to start by selling options on round trips between Paris and New-York, a short term objective is to open other destinations. Second, we have to find a way to estimate the availability of a travel (basically the remaining free places in the plane) in order to sell more than one option at a time (for a given travel).

Bibliography of the second part

Bibliography

P. Brehmer, P. Fernandes, and T. Laloë. Three-dimensional internal spatial structure of young-of-the-year pelagic freshwater fish provides evidence for the identification of fish school species. *Limnology and Oceanography, Methods*, 9:322–328, 2011. URL https://doi.org/10.4319/lom.2011.9.322.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL https://doi.org/10.1023/A:1010933404324.

J. Dunn. Well separated clusters and fuzzy partitions. *Journal on Cybernetics*, 4:95–104, 1974. URL https://doi.org/10.1080/01969727408546059.

J. Dy and C. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5: 845–889, 2004. URL https://doi.org/10.1007/978-1-4419-1428-6_97.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, 2000. URL https://doi.org/10.1007/bfb0103945.

J. Handl, K. Knowles, and D. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21:3201–3212, 2005. URL https://doi.org/10.1093/bioinformatics/bti517.

J. Hartigan and M. Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28:100–108, 1979.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. URL https://doi.org/10.1007/bf01908075.

R. Kass and L. Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90:928–934, 1995. URL https://doi.org/10.2307/2291327.

L. Kaufman and P. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, 1990. URL https://doi.org/10.1002/9780470316801.

J. H. B. Kemperman. The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on the L_1-norm and Related Methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam, 1987.

T. Laloë. A k-nearest neighbor approach for functional regression. *Statistics & Probability Letters*, 78(10): 1189–1193, 2008. URL https://doi.org/10.1016/j.spl.2007.11.014.

T. Laloë. *Sur Quelques Problèmes d'Apprentissage Supervisé et Non Supervisé.* PhD thesis, University Montpellier II, 2009. URL https://tel.archives-ouvertes.fr/tel-00455528.

T. Laloë. L_1 quantizationand clustering in banach spaces. *Mathematical Methods of Statistics*, 19(2):136–150, 2010. URL https://hal.archives-ouvertes.fr/hal-01292694.

T. Laloë and R. Servien. The X-Alter algorithm : a parameter-free method to perform unsupervised clustering. *Journal of Modern Applied Statistical Methods*, 12(1):90–102, 2013. URL https://hal.archives-ouvertes.fr/hal-00674407.

M. Li, M. Ng, Y.-M. Cheung, and J. Huang. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *IEEE transactions on knowledge and data engineering*, 20:1519–1534, 2008. URL https://doi.org/10.1109/tkde.2008.88.

T. Linder. Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002. URL https://doi.org/10.1007/978-3-7091-2568-7_4.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, University of California Press, 1967.

X. Nguyen, J. Epps, and J. Bailey. Information theoretic measures for clustering comparison: Is a correction for chance necessary ? *ICML'09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080, 2009. URL https://doi.org/10.1145/1553374.1553511.

D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, 2000.

T. Pham, S. Dimov, and C. Nguyen. Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219:103–119, 2005. URL https://doi.org/10.1243/095440605x8298.

W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. URL https://doi.org/10.2307/2284239.

Part III

Neuroscience

Chapter 1

Pattern detection

This chapter, related to [Chevallier and Laloë, 2015], is devoted to a collaboration with J. Chevallier in the area of Neuroscience. The idea is to propose a statistical tool to detect synchronisation between the activity of different neurons.

Introduction

The communication between neurons relies on their capacity to generate characteristic electric pulses called action potentials. These action potentials are usually assumed to be identical stereotyped events. Their time of occurrence (called spike) is considered as the relevant information. That is why the study of spike frequencies (firing rates) of neurons plays a key role in the comprehension of the information transmission in the brain [Abeles, 1982, Gerstein and Perkel, 1969, Shinomoto, 2010]. Such neuronal signals are recorded from awake behaving animals by insertion of electrodes into the cortex to record the extracellular signals. Potential spike events are extracted from these signals by threshold detection and, by spike sorting algorithms, sorted into the spike signals of the individual single neurons. After this preprocessing, we dispose of sequences of spikes (called spike trains).

The analysis of spike trains has been an area of very active research for many years [Brown et al., 2004]. Although the rules underlying the information processing in the brain are still under burning debate, the detection of correlated firing between neurons is the objective of many studies in the recent years [Dong et al., 2008, Pillow et al., 2008, Roy et al., 2007]. This synchronization phenomenon may take an important role in the recognition of sensory stimulus. In this chapter, the issue of detecting dependence patterns between simultaneously recorded spike trains is addressed. Despite the fact that some studies used to consider neurons as independent entities [Barlow, 1972], many theoretical works consider the possibility that neurons can coordinate their activities [Hebb, 1949, Palm, 1990, Sakurai, 1999, von der Malsburg, 1981]. The understanding of this synchronization phenomenon [Singer, 1993] required the development of specific descriptive analysis methods of spike-timing over the last decades: cross-correlogram [Perkel et al., 1967], gravitational clustering [Gerstein et al., 1985] or Joint PeriStimulus Time Histogram (JPSTH, [Aertsen et al., 1989]). Following the idea that the influence of a neuron over others (whether exciting or inhibiting) results in the presence (or absence) of coincidence patterns, Grün and collaborators developed one of the most popular and efficient method used this last decade [Grün, 1996, Grün et al., 1999]: the Unitary Events (UE) analysis method and the corresponding independence test, which detects where dependence lies by assessing p-values (a UE is a spike synchrony that recurs more often than expected by chance). This method is based on a binned coincidence count that is unfortunately known to suffer a loss in synchrony detection, but this flaw has been corrected by the multiple shift coincidence count [Grün et al., 1999].

In order to deal with continuous time processes, a new method (Multiple Tests based on a Gaussian Approximation of the Unitary Events Method, MTGAUE), based on a generalization of this count, the delayed coincidence count, has recently been proposed for two parallel neurons (Section 3.1 of [Tuleau-Malot et al., 2014]). The results presented here are in the lineage of this newest method and are applied on continuous point processes (random set of points which are modelling spike trains). Testing independence between real valued random variables is a well known problem, and various techniques have been developed, from the classical chi-square test to re-sampling methods for example. The interested reader may look at [Lehmann and Romano, 2005]. Some of these methods and more general surrogate data methods have been applied on binned coincidence count, since the binned process transforms the spike train in vectors of finite dimension. However, the case of point processes that are not preprocessed needs other tools and remains to study. Although the binned method can deal with several neurons (six simultaneously recorded neurons are analysed in [Grün et al., 2002]), both of the improvements (Multiple Shift and MTGAUE) can only consider pairs of neurons. Thus, our goal is to generalize the method introduced in [Tuleau-Malot et al., 2014] for more than two neurons. Unlike MTGAUE, our test is not designed to be performed on multiple time windows. However it can be multiple with respect to the different possible patterns composed from $n \geq 2$ neurons (see Section 1.3.3).

In Section 1.1, I introduce the different notions of coincidence we consider. In Section 1.2, a test is established and the asymptotic control of its false positive rate is proven. In Section 1.3 the relevance of our method when our main theoretical assumptions are weakened is empirically tested. Finally I present in Section 1.4 an application of our test on real data.

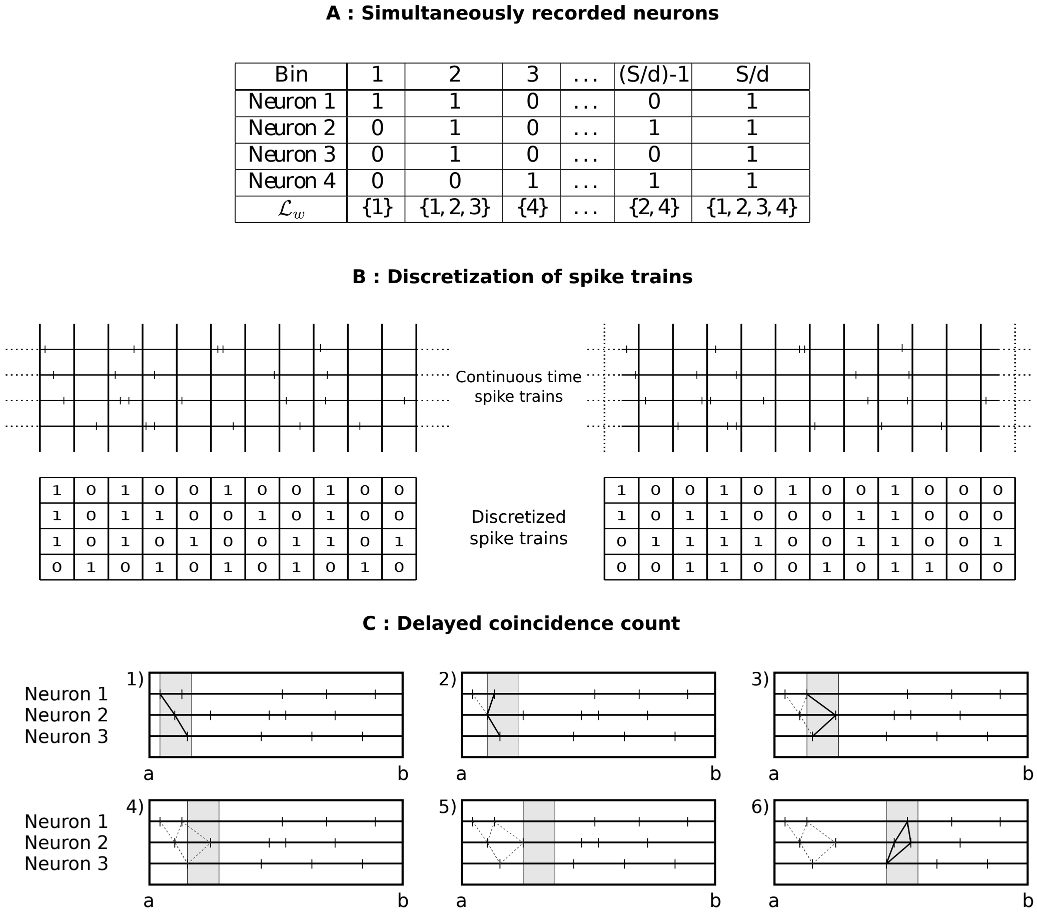
1.1 Notions of coincidence and the classical UE methods

In order to detect synchronizations between the involved neurons, different notions of coincidence can be considered. Informally, there is a coincidence between neurons when they each emit a spike more or less simultaneously. This notion has already been used in UE methods [Grün et al., 2002] and is based on the following idea: a real dependency between $n \geq 2$ neurons should be characterized by an unusually large (or low) number of coincidence [Grammont and Riehle, 2003, Grün, 1996, Tuleau-Malot et al., 2014].

1.1.1 Two notions of coincidence

The UE method considers discretized spike trains at a resolution ℓ of typically 1 or 0.1 millisecond. Therefore, in the discrete-time framework, each trial consists of a set of n spike trains (one for each recorded neuron), each spike train being represented by a sequence of 0 and 1 of length S. Since it is quite unlikely that two spikes occur at exactly the same time at this resolution ℓ, spike trains are binned and clipped at a coarser level. More precisely for a fixed bin size $\Delta = d\ell$ (d being an integer), a new sequence of length S/d of 0 and 1 is associated to each spike train (1 if at least one spike occurs in the corresponding bin, 0 otherwise). For more precise informations on the binning procedure and the link with point processes we refer the interested reader to [Tuleau-Malot et al., 2014].

A constellation or pattern is a vector of size n of 0 and 1 (see Figure 1.1). Of course, there are 2^n different constellations. The UE statistic associated to some constellation w consists in counting the number of occurrences of such w in the set of S/d vectors of size n. However, as shown in Figure 1.1, this method largely depends on the bin choice and it has been proven in [Grün et al., 1999] that this can lead in the case $n = 2$ to up to 60% of loss in detection when Δ is of the order of the range of interaction.



A : Simultaneously recorded neurons

Bin	1	2	3	...	(S/d)-1	S/d
Neuron 1	1	1	0	...	0	1
Neuron 2	0	1	0	...	1	1
Neuron 3	0	1	0	...	0	1
Neuron 4	0	0	1	...	1	1
$\mathcal{L}_w$	{1}	{1, 2, 3}	{4}	...	{2, 4}	{1, 2, 3, 4}

B : Discretization of spike trains

Continuous time spike trains

Discretized spike trains

1	0	1	0	0	1	0	0	1	0	0
1	0	1	1	0	0	1	0	1	0	0
1	0	1	0	1	0	0	1	1	0	1
0	1	0	1	0	1	0	1	0	1	0

1	0	0	1	0	1	0	0	1	0	0	0
1	0	1	1	0	0	0	1	1	0	0	0
0	1	1	1	1	0	0	1	1	0	0	1
0	0	1	1	0	0	1	0	1	1	0	0

C : Delayed coincidence count

Figure 1.1: In **A**, 4 parallel binary processes of length S are displayed. At each time step, the constellation and its corresponding subset of $\{1, 2, 3, 4\}$ are given. For instance, the constellation associated to the first bins is the vector $(1, 0, 0, 0)$ and the corresponding subset is $\{1\}$. In **B**, illustration of the UE method with two different choices of bins of the same size (the results are different, for example the constellation full of 1s is present in the second case and not in the first one). In **C**, an illustration of the six first steps in the dynamical computation of the delayed coincidence count. Here, there are 3 parallel time point processes. We consider the full pattern, i.e. $\mathscr{L} = \{1, 2, 3\}$. The grey rectangle represents the sliding time window of length δ. The bold lines denote the coincidence patterns counted at each step and the grey dashed ones denote the coincidence pattern which have been counted in one the previous steps. At each of steps 1,2 and 3, exactly one coincidence is counted. At steps 4 and 5, no coincidence is detected. And, at step 6, two coincidences are counted.

Then, we focus on another coincidence count that deals with continuous data. This notion of delayed coincidence count is pretty natural and was used in [Borgelt and Picado-Muino, 2013, Muiño and Borgelt, 2014] in a simplified formalism. For sake of simplicity, we use the same formalism of point processes as in [Tuleau-Malot et al., 2014]: Considering $N_1, \ldots, N_n$, some point processes on $[a, b]$, and $\mathscr{L} \subset \{1, \ldots, n\}$ a set of indices of cardinal $L \geq 2$, the *delayed coincidence count* $X_{\mathscr{L}}$ (of delay $\delta < (b - a)/2$) over the neurons of subset $\mathscr{L}$ in the time window $[a, b]$ is given by

$$X_{\mathscr{L}} = X_{\mathscr{L}}(\delta) = \sum_{(x_1,\ldots,x_L)\in\prod_{l\in\mathscr{L}} N_l} \mathbf{1}_{\left|\max\limits_{i\in\{1,\ldots,L\}} x_i - \min\limits_{i\in\{1,\ldots,L\}} x_i\right| \leq \delta}. \tag{1.1}$$

This delayed coincidence count can be explained in the following way (see Figure 1.1):

- Fix some duration parameter δ which is the equivalent of the bin size Δ;
- Count how many times each neuron in $\mathscr{L}$ spikes almost at the same time, modulo the delay δ.

1.1.2 Original UE method

The notion of constellation is closely linked to the binning procedure and is not relevant in the continuous time framework. In this work, we fix some subset of neurons, denoted $\mathscr{L}$, and count how many times the neurons of $\mathscr{L}$ admit nearly simultaneous activity. However, there is a canonical correspondence between constellations and set of indices (see Figure 1.1). Then, in order to harmonize the notations between both methods, let us denote $\mathscr{L}(w)$ the set of indices corresponding to the constellation w.

To detect dependency between neurons, two estimators of the expected coincidence count are compared. The first one is the empirical mean $\bar{m}_w$ of the number of occurrences of a given constellation w through M trials,

$$\bar{m}_w = \frac{1}{M}\sum_{k=1}^{M} m_w^{(k)},$$

where $m_w^{(k)}$ is the number of occurrences of w during the k^{th} trial. This estimator is consistent (that is, converges towards the expected value of the number of occurrences) even with dependency between the spike trains. The second one is consistent only under the independence hypothesis, and is given by

$$\hat{m}_{g,w} = \frac{S}{d} \prod_{l \in \mathscr{L}(w)} \hat{p}_l \prod_{k \notin \mathscr{L}(w)} (1 - \hat{p}_k), \tag{1.2}$$

where $\hat{p}_i$ is the empirical probability of finding a spike in a bin of neuron i.

This enables the construction of the test described in [Grün, 1996, Grün et al., 2002] and based on the comparison between the statistic $M\bar{m}_w$ and a quantile of the Poisson distribution $\mathcal{P}(M\hat{m}_{g,w})$ where M is the number of trials. Most of the time only tests by upper values are computed. However, following the study of [Tuleau-Malot et al., 2014], we have decided to focus on symmetric tests. Hence, the symmetric test based on the UE method rejects the independence hypothesis when $\bar{m}_w$ is too different from $\hat{m}_{g,w}$. However, such a test necessarily makes mistakes. For example, a *false positive* corresponds to an incorrect rejection of the null hypothesis. Hence, an a priori upper bound on the false positive rate, that is the *significance level* (or just *level*), must be given in order to construct a decision rule. The symmetric independence test with level α based on the UE method is governed by the following rule: if

$$M\bar{m}_w \geq q_{1-\alpha/2} \quad \text{or} \quad M\bar{m}_w \leq q_{\alpha/2},$$

where q_x is the x-quantile of the Poisson distribution $\mathcal{P}(M\hat{m}_{g,w})$, then the independence hypothesis is rejected.

The UE method is applied under the hypothesis that the discrete processes modelling the spike trains of neurons are Bernoulli processes. The equivalent in the "continuous" framework is the Poisson process (as it can be seen in [Tuleau-Malot et al., 2014]). This leads to a different estimator of the expected coincidence count and a different test which are defined properly in the next section.

1.2 Study of the delayed coincidence count

Once the notion of coincidence is defined with respect to continuous data (Equation (1.1)) , mathematical tools can be used to construct the desired independence test. The procedure is to provide the expected value and variance of the variable $X_{\mathscr{L}}$ in function of the firing rates. These computations classically imply a Gaussian approximation with respect to i.i.d trials. Unfortunately the firing rates are usually unknown. Thus the final step is to replace the firing rates by their estimator to compute the estimated expected value and variance. This plug-in procedure is known to change the underlying distribution. As in [Tuleau-Malot et al., 2014], the delta method provides the exact nature of this change.

In the continuous framework, a sample is composed of M observations of $N_1, \cdots, N_n$ which are the point processes associated to the spike trains of n neurons on a window $[a, b]$. The goal is to answer the following question:

Given $\mathscr{L}$ a subset of $\{1, \ldots, n\}$, are the processes N_l, $l \in \mathscr{L}$ independent?

To do this, a statistical test comparing the two hypotheses

$$\begin{cases} (\mathcal{H}_0) & \text{The processes } N_l,\ l \in \mathscr{L} \text{ are independent;} \\ (\mathcal{H}_1) & \text{The processes } N_l,\ l \in \mathscr{L} \text{ are not independent;} \end{cases}$$

is proposed.

In this section our test and its asymptotic relevance are introduced. First, let us present and discuss our main assumptions which are the same as in [Tuleau-Malot et al., 2014].

A1 $N_1, \ldots, N_n$ are Poisson processes.

This assumption can be resumed to an assumption of independence of a point process with respect to itself over the time, as Bernoulli processes in discrete settings.

A2 The Poisson processes $N_1, \ldots, N_n$ are homogeneous on $[a, b]$.

Assumption **A2** may also appear very restrictive, but Bernoulli processes considered in [Grün et al., 1999, 2002] have the same drawback. Moreover, if necessary, one can partition $[a, b]$ in smaller intervals on which **A2** is satisfied. For more precise informations on Poisson processes we refer the interested reader to [Kingman, 1993].

These assumptions are necessary in this work in order to obtain an explicit form for the expected number of coincidences (and its variance). Note that there exist some surrogate methods in the literature for which there is no need of a model on the data (see [Grün, 2009, Louis et al., 2010a] for a review). In particular two kind of methods are commonly used: dithering methods (involving random shifts of individual spikes [Louis et al., 2010b, Stark and Abeles, 2009], or random shifts of patterns of spikes [Harrison and Geman, 2009]), and trial-shuffling methods [Pipa and Grün, 2003, Pipa et al., 2003]. However, they are based on binned coincidence count, and there is no equivalent, up to our knowledge, with a delayed coincidence count, due to serious computational issues. Alternative works have also been done in the Bayesian paradigm [Archer et al., 2013]. However, as announced in the introduction, we empirically show in Section 1.3 that the assumptions can be weakened. In particular, point processes admitting refractory periods can be taken into account. Thus, a nice perspective of this work could be to derive theoretical results with these weakened assumptions.

1.2.1 Asymptotic properties

In order to build our independence test, one needs to understand the behaviour of the number of coincidence $X_{\mathscr{L}}$ under the independence hypothesis $\mathcal{H}_0$. In particular, the expected value and the variance of $X_{\mathscr{L}}$ are computed here. In a general point processes framework, these computations are impossible. This is why some restrictive assumptions are needed, such as **A1**, **A2**, or the independence of the processes, as done in the original UE method where independent Bernoulli processes have been considered.

Theorem 1.2.1 *Let $\mathscr{L}$ and $X_{\mathscr{L}}$ be defined as previously. Assume assumptions* **A1** *and* **A2** *and denote by $\lambda_1, \ldots, \lambda_n$ the respective intensities of $N_1, \ldots, N_n$. Under hypothesis $\mathcal{H}_0$, the expected value and the variance of the number of coincidences $X_{\mathscr{L}}$ are given by:*

$$m_{0,\mathscr{L}} := \mathbb{E}\left[X_{\mathscr{L}}\right] = \left(\prod_{l\in\mathscr{L}} \lambda_l\right) I(L,0)$$

and

$$\mathbf{V}\mathrm{ar}(X_{\mathscr{L}}) = m_{0,\mathscr{L}} + \sum_{k=1}^{L-1} \left(\sum_{\substack{\mathcal{J}\subset\mathscr{L}\\ \#\mathcal{J}=k}} \prod_{j\in\mathcal{J}} \lambda_j^2 \prod_{l\notin\mathcal{J}} \lambda_l \right) I(L,k),$$

where the $I(L,k)$ are given by Proposition 1.2.1 below.

The proof, given in [Chevallier and Laloë, 2015], relies on the calculus of the moments of a sum over a Poisson Process and is . The integral $I(L,k)$ can be seen as the contribution of a subset of k neurons to the number of coincidences between the L neurons.

Proposition 1.2.1 *For $b > a \geq 0$ and $0 < \delta < b - a$, define for every k in $\{0, \ldots, L\}$*

$$I(L,k) = \int\limits_{[a,b]^{L-k}} \left(\int\limits_{[a,b]^k} \mathbf{1}_{\left|\max\limits_{i\in\{1,\ldots,L\}} x_i - \min\limits_{i\in\{1,\ldots,L\}} x_i\right| \leq \delta} dx_1 \ldots dx_k \right)^2 dx_{k+1} \ldots dx_L,$$

where the convention $\int\limits_{[a,b]^0} f(x)\,dx = f(x)$ is set. Then, for $L \geq 2$, and k in $\{0, \ldots, L-1\}$,

- $I(L,L) = L^2 (b-a)^2 \delta^{2L-2} - 2L(L-1)(b-a)\delta^{2L-1} + (L-1)^2 \delta^{2L}$,
- $I(L,k) = f(L,k)(b-a)\delta^{L+k-1} - h(L,k)\delta^{L+k}$,
  where $f(L,k) = \dfrac{k(k+1) + L(L+1)}{L-k+1}$,
  and $h(L,k) = \dfrac{-k^3 + k^2(2+L) + k(5+2L-L^2) + L^3 + 2L^2 - L - 2}{(L-k+2)(L-k+1)}$.

Once the behaviour of $X_{\mathscr{L}}$ under $\mathcal{H}_0$ is known, the method to construct an independence test is straight-forward. Suppose that M independent and identically distributed (i.i.d.) trials are given. Denote $N_i^{(k)}$ the spike train corresponding to the neuron i during the k^{th} trial. As for the UE method, the idea is to compare two estimates of the expectation of $X_{\mathscr{L}}$. The first one is the empirical mean of $X_{\mathscr{L}}$:

$$\bar{m}_{\mathscr{L}} = \frac{1}{M} \sum_{k=1}^{M} X_{\mathscr{L}}^{(k)}, \tag{1.3}$$

where $X_{\mathscr{L}}^{(k)}$ is the delayed coincidence count during the k^{th} trial. This estimate converges even if the processes are not independent. More precisely the following asymptotic result is given by the Central Limit Theorem

$$\sqrt{M}\left(\bar{m}_{\mathscr{L}} - \mathbb{E}\left[X_{\mathscr{L}}\right]\right) \underset{M\to\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}\left(0, \mathbf{V}\mathrm{ar}(X_{\mathscr{L}})\right),$$

where $\xrightarrow{\mathcal{D}}$ denotes the convergence of distribution and $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2.

The second estimate is given by Theorem 1.2.1. Indeed, under $\mathcal{H}_0$ the following equality holds

$$\mathbb{E}\left[X_{\mathscr{L}}\right] = m_{0,\mathscr{L}} = \left(\prod_{l\in\mathscr{L}} \lambda_l\right) I(L,0).$$

Replacing each spiking intensity λ_l by

$$\hat{\lambda}_l := \frac{1}{M(b-a)} \sum_{k=1}^{M} N_l^{(k)}([a,b]),$$

where $N_l^{(k)}([a,b])$ denotes the number of spikes in $[a,b]$ for neuron l during the k^{th} trial, gives the following estimator,

$$\hat{m}_{0,\mathscr{L}} = \left(\prod_{l\in\mathscr{L}} \hat{\lambda}_l\right) I(L,0). \tag{1.4}$$

Note that $\bar{m}_{\mathscr{L}}$ is always consistent (that is, converges towards the true parameter) whereas $\hat{m}_{0,\mathscr{L}}$ is consistent only under $\mathcal{H}_0$. This leads to the following independence test: the independence assumption is rejected when the difference between $\bar{m}_{\mathscr{L}}$ and $\hat{m}_{0,\mathscr{L}}$ is too large. More precisely, Theorem 1.2.2 gives the asymptotic behaviour of $\sqrt{M}\left(\bar{m}_{\mathscr{L}} - \hat{m}_{0,\mathscr{L}}\right)$ under $\mathcal{H}_0$.

Theorem 1.2.2 *Under the notations and assumptions of Theorem 1.2.1, and under $\mathcal{H}_0$, the following affirmations are true*

- *The following convergence of distribution holds:*

$$\sqrt{M}\left(\bar{m}_{\mathscr{L}} - \hat{m}_{0,\mathscr{L}}\right) \underset{M\to\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}\left(0,\sigma^2\right),$$

with

$$\sigma^2 = \mathbf{V}\mathrm{ar}(X_{\mathscr{L}}) - (b-a)^{-1}\mathbb{E}\left[X_{\mathscr{L}}\right]^2 \left(\sum_{l\in\mathscr{L}} \lambda_l^{-1}\right).$$

- *Moreover, σ^2 can be estimated by*

$$\hat{\sigma}^2 = \hat{v}\left(X_{\mathscr{L}}\right) - (b-a)^{-1} I(L,L) \prod_{l\in\mathscr{L}} \hat{\lambda}_l^2 \left(\sum_{l\in\mathscr{L}} \hat{\lambda}_l^{-1}\right),$$

where

$$\hat{v}(X_{\mathscr{L}}) = \hat{m}_{0,\mathscr{L}} + \sum_{k=1}^{L-1} \left(\sum_{\substack{\mathcal{J}\subset\mathscr{L}\\ \#\mathcal{J}=k}} \prod_{j\in\mathcal{J}} \hat{\lambda}_j^2 \prod_{l\notin\mathcal{J}} \hat{\lambda}_l\right) I(L,k),$$

and the following convergence of distribution holds:

$$\sqrt{M}\frac{\left(\bar{m}_{\mathscr{L}} - \hat{m}_{0,\mathscr{L}}\right)}{\sqrt{\hat{\sigma}^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The proof of this theorem relies on a standard application of the delta method [Casella and Berger, 2002] and is given in [Chevallier and Laloë, 2015]. Note that the results obtained in Theorems 1.2.1 and 1.2.2 are true for more general delayed coincidence counts. However when one considers more general ways to count coincidences the integrals $I(L,k)$ are harder to compute.

1.2.2 Independence test

The results obtained in Theorem 1.2.2 allow us to straightforwardly build a test for detecting a dependency between neurons:

Definition 1.2.1 (The GAUE test) *For α in $]0,1[$, denote z_α the α-quantile of the standard Gaussian distribution $\mathcal{N}(0,1)$. Then the symmetric test of level α rejects $\mathcal{H}_0$ when $\bar{m}$ and $\hat{m}_{0,\mathscr{L}}$ are too different, that is when*

$$\left|\sqrt{M}\frac{(\bar{m}_{\mathscr{L}} - \hat{m}_{0,\mathscr{L}})}{\sqrt{\hat{\sigma}^2}}\right| > z_{1-\alpha/2}.$$

Note that once a subset is rejected by our test, one can determine if the dependency is rather excitatory or inhibitory according to the sign of $\bar{m}_{\mathscr{L}} - \hat{m}_{0,\mathscr{L}}$. If $\bar{m}_{\mathscr{L}} - \hat{m}_{0,\mathscr{L}} > 0$ (*respectively* < 0) then the dependency is rather excitatory (*respectively inhibitory*).

The result of a test may be wrong in two distinct manners. On the one hand, a false positive is an error in which the test is incorrectly rejecting the null hypothesis. On the other hand, a *false negative* is an error in which the test is incorrectly accepting the null hypothesis. The false positive (respectively negative) rate is the test's probability that a false positive (resp. negative) occurs. Usually, a theoretical control is given only for the false positive rate which is considered as the worst error. The following corollary is an immediate consequence of Theorem 1.2.2 and states the appropriateness of the GAUE test.

Corollary 1.2.1 *Under assumptions of Theorem 1.2.2, the test of level α presented in Definition 1.2.1 is asymptotically of false positive rate α. That is, the false positive rate of the test tends to α when the sample size M tends to infinity.*

1.3 Empirical study: Non-Poissonian framework

In this section, a more neurobiologically realistic framework than the Poisson one is considered (a similar simulation study under the Poisson framework is proposed in [Chevallier and Laloë, 2015]). Indeed, it is interesting to see if our test is still reliable when the Poisson framework is not valid anymore. Our test is confronted to multivariate Hawkes processes, which can be simulated thanks to Ogata's thinning method [Ogata, 1981] inspired by [Lewis and Shedler, 1979]. The use of Hawkes processes in neurobiology was first introduced in [Chornoboy et al., 1988]. With the development of simultaneous neuron recordings there is a recent trend in favour of Hawkes processes for modelling spike trains [Krumin et al., 2010, Pernice et al., 2011, 2012, Pillow et al., 2008, Tuleau-Malot et al., 2014]. Furthermore, Hawkes processes have passed some goodness-of-fit tests on real data [Reynaud-Bouret et al., 2014]. In this model, interaction between two neurons can be easily and in a more realistic way inserted. Note that the homogeneous Poisson process is a particular case of Hawkes processes, with no interaction between neurons.

A counting process N is characterized by its conditional intensity λ_t which is related with the local probability of finding a new point given the past. Informally, the quantity $\lambda_t dt$ gives the probability that a new point on N appears in $[t, t+dt]$ given the past. The process $\left(N^i\right)_{i=1\dots n}$ is a multivariate Hawkes process if there exist some functions $(h_{ij})_{i,j=1\dots n}$ (called interaction functions) and some positive constants $(\mu_i)_{i=1\dots n}$ (spontaneous intensities) such that, for all $j = 1, \dots, n$, λ^j given by

$$\lambda_t^j = \max\left(0, \mu_j + \sum_{i=1}^{n} \int_{s<t} h_{ij}\left(t-s\right) N^i\left(ds\right)\right)$$

is the intensity of the point process N^j, where $N^i(ds)$ is the point measure associated to N^i, that is $N^i(ds) = \sum_{T\in N^i} \delta_T(ds)$ where δ_T is the Dirac measure at point T.

The functions h_{ij} represent the influence of neuron i over neuron j in terms of spiking intensity. This influence can be either exciting ($h \geq 0$) or inhibiting ($h \leq 0$). For example, suppose that $h_{ij} = \beta \mathbf{1}_{[0,x]}$. If $\beta > 0$

(*respectively* $\beta < 0$) then the apparition of a spike on N^i increases (*respectively decreases*) the probability to have a spike on N^j during a short period of time (namely x): neuron i excites (*respectively inhibits*) neuron j. The processes N^i for $i = 1, \dots, n$ are independent if and only if $h_{ij} = 0$ for all $i \neq j$.

Note also that the self-interaction functions h_{jj} can model refractory periods, making the Hawkes model more realistic than Poisson processes, even in the independence case. In particular when $h_{jj} = -\mu_j \mathbf{1}_{[0,x]}$, all the other interaction functions being null, the n-dimensional process is composed by n independent Poisson processes with dead time x, modelling strict refractory periods of length x [Reimer et al., 2012].

All the following tests are computed according to the Framework $\mathbf{F}_1$ below:

- the trial duration of $b - a$ is randomly selected (uniform distribution) between 0.2 and 0.4s;
- the $n = 4$ neurons are simulated with spontaneous intensity $\mu_1, \dots, \mu_4$ randomly selected (uniform distribution) between 8 and 20Hz;
- the non-positive auto interaction functions are given by $h_{i,i} = -\mu_i \mathbf{1}_{[0,0.003s]}$;
- the set of tested neurons is given by $\mathcal{L} = \{1, 2, 3, 4\}$;

$\Bigg\} \ \mathbf{F}_1$

Note that a parameter scan is available in [Chevallier and Laloë, 2015].

1.3.1 Illustration of the level

Before all, one wants to know if Theorem 1.2.2 and Corollary 1.2.1 are still reliable for Hawkes processes. Thus Figure 1.2.A shows the evolution of the Kolmogorov distance $KS(F_{M,1000}, F)$ between the empirical distribution function over the 1000 repetition $F_{M,1000}$ and the standard Gaussian distribution function F:

$$KS(F_{M,1000}, F) = \sup_x |F_{M,1000}(x) - F(x)|.$$

Then we look at the Kolmogorov distance between the empirical distribution function of the p-values and the uniform distribution function to see if one can trust the level of the different tests (Figure 1.2.B). Finally, Figure 1.2.C presents the sorted p-values in function of their normalized rank (for $M = 50$). We see that our test is rather conservative whereas the UE test rejects too many cases.

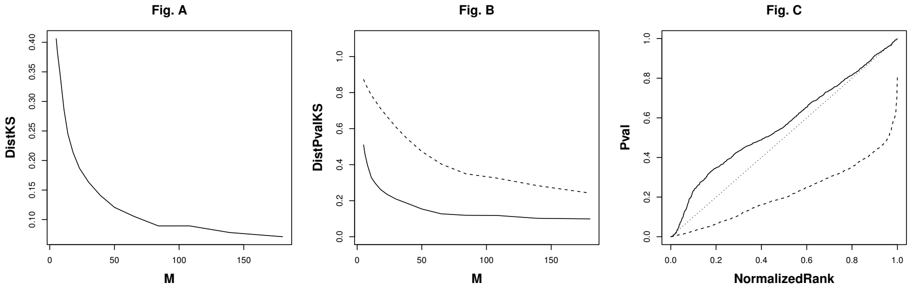


Figure 1.2: Under Framework $\mathbf{F}_1$ (Independence assumption). **Figure A.** Evolution of the Kolmogorov distance (in function of the number of trials) averaged on 1000 simulations between the empirical distribution function of the test statistics and the standard Gaussian distribution function. **Figure B.** Evolution of the Kolmogorov distance averaged on 1000 simulations between the empirical distribution function of the p-values and the uniform distribution function with respect to the number of trials. The plain line stands for our test and the dashed line for the original UE one. **Figure C.** Graphs of the sorted 1000 p-values (for 50 trials) in function of their normalized rank under $\mathcal{H}_0$. The plain line stands for our test, the dashed line for the original UE one and the dotted line for the uniform distribution function.

1.3.2 Illustration of the true positive rate

As said previously, it is more realistic to introduce dependency between Hawkes processes than Poisson processes. Still considering Framework $\mathbf{F}_1$, interaction functions $h_{i,j} = \beta \mathbf{1}_{[0,0.005s]}$, β being randomly selected between 20 and 30 Hz, are added. More precisely, we add five interaction functions: $h_{1,3}$, $h_{2,3}$, $h_{1,4}$, $h_{2,4}$ and $h_{3,4}$ (summarized in Figure 1.3). Moreover, the auto interactions are updated to preserve strict refractory periods : $h_{i,i} = -(\mu_i + m_i.\beta)\mathbf{1}_{[0,0.003s]}$, where m_i is the number of neurons exciting neuron i (for example, $m_4 = 3$). This new framework ($\mathbf{F}_1$ completed by the five interaction function) is referred as Framework $\mathbf{F}_2$.

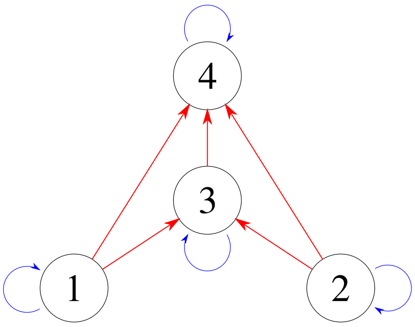


Figure 1.3: Local independence graph. An arrow means a non null interaction function. Blue arrow means inhibition and red arrow means excitation.

As previously we first provide an illustration of the true positive rate of the two tests, associated to a theoretical level of 5%, in function of M (Figure 1.4.A). Then Figure 1.4.B represents the p-values in function of their normalized rank, for $M = 50$. The difference between the true positive rates is smaller than in the Poissonian Case.

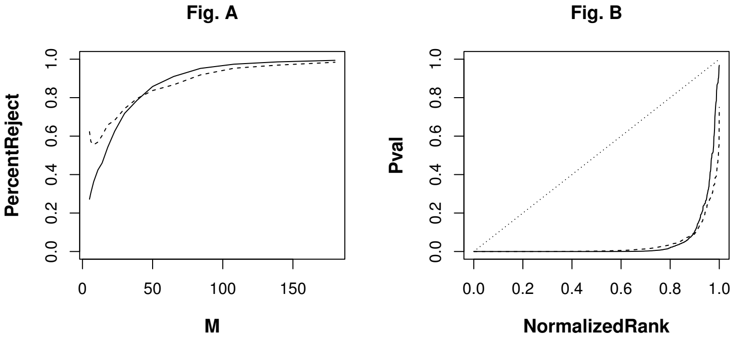


Figure 1.4: Under Framework $\mathbf{F}_2$ (Dependence assumption, see Figure 1.3). **Figure A.** Illustration of the true positive rate of the test, for a theoretical test level of 5%. The curves represent the evolution, with respect to the number of trials, of the true positive rate (averaged on 1000 simulations). The plain line stands for our test and the dashed line for the original UE one. **Figure B.** Graphs of the sorted 1000 p-values for 50 trials. The plain line stands for our test, the dashed line for the original UE one and the dotted line for the uniform distribution function.

1.3.3 Multiple pattern test

In the original MTGAUE method, a multiple testing procedure is applied with respect to 1900 sliding time windows. In our framework, we cannot guarantee the relevance of the multiple test with this high order of multiplicity (due to the default of the Gaussian approximation and, more precisely, to the excess of very small p-values). However, we are able to propose a multiple testing procedure with respect to the different possible patterns. For example, with four neurons there are eleven different possible patterns, which gives a much lower order of multiplicity. So, the multiple test over all the eleven sub-pattern of two, three or four neurons is presented here.

In multiple testing, the notion of false positive rate is not relevant. The closest notion might be the *Family-Wise Error Rate* (FWER) which is the probability to wrongly reject at least one of the tests. This error rate can be controlled using Bonferroni's method but it is too restrictive, in particular when the number K of tests involved is too large. One popular way to deal with multiple testing is the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] which ensures a control of the *False Discovery Rate* (FDR). False discoveries cannot be avoided but it is not a problem if the ratio of F_p (the number of false positives detections) divided by R (the total number of rejects) is controlled. Therefore, the FDR is mathematically defined by $\text{FDR} = \mathbb{E}\left[F_p/R\ \mathbf{1}_{R>0}\right]$. The following procedure, due to Benjamini and Hochberg ensures a small FDR over K tests:

1. Fix a level q ($q = 5\%$ for example);
2. Denote by $(P_1, \ldots, P_K)$ the p-values obtained for all considered tests;
3. Order them in increasing order and denote the increasing vector $(P_{(1)}, \ldots, P_{(K)})$;
4. Note k_0 the largest k such that $P_{(k)} \leq kq/K$;
5. Then, reject all the tests corresponding to p-values smaller than $P_{(k_0)}$.

The theoretical result of [Benjamini and Hochberg, 1995] ensures that if the p-values are upper bounded by a uniform distribution and independently distributed under the null hypothesis, then the procedure guarantees a FDR less than q. The main drawback of this procedure in our case is that one needs to compute p-values that are very small when K is large. For example, if $K \geq 50$ and $q = 5\%$, the upper bound given by kq/K can be smaller than 0.001 and the empirical frequency of very small p-values is greater than expected and therefore the uniform upper bound of the p-values is not guaranteed in our case. However, only 11 tests are considered here and the procedure still returns reliable results. We perform 1000 simulations and count how many times each test rejects the independence. The results, obtained for $M = 50$, are presented in Figure 1.5.

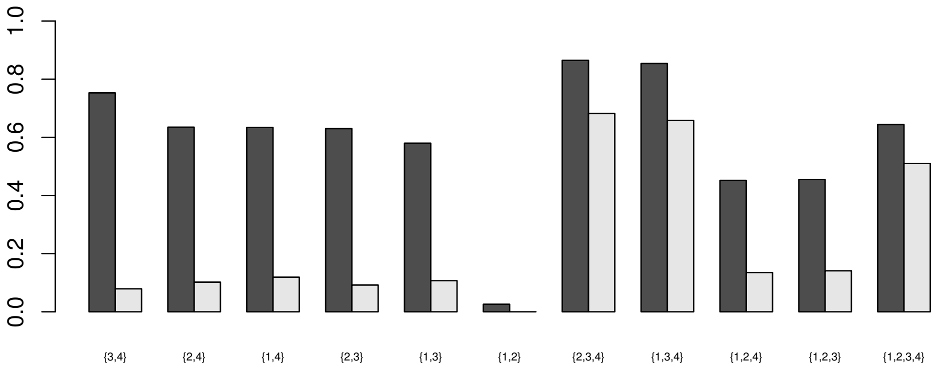


Figure 1.5: Under Framework $\mathbf{F}_2$ (Dependence assumption, see Figure 1.3). Frequency of dependence detections (1000 simulations) for each pattern. Grey for our test, white for the original UE method.

The results show that our test detects all patterns except $\{1,2\}$. This is consistent with the considered framework ($\mathbf{F}_2$) since we simulate connections between all pairs of neurons except $\{1,2\}$. The U.E. test essentially detects the patterns $\{2,3,4\}$, $\{1,3,4\}$, $\{1,2,3,4\}$ and to a lesser extent $\{1,2,4\}$ and $\{1,2,3\}$. Moreover, it misses all the pairs.

1.4 Illustration on real data

We apply our method on real data and show results in agreement with classical knowledge on those data.

1.4.1 Description of the data

The data set considered here is the same as in [Tuleau-Malot et al., 2014] and previous experimental studies [Grammont and Riehle, 2003, Riehle et al., 2000, 2006]. The following description of the experiment is copied from Section 4.1 of [Tuleau-Malot et al., 2014]. These data were collected on a 5-year-old male Rhesus monkey who was trained to perform a delayed multi-directional pointing task. The animal sat in a primate chair in front of a vertical panel on which seven touch-sensitive light-emitting diodes were mounted, one in the centre and six placed equidistantly (60 degrees apart) on a circle around it. The monkey had to initiate a trial by touching and then holding with the left hand the central target. After a fix delay of 500ms, the preparatory signal (PS) was presented by illuminating one of the six peripheral targets in green. After a delay of either 600ms (with probability 0.3) or 1200ms (with probability 0.7), it turned red, serving as the response signal and pointing target. Signals recorded from up to seven micro-electrodes (quartz insulated platinum-tungsten electrodes, impedance: 2-5MΩ at 1000Hz) were amplified and band-pass filtered from 300Hz to 10kHz. Using a window discriminator, spikes from only one single neuron per electrode were then isolated. Neuronal data along with behavioural events (occurrences of signals and performance of the animal) were stored on a PC for off-line analysis with a time resolution of 10kHz. The idea of the analysis is to detect some conspicuous patterns of coincident spike activity appearing during the response signal in the case of a long delay (1200ms). Therefore, we only consider trials where the response signal is indeed occurring after a long delay.

1.4.2 The test

We have at hand the following data: spike trains associated to four neurons (35 trials by neurons). We consider two sub windows: one between 300ms and 500ms (i.e. before the preparatory signal), the other between 1100ms and 1300ms (i.e. around the expected signal). Our idea is that more synchronisation should be detected during the second window. Moreover, we do not only want to test if the four considered neurons are independent (that is perform our test on the complete pattern $\{1,2,3,4\}$). Indeed one can be interested in knowing if neurons in some sub-patterns (for example $\{1,2\}$ or $\{1,3,4\}$ are independent. That is why we use the multiple pattern test procedure defined at the end of Section 1.3 to test all the eleven subsets (of at least two neurons) of the four considered neurons are tested. Thus we use the Benjamini-Hochberg procedure (presented in the previous section) for $K = 22$ tests. Moreover, we took several values for the delay δ between 0.01s and 0.025s and the results remained stable.

The results are presented in Figure 1.6. Note that we saw in section 1.3 that our test is too conservative even for small number of trials. This ensures that the theoretical level of our test can be trusted. We see that synchronizations between the subsets $\{3,4\}$ and $\{1,3,4\}$ appear in the second window. These results suggest that neurons 1, 3 and 4 belong to a neuronal assembly which is formed around the expected signal. This is in agreement with more quantitative results on those data [Grammont and Riehle, 2003, Tuleau-Malot et al., 2014].

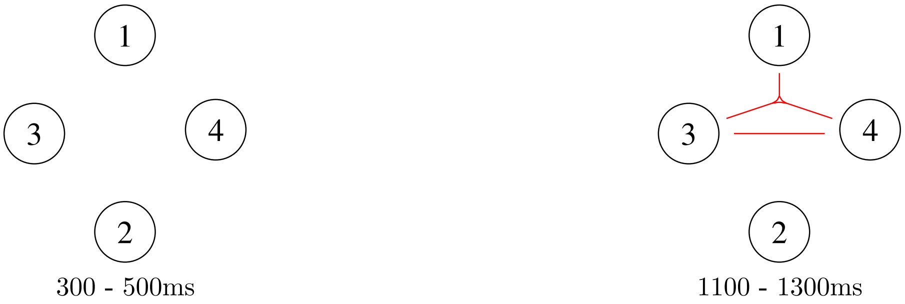


Figure 1.6: Evolution of the synchronization between neurons. The lines indicate the subset for which our test detects dependence. Here we detect an excess of coincidences between neurons $\{1, 3, 4\}$ and $\{3, 4\}$

Conclusion and prospects

Pattern detection

I presented in this chapter a generalization of the delayed coincidence count performed in [Tuleau-Malot et al., 2014] to more than two neurons. This delayed coincidence count leads to an independence test for point processes which are commonly used to model spike trains.

Under the hypothesis that the point processes are homogeneous Poisson processes, the expectation and variance of the delayed coincidence count can be computed (Theorem 1.2.1), and then a test with prescribed asymptotic level is built (Theorem 1.2.2). A simulation study allows us to confirm our theoretical results and to state the empirical validity of our test with a relaxed Poisson assumption. Indeed, we considered Hawkes processes which are a more realistic model of spike trains. The simulation study gives good results, even for small sample size. This allows us to use our test on real data, in order to highlight the emergence of a neuronal assembly involved at some particular time of the experiment.

We achieved the full generalization of the single test procedure introduced in [Tuleau-Malot et al., 2014]. However, we could not achieve the multiple time windows testing procedure mainly because of the default of Gaussian approximation concerning extreme values of the test statistics. More precisely, very small p-values are not distributed as expected. In particular when the sample size M is moderate ($M = 50$), our test returns too many very small p-values. In [Tuleau-Malot et al., 2014], the MTGAUE method is applied simultaneously on 1900 sliding windows. In the present work, in order to apply multiple testing both with respect to the sliding time windows and the subsets, the total number of tests is even larger. Indeed, for each sliding window, there are $2^n - n - 1$ tests to perform, where n is the number of recorded neurons. This would lead to extremely small p-values, for which our test is less reliable.

Even if our test remains empirically reliable under a non Poissonian framework, it could be therefore of interest to explore surrogate data method such as trial-shuffling [Pipa et al., 2003]. A very recent work based on permutation approach for delayed coincidence count with $n = 2$ neurons [Albert et al., 2015] is a first step in this direction but needs to be generalized to more than 2 neurons.

Ongoing PhD in Cognitive science I am also interested in other levels of cognition. On this spirit I am currently co-supervising (with a psychologist : F. Mathy and P. Reynaud-Bouret) the PhD of Giulia Mezzadri.

In cognitive psychology, the literature showed that the manipulated ordering of stimuli seems to determine the efficiency of a learning session for a human being [Kornell and Bjork, 2008, Mathy and Feldman, 2009]. The goal of this PhD is to model in detail the temporal effect of the presenting order. In a more general way, our goal is there to understand how a human being learns a classification rule.

A first step is to define the proper distance between the presented stimuli (taking into account the order of presentation) allowing to efficiently test the initial assumption on the importance of the presentation order. Then we could be able to generalize the General Context Model (GCM) developed by [Nosofsky, 1986] to include the presentation order in the model. Finally, goodness-of-fit test should be developed to establish the relevance of our model.

Bibliography of the third part

Bibliography

M. Abeles. Quantification, smoothing, and confidence limits for single-units' histograms. *Journal of Neuroscience Methods*, 5(4):317–325, 1982. URL https://doi.org/10.1016/0165-0270(82)90002-4.

A. M. Aertsen, G. L. Gerstein, M. K. Habib, and G. Palm. Dynamics of Neuronal Firing Correlation: Modulation of "Effective Connectivity". *Journal of Neurophysiology*, 61:900–917, 1989. URL https://doi.org/10.1152/jn.1989.61.5.900.

M. Albert, Y. Bouret, M. Fromont, and P. Reynaud-Bouret. Bootstrap and permutation tests of independence for point processes. *The Annals of Statistics*, 43(6):2537–2564, 2015. URL https://doi.org/10.1214/15-AOS1351.

E. W. Archer, I. M. Park, and J. W. Pillow. Bayesian entropy estimation for binary spike train data using parametric prior knowledge. In *Advances in Neural Information Processing Systems*, pages 1700–1708, 2013. URL https://arxiv.org/abs/1302.0328.

H. B. Barlow. Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394, 1972. URL https://doi.org/10.1068/p010371.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B.*, 57(1):289–300, 1995. URL http://links.jstor.org/sici?sici=0035-9246(1995)57:1%3C289:CTFDRA%3E2.0.CO.

C. Borgelt and D. Picado-Muino. Finding frequent patterns in parallel point processes. In *Advances in Intelligent Data Analysis XII*, pages 116–126. Springer, 2013. URL https://link.springer.com/chapter/10.1007%2F978-3-642-41398-8_11.

E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 26(7):456–461, 2004. URL http://dx.doi.org/10.1002/nav.3800260304.

G. Casella and R. Berger. *Statistical Inference*. Duxbury, 2002.

J. Chevallier and T. Laloë. Detection of dependence patterns with delay. *Biometrical Journal*, 57(6):1110–1130, 2015. URL https://doi.org/10.1002/bimj.201400235.

E. Chornoboy, L. Schramm, and A. Karr. Maximum likelihood identification of neural point process systems. *Biological Cybernetics*, 59(4-5):265–275, 1988. URL http://dx.doi.org/10.1007/BF00332915.

Y. Dong, S. Mihalas, F. Qiu, R. von der Heydt, and E. Niebur. Synchrony and the binding problem in macaque visual cortex. *Journal of vision*, 8(7):30, 2008. URL http://jov.arvojournals.org/article.aspx?articleid=2193657.

G. L. Gerstein and D. H. Perkel. Simultaneously Recorded Trains of Action Potentials: Analysis and Functional Interpretation. *Science*, 164(3881):828–830, 1969. URL http://www.sciencemag.org/content/164/3881/828.abstract.

G. L. Gerstein, A. M. Aertsen, et al. Representation of cooperative firing activity among simultaneously recorded neurons. *Journal of Neurophysiolology*, 54(6):1513–1528, 1985. URL https://doi.org/10.1152/jn.1985.54.6.1513.

F. Grammont and A. Riehle. Spike synchronization and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time. *Biological Cybernetics*, 88(5):360–373, 2003. URL https://doi.org/10.1007/s00422-002-0385-3.

S. Grün. *Unitary joint events in multiple neuron spiking activity: detection, significance, and interpretation.* PhD thesis, 1996.

S. Grün. Data-driven significance estimation for precise spike correlation. *Journal of Neurophysiology*, 101(3): 1126–1140, 2009. URL https://doi.org/10.1152/jn.00093.2008.

S. Grün, M. Diesmann, F. Grammont, A. Riehle, and A. Aertsen. Detecting unitary events without discretization of time. *Journal of neuroscience methods*, 94(1):67–79, 1999. URL https://doi.org/10.1016/s0165-0270(99)00126-0.

S. Grün, M. Diesmann, and A. Aertsen. Unitary Events in Multiple Single-Neuron Spiking Activity: I. Detection and Significance. *Neural Computation*, 14(1):43–80, 2002. URL https://doi.org/10.1162/089976602753284455.

M. T. Harrison and S. Geman. A rate and history-preserving resampling algorithm for neural spike trains. *Neural Computation*, 21(5):1244–1258, 2009. URL https://doi.org/10.1162/neco.2008.03-08-730.

D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory.* Wiley, 1949.

J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability.* The Clarendon Press Oxford University Press, 1993. URL https://doi.org/10.1002/0470011815.b2a07042.

N. Kornell and R. Bjork. Learning concepts and categories is spacing the enemy of induction? *Psychological science*, 19:585–92, 2008. URL https://doi.org/10.1111/j.1467-9280.2008.02127.x.

M. Krumin, I. Reutsky, and S. Shoham. Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Frontiers in computational neuroscience*, 4, 2010. URL https://doi.org/10.3389/fncom.2010.00147.

E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses.* Springer-Verlag New York Inc, 3rd edition, 2005. URL https://link.springer.com/book/10.1007%2F0-387-27605-X.

P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res. Logist. Quart.*, 26(3):403–413, 1979. URL http://dx.doi.org/10.1002/nav.3800260304.

S. Louis, C. Borgelt, and S. Grün. Generation and selection of surrogate methods for correlation analysis. In *Analysis of Parallel Spike Trains*, pages 359–382. Springer, 2010a. URL https://doi.org/10.1007/978-1-4419-5675-0_17.

S. Louis, G. L. Gerstein, S. Grün, and M. Diesmann. Surrogate spike train generation through dithering in operational time. *Frontiers in computational neuroscience*, 4, 2010b. URL https://doi.org/10.3389/fncom.2010.00127.

F. Mathy and J. Feldman. A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, 16(6):1050–1057, 2009. URL https://doi.org/10.3758/PBR.16.6.1050.

D. P. Muiño and C. Borgelt. Frequent item set mining for sequential data: Synchrony in neuronal spike trains. *Intelligent Data Analysis*, 18(6):997–1012, 2014. URL https://doi.org/10.1002/widm.1074.

R. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115(1):39–57, 1986. URL https://doi.org/10.1037//0096-3445.115.1.39.

Y. Ogata. On Lewis simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1): 23–30, 1981. URL https://doi.org/10.1109/tit.1981.1056305.

G. Palm. Cell assemblies as a guideline for brain research. *Concepts in Neuroscience*, 1(1):133–147, 1990.

D. H. Perkel, G. L. Gerstein, and G. P. Moore. Neuronal Spike Trains and Stochastic Point Processes: II. Simultaneous Spike Trains. *Biophysical Journal*, 7(4):419–440, 1967. URL http://www.sciencedirect.com/science/article/pii/S0006349567865974.

V. Pernice, B. Staude, S. Cardanobile, and S. Rotter. How structure determines correlations in neuronal networks. *PLoS computational biology*, 7(5):e1002059, 2011. URL https://doi.org/10.1371/journal.pcbi.1002059.

V. Pernice, B. Staude, S. Cardanobile, and S. Rotter. Recurrent interactions in spiking networks with arbitrary topology. *Physical Review E*, 85(3):031916, 2012. URL https://doi.org/10.1103/physreve.85.031916.

J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008. URL https://doi.org/10.1038/nature07140.

G. Pipa and S. Grün. Non-parametric significance estimation of joint-spike events by shuffling and resampling. *Neurocomputing*, 52:31–37, 2003. URL https://doi.org/10.1016/s0925-2312(02)00823-8.

G. Pipa, M. Diesmann, and S. Grün. Significance of joint-spike events based on trial-shuffling by efficient combinatorial methods. *Complexity*, 8(4):79–86, 2003. URL https://doi.org/10.1002/cplx.10085.

I. C. Reimer, B. Staude, W. Ehm, and S. Rotter. Modeling and analyzing higher-order correlations in non-Poissonian spike trains. *Journal of neuroscience methods*, 208(1):18–33, 2012. URL https://doi.org/10.1016/j.jneumeth.2012.04.015.

P. Reynaud-Bouret, V. Rivoirard, F. Grammont, and C. Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience (JMN)*, 4(1):1–41, 2014. URL https://doi.org/10.1186/2190-8567-4-3.

A. Riehle, F. Grammont, M. Diesmann, and S. Grün. Dynamical changes and temporal precision of synchronized spiking activity in monkey motor cortex during movement preparation. *Journal of Physiology-Paris*, 94(5): 569–582, 2000. URL https://doi.org/10.1016/s0928-4257(00)01100-1.

A. Riehle, F. Grammont, and W. A. MacKay. Cancellation of a planned movement in monkey motor cortex. *Neuroreport*, 17(3):281–285, 2006. URL https://doi.org/10.1097/01.wnr.0000201510.91867.a0.

A. Roy, P. N. Steinmetz, S. S. Hsiao, K. O. Johnson, and E. Niebur. Synchrony: a neural correlate of somatosensory attention. *Journal of neurophysiology*, 98(3):1645–1661, 2007. URL https://doi.org/10.1152/jn.00522.2006.

Y. Sakurai. How do cell assemblies encode information in the brain? *Neuroscience & Biobehavioral Reviews*, 23(6):785–796, 1999. URL https://doi.org/10.1016/S0149-7634(99)00017-2.

S. Shinomoto. Estimating the Firing Rate. In *Analysis of Parallel Spike Trains*, volume 7 of *Springer Series in Computational Neuroscience*, pages 21–35. Springer US, 2010. URL http://dx.doi.org/10.1007/978-1-4419-5675-0_2.

W. Singer. Synchronization of Cortical Activity and its Putative Role in Information Processing and Learning. *Annual Review of Physiology*, 55:349–374, 1993. URL https://doi.org/10.1146/annurev.ph.55.030193.002025.

E. Stark and M. Abeles. Unbiased estimation of precise temporal correlations between spike trains. *Journal of neuroscience methods*, 179(1):90–100, 2009. URL https://doi.org/10.1016/j.jneumeth.2008.12.029.

C. Tuleau-Malot, A. Rouis, F. Grammont, and P. Reynaud-Bouret. Multiple Tests Based on a Gaussian Approximation of the Unitary Events Method with delayed coincidence count. *Neural Computation*, 26:7, 2014. URL https://doi.org/10.1162/neco_a_00604.

C. von der Malsburg. The Correlation Theory of Brain Function. Internal Report 81-2, Department of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany, 1981. URL https://doi.org/10.1007/978-1-4612-4320-5_2.

independence
level-set
algorithme
dependence
neuroscience
neurones
synchronization
option
ensembles
convergence
multivariate
kernel
processus
detection
estimators
supervisé
design
data
niveau
learning
estimation
function
Statistics
random
risk
CCTE
clustering
statistique
c.d.f
risk
learning
K-means
neurons
Alter
copula
density
apprentissage
algorithm
algorithms
non-supervisé
X-Alter
drawbacks
neuroscience
multivarié
distribution
classification
probability
estimation
X-means
warping
compactness