Contributions to RSSI-based geolocation

Kevin Elgui

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAT047

Thèse de doctorat

Contributions to RSSI-based geolocation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 17 décembre 2020, par

KEVIN ELGUI

Composition du Jury :

Stéphanie Bidon Professeure, ISAE-Supaéro	Président et rapporteur
Maarten Weyn Professor, University of Antwerp, Belgium	Rapporteur
Aurélien Bellet Chargé de recherche, INRIA	Examinateur
Pascal Bianchi Professeur, Télécom Paris	Directeur de thèse
François Portier Maître de conférence, Télécom Paris	Co-directeur de thèse
Maurice Charbit Professeur, Télécom Paris	Invité
Olivier Isson Ingénieur de recherche, Sigfox	Invité

Contents

Remerciements

Je tiens tout d'abord à exprimer ma gratitude envers mes directeurs de thèse. Olivier, pour ton amitié, et parce que tu as su me montrer que l'on pouvait chercher en s'amusant, même si on ne trouvait rien. A Pascal, pour m'avoir guidé ces trois années, pour ta rigueur et ton excellence sans qui cette thèse n'aurait pas été possible. Et enfin à François, qui a été mon bol d'air frais, et qui m'aura permis d'étudier des sujets auxquels je n'aurais jamais pensé. Tu es le phare de ces locaux de Palaiseau ! J'aurai beaucoup appris durant ces années, et je vous le dois en grande partie.

I thank Stéphanie Bidon and Maarten Weyn for accepting to review this thesis. Je remercie également Aurélien Bellet et Maurice Charbit d'avoir accepté de constituer mon jury. Je vous en suis très reconnaissant.

Je tiens aussi à exprimer mes remerciements à la société Sigfox, et en particulier à l'équipe Data. Renaud, Oliv', le Duc: votre goût pour la recherche, votre confiance et votre intérêt inébranlable pour l'avancée de mes travaux ont formé un terreau idéal pour le déroulement de cette CIFRE. J'espère que les algorithmes développés pourront un jour atteindre une précision de localisation au mètre. Je suis convaincu que votre équipe continuera à faire de grandes choses. Merci également à tous les autres collègues de Sigfox Paris, grâce à qui aller au bureau était un plaisir renouvelé chaque vendredi (ou les jours de foot et de petits déjeuner). Donc à Pourya, Antho, Aurélien, Antoine, Laura C., Laura D. et j'en oublie... j'espère sincèrement vous retrouver autour d'un verre, d'un pizza, d'une pâtisserie ou d'un ballon rond !

Certaines rencontres sont bonnes pour la santé. Parmi les camarades qui ont rendu ces années si plaisantes, Alex, Charlito, le Quent, Ricco, Vous avez été, tous à votre manière, une source impérissable de bon air (doit-on dire débonnaire).

Pierre, Mathus, voi due siete i miei due accoliti, non oso pensare come sarebbero stati quegli anni senza di voi. Grazie.

À mes parents et à mes soeurs, qui m'auront toujours encouragé.

À Océane pour qui les mots m'ont toujours manqués, et à Zazou qui est venu illuminer ces derniers jours de doctorat.

Abstract

Network-Based Geolocation is one of the challenges of the 21st century. In the last decades, Internet of Things (IoT) has raised a great deal of attention in very diverse fields such as agriculture or health care. Since most of connected objects will need to be located, the need of designing a low-power consumption system allowing geopositioning without using GPS and GSM has grown.

Amongst geolocation techniques, RSSI-based geolocation stands out for several reasons: the RSSI is sufficiently explanatory and does not constitute a memory burden.

Predicting the location of an emitter from the RSSI can be split in two categories of methods, which are referred to as likelihood-based ones and fingerprinting ones. The first ones, considered in this manuscript, first learn a likelihood model for the conditional probability of the RSSI vector given the position; then, the second stage consists to finding the position which best agrees with the observed RSSI vector, by means of a grid search for instance. In contrast, the second ones directly map the vector of RSSI into a location (radio-map), typically by means of a supervised learning algorithm. In this work, we first propose improvements of methods used for the RSSI-based geolocation problem. The first proposed technique relies on a semi-parametric Nadaraya-Watson estimator of the likelihood, followed by a maximum a posteriori estimator of the object's position. The second technique consists to learning a distance, constructed by means of a Gradient boosting regressor: a k-nearest neighbor algorithm is then used to estimate the position. The proposed methods are compared on two data sets originated from Sigfox network, and an indoor dataset from a three-story building. Experiments demonstrate the interest of the proposed methods, both in terms of location estimation performance, and ability to build radio maps. Results also show that the quality of the prediction is highly related to the chosen distance on the RSSI space. The metric learning problem is therefore a fundamental issue to improve RSSI-based geolocation technique.

Second, we introduce an original objective for learning a similarity between pairs of data points. In this manuscript, we propose to build the similarity by directly minimizing the regression error of an estimator. We thus obtain a task-driven learning objective. To minimize it, the similarity is chosen as a sum of regression trees and is sequentially learned by means of a modified version of XGBoost detailed in this document. This method benefits from the well-known qualities of XGBoost such as its efficiency and its scaling capabilities. Furthermore, our similarity, although non-parametric, does not require a storage of the size of the dataset. Finally, experiments show that our model outperforms other kernel regression models on several benchmark datasets.

Conditional Independence has been broadly used in the RSSI-based geolocation literature in order to decrease complexity of statistical models such as the ones presented in this manuscript. Testing CI is therefore critical for the performance of such estimators.

We introduce the weighted partial copula function for testing conditional independence. The proposed test procedure results from the following ingredients: (i) the test statistic is an explicit Cramér-von Mises transformation of the weighted partial copula, (ii) the regions of rejection are computed using a boot-strap procedure which mimics conditional independence by generating samples. Under CI, the weak convergence of the weighted partial copula process is established and endorses the soundness of our approach. Experiment finally demonstrate the competitive power of our approach compared to recent state-of-the-art methods.

Notation

$\triangleq$	Equal by definition	
$[d]$	Set of integers from 1 to d included	
$\mathcal{Y}^{\mathcal{X}}$	Set of functions from $\mathcal{X}$ to $\mathcal{Y}$	
$\mathbb{R}^{d_1 \times d_2}$	Set of real matrices of size d_1 by d_2	
Id_n	Identity matrix in $\mathbb{R}^{n \times n}$	
$A_{i:}$	i^{th} row of matrix A	
$A_{:j}$	j^{th} column of matrix A	
$\mathrm{Tr}\, A$	Trace of $A \in \mathbb{R}^{d \times d}$	$\mathrm{Tr}\, A = \sum_{i=1}^{d} A_{ii}$
$A^\top$	Transpose of matrix A	
$A^\dagger$	Moore-Penrose pseudo-inverse of matrix A	
$\|\cdot\|$	Euclidean norm on vectors and matrices	
$\mathcal{S}_{++}^n$	Positive definite matrices of size $n \times n$	
$\mathcal{S}_{+}^n$	Semipositive definite matrices of size $n \times n$	
$\|\cdot\|_S$	Mahalanobis matrix norm induced by $S \in \mathcal{S}_{++}^n$	$\|A\|_S = \sqrt{\mathrm{Tr}(A^\top S A)}$
$\|\cdot\|_2$	Spectral norm on matrices	

$a \vee b$	Maximum of real numbers a and b	
$a \wedge b$	Minimum of real numbers a and b	
$(a)_+$	Positive part of $a \in \mathbb{R}$	$a \vee 0$
$\mathbf{0}$	Vector or matrix of zeros	
$\mathbf{1}$	Vector or matrix of ones	
$\mathbb{1}$	Indicator of an event	

1
Motivation and Contributions

"Je suis abasourdi par le nombre de personnes qui veulent 'connaître' l'univers alors qu'il est déjà suffisamment difficile de se repérer dans le quartier chinois de New York."

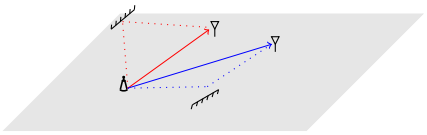
Woody Allen

With the advances in wireless communications and low-power electronics, accurate position location may now be accomplished by a number of techniques which involve commercial wireless services. Emerging position location systems, when used in conjunction with mobile communications services, will lead to enhanced public safety and revolutionary products and services. The fundamental technical challenges and business motivations behind wireless position location systems are described, and promising techniques for solving the practical position location problem are treated.

1.1 Context of this Thesis

This thesis is the result of a CIFRE agreement (Convention Industrielle de Formation et de Recherche - Industrial Training and Research Agreement) between Télécom Paris Saclay and Sigfox, a French telecommunications operator created in 2009 by Christophe Fourtet and Ludovic Le Moan. Sigfox is specialised in Machine to Machine (M2M) via low-speed networks. It contributes to the Internet of Things (IoT) by enabling interconnection via a gateway. Its UNB ("Ultra narrow band") radio technology enables it to build a low-speed, energy-efficient cellular network. This type of network is deployed in the so-called Industrial Scientific and Medical (ISM) radio frequency bands, available worldwide without any license.

In the last few years, Internet of Things (IoT) has raised a great deal of attention in very diverse fields such as agriculture or health care. Experts agree that 30 billions objects will be part of the IoT by 2023 and 40% of these objects will need to be geolocated, e.g. for freight transport (Hatton). One of the most significant challenges for the field is the need for localization. Indeed, numerous applications of sensor networks need to track mobile objects, such as people, animals, cars, etc. To make these applications viable, device cost will need to be low (from a few dollars to a few cents depending on the application) and devices will need to last for years or even decades without battery replacement. Additionally, the network will need to organize without significant human moderation. Moreover, in order to enable connectivity for billions of devices, most of IoT dedicated networks are using long-range and low power communications. Therefore, the challenges for IoT networks are to achieve high scalability to handle massive number of devices, to achieve low cost and to have wide coverage while keeping low energy consumption. Devices which have these demands are difficult to be integrated into



traditional cellular networks. That is why, LPWAN (Low Power Wide Area Network) dedicated technology, such as UNB (Ultra Narrow Band), developed and patented by Sigfox has emerged.

The knowledge of the geolocation of each device is a very valuable resource. Indeed, it allows Sigfox to provide this information to network users that leads to numerous applications such as in logistics or transport of merchandises, to monitor and track in smart buildings or even for proximity marketing and advertising in shopping malls. However, traditional localization techniques such as global positioning system (GPS) are thus not well-suited for the particular needs of the IoT industry. Providing a GPS on each device is cost and energy prohibitive for many applications, and furthermore not suited for indoor applications. All these requirements complicate greatly the localization of these objects.

Alternatively, range-based methods use measurements such as the two introduced here, the Time-of-Arrival (TOA) or Time-Difference-of-Arrival (TDoA) (Ho and Chan, 1993; Cong and Zhuang, 2002), and the Received Signal Strength Indicator (RSSI) to estimate the distance between an emitting device and a receiving antenna. Other ranging methods are commonly used in the literature, such the Angle of Arrival (AOA) (Niculescu and Nath, 2003), or the Frequency Difference of Arrival (FDoA) (Amar and Weiss, 2008) but their study is beyond the scope of this thesis.

This chapter allows us to put the stakes of this thesis into perspective. Its purpose is to provide a better understanding of the singularity of the geolocation problem in a sensor network such as the Sigfox network.

Outline In this chapter, we address the problem of channel-based location estimation techniques. First, we discuss the general problem of geolocation in Appendix B.2. To predict the location of an emitter, state-of-the-art channel-based approaches consist either in the estimation of the time delay from the transmitter to the receiver, from which it is possible to infer the distance between the emitter and the receiver, or to directly predict this latter distance from the observed power decay between the received power and the emission power. A introduction to the wireless channel, in particular the key parameters for model it, and a presentation of the discussed predictors are proposed in Section 1.3. Finally, the performance of both these approaches in the particular context of IoT communication are discussed in Section 1.4.

1.2 The estimation of the geographic location

1.2.1 Geolocation Principle

The terms geolocation and positioning are used to designate the real-world geographic estimation of the location of an object. This problem has entered our society with a massive diffusion. It is used in many applications such as navigation, communication, self-driving vehicles, connected objects and communicating cities, or more recently with issues such as controlling the amount of contamination in a population. It affects a wide variety of scientific fields such as Géolocalisation et Navigation par un Système

de Satellites (GNSS), on which our economy increasingly depends. Experts agree that 30% of the gross domestic product would depend partly on GNSS by 2030, compared to 10% today.

The geolocation of an object refers to the (latitude, longitude)-coordinates, that is the position of the object on the Earth's surface. Sometimes, the word positioning is rather employed when it comes to identify the location of an object in a particular space such as a cell-phones in a shopping mall, or a robot in a building. In this thesis, we shall speak about geolocation when it comes to a localization on a global scale, while the term positioning is generally used in indoor and/or confined areas.

The radio frequency methods are used for most systems of geolocation. This family of methods, also called radiolocation methods make use of characteristics of received radio waves to predict the location of an emitting object. The examples are numerous. The widely used *Global Positioning System* (GPS) (see Figure 1.1) is based on the estimation of the Time-of-Arrival (TOA) of a signal at a satellite. When the time of transmission, the speed of propagation and the position of the satellite are known, the TOA leads to a very good estimator of the distance between the emitting object and the satellite. Combining several TOA leads to the estimated position of the transmitter. The use of multiple receivers to locate a transmitter is known as *multilateration* (illustrated in Figure 1.2). In cellular telephony, the radiolocation is directly performed via the Base Stations (BS) of the cellular network by means of one or a combination of the following features:

- The TOA (or TDoA). In contrast with the GPS, these quantities are estimated w.r.t the BS of the cellular network.
- The Angle of Arrival (AOA) corresponds to the direction from which the signal is received. A practical way to determine the AOA is to consider that this direction is the one of the maximum signal strength during a complete rotation of the BS. Combining several AOA results in the desired position estimation.
- The RSSI of great interest in this thesis. It corresponds to the power of received signal strength minus the emitting signal strength. It yields to a ranging system by means of the Log-distance path loss model described in Section 1.3.2 or by fingerprint-based methods (when the different locations of emission are known to exhibit very different power "signatures"). These methods are studied in detail in Chapter 4.

1.2.2 Network-Based Geolocation

This thesis focuses on network-based geolocation. These methods only use the network infrastructure. Among all the described geolocation methods, the latter are the cheapest and require the least of energy. Methods of this kind have met a tremendous success with the apparition of the Internet of Things (IoT) in the late 1990's. Essentially, the concept of IoT is to provide to any objects the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction (Rouse, 2020). Nowadays, the set of applications for IoT devices is dramatic: smart homes (Samuel, 2016), medical and healthcare applications (Catarinucci et al., 2015), agriculture (Mekala and Viswanathan, 2017) or even transportation systems (Zhou et al., 2012).

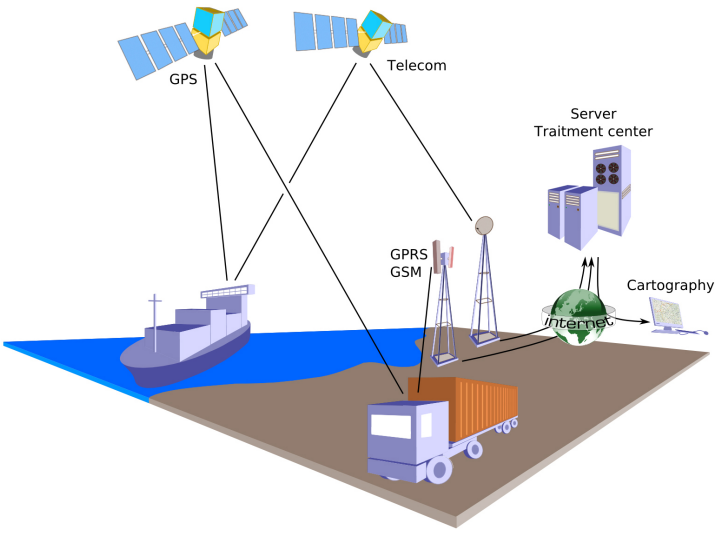


Figure 1.1 – Principles of geolocation using GPS. This image comes from the Geolocation Wikipedia page.

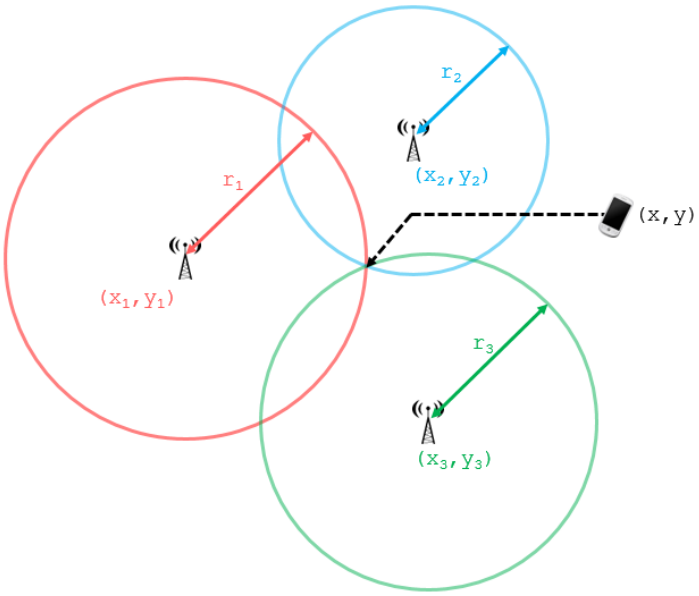


Figure 1.2 – Trilateration principle. Each circle represents all the possible locations of a mobile phone at a given distance (radius) of a cell tower. The aim of a trilateration algorithm is to calculate the (x, y) coordinates of the intersection point of the three circles.

We next move to two concrete geolocation examples encountered in practice.

Logistic Tracking cargo and assets via the network can also result to great benefits in transportation systems. One can send specific alerts when remarkable events occur such as the arrival in a warehouse.

Ecology It has then been argued that the IoT will revolutionize the ecology area. First, in terms of energy management: the connectivity of significant numbers of energy-consuming devices (e.g. lamps, motors, pumps, etc.) can allow them to communicate with utilities not only to balance power generation but also optimize the energy consumption as a whole. Second, environmental monitoring applications of the IoT use sensors to help environmental protection by monitoring as instance air or water quality. Other applications like earthquake or tsunami prediction systems can also be used to provide more effective aid.

A natural idea is to apply the radiolocation methods described above to the Sigfox IoT dedicated network. As stated before, the performance of such methods strongly relies on the network infrastructure. Without going into details here, we propose to give some elements that compromise their use and therefore motivate the subsequent works. First of all, Sigfox BS have a lack of directivity and thus cannot discriminate the incident wave directions. Therefore, the AOA-based methods are irrelevant here. In addition, Sigfox bases its communications on Ultra-Narrow-Band (UNB) Technology which allows to achieve both long range and extended battery life. Thus, each signal has a frequency band of 100 Hz width within the unlicensed frequency band (ranging from 868.0 to 868.6 MHz). This band is popular because it presents a good balance of range, building penetration and the ability to use small antennas. Nevertheless, it is well known that the performance of time-based methods using TOA or TDoA estimation strongly relies on the signal bandwidth. The Cramér-Rao Lower Bound (CRLB) on TOA estimators is thus often exhibited to quantify this effect. It states that the minimal variance of any unbiased estimator of the TOA is inversely proportional to $B^3 \times \text{SNR}$, where B stands for the bandwidth and Signal to Noise Ration (SNR) is the well-known Signal-to-Noise ratio. This bound is then unfavourable to the use of such approaches and yields for instance to a standard deviation of range estimators at least equal to 20 km with Sigfox US infrastructure. Nevertheless, it also constitutes a first benchmark value to which our proposed geolocation methods are compared. It is shown (Boucher and Hassab, 1981) that another disfavourable element is that they are particularly memory demanding since their performance is directly related to the choice of the sampling interval.

1.2.3 RSSI-based Geolocation

The RSSI is a relevant feature for the geolocation task pursued. It is sufficiently explanatory without constituting a memory burden. The related literature is extremely vast and is often classified into two categories of methods: range-based and range-free. They essentially differ in the information used for the localization task. The first category of methods use measurements to predict the range between emitters and receivers (the coordinates of the BS are known in this case). Combining several estimated ranges allows the emitter position estimation by means of some mathematical methods such as trilateration (Thomas and Ros, 2005). There are thus always two phases: a ranging phase and a localization one. In contrast, the second family of methods do not base their predictions upon the range estimation. A simple example of such a method would

estimates the position as the barycenter of the receiving BS. As far as the RSSI is the only measurements at our disposal, the range-based methods quickly show their limits (Chandrasekaran et al., 2009). Indeed, they essentially found upon the log-distance path loss model of which the lack of realism is brought out by the investigation of Section 1.3.2.

In this thesis, we propose another classification of RSSI-based geolocation methods: likelihood-based and fingerprint-based methods. The first ones, that encapsulate the range-based methods already discussed, consist of learning (on a dataset) a model for the RSSI at a BS (denoted by X in the sequel) given the position (denoted by Y in the sequel). This learning phase is also referred to as the "calibration phase" in the literature. Once this model has been inferred, one can predict the emitter position as the one that best agrees the measured RSSI. The second ones are the fingerprint-based methods. The latter directly map the RSSI to the position by means of a function that has been previously learned on a dataset.

The recent advances in machine learning and its successes in a wide range of areas have driven the IoT community to apply these methods to RSSI-based geolocation. Let us now present the contributions we have developed to address this issue.

1.3 State-of-the-Art Gelocation Techniques

A formal introduction to the wireless channel and to the physical parameters involved in its modeling is necessary to understand the location estimation techniques developed in this chapter. Following Tse and Viswanath (2005), we start with the physical modeling of the wireless channel in terms of electro-magnetic waves and then define some important physical parameters. This will allow a better understanding of the foundations of the two geolocation methods discussed in this chapter.

1.3.1 Physical Modeling for Wireless Channel

First, let consider a fixed emitter. If the receiving antenna is sufficiently far away from this emitter (which we refer as the far field) the electric field and magnetic field at any given location are perpendicular both to each other and to the direction of propagation. In response to a transmitted sinusoid $\cos 2\pi f t$, the electric far field at time t can be written as follows:

$$E(f, t, (r, \vartheta, \psi)) = \frac{\alpha_s(\vartheta, \psi, f) \cos 2\pi f \left(t - \frac{r}{c}\right)}{r}, \tag{1.1}$$

where, as shown in Figure 1.3, (r, ϑ, ψ) represents the polar coordinates of the point M at which the electric field is measured; r is then the distance from the emitter to M, (ϑ, ψ) are respectively the vertical and horizontal angles from the emitter to M. The constant c is the speed of light, and $\alpha_s(\vartheta, \psi, f)$ is the radiation pattern of the sending antenna at frequency f in the direction (ϑ, ψ).

We observe that, as the distance r increases, the electric field decreases as r^{-1} and thus the power per square meter in the free space wave decreases as r^{-2} as illustrated in Figure 1.4. This attenuation is referred in the literature as the *large-scale fading*.

Nevertheless, we will see that this r^{-2} reduction of power with distance is not valid as soon as there are obstructions or reflections in the propagation space.

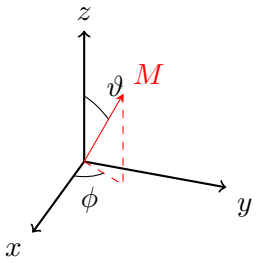


Figure 1.3 – Spherical coordinates (r, ϕ, φ): radial distance r, azimuthal angle ϕ, and polar angle φ.

Thereupon, two methods to estimate the unknown distance from the emitter emerge:

- First, a straightforward method that consists in mapping the received power to the desired distance. In free space, the signal power decays proportionally to r^{-2}. In real-world channels, multi-paths (due to reflectors) and shadowing are two major sources of environment-dependence in the measured power. These phenomena are generally modeled as random. As a consequence, the power decay is rather chosen proportionally to $r^{-\alpha}$, where α is the so-called path-loss exponent. Once this path-loss exponent has been estimated (through the *calibration* phase), it is possible to provide a range estimate from the measured power signal strength. The procedure is described in the sequel.

- On the other hand, an intuitive method is to estimate the delay between the emission and the reception of the signal. To provide this Time Of Arrival (TOA) estimate, we suppose that the signal emanating from the source is known, and that the received signal is a superposition of an attenuated and delayed replica of the known signal plus noise. The remaining task is thus to estimate the arrival times of the replica and its attenuation amplitude. This is the well-known Time-of-Arrival estimation problem. There is a vast literature dealing with this problem. The most well-known estimators are the matched filter approach described in Ehrenberg et al. (1978), or the maximum likelihood (ML) estimator introduced in the work of Knapp and Carter (1976). This latter estimator is obtained as the peak of the cross-correlation function of both the emitted and the received signal. The Cramér-Rao Lower Bound (CRLB) (see Chapter 2) is taken as the standard of reference, because it should give a realistic indication of the attainable mean square error of the estimators. This problem is close to the problem of Time-Delay of Arrival (TDoA) or simply Time-Delay estimation. This latter aims to predict the time delay between the receptions of the signal at two different antennas. It allows to circumvent the imperfect knowledge of the emission time due to the lack of both precision and synchronisation of the emitting devices inner clocks. As far as location is concerned, the first method requires solving circles intersection equation while the second an hyperbolas intersection equation. We thus chose to present this problem here-below rather than the ToA problem.

In the sequel, we will discuss this two ranging-methods, as well as their theoretic reachable precision through their CRLB. For this purpose, we first introduce the singular context of IoT dedicated networks, particularly the Ultra Narrow Bound technology and its impact on the lower bound of the range estimate.

1.3.2 Location Estimation Techniques

The Time-Delay problem

The parameters of primary interest are the relative delays between receiving antennas which determine range (as well as the bearing) to the source. We assume hereafter that the transmission medium is homogeneous so that the signal wave-front is perfectly coherent over the receivers: the signal components received by various sensors are then delayed replicas of each other. The noise is considered to be additive, white, Gaussian, and incoherent from sensor to sensor.

Following the seminal work of Carter (1987), we can derive a location estimation technique based on the time delay estimation. Let first consider the case of two receiving antennas at which the received signal can be modeled as follows:

$$x_1(t) = s(t) + n_1(t), \tag{1.2}$$
$$x_2(t) = As(t - D) + n_2(t), \tag{1.3}$$

where it is assumed that $s(t), n_1(t)$, and $n_2(t)$ are real and jointly stationary random processes, $s(t)$ is furthermore independent from $n_1(t)$ and $n_2(t)$.

The maximum likelihood estimator of the time delay maximizes the following cross-correlation function (CCF).

$$R_{x_1,x_2}(\tau) \triangleq \int x_1(t)x_2(t-\tau)\mathrm{d}t, \tag{1.4}$$

We obtain the following Cross-Correlation (CC) estimator:

$$\hat{D}_{CC} \triangleq \arg\max_{\tau} R_{x_1,x_2}(\tau), \tag{1.5}$$

Carter and Knapp (1976) proposes a improved version of the CC estimator by introducing the so-called generalized cross-correlation (GCC) function. This function is obtained by applying pre-filters to amplify spectral components of the signal that have little noise and attenuate components with large noise. As such, the GCC requires knowledge (or estimates) of the signal and noise power spectra. There is a number of algorithms in the GCC family depending on the choice of the filters. Commonly used weighting functions include the constant weighting (in this case, the GCC becomes a frequency domain implementation of the cross-correlation method), the smoothed coherence transform (SCOT) (Carter et al., 1973), the Hassab-Boucher transform (Hassab and Boucher, 1981) or the maximum-likelihood (ML) (Knapp and Carter, 1976). It is well known that the ML estimator obtained in the ideal propagation situation is optimal from a statistical point of view since the estimation variance can achieve the CRLB which has the rather simple form :

$$\sigma^2_{\mathrm{CRLB}} = \frac{3}{8\pi^2}\frac{1+2\mathrm{SNR}}{\mathrm{SNR}^2}\cdot\frac{1}{B^3T}, \tag{1.6}$$

where B is the signal bandwidth, T is the observation time, and SNR is the Signal-to-Noise Ratio defined as the ratio of the power of a signal (meaningful input) to the power of background noise (meaningless or unwanted input). In practice, evaluating this expression on the values observed in the Sigfox network in the U.S and Europe (see

also Sallouha et al. (2017)) leads to

$$\sigma_{\text{CRLB, Sigfox US}} \simeq 8.46 \cdot 10^{-5}\text{s}, \tag{1.7}$$

$$\sigma_{\text{CRLB, Sigfox Eu}} \simeq 5.07 \cdot 10^{-4}\text{s} \tag{1.8}$$

Multiplying by speed of light gives the following ranging error variances:

$$\sigma_{\text{Sigfox US}} \simeq 25\text{km}, \tag{1.9}$$

$$\sigma_{\text{Sigfox Eu}} \simeq 150\text{km}. \tag{1.10}$$

However, the simulation results of Hassab and Boucher (1979); Scarbrough et al. (1981) show that ML estimator actual performance can be much worse for a given SNR and observation time. More specifically, when the bound is plotted as a function of signal-to-noise ratio one observes a distinct threshold. The work of Chow and Schultheiss (1981) states that below this threshold, the true bound can exceed the CRLB by large factors of the order of $\left(f_s/B\right)^2$.

In other words, although the CRLB is a promising bound for the variance of our time delay estimator, it is not reachable in practice because it is obtained by too idealistic assumptions: the observation sample space has to be large enough; the environment should be free of reverberation; and the spectra of noise signals have to be known a priori. In practice, other problems may also arise. All the techniques described above measure the analog processing of the time delay parameter, while discrete signal processing methods are more commonly used. In the sequel, we briefly discuss the case when we only have at our disposal sampled signals. The problem remains the same: finding an estimator $\hat{D}$ of the true delay D using a finite set of samples of $x_1(t)$ and $x_2(t)$ using as in the continuous case, the peak of the CCF. A normal practice is to replace the CCF defined in Equation (1.5) by its time-averaged estimate:

$$\hat{R}_{x_1,x_2}(\tau) = \frac{1}{N}\sum_{k=1}^{N} x_1(kh)x_2(kh-\tau)\mathrm{d}t, \tag{1.11}$$

where h is the sampling interval, and $(N-1)h$ is the estimation window width. The so-called Direct Correlator (DC) estimator of the time delay is defined as the maximizer of the latter quantity w.r.t. τ. It is thus necessary to interpolate the CCF in the neighborhood of the peak. A commonly used approach (Boucher and Hassab, 1981) is to fit a parabola in the neighborhood of this maximum using 3 samples of $\hat{R}_{x_1,x_2}(\cdot)$. Thus, near its peak the CCF can be approximated by the following convex parabola:

$$\hat{R}_{x_1,x_2}(\tau) \simeq a\tau^2 + b\tau + c,$$

where (a, b, c) are the fitted parameters. The apex of the parabola is then used as a proxy of the time delay:

$$\hat{D}_{\text{CC}} = -\frac{b}{2a}. \tag{1.12}$$

It is shown that this methods yields a biased estimator of the time delay D (Jacovitti and Scarano, 1993). In general, this bias can be decomposed into a systematic bias, due to parabolic estimation; and another one resulting from the noise and a finite width of signals observation.

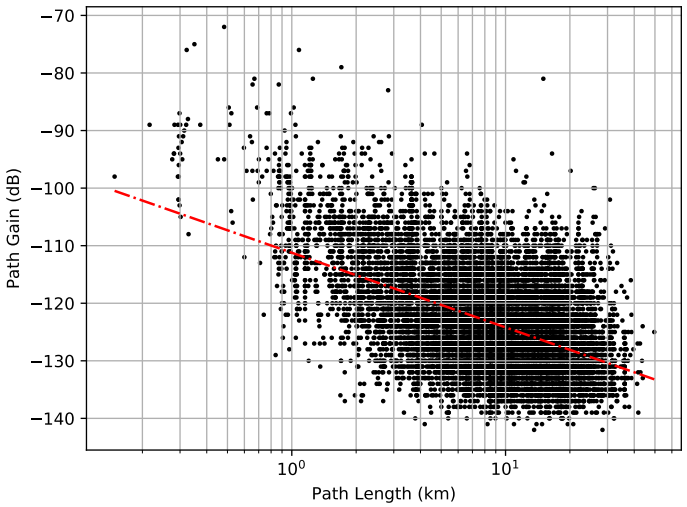


Figure 1.4 – Measured wideband path gain as a function of path length. Linear fit (–) with r^{-2} reduction.

The results of these works are enlightening in many ways. First because they state that in the range of SNR that are observed in the Sigfox network, the second bias is negligible with respect to the first one. The parabola misfit is therefor the keystone of these discrete methods. Moreover, the found expression for the bias is the product of two terms. A first term which depends linearly to the the sampling interval h. Let us recall that, for Sigfox, it is chosen as $h = (1/2B) = 0.005$ s. A second term is a function which depends only on the reference signal. Simulations of theoretical bias provided in these two works are shown in the Jacovitti and Scarano (1993) where $\mathbb{E}[\hat{D}_{CC}]$ is computed for various D. As expected, the bias is the higher when the true TD lies between two samples points. The simulation results, obtained using the quantities observed in the Sigfox network invalidate these approaches because results to low time resolution.

The Received-Signal Strength and Log-distance Path Loss (PL) model

In telecommunications, Received Signal Strength Indicator (RSSI) is a measurement of the power present in a received radio signal. Wireless sensors communicate with neighboring sensors, and RSSIs can be measured by each receiver without any further bandwidth or energy requirements. Because RSSIs are relatively inexpensive and simple to implement in hardware, they are an important and popular topic of localization research. Typically, the ensemble mean received power in a real-world, obstructed channel decays proportional to $r^{-\alpha}$, where α is the "path-loss exponent", typically between 2 and 4 as acknowledged in Rappaport et al. (1996). Log distance Path Loss (PL) model is an extension to the Friis free space model introduced in Section 1.3.1. It is used to predict the propagation loss for a wide range of environments and to therefore overcome the limitation of the Friis free space model. The model encompasses random shadowing effects due to signal blockage by trees, buildings *etc.* This model is a parametric model of the RSSI given the distance d. The vector of parameters $\boldsymbol{\theta} = (P_0, \alpha, d_0)$ is such that d_0 is some reference distance and the parameter P_0 represents the power in dBm at distance d_0.

Using loose notation, given the distance d between the emitter and the receiving antenna, the RSSI X at this receiver is distributed as follows (Mao et al., 2007):

$$X[\text{dBm}] \sim \mathcal{N}\left(P_0 - 10\alpha \log \frac{d}{d_0}, \sigma^2_{\text{dB}}\right). \tag{1.13}$$

Estimating the range from the RSSI The estimated distance between the receiver and the emitter, can be thus estimated from the realization of the variable X. We recall that the log-likelihood of X given d is,

$$\log p(X|\boldsymbol{\theta}, d) = c - \frac{\left(X - P_0 + 10\alpha \log \frac{d}{d_0}\right)^2}{2\sigma^2_{\text{dB}}}, \tag{1.14}$$

where c is independent of $\boldsymbol{\theta}$. The distance which maximizes the likelihood is

$$\hat{d}_{\text{ML}}(X) = d_0 10^{\frac{P_0 - X}{10\alpha}}. \tag{1.15}$$

We can prove that the bias of the ML estimator is as follows:

$$\mathbb{E}\left[d_0 10^{\frac{P_0 - X}{10\alpha}}\right] = \underbrace{\exp\left(\frac{10\alpha}{\sqrt{2}\sigma_{\text{dB}} \log 10}\right)^2}_{C} d.$$

In typical channels studied e.g. in Rappaport et al. (1996), $C \simeq 1.2$, and the range is then over estimated by a factor 20%. It is therefore preferable to use the bias-corrected estimator as follows:

$$\hat{d}_{\text{CML}}(X) = d_0 \exp - \left(\frac{10\alpha}{\sqrt{2}\sigma_{\text{dB}} \log 10}\right)^2 10^{\frac{P_0 - X}{10\alpha}}. \tag{1.16}$$

Remark 1.1. *Note that the variance of the bias-corrected estimator is proportional to the actual range. This result is however not surprising, and above all, informs on the need to have a very dense network in order to have a quality estimation of the emitter.*

1.4 Introduction to Sigfox Radio System

Two of the most commonly used techniques for geolocation have been discussed above. This section provides a better understanding of the singularity of the geolocation problem in a sensor network such as the Sigfox network. The performance of the two described geolocation approaches in Sigfox network are studied.

1.4.1 Internet of Things, a new usage for the radio-communications

There is no consensus on the definition of the Internet of Things (IoT), although its initial use has been attributed to Kevin Ashton. In fact, many groups including academicians, researchers, developers and corporate people have their own. Yet, all the

definitions have in common the idea that the first version of the Internet was about data created by people, while the next version is about data created by things. We can therefore cite Madakam et al. (2015) and give the following definition: "*An open and comprehensive network of objects that have the capacity to auto-organize, share information, data and resources*". The connection between objects is not to be understood in terms of servers, computers or smartphones. In the context of IoT, sensors embedded in physical objects (from roadways to pacemakers) are linked through wireless networks. This implies that the implementation of IoT communication must be cost-effective and energy frugal since the object may not have been natively designed with an energy source. As a consequence, IoT comes with several constraints for the radio-communication system:

- The communication function of a connected device must have a marginal cost (when it is compared to the cost of the device itself). For ease of reference, the target price of a device for mass production should be less that 0.2 $.
- The volume of transmitted data by the connected devices must be low. Indeed, devices transmit information originated from sensors, alarms or GPS. trackers which communicate through small "packets". In Sigfox's network, the size of the messages must not exceed 12 bytes (see Section 1.4.2 for details).
- The volume of connected objects might be huge. Due to the low-cost and simplicity of use, it is possible to connect a large number of objects. This may result to a very high density of connections by square kilometer. Typically, in Sigfox network, the density of connected objects reaches $50,000$ per square kilometers (that is much higher than in cellular networks).

We thus note that in order to make use of IoT, many requirements come into play. The radio technology that can support all these requirements is encapsulated into the name of Low-Power Wide-Area Network (LPWAN[1]), where we refer to Raza et al. (2017) for a in-depth study.

1.4.2 Ultra-narrowband benefits for LPWAN

While IoT is an expression to describe the usages, LPWAN rather encapsulates the underlying radio technology. The aim of this technology is to meet two contradictory needs: the low-energy consumption and the wide-area network. Therefore, LPWANs have the following characteristics:

- To reduce the infrastructure cost, the base stations of the network must be limited in number. The network counts a very large number of connected devices compared to the number of BS. Thus, one BS can receive numerous messages originated from multiple objects (sometimes at the same time).
- To reduce the energy cost of communication, the connected objects must transmit at very low power. They can therefore be endowed with small batteries (or even no battery at all thanks to energy harvesting). High sensitivity of the BS is thus a need to achieve a wide-area coverage with one-hop communication.

[1] Machina Research, 2013

We now discuss briefly about Ultra-Narrow Band (UNB) and how this technology can meet the requirements described above.

UNB Technnology

Ultra Narrow-Band systems are those in which the channel has a very narrow bandwidth. This band is significantly smaller than the total available frequency resource, typically of few hundred Hertz. Essentially, Sigfox's technology is such that the receiver listens for a large spectrum range with a high dynamic range and the software is continuously looking for the signal of interest and tracks its center frequency, as a human radio operator would do (Chaudhari et al., 2020). In the Sigfox's network, the transmitted signal of UNB occupies a band of 100 Hz (in Europe, and 600 Hz in USA), inside a typical possible band of 192 kHz to 2 MHz. In Europe, the frequency band ranges from 868.0 MHz to 868.6 MHz and thus brings up to 6000 sub-channel of 100 Hz width each. Objects now can randomly select carrier frequency, without excessive collision rate. We assume that all objects have the same behavior and transmit a message of duration T seconds at a frequency f_s. For the time-slotted case, any active node randomly selects a time-slot, since the temporal resource is separated by slots. Thus in the time domain, the transmissions either do not collide at all (when they choose different time-slots), or collide for the whole duration (when they choose the same time-slot). All transmissions are performed within a dedicated band, which has band-width B Hz. Each transmission occupies a bandwidth b Hz which represents a small portion of the total channel bandwidth B Hz. From the base-station point of view, the total dedicated band contains, from time to time, transmitted ultra narrow signals at random carrier frequencies. For each detected transmission, the BS extracts the signal at the estimated frequency of interest, and decodes the packet. Such a detection and estimation can be done as described in a Sigfox patent (Artigue, 2017).

1.4.3 Campaign of validation of the PL model in Sigfox Network

We now conclude this section with a numerical investigation. We give numerous practical examples that invalidate the use of the log-distance path loss model in rural environment context of our concern. First, the log-normal distribution of the RSSI measurements is verified by examining the residuals $X-(P_0+10\alpha \log \frac{d}{d_0})$ using quantile-quantile plots shown in Figure 1.5. This figure shows that the observed data have a heavier tail than the one of the normal distribution: the quantiles move away from the diagonal outside a certain range. In that respect, the next numerical experiment investigate the lack of realism of the studied model. More specifically, the non-isotropy of the propagation model is highlighted. Indeed, these experiments show the high dependency of the RSSI upon the direction of the arrival signal. This is shown in Figure 1.6. This model of propagation, assuming that there is a free propagation space, and thus omitting the presence of obstacles in certain directions, is vain. A first extension of the model, is to encapsulate this dependence of the loss by making the path loss exponent $\alpha(\vartheta, \phi)$ direction-dependent. This approach is however interesting because it amounts to learning a path-loss exponent for every direction (ϑ, ϕ). Instead, a very general semi-parametric model is introduced in Section 3.3.1. This model provides an estimation of the expected RSSI at a BS given the emission position (x, y). This enables to free ourselves from the knowledge of the positions of the BSs, which would have been needed if the dependence upon the direction between the emitter and the BS was modeled.

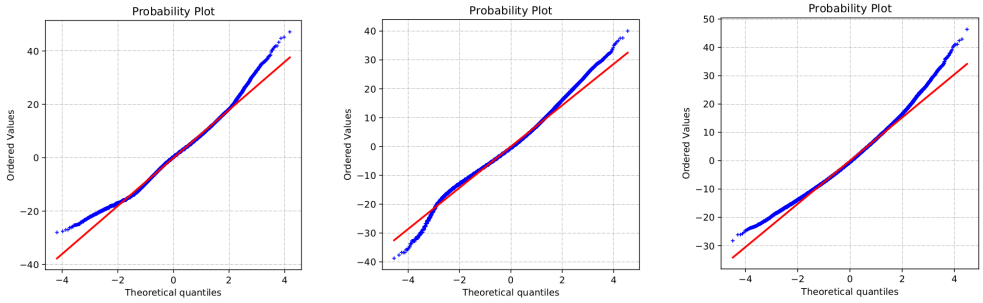


Figure 1.5 – Q-Q plot (w.r.t. normal distribution) of the residuals $X - (P_0 + 10\alpha \log \frac{d}{d_0})$ shown for 3 BS in Paris.

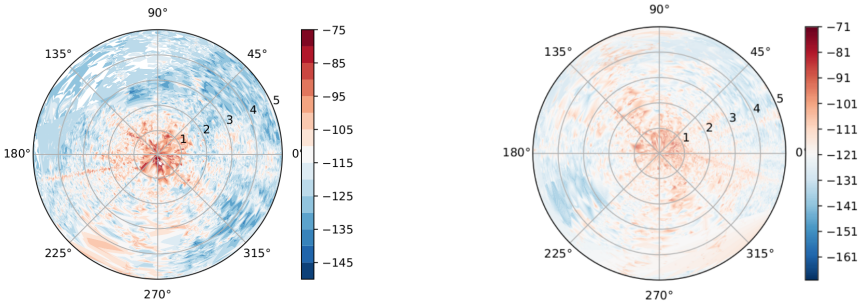


Figure 1.6 – For each plots, we show the empirical mean of the RSSI at any point (r, ϑ), measured by the BSs placed at the center $(0, 0)$. We notice that some directions (e.g. around 190° for the left figure) show high values of RSSI and no loss of power in this directions. These behaviours correspond to major traffic routes.

This study has put into light how the context of IoT network jeopardizes the precision of the two presented methods of geolocation. It also identify more clearly the interests of using recent methods of machine learning to the problem of network-based geolocation. We therefore dedicate the future chapters to an intensive investigation of these approaches.

1.5 Geolocation as a Prediction Learning Task

The network is composed of d fixed BS, say $(\mathrm{BS}_1, \ldots, \mathrm{BS}_d)$, whose respective coordinates $(y_1, \ldots, y_d)$ in the complex plane are fixed but not necessarily known (unless it is specified).

We consider a connected device whose position Y is a random variable in some given subset $\mathcal{Y}$, typically an open subset of $\mathbb{R}^2$. The device sends packets/messages which are collected by the neighboring BS. For a given message, each BS k $(k = 1, \ldots, d)$ computes a RSSI X_k as the temporal mean of the received signal strength. The RSSI X_k is typically real-valued in a certain subset $\mathcal{X} \subset \mathbb{R}$. However, some messages may not be detected by some BS, in which case we just set $X_k = \texttt{NaN}$, where `NaN` stands for an unobserved value. Thus, for every $k = 1, \ldots, d$, X_k is a random variable in the set $\tilde{\mathcal{X}} \triangleq \mathcal{X} \bigcup \{\texttt{NaN}\}$.

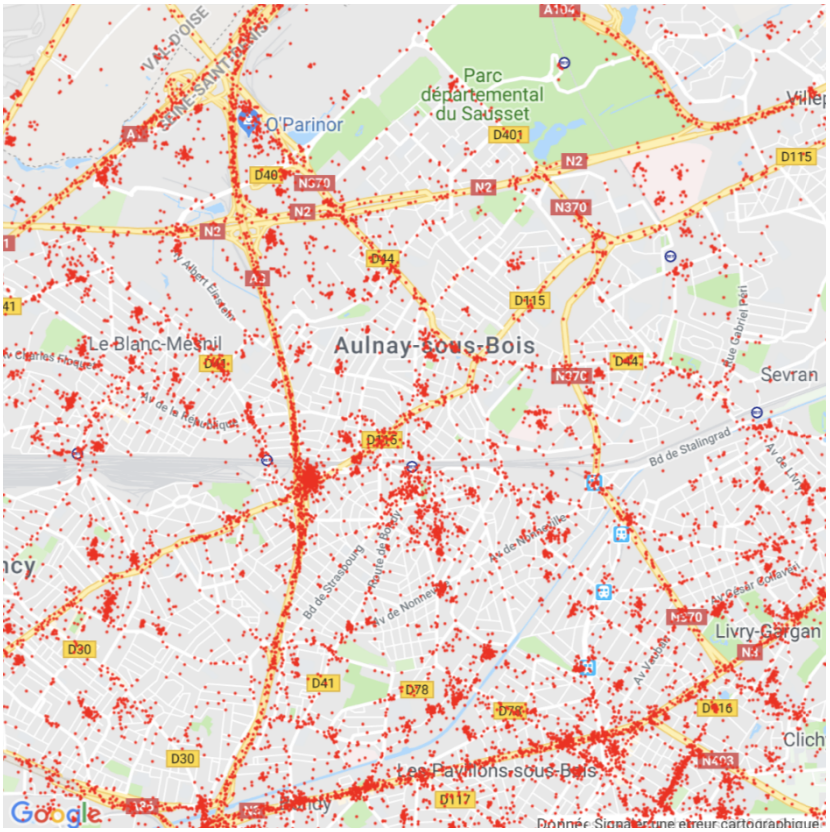


Figure 1.7 – Scatter plots of a sub-sample of the Sigfox training dataset. A red dot corresponds to an emitting position of a device in the Sigfox dataset.

The aim of location estimate is then to predict the unknown position Y from the observation of the RSSI-vector

$$\boldsymbol{X} \triangleq (X_1, \ldots, X_d)\,.$$

It is the general prediction learning task: we wish to infer the relationship between the random variable $\boldsymbol{X}$ and a target random variable Y, taking values in sets $\tilde{\mathcal{X}}^d$ and $\mathcal{Y}$ respectively. For a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ that defines a discrepancy on $\mathcal{Y}$ (in typical settings, $\ell(y, \hat{y}) = \|y - \hat{y}\|^2$), the generic writing of this task is then to find a function $h^\star : \tilde{\mathcal{X}}^d \to \mathcal{Y}$ that verifies:

$$h^\star = \arg\min_{h \in \mathcal{H}} \mathbb{E}\left[\ell(h(\boldsymbol{X}), Y)\right], \tag{1.17}$$

where $\mathbb{E}$ is the expectation w.r.t. the distribution of the vector $(\boldsymbol{X}, Y)$, and $\mathcal{H}$ is a subset of $\mathcal{Y}^{\tilde{\mathcal{X}}^d}$, reflecting a knowledge about the dependency between $\boldsymbol{X}$ and Y (Hastie et al., 2009). However, the latter requires the knowledge of the distribution of $(\boldsymbol{X}, Y)$, that we cannot access in practice. A more practical setting consists in replacing this true risk by an empirical average computed on a dataset $\mathcal{Z}_n \triangleq \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ of n independent and identically distributed (i.i.d) samples drawn from the same distribution as $(\boldsymbol{X}, Y)$. This dataset is built by gathering observed RSSI's of devices equipped with GPS. As represented in Table 1.1, every row of the dataset corresponds to a message. The features are the RSSI's at the receiving BS's and the label is the GPS coordinates of the transmitting device at the instant when the packet is sent.

BS 1	BS 2	...	BS d	Lat	Long
−102	NaN	...	−83	49.15434	2.24928
NaN	−98	...	NaN	48.865584	2.44567

Table 1.1 – Sample from the Sigfox dataset.

The problem defined in Equation (1.17) then becomes the well-known Empirical Risk Minimization (ERM):

$$h^\star = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} \ell(h(\boldsymbol{x}_i), y_i). \tag{1.18}$$

Equipped with this mathematical background, we now move to our application focus: the problem of network-based geolocation.

1.6 Contributions

The organization of this manuscript is as follows. Each chapter can be read independently.

- Chapter 2 introduces the theoretical elements needed to the reading of this thesis. First, elements of statistical inference are introduced. Then, the notion of risk, and a lower bound on the so-called quadratic risk on an estimator is discussed. These notions will prove themselves useful to highlight that the parametric family of estimators are not well suited when applied to RSSI-based geolocation, and that we need to go beyond simple parametric estimation. This last point is therefore studied in this chapter as well.

- Chapter 3 investigates machine learning approaches addressing the problem of geolocation. First, we review some classical learning methods to build a radio map. These methods are split in two categories, which we refer to as likelihood-based methods and fingerprinting methods. Then, we provide a novel geolocation approach in each of these two categories. The first proposed technique relies on a semi-parametric Nadaraya-Watson (NW) estimator of the likelihood, followed by a maximum a posteriori (MAP) estimator of the object's position. The second technique consists in learning a proper metric on the dataset, constructed by means of a Gradient boosting regressor: a k-nearest neighbor algorithm is then used to estimate the position. The proposed methods are compared on two data sets originated from Sigfox network, and an indoor dataset performed in a three-story building. Experiments show the interest of the proposed methods, both in terms of location estimation performance, and ability to build radio maps.

- Chapter 4 is dedicated to the similarity learning. Indeed, the choice of the similarity (or distance) is known to be critical for the performance of neighbors-based predictor such as k-NN regressor. That is why there is an increasing interest for optimizing distance and similarity functions. However, in most prior works, no link is made between the learned metrics and the estimator performance. In this chapter, we propose to build the metric by directly minimizing the regression error of our estimator, and thus obtain an ad-hoc learning objective. To minimize

this objective, we propose a modified version of the eXtreme Gradient Boosting algorithm (XGBoost). The soundness of our approach is finally endorsed by conclusive numerical experiments on numerous datasets.

- Chapter 5 introduces the *weighted partial copula* function for testing conditional independence (CI). The CI has been broadly used in the RSSI-based geolocation literature in order to decrease the statistical models complexity. Testing this assumption can leads to a better understanding of the likelihood based models proposed herein. The proposed test procedure results from the following ingredients: (i) the test statistic is an explicit Cramer-von Mises transformation of the *weighted partial copula*, (ii) the regions of rejection are computed using a bootstrap procedure which mimics conditional independence by generating samples from the product measure of the estimated conditional marginals. Under conditional independence, the weak convergence of the *weighted partial copula process* is established when the marginals are estimated using a smoothed local linear estimator.

1.7 Publications

The works presented in this manuscripts have resulted in several accepted publications and preprints, that are listed here in chronological order:

- Elgui, K., Bianchi, P., Portier, F. & Isson, O. (2019, September). Learning Methods for RSSI-based Geolocation: A Comparative Study. In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.
- Elgui, K., Bianchi, P., Isson, O., Portier, F. & Marty, R. (2020, April). Metric Learning for Fingerprint RSSI-Localization. In 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS) (pp. 1036-1042). IEEE.
- Elgui, K., Bianchi, P., Portier, F. & Isson, O. Learning Methods for RSSI-based geolocation: A comparative study, Pervasive and Mobile Computing (2020).
- Bianchi, P., Elgui, K., & Portier, F. (2020). Conditional independence testing via weighted partial copulas. arXiv preprint arXiv:2006.12839.
- Elgui, K., Bianchi, P., Portier, F. & Isson, O. Similarity Learning with XGBoost. (To be submitted)

2

Preliminary Background

"L'intelligence artificielle se définit-elle comme le contraire de la bêtise naturelle?"

In this chapter, we introduce the theoretical elements needed to the reading of this thesis. First, we properly define elements of statistical inference that are afterwards used for RSSI-based methods of geolocation. Then, the notion of risk, and a lower bound on the so-called quadratic risk on an estimator is discussed. These notions will prove themselves useful to highlight the irrelevance of the parametric family of estimators (when applied to RSSI-based geolocation), and the need to go beyond simple parametric estimation. This last point is therefore studied in this chapter as well.

2.1 Elements of Statistical Inference

2.1.1 Introduction

The problem of estimating statistical models depending on a finite number of parameters goes back to Fisher (Fisher, 1925) and has met with tremendous success ever since (James et al., 2013).

Parametric approaches have the advantage of being very easy to compute, hence their predominance in the related literature. Furthermore, powerful tools are available for the statistical analysis of such estimators, such as a lower bound of their variance. This kind of knowledge, necessary to assess the quality of an estimator, is unfortunately difficult to obtain when going beyond parametric model, and often requires complex simulation methods. However, parametric models are limited in their expressive power, as they only provide an approximation, often imprecise, of the underlying statistical structure. Statistical models that explain the data in a more consistent way are often more complex; unknown elements in these models are usually, instead of scalar parameters, functions having certain smoothness properties. In this thesis, we shall indeed see that parametric estimation does not provide sufficiently satisfactory results when applied to RSSI-based location estimation. Hence, semi-parametric or non-parametric methods will often be preferred. This section is structured as follows.

Outline First, we set in Section 2.1.2 the framework of parametric estimation. This framework allows to introduce the Cramér-Rao Lower Bound of an estimator. This naturally leads to the introduction in Section 2.2 of the non-parametric estimation framework through the two main problems. Practical aspects of the regression problem are then investigated, and Tree-Based methods are thus introduced in Section 2.2 as a conclusion of this chapter.

2.1.2 Parametric Estimation

Statistical Model

We first consider a family $\mathcal{P}$ of probabilities over the space of observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We refer to the two first chapters of Shorack and Shorack (2000) for a thorough definition of these mathematical objects. The family $\mathcal{P}$ is called the statistical model.

Then, we can define a parametrization of the model by attaching each $P \in \mathcal{P}$ with a parameter $\theta \in \Theta$, where Θ is any open subset of $\mathbb{R}^d$. One will then write P_θ to designate the law thus labeled.

Definition 2.1 (Statistical model)**.** *We call a statistical model a family $\mathcal{P}$ of probabilities P over the space of observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. If there exists a set Θ set such that*

$$\mathcal{P} = \left\{P_\theta, \theta \in \Theta\right\},$$

then, Θ is called the parameter space, and $\mathcal{P}$ a parametric model.

A model $\left\{P_\theta, \theta \in \Theta\right\}$ is said to be *dominated* if any $P_\theta \in \mathcal{P}$ is absolutely continuous with respect to a common reference measure μ. In the sequel, we place ourselves in the case where the model is dominated. As a consequence, for all $\theta \in \Theta$, P_θ always admits a probability density function p_θ w.r.t the reference measure μ.

This fact allows to directly work on a family of densities rather than on a family of distributions. This leads to the introduction of the notion of *likelihood*, that we define as follows:

Definition 2.2 (Likelihood)**.** *The likelihood of the observation x is the mapping $\theta \mapsto p_\theta(x)$. This likelihood is also denoted $p(x; \theta)$.*

The likelihood is a central element of a family of inference methods (*e.g.* maximum likelihood, or likelihood methods), for which the intuition is clear. Considering an observation x, the likelihood "assesses" how likely it is that this observation has been generated under P_θ. The definition of the Maximum Likelihood estimator follows naturally:

Definition 2.3 (Maximum Likelihood)**.** *We consider a dominated statistical model $\mathcal{P} = \left\{P_\theta, \theta \in \Theta\right\}$, $p(\cdot; \theta)$ the probability density of P_θ w.r.t. the reference measure μ. Suppose that we observe x drawn from P_θ. A Maximum-Likelihood (ML) estimator of θ, is any estimator $\hat{\theta}(x)$ that verifies:*

$$p(x; \hat{\theta}) \geq \sup_{t \in \Theta} p(x; t). \tag{2.1}$$

Applications of these estimators are considered in Chapter 3, in order to estimate the time-delay given the observation of both the emitted and received signals, or to estimate the range given the observation of the RSSI.

Risk and Cramér-Rao Lower Bound

We first recall the notion of risk for a parameter estimate. Then, lower bounds on the quadratic risk of unbiased estimators are established. In this thesis we focus on the estimation of parameters of interest that write as $g(\theta)$ for some function $g : \Theta \to \mathbb{R}$.

An example is the Time Delay estimation in Section 1.3.2. Let now $\hat{g}$ be an estimator of the parameter of $g(\theta)$. The risk evaluated at $\hat{g}$ is defined as follows:

$$R(\theta, \hat{g}) \triangleq \mathbb{E}_\theta \left[\ell \left(g(\theta), \hat{g}(X) \right) \right] \quad , \tag{2.2}$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is some loss function. In typical settings, ℓ is the quadratic loss, and the corresponding risk is then called the *quadratic risk*, or Mean Squared Error (MSE). The MSE can be decomposed as follows:

$$\begin{aligned} \mathrm{MSE}(\theta, \hat{g}) &\triangleq \mathbb{E}_\theta \left[(g(\theta) - \hat{g}(X))^2 \right] \\ &= \mathbb{E}_\theta \left\{ \left[(\hat{g}(X) - \mathbb{E}_\theta \hat{g}(X)) + (\mathbb{E}_\theta \hat{g}(X) - g(\theta)) \right]^2 \right\} \\ &= \mathbb{E}_\theta^2 \left[\hat{g}(X) - g(\theta) \right] + \mathbb{E}_\theta \left[(\mathbb{E}_\theta \hat{g}(X) - g(\theta))^2 \right] \\ &\triangleq b^2(\theta, \hat{g}) + \mathrm{Var}_\theta \left[\hat{g}(X) \right] \quad . \end{aligned} \tag{2.3}$$

This formula is known as *bias-variance* decomposition. Its simplicity is only provided by the use of the quadratic loss. The family of estimators that satisfy $b(\theta, \hat{g}) = 0$ for all $\theta \in \Theta$ is called the class of unbiased estimators.

In the sequel, we discuss the quadratic risk of such unbiased estimators via the notion of Fisher Information. We begin with the case where $d = 1$, that is $\theta \in \Theta \subset \mathbb{R}$, to simplify the presentation. We consider an unbiased estimator of $g(\theta)$, i.e., a statistic $\hat{g} : \mathcal{X}^n \to \mathbb{R}$ such that $\mathbb{E}[\hat{g}(X)] = g(\theta)$ for all $\theta \in \Theta$. The Cramér-Rao bound gives a lower bound on the variance of $\hat{g}$, hence on its quadratic risk (since $\hat{g}$ is unbiased). This theorem is defined under certain technical conditions on the statistical model at stake. These *regularity conditions* of the model are provided e.g. in Shao (2006). When these conditions are met, the model is said to be regular. We then define the *Fisher information quantity* (Dembo et al., 1991):

$$I(\theta) \triangleq \left\{ \left(\frac{\partial \log p}{\partial \theta}(X; \theta) \right)^2 \right\}. \tag{2.4}$$

Theorem 2.4. *Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ be a regular model, $g(\theta)$ the parameter of interest and let $\hat{g}(X)$ be a statistic such that $\mathbb{E}_\theta[\hat{g}(X)] = g(\theta)$ and $\mathrm{Var}_\theta[\hat{g}(X)] < \infty$ for all $\theta \in \Theta$. Suppose furthermore that $0 < I(\theta) < \infty$. Then,*

$$\mathrm{Var}_\theta[\hat{g}(X)] \geq \frac{g'(\theta)^2}{I(\theta)}. \tag{2.5}$$

The proof is beyond the scope of this section, but we refer the interested reader to Kay (1993).

Corollary 2.5. *Suppose that the conditions of Theorem 2.4 are satisfied and that $\hat{g}$ is an unbiased and regular estimator of the parameter. Then,*

$$\mathrm{Var}_\theta[\hat{g}(X)] \geq I^{-1}(\theta), \forall \theta \in \Theta. \tag{2.6}$$

This bound is called Cramér-Rao Lower Bound (CRLB) or Darmois-Fréchet bound.

This bound is applied to the case when θ is a scalar. The following proposition is an extension to the multi-dimensional case.

Proposition 2.6. *Let $\hat{g}$ be a statistic such that for every $\boldsymbol{\theta} \in \Theta$, $\mathrm{Var}_{\boldsymbol{\theta}}[\hat{g}(X)] < \infty$, with $0 < I(\boldsymbol{\theta}) < +\infty$. We furthermore denote by $g(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\hat{g}]$. Then, $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is differentiable and:*

$$\mathrm{Var}_{\theta}[\hat{g}(X)] \geq \nabla_{\theta} g(\theta)^{\top} I(\theta)^{-1} \nabla_{\theta} g(\theta). \tag{2.7}$$

2.2 Non-Parametric Estimation

The problem of nonparametric estimation consists in estimating, from some observations, an unknown function belonging to a class of functions. We focus on two nonparametric estimation problems in this thesis: the estimation of a density and that of nonparametric regression Chapters 3 and 5.

An introduction to nonparametric methods is therefore proposed herein.

2.2.1 Kernel Density Estimation

Let $X_1, \ldots, X_n$ be n i.i.d. real-valued random variables whose common distribution is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$. The problem of this section is to estimate the density of the common (unknown) distribution p. The estimate of p is thus a measurable application w.r.t. the observations $(X_1, \ldots, X_n)$, $\hat{p}_n : x \mapsto \hat{p}_n(x; X_1, \ldots, X_n) = \hat{p}_n(x)$. It happens that some prior can be known about this density function, and we can look for p in a parametric family. This case has been discussed in Section 2.1.2. But most of the time, this prior is not available and one may assume that p belongs to an hypothesis set of function $\mathcal{H}$. This set can be the set of all Lipschitz continuous probability densities for instance (Tsybakov, 2009). In this thesis, we focus particularly on kernel density estimation methods. The kernel density estimate of p, also called the Parzen window estimate (Parzen, 1962), is a non-parametric estimate given by:

$$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{2.8}$$

where $K_h : \mathbb{R}^d \to \mathbb{R}_+$ is defined as $K_h(x) = h^{-1} K(h^{-1} x)$ with $h > 0$, and $K : \mathbb{R}^d \to \mathbb{R}_+$ a kernel function, satisfying the condition

$$\int_{\mathbb{R}^d} K(u) \mathrm{d}u = 1. \tag{2.9}$$

Popular kernels include for example:

- the rectangle kernel: $K(u) = \frac{1}{2} \mathbb{1}_{[-1,1]}(|u|)$.
- the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)$.
- the Epanechnikov kernel: $K(u) = \frac{3}{4}(1 - u^2) \mathbb{1}_{[-1,1]}(|u|)$.

Multidimensional estimator can be easily obtained from the latter one-dimensional kernels. Consider for instance the case $d = 2$: we suppose that we have a sample of n i.i.d. pairs of r.v. $(U_1, V_1), \ldots, (U_n, V_n)$, with a common density $p(u, v)$ in $\mathbb{R}^2$. Then,

the kernel estimator of $p(u, v)$ is given as follows:

$$\hat{p}_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(u - U_i) K_{h_2}(v - V_i),$$

where h_1 and h_2 are the bandwidths associated to the "selection" of the neighbourhood of u and v respectively.

2.2.2 Nadaraya-Watson Estimator

We now study the case of non-parametric regression. We introduce the methods of construction of well-known Nadaraya-Watson estimators (Nadaraya, 1964), and to give guarantees on the risk of such estimators.

We consider a pair $(\boldsymbol{X}, Y)$ of real-valued random variables, where Y is integrable: $\mathbb{E}|Y| < \infty$. We call regression function of Y given $\boldsymbol{X}$ the function $m : \mathbb{R}^d \to \mathbb{R}$:

$$m(x) = \mathbb{E}[Y | \boldsymbol{X} = x]. \tag{2.10}$$

We have at our disposal a sample $(\boldsymbol{X}_i, Y_i)_{i=1\dots n}$ of n, independent and identically distributed (i.i.d) pairs of random variable, with the same distribution as $(\boldsymbol{X}, Y)$. The aim is to build from this sample an estimator of the regression function m.

Given a kernel K and a positive bandwidth $h > 0$, for $x \in \mathbb{R}^d$, put $K_h(x) = h^{-1} K(h^{-1} x)$. It follows easily that K_h still verifies Equation (2.9). The Nadaraya-Watson estimator of the regression function m reads as follows:

$$m_n^{\mathrm{NW}}(x) = \frac{\sum_{i=1}^{n} Y_i K_h(\boldsymbol{X}_i - x)}{\sum_{i=1}^{n} K_h(\boldsymbol{X}_i - x)}, \text{ if } \sum_{i=1}^{n} K_h(\boldsymbol{X}_i - x) \neq 0 \ , \tag{2.11}$$

and $m_n^{\mathrm{NW}}(x) = 0$ otherwise.

To conclude this section, we briefly discuss the intuition beyond the expression of this estimator. We assume that the distribution of the random variable $(\boldsymbol{X}, Y)$ admits a density $p(x, y)$ w.r.t. the Lebesgue measure, and that $p(x) = \int_{\mathbb{R}} p(x, y) \mathrm{d}y > 0$ a.s. We can write

$$\mathbb{E}[Y | \boldsymbol{X} = x] = \int_{\mathbb{R}} y p_{Y|X}(y|x) \, \mathrm{d}y = \int_{\mathbb{R}} \frac{y p(x, y)}{\int p(x, y) \mathrm{d}y} \, \mathrm{d}y. \tag{2.12}$$

The following proposition shows that the Nadaraya-Watson estimator of the regression function is simply obtained by replacing in Equation (2.12) the unknown probability density functions $p(x)$ and $p(x, y)$ by their estimated versions $\hat{p}_n(x)$ and $\hat{p}_n(x, y)$ introduced in Section 2.2.1.

Proposition 2.7. *Let $\hat{p}_n(x)$ and $\hat{p}_n(x, y)$ be the kernel estimates of $p(x)$ and $p(x, y)$ respectively as defined in Equation (2.8), then:*

$$m_n^{\mathrm{NW}}(x) = \int_{\mathbb{R}} \frac{y \hat{p}_n(x, y)}{\hat{p}_n(x)} \mathrm{d}y.$$

The proof can be found in Tsybakov (2009).

Remark 2.8. *The Nadaraya-Watson estimator m_n^{NW} satisfies the following property:*

$$m_n^{\mathrm{NW}}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - a)^2 K_h(\boldsymbol{X}_i - x). \tag{2.13}$$

Thus m_n^{NW} is obtained by a local constant least squares approximation of the outputs Y_i.

We are interested in the point-wise bias of such estimate $\mathbb{E}[m^{\mathrm{NW}}(x)] - m(x)$.

Remark 2.9. *We draw to your attention that we chose to conserve the writing in its standard form, that is for the estimation of $\mathbb{E}[Y|X = \cdot]$. However, in Section 3.3.1, Nadaraya-Watson estimators are in fact, employed to estimate $\boldsymbol{X}$ given the position Y. Indeed, the key question is whether it is possible to choose an appropriate kernel on $\mathcal{X} \times \mathcal{X}$. One can either build an ad-hoc kernel on the space of the RSSI (see our proposed solution in Chapter 4), or leverage the fact that the choice of kernel is much easier on $\mathcal{Y} \times \mathcal{Y}$. The latter approach implies the use of Maximum a Posteriori (MAP) estimator of Y, described in the next section.*

2.2.3 Maximum A Posteriori estimators

We focus in this section on the Maximum A Posteriori (MAP) estimator of the target Y given $\boldsymbol{X}$. In this thesis, this approach is at the center of the contribution proposed in Section 3.2.1.

First, we denote by p_y, a prior on Y. In the case of a dominated model, one can determine the distribution a posteriori by explicitly writing its density. Suppose that $\left\{P_y, y \in \mathbb{R}\right\}$ is a dominated model with $P(\mathrm{d}x|y) = p(x|y)\mu(dx)$ and let ν be a dominant measure of the distribution $p(\cdot|x)$, and continue to note p_Y the density, $p(\mathrm{d}y) = p(y)\nu(\mathrm{d}y)$. The joint density of the random vector $(\boldsymbol{X}, Y)$ with respect to the product measure $\mu^d \otimes \nu$ is then given by :

$$p(x, y) = p_Y(y)p(x|y). \tag{2.14}$$

Thus the density a posteriori is given (under the domination assumptions) as follows:

$$p(y|x) = \frac{p_Y(y)p(x|y)}{\int_{\mathbb{R}} p(x|u)p_Y(\mathrm{d}u)}. \tag{2.15}$$

This formalism has the advantage of making the density of the distribution of $X|Y$ the one to be estimated. This approach has numerous benefits. It allows us to introduce propagation models as in Section 1.3.2. These propagation models are based on physical considerations and then provide relevant parametrization for the distributions of interest. Furthermore, assumptions (as the one in Section 3.2.1) are usually made and drastically decrease the algorithmic complexity of the parametric model inference.

We refer to the MAP estimator of Y the mapping $\hat{Y}^{\mathrm{map}}$ defined as follows:

$$\hat{Y}^{\mathrm{map}}(x) = \arg\max_{z\in\mathbb{R}} p_Y(z)p(x|z). \tag{2.16}$$

One can also propose an estimator of the expectation of Y under the a posteriori distribution, noted $\hat{Y}^{\mathrm{mmse}}$:

$$\hat{Y}^{\mathrm{mmse}}(x) = \int_{\mathbb{R}} z \frac{p_Y(z)p(x|z)}{\int_{\mathbb{R}} p(x|u)p_Y(\mathrm{d}u)} \nu(\mathrm{d}z). \tag{2.17}$$

2.3 Tree-Based Methods

Tree boosting is a highly effective and widely used machine learning method (Friedman, 2002; Maimon and Rokach, 2014). Among the machine learning methods used in practice, gradient tree boosting is one technique that shines in many applications. Recent well-optimized implementations of tree boosting (such as GBM or XGBoost) allow its incorporation into real-world production pipelines. This type of predictor is thus a relevant choice for practical application this thesis focuses on. We present hereafter the basics to define the eXtreme Gradient Boosting (XGBoost Chen and Guestrin (2016)) that will be essential for a good reading of this thesis, in particular the Chapters 3 and 4. The general framework is the regression problem in which an input random vector $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ is observed, and the goal is to predict a random response $Y \in \mathbb{R}$ by estimating the regression function $m(x) = \mathbb{E}\left[Y|\mathbf{X} = \boldsymbol{x}\right]$. As usual, the objective is to use the data set $\mathcal{Z}_n$ to built an estimate $\hat{m}_n : \mathcal{X} \to \mathbb{R}$ of the function m.

2.3.1 Regression Trees

The basics of a regression tree

Decision tree can be both applied to classification and regression problem. A regression tree can be viewed as a tuple $(q, \boldsymbol{\omega})$ where $q : \mathbb{R}^d \to \{1, \dots, L\}$ is the structure of the tree that maps an input to a leaf index, $\boldsymbol{\omega} \in \mathbb{R}^L$ is the vector containing the leaves weights and L is the number of leaves. For a observed input $\boldsymbol{x}$, we use the decision rule given by q to map this input into a leaf and define the output as the weight of the leaf as illustrated in Figure 2.1.

We now briefly discuss the process of learning a regression tree. As mentioned above, a tree divide the feature space into non-overlapping regions (via its structure q). Each region is associated to a specific leaf. Let us denote these regions $R_1, \dots, R_L$. To every input that falls into a region, let say R_j, the model prediction will be the mean of the target values for the training inputs in R_j. Although these regions could have any shape in theory, it is preferred to give them a rectangular shape to ease the learning of the tree.

Recursive Binary Splitting

This greedy approach consists in successively splitting the predictor space from the top of the tree (the root) to the leaves. At each step, the algorithm chooses the direction of split j, and the cutpoint s such that dividing the predictor space into the regions $\{\boldsymbol{x} \mid \boldsymbol{x}_j < s\}$ and $\{\boldsymbol{x} \mid \boldsymbol{x}_j \geq s\}$ leads to the greatest possible reduction of the regression error. That is, we consider all predictors directions $1, \dots, d$, and all possible values of the cutpoint s for each of these directions, and then choose the direction and cutpoint such that the resulting tree has the lowest regression error. Next, we repeat this process, looking for the best direction and cutpoint in order to minimize the regression error within each of the resulting regions. The process continues until it reaches a stopping criterion; common stopping criteria are that the resulting regions contains a minimum number of observations, or that the tree contains a maximum number of leaves.

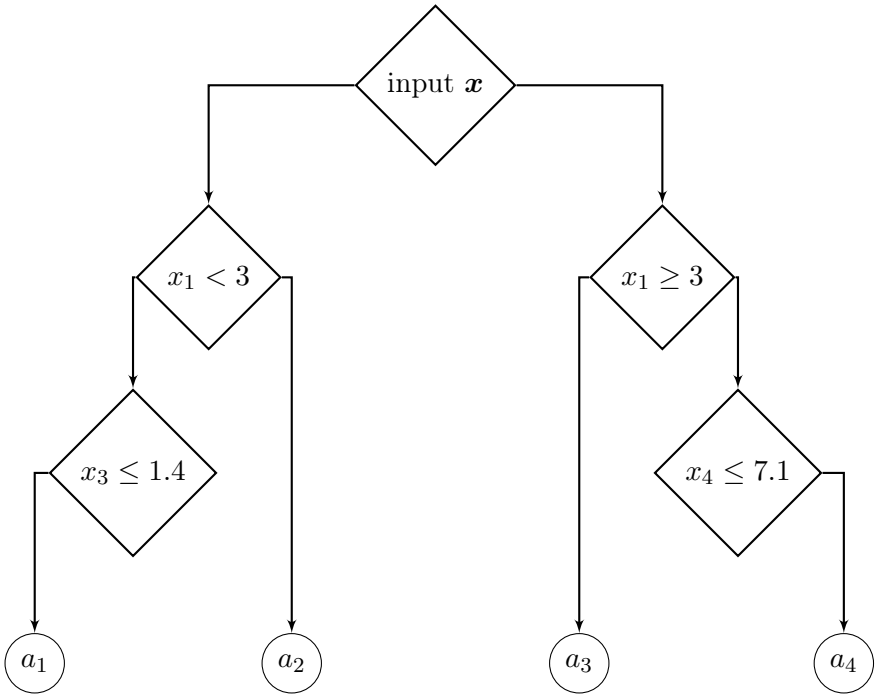


Figure 2.1 – Generic representation of a decision tree. The decision (output) of the tree given the input $\boldsymbol{x}$ is one of the weights $(a_i)_{i=1,\dots,6}$. For instance, the split at the top of the tree results in two large branches. The left-hand branch corresponds to $x_1 < 3$, and the right-hand branch corresponds to $x_1 \geq 3$. The tree has four terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there. As an illustration, the prediction of this tree given the input $\boldsymbol{x} = (7, 1, 3.1, 5, -1)$ is a_4: $x_1 = 7 \geq 3$, and $x_4 = 5 \leq 7.1$.

Pros and Cons of Trees

Regression trees play a central role in this thesis and are at the center of the main contribution (see Chapter 4). It is therefore appropriate to recall here that they have a number of advantages over the more classical approaches developed in Chapter 3. They also suffer from a lack of robustness: a small change in the data can substantially modify the final estimator. To overcome this lack of robustness, we discuss in the next section the concept of boosting.

- △ Trees can easily handle qualitative predictors without the need to create dummy variables (see the practical case herein).
- △ Trees are highly interpretable and can be furthermore graphically displayed.
- ▽ Trees can be highly non-robust, a small change in the data can cause a large change in the final estimated tree.

However, the aggregation of many decision trees allows to obtain a final estimator, called ensemble model, with a reduced sensitivity to a change of data (Breiman, 1996), and with better predictive performance at the price of losing interpretation. This aggregation procedure is discussed in the next section.

2.3.2 Boosting

There exists three methods for constructing ensembles of decision trees: *bagging, boosting*, and *randomization*. In this sequel, we mainly focus on the boosting method (Freund et al., 1999, 1996). This method is known to give better results as shown in Dietterich (2000) where the experiment section show that over a set of 33 tasks, the Adaboost algorithm gives the best results in most cases. However, this latter algorithm has already been proved to be sensitive to noise. Therefore, two of the most recent boosting algorithms (which perform better than Adaboost) are preferred hereinafter: eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), and Generalized Boosted Models (GBM) (Ridgeway, 2007).

First, we define an ensemble tree model as a predictor consisting of a collection of T randomized regression trees. Let consider the t^{th} tree, and denote by $f_t(\boldsymbol{x})$ the predicted value of this tree at the query point $\boldsymbol{x}$. The trees are combined in an additive manner and thus form the forest estimate:

$$\hat{m}_n^T(\boldsymbol{x}) \triangleq m_n(\boldsymbol{x}; f_1, \ldots, f_T) = \sum_{t=1}^{T} f_t(\boldsymbol{x}).$$

The idea of the boosting procedure is to learn the trees sequentially. Each new tree is learned with respect to the residual of the previous model rather that to the outcome Y as shown in Algorithm 2.2. This means that at each step a basis function that leads to the largest reduction of empirical risk is added into the estimator. In case of tree boosting, since a tree corresponds to split of the space of inputs into regions (see Section 2.3.1), the model is then improved in areas where it does not perform well. Note that in boosting, unlike in bagging, the construction of each tree strongly depends on the trees that have already been previously built. In general, boosting procedure can be tuned by means of parameters:

- The number of trees, noted $T \in \mathbb{N}^*$ herein and corresponds to the parameter `n_estimators` in `sklearn.ensemble.GradientBoostingRegressor` class. Unlike bagging and random forests, boosting can overfit if T is chosen too big. We therefore use a validation set to select T.
- The shrinkage parameter $\nu > 0$, a positive scalar. This parameter controls the rate at which boosting learns. At each stage, the contribution of each tree is shrunk by ν. This parameter is called `learning_rate`.
- The complexity of the tree, which can be controlled using different parameters such as the maximum depth of the trees, the number of leaves in a tree, or the minimum number of samples required to be at a leaf node denoted respectively by `max_depth`, `max_leaf_nodes` and `min_samples_leaf`.

In the literature James et al. (2013), it is often recommended to use small trees, with just a few terminal nodes. For this plot, we consider the well-known prediction problem in which we want to predict the housing values based on 13 features such as per capita crime rate by town (`crim`), the average number of rooms per dwelling (`rm`) or the nitrogen oxides concentration (`nox`). This dataset can be found in the `sklearn.datasets`[1].

[1] sklearn Boston Dataset

In particular, simple stumps with a depth of one perform well if enough of them are included. This model can even outperform the depth-two model or a random forest. This highlights one difference between boosting and random forests: in boosting, because the growth of a particular tree takes into account the other trees that have already been grown, smaller trees are typically sufficient.

GBM

We begin this section on tree-based methods with a brief description of the recent Boosting implementation known as GBM. This implementation can be found in the gbm package (LightGBM). The gbm package takes the approach described in Friedman (2002). Some of the terminology somehow differs for generality. In addition, the gbm package implements boosting for models commonly used in statistics but not commonly associated with boosting. The gbm implementation of the boosting procedure is described in Algorithm 2.1

Algorithm 2.1 Boosting as implemented in `gbm()`

Result: $\hat{m}_n = \sum_t f_t$

$m_n = 0,\ t = 1,$ and $r_i = y_i\ \forall i \in \{1, \ldots, n\}$, a loss function $\mathcal{L}$.

while *Stopping criterion is* ***False*** **do**

  Compute the negative gradient as the response.

  $r_i = \frac{\partial}{\partial m(\boldsymbol{x}_i)} \mathcal{L}\left(y_i, m(\boldsymbol{x}_i)\right)\Big|_{m(\boldsymbol{x}_i)=\hat{m}_n(\boldsymbol{x}_i)}$

  Randomly select $p \times k$ cases from the dataset.

  Fit a regression tree f_t w.r.t. those randomly selected observations.

  $\hat{m}_n \leftarrow \hat{m}_n + f_t$

  $t \leftarrow t + 1$

XGBoost Algorithm

We now discuss in details the XGBoost algorithm, and conclude this section with a presentation of Gradient tree boosting also known as Gradient Boosted Regression Tree (GBRT) and specifically a scalable end-to-end tree boosting system called eXtreme Gradient Boosting (XGBoost). Boosting for regression trees is described in Algorithm 2.2.

Algorithm 2.2 Pseudo-code of Boosting for Regression Trees

Result: $\hat{m}_n = \sum_t f_t$

$\hat{m}_n = 0,\ t = 1,$ and $r_i = y_i\ \forall i \in \{1, \ldots, n\}$.

while *Stopping criterion is* ***False*** **do**

  fit a tree f_t to the training data $\left\{(\mathbf{x}_i, r_i)\right\}_{i=1,\ldots,n}$.

  $\hat{m}_n \leftarrow \hat{m}_n + f_t$

  $\forall i \leq n : r_i \leftarrow r_i - f_t(\boldsymbol{x}_i)$

  $t \leftarrow t + 1$

The main contribution of GBRT is to use a second-order approximation of the objective in order to ease its minimization. Let ℓ be a regression loss function, typically the L_2-norm, and Ω a regularizer to be defined later, and consider the following *regularized*

objective:

$$\mathcal{L}(\hat{m}_n) \triangleq \sum_{i=1}^{n} \ell\left(\hat{m}_n(\mathbf{x}_i), y_i\right) + \sum_s \Omega(f_s). \tag{2.18}$$

Let $\hat{m}_n^{(t-1)}$ be the obtained predictor at the $(t-1)^{\text{th}}$ iteration:

$$\hat{m}_n^{(t-1)}(x) = \sum_{s=1}^{t-1} f_s(x). \tag{2.19}$$

We need to add f_t to minimize the following objective.

$$\mathcal{L}^{(t)}(f) \triangleq \sum_{i=1}^{n} \ell\left(\hat{m}_n^{(t-1)}(\boldsymbol{x}_i) + f(\boldsymbol{x}_i), y_i\right) + \Omega(f). \tag{2.20}$$

To ease the minimization of the objective, second-order approximation originated from Friedman et al. (2000) can be used:

$$\mathcal{L}^{(t)}(f) \simeq \sum_{i \leq n} \left[\ell\left(\hat{m}_n^{(t-1)}(\boldsymbol{x}_i), y_i\right) + g_i f(\boldsymbol{x}_i) + \frac{1}{2} h_i f^2(\boldsymbol{x}_i)\right] + \Omega(f), \tag{2.21}$$

where $g_i = \partial_{\hat{m}_n^{(t-1)}(\boldsymbol{x}_i)} \ell(\cdot, y_i)$, and $h_i = \partial^2_{\hat{m}_n^{(t-1)}(\boldsymbol{x}_i)} \ell(\cdot, y_i)$ are the first and second order gradient of the loss function ℓ w.r.t. the first coordinate. Since the term $\ell(\hat{m}_n^{(t-1)}(\boldsymbol{x}_i), y_i)$ is a constant of f, we can remove it to obtain a simplified version of the objective as follows:

$$\tilde{\mathcal{L}}^{(t)}(f) = \sum_{i \leq n} \left[g_i f(\boldsymbol{x}_i) + \frac{1}{2} h_i f^2(\boldsymbol{x}_i)\right] + \Omega(f), \tag{2.22}$$

It is possible to give a closed-form expression of the leaves weights $\boldsymbol{\omega}$ for a fixed structure q of the predictor f. To this aim, we need to reindex the sum in Equation (2.22). Let l be an index of a leaf, $(q, \boldsymbol{\omega})$ the tuple representing the tree regressor f and $\mathcal{I}_l \triangleq \{i \leq n : q(\boldsymbol{x}_i) = l\}$ the set of indices that falls into the leaf l. Now, on the set of indices $\mathcal{I}_l$, the tree regressor f is constant and equal to ω_l. Consequently, noting L the size of $\boldsymbol{\omega}$ we have

$$\tilde{\mathcal{L}}^{(t)}(f) = \sum_{l \leq L} \sum_{i \in \mathcal{I}_l} \left[g_i \omega_l + \frac{1}{2} h_i \omega_l^2\right] + \lambda \|\boldsymbol{\omega}\|_2^2 + \gamma |L|, \tag{2.23}$$

where we expanded the regularizer $\Omega(f) = \lambda \|\boldsymbol{\omega}\|_2^2 + \gamma L$.

For a fixed structure q (L is constant), the optimal values of the leaves weights $\omega_l^\star(q)$ are thus given for all $l \leq L$ by :

$$\omega_l^\star(q) = \frac{\sum_{\mathcal{I}_l} g_i}{\sum_{\mathcal{I}_l} h_i + \lambda}. \tag{2.24}$$

We can compute the value of the objective at the point $\left(q, \boldsymbol{\omega}^\star(q)\right)$ and use it as a scoring function to measure the quality of a tree structure q. To lighten notation, we will simply write $\omega_l^\star$ instead of $\omega_l^\star(q)$. It comes directly that this optimal value is given

by

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{l\leq L}\frac{\left(\sum_{\mathcal{I}_l} g_i\right)^2}{\sum_{\mathcal{I}_l} h_i + \lambda} + \gamma L. \tag{2.25}$$

Now that we have a score to evaluate the quality of a tree structure (like the impurity score for evaluating decision trees (Shalev-Shwartz and Ben-David, 2014)) we can apply the greedy procedure introduced in Section 2.3.1 by starting from the root and iteratively adds branches to the tree.

Outline of the thesis

We would like to recall the place that the elements seen in this chapter will have in the sequel. Section 1.3.2 focuses on the problem of channel-based location estimation techniques, by means of two well-known techniques: the time delay estimation, and the RSSI-based range estimation. Both of these approaches involve a parameter estimation as well as the notion of CRLB. Chapter 3 provides a review of methods of geolocation. It also proposes a original classification of these methods which are referred to as fingerprinting methods and likelihood-based methods. While the first family of methods is quite self-contained, the last one relies on the estimation of the likelihood by means of Nadaraya-Watson estimates described above. In Chapter 5, we propose to assess the conditional independence between variables. The proposed test deeply relies on the conditional distributions estimation; we besides provide a practical example based on the Nadaraya-Watson estimates. Finally, in Chapter 4 we propose to adapt the boosting methods seen in Section 2.3 to metric learning. The learned metric has the particularity to be optimal w.r.t. the regression task pursued.

3

RSSI-based Methods for Location Estimation

"Adam, où es-tu?"

Genèse 3:9

In this chapter, we investigate machine learning approaches addressing the problem of geolocation. First, we review some classical learning methods to build a radio map. These methods are split in two categories, which we refer to as likelihood-based methods and fingerprinting methods. Then, we provide a novel geolocation approach in each of these two categories. The first proposed technique relies on a semi-parametric Nadaraya-Watson (NW) estimator of the likelihood, followed by a maximum a posteriori (MAP) estimator of the object's position. The second technique consists in learning a proper metric on the dataset, constructed by means of a Gradient boosting regressor: a k-nearest neighbor algorithm is then used to estimate the position. The proposed methods are compared on two data sets originated from Sigfox network, and an indoor dataset performed in a three-story building. Experiments show the interest of the proposed methods, both in terms of location estimation performance, and ability to build radio maps.

This chapter covers the following publications:

- Elgui, K., Bianchi, P., Portier, F. & Isson, O. (2019, September). Learning Methods for RSSI-based Geolocation: A Comparative Study. In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.
- Elgui, K., Bianchi, P., Isson, O., Portier, F. & Marty, R. (2020, April). Metric Learning for Fingerprint RSSI-Localization. In 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS) (pp. 1036-1042). IEEE.
- Elgui, K., Bianchi, P., Portier, F. & Isson, O. Learning Methods for RSSI-based geolocation: A comparative study, Pervasive and Mobile Computing (2020).

3.1 Introduction

In the considered IoT dedicated networks, with low power and bandwidth devices, localization is challenging. Standard methods such as famous range-based methods by means of TOA estimation as instance are not well suited for this application. Therefore, machine learning methods for localization have met a great deal of attention these past ten years. Besides they have shown their value for indoor scenarios as in Farjow et al. (2011); Wang et al. (2018). Nevertheless, the accuracy of such methods might be jeopardized in LPWAN (Farid et al., 2013). In this chapter we focus on a baseline probabilistic RSSI-only localization algorithm. The main challenge comes from the large fluctuations of the observed RSSI values, for a given source location (see Fig. 3.1).

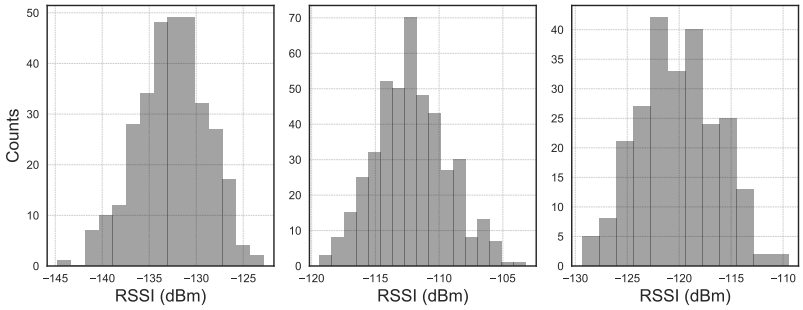


Figure 3.1 – Histograms from Sigfox network. Each plot shows histogram of the RSSI at a BS for three emitters of which the positions are fixed (the empirical standards deviations are resp. 13dBm, 7dBm and 13dBm). Given a fixed position of an emitter, the received RSSI's show large fluctuations.

In such data, the observed signals can be very noisy, especially in urban environment (RSSI based methods are often assisted with accelerometers, gyroscopes or Bluetooth beacons to improve their accuracy Yiu et al. (2017)). It may also happen that, due to range limitation or network sensitivity, some messages are not detected by some BSs. In the network of interest for instance, the lowest observed value of the RSSI is -166 dBm. Below this value, the signal is not collected by the BSs, either because of a lack of sensitivity, or because of an SNR that is too low. In certain cases however, it can also be characteristic of a signal emitted from too far away. Despite the fact that experience has shown the benefits of taking into account the information of non-reception, this information is still, rarely used in the literature.

Related literature. Machine learning techniques have been extensively considered for RSSI based localization, for instance to track customers Oosterlinck et al. (2017), or in application to autonomous vehicles Campbell et al. (2018). The first fingerprint-based method was proposed in Bahl et al. (2000). In this prior work, k-Nearest Neighbor (k-NN) is used to tackle the task of predicting the location given RSSI with *RADAR*, a system for locating and tracking users inside a building. Let us also mention Jedari et al. (2015) which compares the performance of k-NN with random forest to predict the indoor location of using RSSI-based fingerprinting method. This work emphasizes the importance of the distance function used by k-NN. Besides, different distance functions such as Euclidean, Manhattan and Minkowski are compared in terms of accuracy of the resulting predictor. In Xie et al. (2016), the authors exploit the Spearman rank correlation of RSSI measurements to improve k-NN. However, the authors of these prior works do not study the choice of the optimal metric. In contract, we propose to learn a (non parametric) metric function, by means of a Gradient Boosting algorithm. This metric is learned such that, when evaluated on a pair of RSSI vectors, it gives a good approximation of the Vincenty distance between the two emitters.

In recent years, neural networks have also been widely used for RSSI-based localization, in particular for indoor use cases Iqbal et al. (2018); Ahmadi and Bouallegue (2015); Zhang et al. (2016). For instance, Zhang et al. (2016) proposes to deal with the high variability of the RSSI by using a four-layer deep neural network structure to extract relevant features.

Another approach is to model the signal strength as a function of the position. In the literature, these methods are generally referred to as *range-based* methods because they use a signal propagation model that maps an observed RSSI value to a distance (range) estimate. The localization procedure then tries to find the location which best agrees with the measured signal strengths. These methods then require the introduction of a likelihood model, which can be parametric or semi-parametric. We therefore refer to these approaches as likelihood-based methods. A commonly used parametric model is the so-called *log distance path-loss* model Wang et al. (2012); Xu et al. (2014). This model is a radio propagation model that provides a mapping between the received signal strength and the distance between the emitter and the receiver. The main advantage of parametric models is that they only require the calibration of few parameters, which helps the training phase of the model and do not require a large training set. However, they often fail to represent arbitrary distributions Yiu et al. (2017), which could deteriorate the geolocation accuracy (we refer to Zhou et al. (2018); Botteron (2003) for in-depth studies of the geolocation accuracy of log-loss model). To remedy this, semi-parametric models have been introduced. The latter benefit from the advantages of parametric models while still being able to represent arbitrary distributions. To build a likelihood in a semi-parametric fashion, many methods have been employed such as kernel methods Mirowski et al. (2012); Mahfouz et al. (2013, 2015) and Gaussian Processes (GPs) Schwaighofer et al. (2004); Hähnel and Fox (2006) or more recently in Bisio et al. (2017) with *Smart*^{2}. First, as far as we know, in the related literature, few of works (Piórkowski and Grossglauser, 2006) have payed attention on the variable representing the reception/non reception of the signal at BSs discussed above. In most works, an arbitrary low RSSI value is attributed whenever a signal is not received as in Dashti et al. (2015); Janssen et al. (2020). Moreover, almost all of those found in the literature give a statistical relationship between the RSSI and the distance between emitter and receiver through a propagation model involving physical considerations. This implies that they therefore disregard the non-isotropy of the environment.

The methods proposed in this chapter, in contrast, take into consideration both the non-isotropy and the information of non-reception. The detailed contributions of this chapter are the following.

Contributions.

- We build a relevant metric for RSSI based geolocation. In order to achieve this, we propose to learn a non parametric metric function by means of a Gradient Boosting algorithm. The proposed method benefits both from the simplicity and robustness of k-NN and the regression performance of XGBoost, and more specifically from its ability to deal with complex and high dimensional data such as RSSI vectors. The idea is to learn the metric used by the k-NN explicitly to enhance a predictor estimation Bellet et al. (2013). Herein, the metric is built to compare two RSSI's vectors, such that the k-NN regressor provides the most appropriate neighbors for the geolocation. The main idea is to learn this metric d such that for a couple of RSSI's vectors $(\boldsymbol{x}, \boldsymbol{x'})$, $\mathsf{d}(\boldsymbol{x}, \boldsymbol{x'})$ is a relevant predictor of the Euclidean distance between the two emitters' locations $\|y - y'\|_2$. We propose to learn d as a sum of T regression trees. Those trees are obtained through XGBoost algorithm (see Section 2.3). The benefits w.r.t. a classic k-NN regressor are twofold. First, the estimate benefits from the information of reception/non reception of the signal at a BS. Second, it improves the model by optimizing the metric explicitly for the

task of geolocation. This leads to higher accuracy of the model as demonstrated in Section 3.4.

- We propose a method that takes advantage of the Boolean variable representing the reception/non reception of the signal at BSs. Our proposed model relies on a specific likelihood model of the RSSI's given the object's position. The expression of the likelihood is based on a model assumption of Naive Bayes type: given the emitter's position, the coordinates of the RSSI vector are assumed independent. It allows to build an estimator whose implementation is practical. The distribution of a RSSI at a given BS given the location of the emitter, is set as a Gaussian distribution. The mean and the standard deviation of this distribution are obtained by a non-parametric Nadaraya-Watson estimator Section 2.2.2. The final location is obtained using a Maximum-A-Posteriori (MAP) estimator. As will be shown in Section 3.4.1, one of the assets of the method is that it performs well, even on small training data sets. Finally, as a generative model, it allows to compute useful statistical information about our estimate such as density level sets and confidence regions unlike standard machine learning methods such as the k-NN estimator described below.

- We discuss the computational complexity as well as the memory cost of the methods founds in the related literature.

- We provide detailed experiments using real data originated from the Sigfox network, and an Indoor Positioning Dataset from Zsolt Tóth (2016).

Outline The rest of the chapter is organized as follows. Section 3.2 investigates several popular geolocation techniques of the literature. Then, in Section 3.3, we introduce the two proposed predictors. Finally, Section 3.4 is devoted to the numerical experiments and discussions.

3.2 Machine Learning Methods for Geolocation

In this section, we discuss different off-the-shelf predictors which can be used to solve the geolocation task introduced above.

3.2.1 Likelihood-based Methods

We refer to as Likelihood-based methods the methods which learn from the dataset a likelihood model $p(\boldsymbol{x}|y)$ for the conditional probability of the RSSI vector $\boldsymbol{X}$ given the position Y. These methods are presented in the Section 2.1.2.

One first learns from the observed data $\mathcal{Z}_n$ a mapping $p(\boldsymbol{x}|y)$ which represents the conditional pdf of $\boldsymbol{X}|Y$. One way is to introduce a parametric likelihood model as presented in Section 2.1.2, such as the path-loss model discussed at the end of this paragraph (and previously introduced in Section 1.3.2). The parameters of the model are then learned from the dataset. Non-parametric methods can be used as well (see Section 3.3). One of the main advantages is that some prior hypotheses on the form of the likelihood $p(\boldsymbol{x}|y)$ can be, based on physical considerations, easily introduced. In the sequel, we will assume as in Kaemarungsi and Krishnamurthy (2004); Mazuelas et al. (2009); Li (2006); Bshara et al. (2010) that the conditional pdf admits the following

expression:

$$p(\boldsymbol{x}|y) = \prod_{k=1}^{d} p_k(x_k|y)\,, \tag{3.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and where $p_1, \ldots, p_d$ are the conditional marginals to be learned from the training set. This implies that the components $X_1, \ldots, X_d$ of the random vector $\mathbf{X}$ are independent conditionally to Y.

Assume now that a new message arises from the unknown position Y with an observed RSSI vector $\boldsymbol{x}$. we can then make the choice to use the MAP estimator presented in Section 2.2.

$$\begin{aligned}\hat{Y}_{MAP}(\boldsymbol{X}) &\triangleq \arg\max_{y\in\mathcal{Y}} p(y|\boldsymbol{X}) \\ &= \arg\max_{y\in\mathcal{Y}} \sum_{k=1}^{d} \log p_k(X_k|y) + \log p_Y(y). \end{aligned} \tag{3.2}$$

The remaining task is to provide a model for p_k's, which can be learned from the training set. This step is commonly referred to as the *calibration step* and is described in the preliminary Chapter 2.

Let us now provide a practical example with the broadly used *log-loss* (or *path-loss*) parametric model (Yiu et al., 2017; Bshara et al., 2010; Barsocchi et al., 2009) described in Section 1.3.2. The conditional distribution $p_k(x_k|y)$ of $X_k|Y$ is supposed to have the form $p_{\theta_k}(x_k|y)$ where $\theta_k = (P_{0,k}, \nu_k, \sigma_k^2)$ is a triplet of parameters, $p_{\theta_k}(\,.\,|y)$ is a Gaussian distribution of variance σ_k^2 and mean $P_{0,k} - 10\nu_k \log_{10} d_v(y, y_k)/d_0$. Here, d_0 is some reference distance and d_v stands for the Vincenty distance Vincenty (1975), the parameters $P_{0,k}, \nu_k$ respectively represent the power in dBm at distance d_0 and ν_k is the so-called path-loss exponent.

The parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ is then estimated from the dataset $\mathcal{Z}_n$ using a standard maximum likelihood approach as described in Chapter 2. Once the parameter $\boldsymbol{\theta}$ has been estimated, the second phase consists in replacing in Equation (3.2) the unknown distributions p_k's by their estimates and thus obtain our estimator.

3.2.2 Fingerprinting Methods

Fingerprinting methods directly map the vector $\boldsymbol{X}$ into a location $\hat{Y}$ (radio-map), typically by means of a supervised learning algorithm. There have been extensively studied Yiu et al. (2017); Brunato and Battiti (2005); Torres-Sospedra et al. (2015); Honkavirta et al. (2009). We briefly present some popular learning algorithms to build this mapping.

k-Nearest Neighbours

This method is used in Yiu et al. (2017); Patwari (2005) in the context of outdoor geolocation. As discussed in Section 1.5, some messages may not be detected by some BS, in which case we just set the corresponding RSSI value to `NaN`. In this section, we endow the space of RSSI's vectors with the Euclidean distance. For this aim, we need to replace all the NaN values. We suggest, as in Yiu et al. (2017), to replace all the `NaN` values either by the lowest RSSI amongst all observed RSSI (in this paper -200

dBm will be used), or by an arbitrary value (the value −110 dBm is used in Yiu et al. (2017)).

For every D-dimensional RSSI vector $\boldsymbol{X}$, we let $(\boldsymbol{x}_{(1)}, y_{(1)}), \ldots, (\boldsymbol{x}_{(n)}, Y_{(n)})$ be a reordering of the dataset $\mathcal{Z}_n$ such that $\|\boldsymbol{X} - \boldsymbol{x}_{(1)}\| \leq \cdots \leq \|\boldsymbol{X} - \boldsymbol{x}_{(n)}\|$. The unknown position Y is finally estimated as follows:

$$\hat{Y} = \frac{1}{k} \sum_{i=1}^{k} y_{(i)}, \tag{3.3}$$

where the integer k is an hyperparameter (see Hall et al. (2008) for a discussion on the choice of k).

Ensemble Trees Methods

Two important classes of ensemble methods are bagging methods such as random forests Breiman (2001), and boosting methods such as Gradient Tree boosting Friedman (2002). A Random Forest model has been applied as a classifier for a indoor-context geolocation Jedari et al. (2015) in which it gets better accuracy than a k-NN based method. In Li et al. (2018) used both TDOA and RSSI as input to indoor location estimation. The authors proposed a combination of a k-NN (to remove outliers) and a Random Forest Regressor.

Multi-Layers Perceptron

In Ahmad et al. (2006); Dai et al. (2016), the authors propose to use a Multi Layer Perceptron (MLP) approach to reduce the uncertainty in an *indoor* location estimation system. In the present section, we suggest to use a two hidden layers perceptron (see Fig. 3.2). The first hidden layer is composed by $n_h = 250$ nodes. Each of these node computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights, and puts this output through the logistic sigmoid activation function. The second hidden layer is similar to the first, but has no activation function. For every node $l \leq 250$:

$$h_l^{(1)}(\boldsymbol{X}) = \sigma\left(\sum_d \alpha_{l,d} X_d + \beta_l\right) = \sigma\left(\langle \boldsymbol{\alpha}_l, \mathbf{X} \rangle + \beta_l\right), \tag{3.4}$$

where σ is the logistic sigmoid function $\sigma(x) = e^x/(1+e^x)$. For every node $l' \leq 250$:

$$h_{l'}^{(2)}\left(\mathbf{h}^{(1)}(\boldsymbol{X})\right) = \sum_p \gamma_{l',p} h_p^{(1)}(\boldsymbol{X}) + \delta_{l'} = \sigma\left(\left\langle \boldsymbol{\gamma}_{l'}, \mathbf{h}^{(1)}(\mathbf{X}) \right\rangle + \delta_{l'}\right), \tag{3.5}$$

where $\mathbf{h}^{(1)}(\boldsymbol{X}) \triangleq \left(h_1^{(1)}(\boldsymbol{X}), \ldots, h_{250}^{(1)}(\boldsymbol{X})\right)^\top$. The final estimator of the position Y is as follows:

$$\hat{Y} = \sum_{l'} \zeta_{l'} h_{l'}^{(2)}\left(\mathbf{h}^{(1)}(\boldsymbol{X})\right) + y_0. \tag{3.6}$$

The parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\zeta}, y_0)$ are learned by minimizing the quadratic loss on the examples of $\mathcal{Z}_n$.

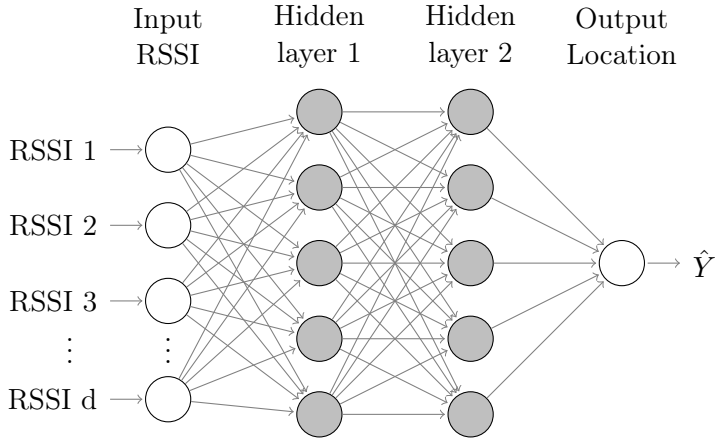


Figure 3.2 – Figure of a two hidden layers perceptron

Method	Training complexity	Memory cost	Prediction complexity	Non-Reception
Likelihood	$\mathcal{O}(1)$	$\mathcal{O}(dn)$	$\mathcal{O}(d\lvert\mathcal{Y}\rvert)$	Yes
k-NN	$\mathcal{O}(1)$	$\mathcal{O}(dn)$	$\mathcal{O}(dnk)$	No
Ensemble Trees	$\mathcal{O}(dTn\log n)$	$\mathcal{O}(2^qT)$	$\mathcal{O}(qT)$	Yes
Perceptron	$\mathcal{O}(nn_h)$	$\mathcal{O}(n_h)$	$\mathcal{O}(n_h)$	No

Table 3.1 – Criteria of the four presented methods. We denoted by d the dimension of the RSSI vector $\boldsymbol{X}$, n_h the number of neurons in the perceptron model. At last, T stands for the number of trees of the ensemble trees regressor, and q the depth of a single tree.

3.2.3 Discussion

In this part, we discuss technical aspects of the methods described above. This discussion is structured around four criteria:

- The training complexity of the model training.
- The memory cost of the predictor.
- The prediction complexity of the model.
- The ability to take into account the non-reception.

In Table 3.1, we illustrate the different criteria for the learning methods introduced in Section 3.2. We denote n the size of the training set, d the dimension of the RSSI vector $\boldsymbol{X}$. In this table, we also denote n_h the number of neurons in the perceptron model. Finally, T stands for the number of trees of the ensemble trees regressor, and q the depth of a single tree.

The likelihood based method is one of the cheapest method in term of both training complexity and memory cost. Indeed, this method only requires to stack into memory the conditional marginals $p_1, \ldots, p_d$. Thus, the memory space of these methods depends on the number of parameters. Non-parametric methods can be used as well (see

Section 3.3) to estimate this conditional marginals, in which case we have a space complexity of $\mathcal{O}(dn)$. However, the prediction is expensive because it requires to compute the arg max of the likelihood over the set $\mathcal{Y}$.

Finally, while ensemble trees methods and perceptron offer lower both memory cost and prediction complexity, they require a long training phase. The time complexity of a method training is yet crucial. If this complexity is too high, one may not be able to retrain the model whenever the network changes. In that sense, the k-NN and the likelihood based methods should be preferred.

3.3 Proposed Geolocation Methods

In this section, we present two new geolocation approaches. The first relies on a semi-parametric Nadaraya-Watson (NW) estimator of the likelihood, followed by a maximum a posteriori (MAP)estimator of the object's position. The second technique consists in learning a proper metric, constructed by means of a Gradient boosting regressor: a k-nearest neighbor algorithm is then used to estimate the position.

3.3.1 Semi-Parametric Likelihood-Based Method

In the sequel, we propose a semi parametric likelihood-based method. Due to the `NaN`-values, we modify the classical Gaussian Process model used Hähnel and Fox (2006) in order to leverage the information of non reception. As in Yiu et al. (2017); Hähnel and Fox (2006) we generate a likelihood model for signal strength measurements using a semi-parametric framework. We propose the following likelihood model for the conditional density $p(\boldsymbol{x}|y)$ of $\boldsymbol{X}$ given Y. Using the likelihood form presented in Section 3.2, it is sufficient to provide a model for the marginal conditional distributions $p_k(x_k|y)$ of X_k given Y, for every $k = 1, \dots, d$.

Here, we recall that, unlike in Hähnel and Fox (2006); Yiu et al. (2017), X_k is a random variable over the set $\mathbb{R} \cup \{\texttt{NaN}\}$. Densities are thus considered w.r.t. the reference measure $\lambda + \delta_{\texttt{NaN}}$ where λ is the Lebesgue measure and $\delta_{\texttt{NaN}}$ is the Dirac measure at the `NaN`-value. We define $\pi_k : \mathcal{Y} \to [0, 1]$ as

$$\pi_k(z) \triangleq \mathbb{P}(X_k = \texttt{NaN}|Y = y) = \mathbb{E}\left[\mathbb{1}_{\{\texttt{NaN}\}}(X_k)|Y = y\right]$$

and we constrain the model by assuming that, given Y and given that $X_k \neq \texttt{NaN}$, X_k follows a Gaussian distribution whose mean and variance are respectively denoted by $m_k(y)$ and $\sigma_k^2(y)$:

$$\begin{aligned} m_k(y) &\triangleq \mathbb{E}(X_k|Y = y, X_k \neq \texttt{NaN}), \\ \sigma_k^2(z) &\triangleq \mathrm{Var}(X_k|Y = y, X_k \neq \texttt{NaN}). \end{aligned}$$

We denote by $\Phi(x; m, \sigma^2) = (\sqrt{2\pi}\sigma)^{-1} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$ the normal density of mean m and variance σ^2. We summarize our model as follows:

1. $X_1, \dots, X_d$ are independent given Y.
2. For every k,

$$\mathbb{P}(X_k \in \mathrm{d}x|Y) = \pi_k(Y)\delta_{\texttt{NaN}}(\mathrm{d}x) + \Big(1 - \pi_k(Y)\Big)\, \Phi\Big(x; m_k(Y), \sigma_k^2(Y)\Big)\, \mathrm{d}x. \quad (3.7)$$

Based on this model, the likelihood $p(\boldsymbol{x}|y)$ is fully determined by the mappings π_k, m_k and σ_k^2 for all $k = 1, \dots, d$. The remaining task is to estimate these quantities using our dataset $\mathcal{Z}_n$. To this end, we propose to use a non-parametric approach, and to replace these mappings with their Nadaraya-Watson estimates introduced in Chapter 2.

The Nadaraya-Watson Estimator for Non-Parametric regression. Let $K : \mathcal{Y} \to \mathbb{R}_+$ be a kernel, i.e., non-negative, symmetric function integrating to one, and let $h > 0$ be a scalar (the so-called *bandwidth*). Define $K_h(y) = h^{-1}K(h^{-1}y)$ for all $y \in \mathcal{Y}$.

The Nadaraya-Watson estimates are respectively given for every $k \leq d$ by:

$$\hat{\pi}_k(y) \triangleq \frac{\sum_{i=1}^n \mathbb{1}_{\texttt{NaN}}(x_{i,k}) K_h\left(y_i - y\right)}{\sum_{i=1}^n K_h\left(y_i - y\right)}, \tag{3.8}$$

$$\hat{m}_k(y) \triangleq D_k(y)^{-1} \sum_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_{i,k}) x_{i,k} K_h\left(y_i - y\right), \tag{3.9}$$

$$\hat{\sigma}_k^2(y) \triangleq D_k(y)^{-1} \sum_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_{i,k}) (x_{i,k} - m_k(y))^2 K_h\left(y_i - y\right), \tag{3.10}$$

where $D_k(y) \triangleq \sum_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_{i,k}) K_h\left(y_i - y\right)$.

Under standard technical conditions, $\hat{\pi}_k$, $\hat{m}_k$ and $\hat{\sigma}_k^2$ converge uniformly towards π_k, m_k and σ_k^2 as $n \to \infty$ and $nh \to \infty$ (Tsybakov, 2013). Finally, the MAP location estimator can be written as:

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \sum_{k \in \mathcal{I}_{\boldsymbol{X}}} (1 - \hat{\pi}_k(y)) \log \Phi(X_k; \hat{m}_k(y), \hat{\sigma}_k^2(y)) + \sum_{k \in \mathcal{I}_{\boldsymbol{X}}^c} (1 - \hat{\pi}_k(y)) + \log \hat{p}_Y(y), \tag{3.11}$$

where $\mathcal{I}_{\boldsymbol{X}} \triangleq \{k = 1, \dots, d : X_k \neq \texttt{NaN}\}$ stands for the set of the receiving BS's, and where $\hat{p}_Y(y)$ stands for an estimation of the prior on Y, which we suggest to estimate from the dataset $\mathcal{Z}_n$ through the kernel density estimator:

$$\hat{p}_Y(y) = n^{-1} \sum_{i=1}^n K_h\left(y_i - y\right).$$

The training step of this algorithm is very efficient. Thus, the model can be re-fitted very quickly. This can be useful if a BS is removed or shifted for example. This method enables us, in a very simple manner, to take into account the Boolean variable representing the reception/non reception of the signal at BS's. In addition, as a generative model, it allows to compute useful statistical guaranties on the location estimation such as confidence level sets (see Section 3.4).

With this choice of estimates, we have for every $k \leq d$:

$$\hat{\pi}_k(y) = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^n \left(\mathbb{1}_{\texttt{NaN}}(x_{i,k}) - a\right)^2 K_h\left(y_i - y\right), \tag{3.12}$$

$$\hat{m}_k(y) = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^n \left(x_{i,k} - a\right)^2 \mathbb{1}_{\mathbb{R}}(x_{i,k}) K_h\left(y_i - y\right). \tag{3.13}$$

Proof The result follows immediately from taking the value that cancels out the gradient. ■

Thus, these estimates are obtained by a local constant least squares approximation of the target. The locality is determined by a kernel K_h that downweights all the Y^i that are not close to y whereas a plays the role of a local constant to be fitted. Because of the use of the conditional probability distribution of the components of $\boldsymbol{X}$ given Y instead of the one of Ygiven $\boldsymbol{X}$, the choice of an appropriate kernel is easy.

We finally discuss broadly the Cross-validation procedure for bandwidth selection. The choice of kernel bandwidth h is a key element for a kernel regression. Therefore, we suggest a K-Fold Cross-Validation selection of bandwidth h. Hereinafter, we provide a criterion which consists in assessing the regression performance of the reception given the position Y. We randomly divide the set of observations into K groups of equal size. These groups are denoted by $\{I_k\}_{k=1}^K$. We denote by $MSE_{(k)}(h) = \frac{1}{|I_k|}\sum_{i\in I_k} \|\mathbb{1}_{\texttt{NaN}}(\boldsymbol{x}_i) - \hat{\boldsymbol{\pi}}_h(y_i)\|^2$ the mean squared error computed on the observations on I_k, where $\hat{\boldsymbol{\pi}}_h(y_i)$ stands for the kernel regressor of $\mathbb{1}_{\texttt{NaN}}(\mathbf{X})|Y$ trained on $\{1,\dots,n\}\backslash I_k$. Finally, we chose $h^\star$ as $h^\star = \arg\min_h \frac{1}{K}\sum_k MSE_{(k)}(h)$.

3.3.2 A new metric for k-NN

To enhanced the k-NN method, we can learn a appropriate metric Xie et al. (2016); Torres-Sospedra et al. (2015)) on $\mathcal{X}^d$. The impact of the distance on the prediction of k-NN regressor is widely discussed in Honkavirta et al. (2009), where several $p-$norms are compared. Nevertheless, there is no mention about learning such a distance. Let recall that `NaN` values are here replaced by a fixed real value as discussed in Section 3.2.2. Our contribution is to build a mapping $\mathsf{d} : \mathcal{X}^d \times \mathcal{X}^d \to [0, +\infty)$ such that *close RSSI* (w.r.t. to the metric d) *correspond to close object positions* (w.r.t. to the Vincenty distance d_v on $\mathcal{Y}$). In that sense, a "good" metric is a mapping d for which the empirical risk

$$R_n(\mathsf{d}) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\mathsf{d}(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_v(y_i, y_j)\Big)^2$$

is small. The main trick, is to search for a mapping d minimizing $R_n(\mathsf{d})$ within a relevant hypothesis class. We suggest to search for d under the form

$$\mathsf{d}(\boldsymbol{x}, \boldsymbol{x\prime}) \triangleq \sum_{t=1}^{T} f_t(\Phi(\boldsymbol{x}, \boldsymbol{x}'))\,,$$

where $f_1, \dots, f_T$ is a collection of T regression trees, and where $\Phi : \mathcal{X}^d \times \mathcal{X}^d \to \mathbb{R}^d \times \mathbb{R}^d$ is given by:

$$\Phi(\boldsymbol{x}, \boldsymbol{x}') \triangleq \begin{pmatrix} |\boldsymbol{x} - \boldsymbol{x}'| \\ \frac{1}{2}(\boldsymbol{x} + \boldsymbol{x}') \end{pmatrix}. \tag{3.14}$$

The first part of $\Phi(\boldsymbol{x}, \boldsymbol{x}')$ encodes the relative position of the RSSI vectors and the second part their absolute position, as opposed to the implicit mapping of the Euclidean distance which only encodes relative information. In practice, the minimization of $ER_n(\mathsf{d})$ w.r.t. $f_1, \dots, f_T$ is intractable. An alternative is to use a Random Forest or an XGboost regressor, which separately optimizes the T regression trees. In practice, the learning stage is thus as follows:

- Compute the pairwise features $\Phi(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for all couples (i, j) in the dataset.

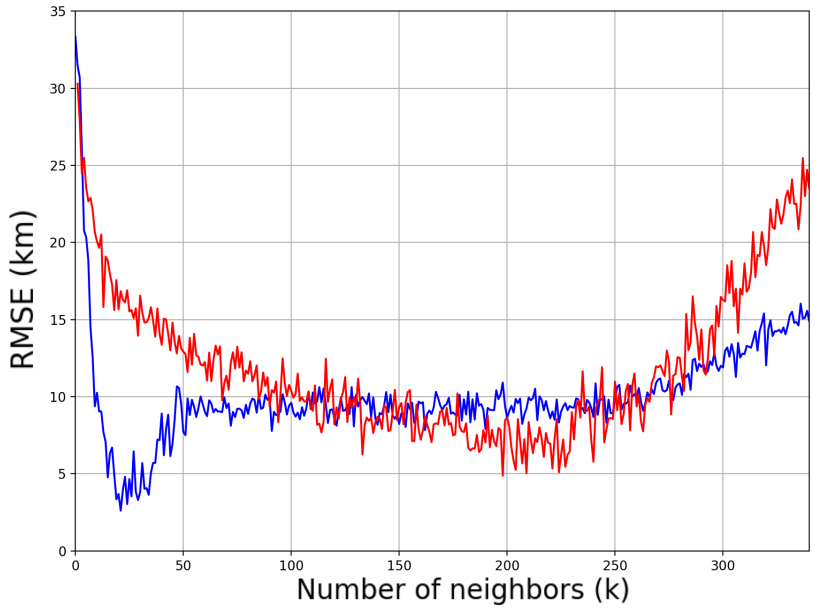


Figure 3.3 – Plots of RMSE (km) on Sigfox dataset vs. the number of neighbors k for the classic k-NN in (—), and the enhanced k-NN in (—). The RMSE is chosen as a criterion for the choice of k, as mentioned in Sec. 3.4.1

- Use a regression tree ensemble method to predict the labels $d_v(Y_i, Y_j)$ based on the features $\Phi(\boldsymbol{X}_i, \boldsymbol{X}_j)$.

Note that the obtained mapping d, though symmetric, is not mathematically speaking a metric. This point is however irrelevant regarding the application of interest. Given the obtained metric and given an observed RSSI vector $\boldsymbol{X}$, the k-NN estimate of Y is computed as in Equation (3.3).

3.4 Numerical experiments

3.4.1 Performance Analysis of Location Estimation Methods

Datasets

To compare the performance of the different methods, we have used two datasets.

UCI Dataset An indoor dataset from Zsolt Tóth (2016), composed of 1540 measurements performed in a three-story building. The measurements are made at the 32 base stations set in the building, and their positions are known. The measurements were recorded by the same kind of Android devices in order to reduce the effect of the variety of the hardware. The recording was preformed at weekend to reduce the noise of the environment.

Figure 3.4 – Comparisons of the presented methods in terms of their cdf of errors on the two datasets. left: Sigfox, right: UCI

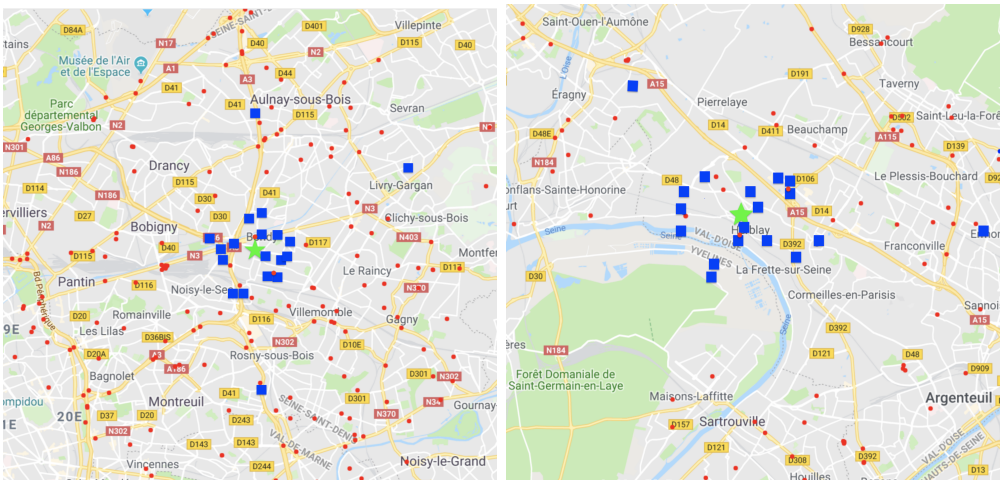


Figure 3.5 – In red (resp. in blue), some of the locations of the 200 (resp. 25) nearest neighbors according to the euclidean distance (resp. learned metric). In green stars, the true position of the emitter. The standard deviation of the blue dots around the true position is much lower the the standard deviation of the red dots. Red dots being scattered across a wide area, only a few of them are displayed in the figures.

Sigfox Dataset The second one, is a private dataset originated from the Sigfox network. It is composed of $n = 1.5 \cdot 10^6$ observations. These measurements are collected at the 164 base stations (see Figure 3.8).

These datasets are randomly split in a training set (90%) and a test set (10%). The training set was used to perform cross-validation (each fold containing 10% of the training set) in order to find the optimal parameters of our algorithms.

Results and discussions

To compute the errors we employ the Vincenty distance between estimated and actual location. Figure 3.4 shows the cumulative distribution function (cdf) of the prediction error for all the presented methods. Our proposed metric learning based k-NN turns out to outperform the other methods of the paper. It is particularly interesting to notice how the learned metric can improve a k-NN. Between these two methods, the optimal choice for the number of neighbours k is different. This parameter is chosen according a cross-validation (see Figure 3.3). Figure 3.3 shows the empirical risk of the Euclidean k-NN, and the proposed enhanced k-NN with respect to k. The main conclusion of this figure is the optimal value of k: 200 (resp. 25) neighbors minimize the empirical risk of the Euclidean k-NN (resp. the proposed k-NN). However, as k increases, the Euclidean k-NN inability to incorporate position information causes its performance to degrade. Our method, by comparison, shows fewer signs of degradation as k increases. In Figure 3.5 we display the locations of the k nearest neighbours for these two methods. As expected, the standard deviation of the k nearest neighbours around the true position is much lower for our proposed metric learning based k-NN than for the k-NN.

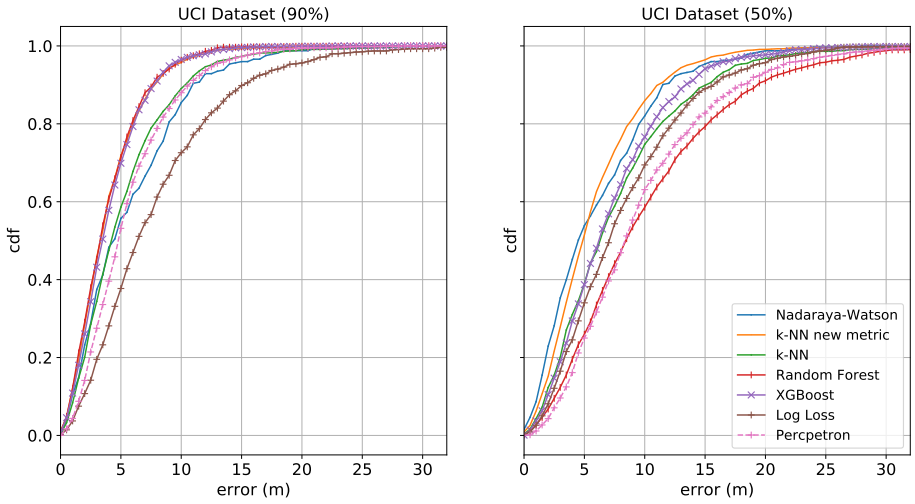


Figure 3.6 – The c.d.f. of errors on UCI dataset by the different presented methods assuming 90% (left) and 50% (right) of the database was available.

By contrast, the Log Loss model is neither relevant for the noisy urban dataset nor for the UCI indoor one. The main assumption made by this model is that the signal is propagating through free space. This assumption cannot be valid in both of these datasets. Hence, the poor performance of the log-loss model.

Finally, notice that the presented likelihood-based method's location errors are relatively stable when the number of training points decreases (see Figure 3.6). This point is emphasized in Yiu et al. (2017) where the impact of the density of the RSSI radio map is discussed.

Heat Map estimation

A major benefit of the Semi-Parametric Likelihood-Based method is that density level sets can be computed easily. Thanks to the statistical framework, this method is able to evaluate the probability density of $Y|\boldsymbol{X}$ at all $y \in \mathcal{Y}$ which can lead us to build density level sets in which Y is most likely to lie given the observation of $\boldsymbol{X}$. This is shown in Figure 3.7 where further information is provided on the uncertainty of the estimation.

3.4.2 PoI Prediction Case Study and Metric Learning Motivation

Geolocation is usually described as a regression problem when the goal is to estimate the location *(latitude, longitude)*. The advantage of this approach lies in its generality as it allows to deal with all situations. Still, in various business use cases and industrial applications, device's movements are constrained/limited and they can only move within a set of places. In IoT, logistic is a major application that fits well with this description: devices can only move within a set of *a priori* unknown specific locations such as warehouses, logistic hubs, etc.

By clustering the devices' locations in the original data set, the location diversity, using a positioning solution such as GPS or WIFI, reduces to a finite set of values. In this context, a Point of Interest (PoI) represents a cluster of *true measured* locations. A PoI must contain the spatial location of the cluster (usually defined as its center) but it can

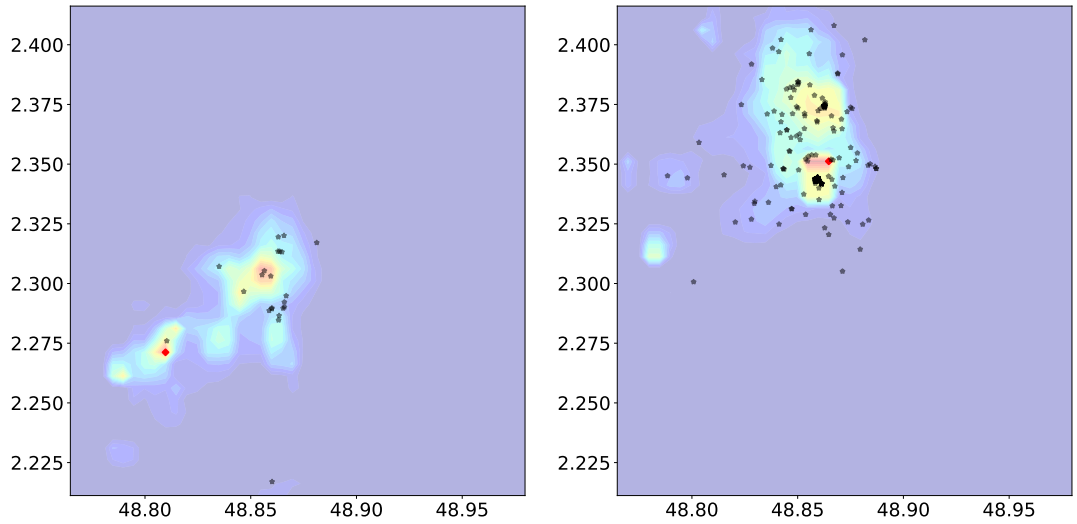


Figure 3.7 – Heat map of the position $Y|\boldsymbol{X} = \boldsymbol{x}$ for two different observations $\boldsymbol{x}$. The axis are given in latitude and longitude scale. The red dots represent the true positions. The black dots represent the positions corresponding for the same observations of $\boldsymbol{x}$ in the test set.

also be enriched with other features including the spatial extension of the cluster and the density of points within the cluster for example.

Note that a PoI's location can be known, partially known or unknown depending on the application and customer. In our pre-treatment, PoIs are supposed to be unknown and must be discovered. Their detection is a well studied problem that can be tackled by means of unsupervised machine learning techniques such as clustering methods. Density-based spatial clustering of applications with noise (DBSCAN (Birant and Kut, 2007)) has proven to be efficient for this task as PoIs are defined as areas with high device or message density.

Attention should be paid to the fact that the set of PoIs may vary over time. It is thus necessary to maintain an up-to-date list of such PoIs by performing clustering steadily and continuously.

This is illustrated in Figure 3.8 where some raw Sigfox devices locations are displayed in a map (a) and are then reduced to a few PoIs (b) after a clustering step.

Once the PoIs have been identified, the next step consists in predicting the unknown PoI belonging of a device, using a measured RSSI-vector. To achieve this goal, a data set made of fully supervised examples will be used: this data set is a collection of observed RSSI-vectors for various devices, each vector being mapped to one of the previously discovered PoIs. In this section only, a "good" metric (see Figure 3.10) is a mapping σ that minimizes the leave-one-out error:

$$R_n(\sigma) \triangleq \sum_{i=1}^{n} \sum_{i' \neq i} \ell\Big(\sigma(\Phi(\boldsymbol{x}_i, \boldsymbol{x}_{i'})), \mathbb{1}_{y_i = y_{i'}}\Big) \tag{3.15}$$

where ℓ is a certain loss function, typically the $0-1$ loss function $\ell(y, y') = \mathbb{1}_{y \neq y'}$ and $\Phi(x, x') \triangleq (|x - x'|, \frac{1}{2}(x + x'))^\top$. One may search for a mapping σ that minimizes within any relevant hypothesis class such as *logit* model, decision tree, ensemble methods,

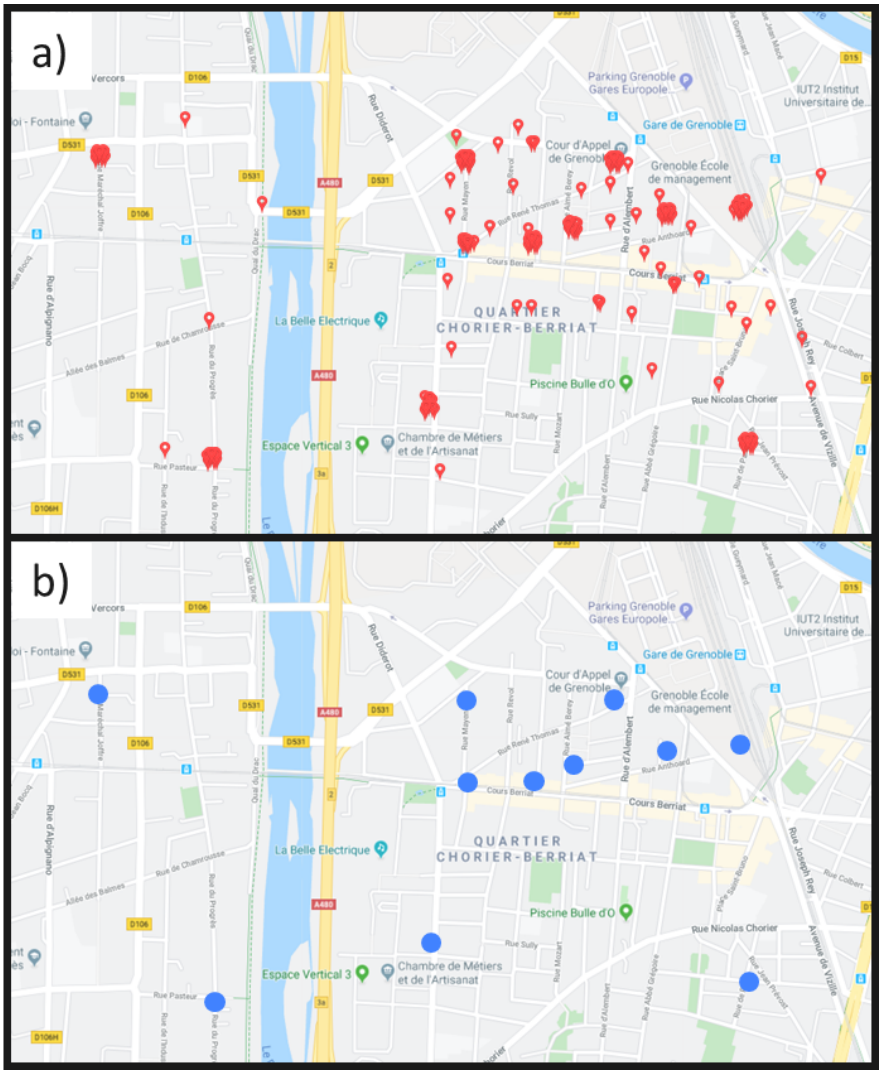


Figure 3.8 – a) Scatter plot of a sub-sample of the Sigfox training data set. A red bubble corresponds to an emitting position of a device in the Sigfox data set. b) PoIs obtained from the training data set displayed in a).

support vector machines (SVM). Note that the obtained mapping σ, though symmetric, is not mathematically speaking a metric (no triangle inequality). This point is however irrelevant regarding the application of interest. Given the learned $\hat{\sigma}_n$ and an observed vector $\boldsymbol{X}$, the *plug-in* predictor of the target Y is as following:

$$\hat{Y} = \arg \max_{j=1,\dots,J} \sum_{i \in \mathcal{I}_j} \hat{\sigma}_n\Big(\Phi(\boldsymbol{x}_i, \boldsymbol{X})\Big), \tag{3.16}$$

where we recall that $\mathcal{I}_j = \{i \leq n : y_i = j\}$.

Similarity performance analysis

In this section, different classifier models are compared for the choice of the similarity function: a logit model, an Euclidean distance based model, a random forest and a Gradient Boosting. The Euclidean distance based model computes a similarity value $1/(1 + ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2)$ for all couples of RSSIs vector $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in the data set. To analyse

their performance, a dataset composed of $n = 50000$ messages received on the Sigfox network was considered. Instead of forming the $n(n-1)$ possible couples as in Equation (3.15), we propose, for the sake of a lower computational cost, to build $n+m$ random couples of RSSIs vectors $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with their corresponding labels $\mathbb{1}_{y_i=y_j}$. In our experiment, we considered $n+m = 1000000$ couples of messages.

The obtained dataset was split into a train dataset of size n and a test dataset of size m. The training set is used to learn the similarity function $\hat{\sigma}_n$. On the other hand, the test set is employed to assess the methods' performance using their ROC curves as shown in Figure 3.9.

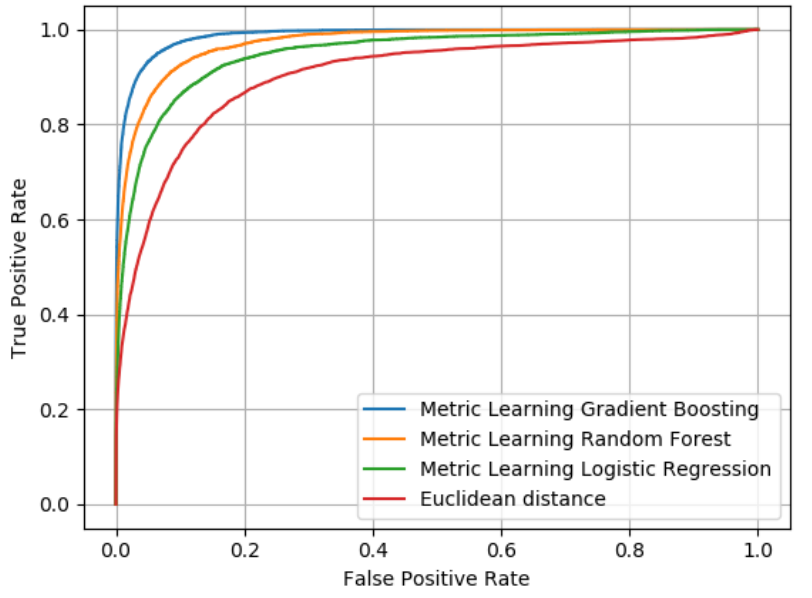


Figure 3.9 – ROC curves measured for different classification model based metric.

For our dataset, the gradient boosting model shows the best performance in predicting whether or not two RSSI vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ belong to the same PoI. In contrast, the Euclidean distance between two vectors of RSSI $\boldsymbol{x}$ and $\boldsymbol{x}'$ fails to correctly represent the underlying distribution of the data. Indeed, close $\boldsymbol{x}$ and $\boldsymbol{x}'$ (w.r.t. the Euclidean distance) does not correspond to the equality of labels. To illustrate this point, we computed the similarities' distributions for couples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ given whether or not $y_i = y_j$ in Figure 3.10. It shows that similarity provided by classical k-NN when $\boldsymbol{x}$ and $\boldsymbol{x}'$ share the same class (target = 1) are very likely to be null in comparison to the ones given by the gradient boosting model. Note that this figure also highlights that all classifiers are much more confident when predicting target 0 than target 1. This result was expected as it is fairly easy for any method to predict that many couples $\boldsymbol{x}$ and $\boldsymbol{x}'$ do not belong to the same POI since their respective RSSI signatures are very different (such as couples of RSSIs with no common received BSs).

An other interesting metric to compare the methods is the lift (Figure 3.11). The lift is a measure of the performance of a classifier as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. A targeting model is good if the response within the target is much better than the average for the population as a whole. The lift is simply the ratio of these values: target response divided by average response.

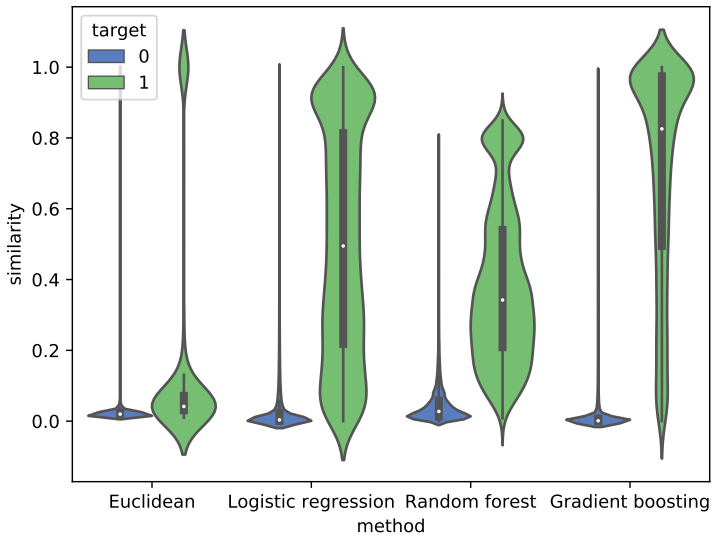


Figure 3.10 – Distribution of the similarity values predicted by the different classification models. In blue (resp. in green), all the couples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ such that $y_i \neq y_j$ (resp. $y_i = y_j$) are displayed.

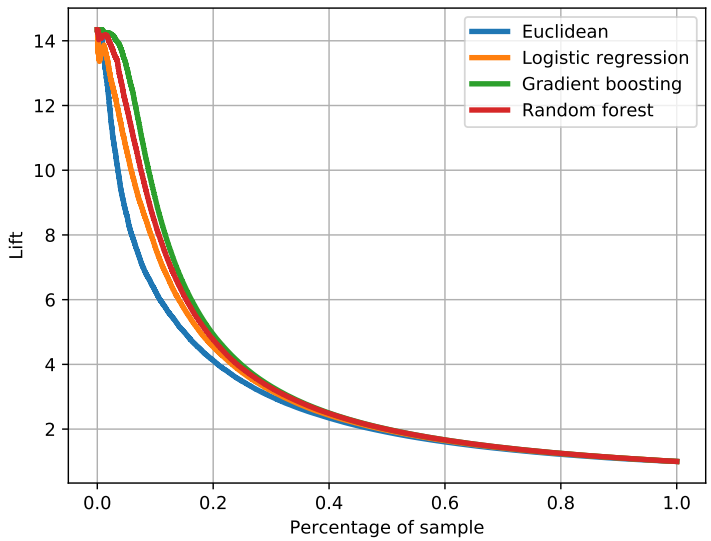


Figure 3.11 – Lift curves for the methods in scrutiny.

Figure 3.11 shows for instance that the 10% highest similarities provided by the Gradient Boosting classifier correspond to couples that are 9.2 times more likely to share the same PoI than an average couple. This experiment shows that the Gradient Boosting based similarity outperforms the other methods in competition.

In the sequel, the similarity function is thus learned as a sum of T classification trees through Gradient Boosting algorithm.

Method	Euclidean k-NN	Gradient boosting k-NN	Random Forest
Accuracy	0.877	**0.907**	0.899

Table 3.2 – Accuracy of POI classifiers computed as the Exact Match Ratio (see Equation (3.17)).

PoI prediction performance analysis

In this section, we focus on the original task that consists in predicting the device's PoI. To do so, we consider three different approaches: a classical k-NN (with Euclidean distance), a metric learning based k-NN (with Gradient Boosting based similarity) and a random forest classifier that provides state of the art performance on large feature space classification problems. To compare the performance of the methods in competition to predict the POI affiliation of messages, we have at our disposal the Sigfox dataset which is split in two, a training set $\mathcal{Z}_n$, and a testing set $\mathcal{Z}_m$. Note that to avoid overfitting, it is important that $\mathcal{Z}_n$, and $\mathcal{Z}_m$ are built using different devices: some devices are used to learn and some other devices are used to predict.

For this problem, the training set was used to either learn a similarity function $\hat{\sigma}_n$ (as presented above) or directly learn the classification model (random forest in our case). On the other hand, the test subset $\mathcal{Z}_m$ was employed to evaluate the algorithms' performance.

First, the Exact Match Ratio of the predicted PoI (see Equation (3.16)), that can be written as follows

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[y_i = \hat{Y}(\boldsymbol{x}_i)] \tag{3.17}$$

is considered.

The results for the different methods are summed up in Table 3.2. This result shows that the gradient boosting metric learning method got the highest classification score.

In addition, the Vincenty distance between the predicted POI and the "true" position (given by a GPS or Wifi apparatus) was also computed. In Figure 3.12, the cumulative distribution function of the spatial error is compared for the different predictive methods.

This result also demonstrates that our proposed local non linear metric learning method outperforms the other considered methods on our data set.

Plotting the cumulative distribution also allows for a deeper interpretation of the observed performance. Indeed, the curves' shapes can be divided in three parts as follows:

- A steep rise up to $150 - 200$ meters that correspond to messages that can be correctly mapped to PoIs, which can also be interpreted as a measure the expressiveness of the algorithms. Among them, messages that originate from isolated PoIs having specific RSSIs' signatures are common to the different classifiers.
- A plateau up to 700 meters followed by a small jump for all the methods. This jump mostly corresponds to errors when PoIs are extremely dense (distance between PoIs lower than 1 km). The presence of this jump show the limitation of the tested

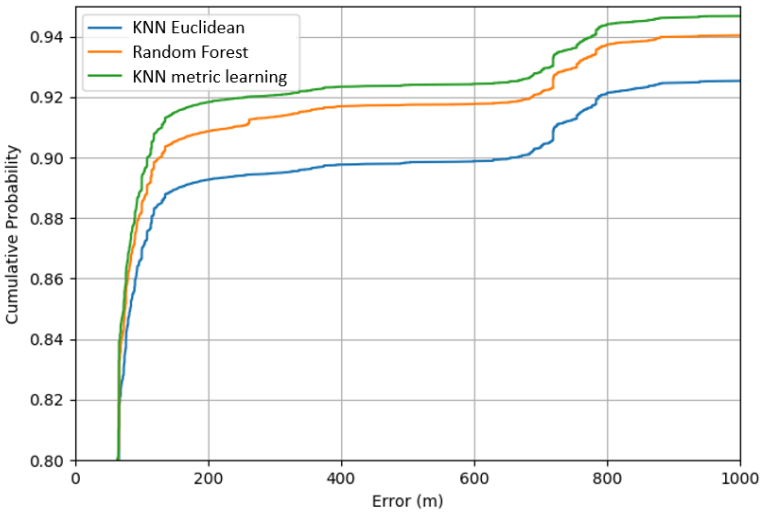


Figure 3.12 – PoI prediction performance: cumulative distribution function of the error measured as the distance between the predicted PoI location and the "true" device location given by GPS or Wifi.

algorithms.

- A slow convergence before reaching a cumulative probability of 1 after several kilometers.

This analysis shows that a gradient boosting metric learning coupled with a k-NN model enables to outperform a classical Euclidean metric k-NN as well as state of the art classification predictors.

Conclusion

In this chapter, we investigated machine learning approaches addressing the problem of geolocation. We presented most popular methods found in the literature. We proposed two new techniques: one based on a likelihood model of the RSSI vector $\boldsymbol{X}$ given Y and an other based on a enhanced k-NN regressor. To compare these methods, 1.5M observations were collected from the Sigfox network, as well as 1540 measurements from the UCI dataset. Results have shown that our metric learning method has the highest accuracy on both of the datasets, with a 90th-percentile of errors of 5 km on Sigfox dataset, and of 7 m on UCI dataset. As for the semi-parametric method, in addition to being very easy to fit, it goes beyond the simple estimation by providing heat maps and level sets, making it a suitable method for industrial applications. Moreover, experiences have shown that the two proposed methods are the most accurate when we provide a reduced training set. They are robust when only a small training set is available. Furthermore, results show that the k-NN with the learned metric reaches the highest accuracy in the particular context of localization within PoIs, which therefore demonstrates the interest of the proposed method. Metric learning methods therefore deserves a great deal of attention. An interesting topic would be to build the metric by directly minimizing the regression error of the estimator. This would differ from the present chapter in which we provide a supervised objective for the learned metric.

4

Extreme Gradient Boosting for Similarity Learning

"Si, marchant dans une forêt aléatoire, tu rencontres deux fois le même arbre, c'est que tu es perdu."

Most of the fingerprint methods presented in Chapter 3 assume that positions corresponding to similar fingerprints are close to each others. By comparing the fingerprints through a distance (or similarity) function, the estimate is thus defined as the position of the nearest fingerprint found in the database. The quality of the prediction is then highly related to the chosen distance on the RSSI space. The metric learning problem is therefore a fundamental issue to improve RSSI-based geolocation techniques. This chapter addresses this problem. However, the results are general and go beyond the framework of the geolocation. The metric learning problem has been shown to improve regression methods that rely on distances or similarities. There is a surge of interest for optimizing distance and similarity functions. However, in most prior works, no link is made between the learned similarity and the estimator performance. In this chapter, we propose to build the similarity by directly minimizing the regression error of an estimator, and thus obtain an ad-hoc learning objective. To minimize this objective, we propose a modified version of the eXtreme Gradient Boosting algorithm (XGBoost) presented in Section 2.3. Experiments show that our model outperforms other kernel regression models on several benchmark datasets including one that proposes to locate emitting devices given the RSSI.

4.1 Introduction

The need for an appropriate way to measure distance or *similarity* between data points is ubiquitous in machine learning. The problem of learning such similarity functions is difficult and has attracted a lot of interest for the past twenty years, in particular with the seminal work of Xing et al. (2003). This led to the emergence of the metric or *similarity* learning field, that aims to automatically learn a function from the data to assess the similarity or the distance between pairs. As stated in Kulis et al. (2012), this practice has besides proven useful when used in conjunction with nearest-neighbors methods or with any other regression method that relies on a distance such as the ones presented in Section 2.2.2. This effect is all the more noteworthy as the dimension of the data increases, as shown for instance in the work of Chopra et al. (2005) that demonstrates the benefits of metric learning for the task of face recognition. In addition, the quality of the geolocation is highly related to the chosen distance on the RSSI space. The metric learning problem is therefore a fundamental issue to improve RSSI-based geolocation techniques.

Nevertheless, in most previous works, the objective used to learn a similarity is generally not linked to the accuracy of the final predictor. In this chapter, we address this problem by proposing a general method to learn the similarity by directly minimizing the regression error of the final predictor. Consequently, the regression error is optimized with respect to the similarity function. This approach then requires the minimization of a particular learning objective. In previous works, such as Weinberger and Tesauro (2007); Keller et al. (2006), the choice of similarities is restricted to Mahalanobis distance metrics, which can be represented by symmetric positive semi-definite matrices. This parametrization allows a simple minimization stage by means of gradient based optimizers such as delta-bar-delta (Jacobs, 1988).

In contrast, we propose here to learn a non-parametric similarity. More specifically, the similarity is chosen as a sum of regression trees and is sequentially learned by means of a modified version of XGBoost. This approach is shown to be well adapted to minimize the objective function of interest and furthermore benefits from the well-known qualities of XGBoost. The advantages of our methods are many:

- XGBoost is very efficient and scales well. It has been shown to be one of the most powerful and scalable supervised method for handling high-dimensional data.
- In contrast to local metric methods discussed subsequently, the storage requirement of our method is independent of the size of the input data.
- Our learned metric is non-parametric. This means that it can handle specific problem such as the XOR example discussed below.

As far as we know, learning a similarity that is used to improve regression by means of tree boosting methods is new.

Related literature. Methods that learn a similarity have been widely studied in recent years e.g. in Kar and Jain (2011); Schultz and Joachims (2004); Nguyen and Guo (2008); Bellet et al. (2013) and are subsequently discussed. Essentially, similarity learning aims at finding the parameters of a similarity function given pair-based constraints: typically "example $\boldsymbol{x}$ should be similar to example $\boldsymbol{x}'$". As stated in Kulis et al. (2012), existing metric learning methods can be divided into two categories: linear and nonlinear. A surge of recent research has focused on learning a Mahanalobis distance defined by $d^2_{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{x}') \triangleq (\boldsymbol{x} - \boldsymbol{x}')^\top \mathbf{M}(\boldsymbol{x} - \boldsymbol{x}')$, with the Positive Semi-Definite (PSD) matrix $\mathbf{M}$ as parameter. This (pseudo)-distance corresponds to computing the Euclidean distance after a linear transformation of the data. These methods differ by the choice of the objective or by the regularizer used to learn $\mathbf{M}$. We can cite as instance Davis et al. (2007) in which a *Burg divergence* is used as regularization for $\mathbf{M}$; in Xing et al. (2003), the learning objective is to minimize the distance between points sharing the same class subject to the constraint that dissimilar points are separated (*i.e* the distance between those points is maximized). However, learning a linear metric such as a Mahanalobis distance can be too restrictive due to the following problems: first, because distance properties such as the triangle inequality must be violated (see Xiong et al. (2012)). Second, because they demonstrate their limitation in simple case such as the XOR example as in (Kulis et al., 2012). Kernel methods are used to extend linear methods to the nonlinear case, and this is achieved by writing algorithms in terms of inner products and replacing these inner products with kernel functions. The resulting distance is then a "kernelized" distance between pairs of examples. Numerous works

employ such a kernelization of linear methods, in particular in computer vision. For example, the pyramid match kernel (Grauman and Darrell, 2007), or face identification (Guillaumin et al., 2009). Although some algorithms have been shown to be kernelizable, a new formulation of the problem must generally be derived, and sometimes leads to insurmountable implementation problems Kulis et al. (2012).

The work of Xiong et al. (2012) applies Random Forests to learn a non-parametric metric called Random Forest Distance (RFD) in order to improve a classifier. This work emphasizes the benefits of using a nonlinear similarity in the classification task pursued. Kedem et al. (2012) proposes the GB-LMNN algorithm that applies gradient-boosting to learn non-linear mappings directly in function space. Neural network-based approaches offer the flexibility of learning arbitrarily complex nonlinear mappings (Chopra et al., 2005). However, they often demand high computational expense, not only in parameter fitting but also in model selection and hyper-parameter tuning. Recently, metric learning has also shown its benefits in feature selection (Navot et al., 2006) or to improve classification performance (Goldberger et al., 2005). In this chapter, we focus on the problem of learning a task-specific similarity that improves regression.

Learning a metric that is later used in regression has been introduced in Goldberger et al. (2005) (Neighborhood Component Analysis (NCA)). In this work, authors propose to learn a global Mahanalobis distance measure to be used in a k-NN classifier by optimizing the expected classification leave-one-out error. The main innovation of this work is to replace the actual leave-one-out classification error, a discontinuous function of the matrix parameter $\mathbf{M}$ by a differentiable cost function based on stochastic ("soft") neighbour assignments in the transformed space. This approach has been extended for regression in Weinberger and Tesauro (2007). In this work, in addition to performing regression, the authors show that their algorithm can also be viewed as an algorithm for dimensionality reduction. Note also Noh et al. (2017) that analyzed the effect of the metric on the asymptotic bias of the kernel estimator, and proposed a Mahanalobis distance to alleviate the bias of the estimator. The main differences between the above-mentioned articles is that our approach makes use of a non-linear similarity instead of learning a Mahalanobis matrix that is used in kernel regression. The work of Xiong et al. (2012) proposes to learn a non-parametric metric called Random Forest Distance (RFD) in order to improve a classifier. The idea of using tree-based methods to build a similarity function is a central element of our contribution.

Outline. The chapter is organized as follows: in Section 4.2 we present our learning formulation for kernel regression. Then, we derive our proposed XGBoost-inspired algorithm to learn the similarity. We show that despite a complex learning objective the tree gradient boosting offers a suitable framework to minimize it. Finally, in Section 4.4 we compare our estimator to state-of-the-art regression models on both synthetic and real datasets.

4.2 Tree Boosting for Metric Learning

In this section, we propose to extend the XGBoost algorithm, detailed at length in Section 2.3 for a regression purpose, to the problem of metric learning. First of all, we provide the similarity learning objective corresponding to the unsupervised setting introduced in the introduction. Given the objective, we propose to chose the similarity function as a sum of regression trees. Each tree is then learned sequentially. This new objective induces drastic changes in the boosting algorithm that are presented in this

section.

4.2.1 Problem Setting

Let $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \mathbb{R}^p$ be a random vector. We consider $\mathcal{Z}_N = \{\boldsymbol{x}_j, y_j\}_{j=1}^N$, $N = m + n$ i.i.d. realizations of $(\boldsymbol{X}, Y)$. By splitting the dataset $\mathcal{Z}_N$ in two, we obtain $\mathcal{Z}_n = \{\boldsymbol{x}_j, y_j\}_{j=1}^n$ to learn the similarity and $\overline{\mathcal{Z}}_m = \{\bar{\boldsymbol{x}}_j, \bar{y}_j\}_{j=n+1}^N$ to learn the regressor. Let $\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a mapping that we suppose symmetric and non-negative. We consider the following estimator of $m(\boldsymbol{x}) = \mathbb{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right]$:

$$\hat{m}_\sigma(\boldsymbol{x}) \triangleq \sum_{j=1}^{m} \sigma(\bar{\boldsymbol{x}}_j, \boldsymbol{x})\bar{y}_j, \tag{4.1}$$

The objective we minimize with respect to σ is the following regularized regression error of the predictor $\hat{m}_\sigma$ on $\mathcal{Z}_n$:

$$\mathcal{L}(\sigma) = \sum_{i=1}^{n} \ell\left(y_i, \hat{m}_\sigma(\boldsymbol{x}_i)\right) + \Omega(\sigma), \tag{4.2}$$

where ℓ is a regression loss function and Ω a regularizer on σ to be defined later on. This approach is original since it is driven by the minimization of a regressor error.

4.2.2 Learning the similarity with XGBoost

We propose to learn σ as a sum of T regression trees f_s, $s = 1, \ldots, T$. As the similarity takes a couple $(\boldsymbol{x}, \boldsymbol{x}')$ as input, it is therefore necessary to make adjustments in the various functions involved in the definition of σ:

- Let Φ be a feature transform of a couple $(\boldsymbol{x}, \boldsymbol{x}')$ (see Section 3.4 for a practical example).
- Let f_s be a regression tree, which can be viewed as a tuple $(q, \boldsymbol{\omega})$ where $q : \mathbb{R}^d \times \mathbb{R}^d \to \{1, \ldots, L\}$ is the structure of the tree that maps an input to a leaf index, $\boldsymbol{\omega} \in \mathbb{R}^L$ is the vector containing the leaves weights and L is the number of leaves. For a given tuple $(\boldsymbol{x}, \boldsymbol{x}')$, we first map it into a feature space via Φ that is the input of f_s. Then, we use the decision rule given by q to map this input into a leaf and define the output of the tree as the weight.
- The output $\sigma(\boldsymbol{x}, \boldsymbol{x}')$ is finally the sum of the T weights. This procedure is illustrated in Figure 4.1.

We propose, following the idea of XGBoost, to build $\hat{\sigma}$ sequentially and we denote by $\hat{\sigma}^{(t)}$ the similarity at the t^{th} iteration. We have for any pair $(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{R}^d \times \mathbb{R}^d$:

$$\hat{\sigma}^{(t)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{s=1}^{t} f_s\left(\Phi(\boldsymbol{x}, \boldsymbol{x}')\right).$$

Consequently, the predictor of $m(\boldsymbol{x})$ at the t^{th} iteration becomes:

$$\hat{m}_{\hat{\sigma}^{(t)}}(\boldsymbol{x}) \triangleq \sum_{j=1}^{m} \hat{\sigma}^{(t)}(\bar{\boldsymbol{x}}_j, \boldsymbol{x})\bar{y}_j. \tag{4.3}$$

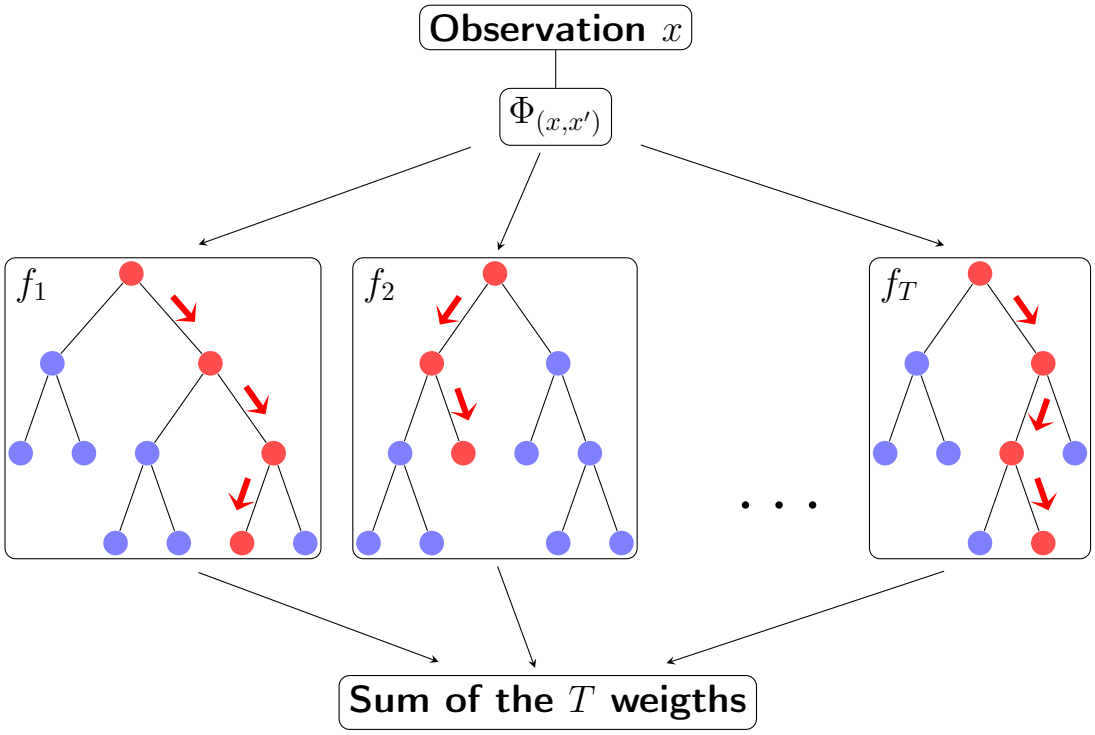


Figure 4.1 – Illustration of the similarity prediction $\sigma^{(T)}(x, x')$.

To lighten notation, we will simply write $\hat{m}^{(t)}$ instead of $\hat{m}_{\hat{\sigma}^{(t)}}$ so that:

$$\hat{m}^{(t)}(\boldsymbol{x}) = \hat{m}^{(t-1)}(\boldsymbol{x}) + \sum_{j=1}^{m} f_t\left(\Phi(\bar{\boldsymbol{x}}_j, \boldsymbol{x})\right) \bar{y}_j. \tag{4.4}$$

The function f_t added at the iteration t is defined as the minimizer of the following objective:

$$\mathcal{L}^{(t)}(f) = \sum_{i=1}^{n} \ell\left(y_i, \hat{m}^{(t-1)}(\boldsymbol{x}_i) + \boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{y}}\right) + \Omega\left(f\right), \tag{4.5}$$

where $\boldsymbol{f}(\boldsymbol{x}) = \left(f\left(\Phi(\bar{\boldsymbol{x}}_1, \boldsymbol{x})\right), \ldots, f\left(\Phi(\bar{\boldsymbol{x}}_m, \boldsymbol{x})\right)\right)^\top$, $\bar{\boldsymbol{y}} = \left(\bar{y}_1, \ldots, \bar{y}_m\right)^\top$, ℓ is now a differentiable regression loss function defined on $\mathbb{R}^p \times \mathbb{R}^p$; and for a tree f with L leaves and weights $\boldsymbol{\omega}$, $\Omega(f) = \frac{1}{2}\lambda\|\boldsymbol{\omega}\|_2^2 + \gamma L$.

Using a second-order approximation of $\ell(y_i, \cdot)$ at the point $\hat{m}^{(t-1)}(\boldsymbol{x}_i)$ it follows that:

$$\mathcal{L}^{(t)}(f) \simeq \sum_{i=1}^{n} \ell\left(y_i, \hat{m}^{(t-1)}(\boldsymbol{x}_i)\right) + g_i^\top \boldsymbol{f}(\mathbf{x}_i)^\top \bar{\boldsymbol{y}} + \frac{1}{2}\bar{\boldsymbol{y}}^\top \boldsymbol{f}(\boldsymbol{x}_i) H\left(\ell(y_i, \cdot)\right) \boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{y}} + \Omega\left(f\right),$$

where $g_i \triangleq \nabla_{\hat{m}^{(t-1)}} \ell(y_i, \cdot) \in \mathbb{R}^p$ and $H\left(\ell(y_i, \cdot)\right) \in \mathbb{R}^{p \times p}$ are respectively the Jacobian and the Hessian of the function $\ell\left(y_i, \cdot\right)$ at $\hat{m}^{(t-1)}(\boldsymbol{x}_i)$. The term $\ell\left(y_i, \hat{m}^{(t-1)}(\boldsymbol{x}_i)\right)$ being a constant of f, we can remove it and define a simplified version of the objective as follows:

$$\tilde{\mathcal{L}}^{(t)}(f) = \sum_{i=1}^{n} g_i^\top \boldsymbol{f}(\mathbf{x}_i)^\top \bar{\boldsymbol{y}} + \frac{1}{2}\bar{\boldsymbol{y}}^\top \boldsymbol{f}(\boldsymbol{x}_i) H\left(\ell(y_i, \cdot)\right) \boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{y}} + \Omega\left(f\right), \tag{4.6}$$

Proposition 4.1. *Denoting as in Section 2.3 $(q, \boldsymbol{\omega})$ the tuple such that $f_t(\bar{\boldsymbol{x}}, \boldsymbol{x}) = \omega_{q(\Phi(\bar{\boldsymbol{x}}, \boldsymbol{x}))}$. We suppose furthermore that the current number of leaves in the structure q is L. First denoting by*

$$\mathcal{I}_l \triangleq \left\{ i \leq n, j \leq m : q\left(\Phi(\bar{\boldsymbol{x}}_j, \boldsymbol{x}_i)\right) = l \right\}, \tag{4.7}$$

the pairs of indices for which the couple $(\bar{\boldsymbol{x}}_j, \boldsymbol{x}_i)$ falls into the leaf l and

$$\mathcal{J}_{l,l'} \triangleq \left\{ i \leq n, j \leq m, k \leq m \ : q\left(\Phi(\bar{\boldsymbol{x}}_j, \boldsymbol{x}_i)\right) = l \text{ and } q\left(\Phi(\bar{\boldsymbol{x}}_k, \boldsymbol{x}_i)\right) = l' \right\} \tag{4.8}$$

the triplets (i, j, k) for which the pairs $(\bar{\boldsymbol{x}}_j, \boldsymbol{x}_i)$ and $(\bar{\boldsymbol{x}}_k, \boldsymbol{x}_i)$ fall respectively into the leaves l and l'. Defining accordingly

$$\mathbf{G} = \left(\sum_{\mathcal{I}_1} g_i^\top \bar{y}_j, \ldots, \sum_{\mathcal{I}_L} g_i^\top \bar{y}_j \right)^\top \in \mathbb{R}^L \text{and} \tag{4.9}$$

$$\mathbf{H} = \begin{pmatrix} \sum_{\mathcal{J}_{1,1}} \bar{y}_j^\top H_i \bar{y}_k & \cdots & \sum_{\mathcal{J}_{1,L}} \bar{y}_j^\top H_i \bar{y}_k \\ \vdots & \ddots & \vdots \\ \sum_{\mathcal{J}_{L,1}} \bar{y}_j^\top H_i \bar{y}_k & \cdots & \sum_{\mathcal{J}_{L,L}} \bar{y}_j^\top H_i \bar{y}_k \end{pmatrix} \in \mathbb{R}^{L \times L}. \tag{4.10}$$

Then, we have the following:

$$\tilde{\mathcal{L}}^{(t)}(q, \boldsymbol{\omega}) = \boldsymbol{\omega}^\top \mathbf{G} + \frac{1}{2} \boldsymbol{\omega}^\top \left(\mathbf{H} + \lambda \mathbf{I}_L \right) \boldsymbol{\omega} + \gamma L. \tag{4.11}$$

Proof Let us re-index the sum in Equation (4.6), such that the objective can easily be minimized with respect to the weights of the leaves $\boldsymbol{\omega}$ of the tree f.

$$\tilde{\mathcal{L}}^{(t)}(f) = \sum_{l=1}^{L} \sum_{\mathcal{I}_l} g_i^\top \boldsymbol{f}(\mathbf{x}_i)^\top \bar{y}_j + \frac{1}{2} \sum_{l'=1}^{L} \sum_{\mathcal{J}_{l,l'}} \bar{y}_j^\top \boldsymbol{f}(\boldsymbol{x}_i) H_i \boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{y}_k + \Omega\left(f\right) \tag{4.12}$$

$$= \sum_{l=1}^{L} \left(\sum_{\mathcal{I}_l} g_i^\top \omega_l \bar{y}_j + \frac{1}{2} \sum_{l'=1}^{L} \sum_{\mathcal{J}_{l,l'}} \omega_l \omega_{l'} \bar{y}_j^\top H_i \bar{y}_k \right) + \frac{1}{2} \lambda \|\boldsymbol{\omega}\|_2^2 + \gamma L. \tag{4.13}$$

Finally, with the definitions of Equation (4.9), and Equation (4.10), $\tilde{\mathcal{L}}^{(t)}$ can then be rewritten in matrix form:

$$\tilde{\mathcal{L}}^{(t)}(q, \boldsymbol{\omega}) = \boldsymbol{\omega}^\top \mathbf{G} + \frac{1}{2} \boldsymbol{\omega}^\top \left(\mathbf{H} + \lambda \mathbf{I}_L \right) \boldsymbol{\omega} + \gamma L, \tag{4.14}$$

. ■

The next lemma states that the so defined matrix $\mathbf{H}$ is positive semi-definite (PSD).

Lemma 4.2. *The matrix $\mathbf{H}_\lambda \triangleq \left(\mathbf{H} + \lambda \mathbf{I}_L \right)$ is positive definite for any $\lambda > 0$. As a consequence, for a fixed structure q, the optimal values of the leaves weights $\boldsymbol{\omega}^\star$ are thus given by:*

$$\boldsymbol{\omega}^\star = -\left(\mathbf{H} + \lambda \mathbf{I}_L \right)^{-1} \mathbf{G}. \tag{4.15}$$

Proof

First, let us recall the existing relations between the sets that are considered. We assume that we are at the leaf and that the tree that is build contains $L-1$ leaves. This current leaf is denoted "parent" in the sequel. By definition, $\mathcal{I}_{\text{parent}}$ is the set of pairs $(i,j) \in \{1,\dots,n\}\times\{1,\dots,m\}$ such that the couple (x_i, x_j) falls into the "parent" leaf. We also have $\mathcal{J}_{\text{parent, other}}$ that is the set of triplets $(i,j,k) \in \{1,\dots,n\}\times\{1,\dots,m\}\times\{1,\dots,m\}$ such that $(i,j) \in \mathcal{I}_{\text{parent}}$ and $(i,k) \in \mathcal{I}_{\text{other}}$. We suppose from now on that "parent" is split into "left child" and "right child". We therefore have the following relations:

$$\mathcal{I}_{\text{parent}} = \mathcal{I}_{\text{left child}} \sqcup \mathcal{I}_{\text{right child}}, \tag{4.16}$$

$$\mathcal{J}_{\text{parent,other}} = \mathcal{J}_{\text{left child,other}} \sqcup \mathcal{J}_{\text{right child,other}} \qquad \text{if other} \neq \text{parent} \tag{4.17}$$

$$\mathcal{J}_{\text{parent,parent}} = \mathcal{J}_{\text{left child,left child}} \sqcup \mathcal{J}_{\text{right child,right child}} \sqcup \mathcal{J}_{\text{left child,right child}} \sqcup \mathcal{J}_{\text{right child,left child}} \qquad \text{o.w,} \tag{4.18}$$

where the symbol $\sqcup$ stands for the disjoint union.

Let us denote by $\mathsf{Y}^{(l)} \in \mathbb{R}^{n\times m\times m}$ the vector defined as follows:

$$\mathsf{Y}^{(l)}[i,j,k] = \begin{cases} \bar{y}_j & \text{if } (i,j) \in \mathcal{I}_l \\ \bar{y}_k & \text{if } (i,k) \in \mathcal{I}_l \\ \sqrt{\bar{y}_j \bar{y}_k} & \text{if } (i,j) \in \mathcal{I}_l \text{ and } (i,k) \in \mathcal{I}_l \\ 0 & \text{o.w.} \end{cases} \tag{4.19}$$

It follows that for any pairs of leaves indices (l, l') we have:

$$H_{l,l'} = \sum_{i\leq n}\sum_{j\leq m}\sum_{k\leq m} \mathsf{Y}^{(l)}[i,j,k]\mathsf{Y}^{(l')}[i,k,j] = \langle \mathsf{Y}^{(l)}, \mathsf{Y}^{(l')} \rangle. \tag{4.20}$$

The matrix in scrutiny is then a Gram matrix. It is therefore PSD. For any $\lambda > 0$, the matrix H_λ is positive-definite. Let us note that the fact that the matrix Y is defined in terms of $\sqrt{\bar{y}_j \bar{y}_k}$ requires that the $\bar{y}_j \geq 0$ for any $j \leq m$, and $\bar{y}_k \geq 0$ for any $k \leq m$. ■

It follows from Proposition 4.1 and Lemma 4.2 that the optimal value of the objective $\tilde{\mathcal{L}}^{(t)}$ for a given structure q has a simple expression that is provided in the following proposition.

Proposition 4.3. *For a fixed structure q (L is constant), the optimal value of the objective function is given as follows:*

$$\tilde{\mathcal{L}}^{(t)}(q, \boldsymbol{\omega}^\star) = -\frac{1}{2}\mathbf{G}^\top \left(\mathbf{H} + \lambda \mathbf{I}_L\right)^{-1} \mathbf{G} + \gamma L. \tag{4.21}$$

Proof The optimal value of the objective is obtained by evaluative the value of $\tilde{\mathcal{L}}^{(t)}(q,\cdot)$ at point $\boldsymbol{\omega}^\star$ as defined in Equation (4.15). We plug (4.15) into (4.14) to obtain :

$$\tilde{\mathcal{L}}^{(t)}(q, \boldsymbol{\omega}^\star) = \boldsymbol{\omega}^{\star\top}\mathbf{G} + \frac{1}{2}\boldsymbol{\omega}^{\star\top}\left(\mathbf{H} + \lambda\mathbf{I}_L\right)\boldsymbol{\omega}^\star + \gamma L \tag{4.22}$$

$$= -\mathbf{G}^\top\left(\mathbf{H} + \lambda\mathbf{I}_L\right)^{-1}\mathbf{G} + \tag{4.23}$$

$$\frac{1}{2}(\left(\mathbf{H} + \lambda\mathbf{I}_L\right)^{-1}\mathbf{G})^\top \underbrace{\left(\mathbf{H} + \lambda\mathbf{I}_L\right)\left(\mathbf{H} + \lambda\mathbf{I}_L\right)^{-1}}_{\mathbf{I}_L}\mathbf{G} + \gamma L \tag{4.24}$$

$$= -\frac{1}{2}\mathbf{G}^\top\left(\mathbf{H} + \lambda\mathbf{I}_L\right)^{-1}\mathbf{G} + \gamma L. \tag{4.25}$$

■

This optimal value of the objective is used to measure the quality of the structure q. The procedure is the same employed in Section 2.3: we start from the root, and iteratively add branches to the tree by splitting the nodes until the maximal number of leaves is reached.

Remark Attention should be paid to the fact that, in contrast with the XGBoost algorithm presented in Section 2.3 the leaves weights $(\omega_l)_{l=1,\dots,L}$ can not be computed separately. The value of a single ω_l will influence the value of the other weights because of the correlation terms $\omega_l\omega_{l'}\bar{y}_j^\top H_i\bar{y}_k$. As a consequence, the computation can not be parallelized over nodes. The algorithm of the modified XGBoost is given in Section 4.3.

L_1 regularization on $\boldsymbol{\omega}$ Consider the case when we add a L_1 regularization, that is $\Omega(f) = \lambda\|\boldsymbol{\omega}\|_2^2 + \mu\|\boldsymbol{\omega}\|_1 + \gamma|L|$. This kind of regularization promotes sparsity so that most of the similarities between pairs of observations end up with a null weight. In this case, Equation (4.14) becomes:

$$\tilde{\mathcal{L}}^{(t)}(q, \boldsymbol{\omega}) = \boldsymbol{\omega}^\top\mathbf{G} + \frac{1}{2}\boldsymbol{\omega}^\top\left(\mathbf{H} + \lambda\mathbf{I}_L\right)\boldsymbol{\omega} + \mu\|\boldsymbol{\omega}\|_1 + \gamma L. \tag{4.26}$$

Denoting by $\tilde{X} \in \mathbb{R}^{L\times L}$ the unique PSD matrix verifying $\tilde{X}^\top\tilde{X} = \left(\mathbf{H} + \lambda\mathbf{I}_L\right)$ and $\tilde{y} \in \mathbb{R}^L$ a solution vector to $\tilde{X}^\top\tilde{y} = \mathbf{G}$, a minimizer of Equation (4.26) is also a[1] solution to:

$$\boldsymbol{\omega}^\star \in \arg\min_{\mathbb{R}^L} \underbrace{\frac{1}{2}\|\tilde{y} - \tilde{X}\boldsymbol{\omega}\|_2^2 + \mu\|\boldsymbol{\omega}\|_1}_{\mathcal{P}(\boldsymbol{\omega})}. \tag{4.27}$$

The optimization can be directly performed using solvers such as Celer (Massias et al., 2018).

4.3 Implementation of the Metric Learning Algorithm

We now discuss in details the XGBoost metric algorithm and give the practical elements needed for its implementation. We provide as well the pseudo code and a discussion on the algorithm complexity. For simplicity of the presentation, we now assume that the loss function $\ell(y, y') = \|y - y'\|_2^2$.

[1] recall that the solution might not be unique

Let us now suppose that we have built so far $t-1$ trees $f_1, \dots, f_{t-1}$ from the datasets $\mathcal{Z}_n$ and $\bar{\mathcal{Z}}_m$, so that at this point $g_i = \frac{1}{2}(y_i - m^{(t-1)}(\boldsymbol{x}_i))$ and $H_i = 2\,\mathrm{Id}_p$ for all $i \leq n$; and that want to build the next one. Starting from the root, we have $\mathcal{I}_{\text{root}} = \{1, \dots, n\} \times \{1, \dots, m\}$, and $\mathcal{J}_{\text{root,root}} = \{1, \dots, n\} \times \{1, \dots, m\} \times \{1, \dots, m\}$. The algorithm now goes through a greedy procedure to select jointly the best direction and value of a split.

For a given couple (direction, value of split), we obtain two children denoted subsequently by "left child" and "right child" and compute accordingly the sets $\mathcal{I}_{\text{left}}$, $\mathcal{I}_{\text{right}}$, $\mathcal{J}_{\text{left,left}}$, $\mathcal{J}_{\text{left,right}}$, $\mathcal{J}_{\text{right,right}}$, and the corresponding matrices $\mathbf{G}'$ and $\mathbf{H}'$. Then, the score of the couple (direction, value of split) can be computed as in Equation (4.21). The couple with the higher reduction of the objective function is selected and the root leaf is split accordingly. The outcome of this procedure is then a stump with 2 new leaves. The latter procedure is repeated until a stopping criterion is reached as described in Algorithm 4.2.

Algorithm 4.1 Find best split of leaf l for t^{th} tree.

Require: $\mathcal{I}_l, \{\mathcal{J}_{l,l'}\}_{l' \in \text{leaves}}, \text{score}_q$
  score $\leftarrow$ score$_q$
  $s \leftarrow []$
  for $k \in \{1, \dots, 2d\}$(Coordinates of $\Phi(x, x')$) **do**
    for $s \in$ Splits of $\Phi[k]$ **do**
      Compute $\mathcal{I}_{\text{left}}, \mathcal{I}_{\text{right}}$
      Compute $\mathbf{G}', \mathbf{H}'$
      score $\leftarrow$ score $\vee\ \mathbf{G}'^{\top} \left(\mathbf{H}' + \lambda I_{L+1}\right)^{-1} \mathbf{G}'$
    end for
  end for
  return s with max score, score.

Practical cost of greedy procedure The main computation cost lies in finding the best split. In order to do so, a split finding algorithm enumerates over all the possible splits on all the dimensions of the feature space. When a split is proposed, two new leaves are added to a former leaf (a left and a right child), the sets $\mathcal{I}_{\text{child}}$ and $\mathcal{J}_{l,\text{child}}$ for $l \in \{1, \dots, L\}$ and child $\in \{\text{left}, \text{right}\}$ are built and the corresponding split score is computed. The cost of computing the score of a split is small, since the matrix is only of size of the current number of leaves, which is in practice lower that 8. Supposing that the tree is well-balanced, that is the number of observations is equi-distributed over the leaves. Each leaf contains consequently approximately $\frac{n \times m}{L}$ pairs of observations. The complexity of choosing the best split is then $\mathcal{O}\left(2d \times \#\text{splits} \times n \times m \times L^3\right)$. The exact greedy algorithm is very powerful since it enumerates over all possible splitting points. However, it is impossible to efficiently do so when the data does not fit entirely into memory. To support effective gradient tree boosting, an approximate algorithm is needed. To summarize, the algorithm first proposes candidate splitting points according to percentiles of feature distribution. The algorithm then maps the continuous features into buckets split by these candidate points, aggregates the statistics and finds the best solution among proposals based on the aggregated statistics.

Algorithm 4.2 Pseudo-code: f_t-builder

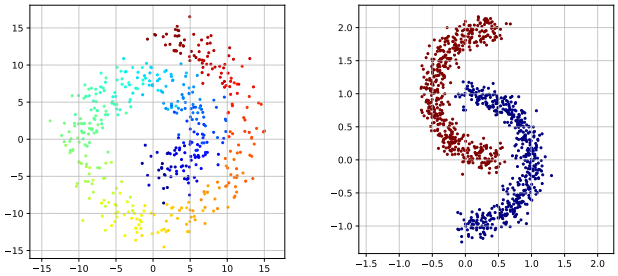
Require: $\mathcal{X}_n, \bar{\mathcal{X}}_m, \ell, T, \max_{N_{\text{leaves}}}, (g_i, H_i)_{i=1\dots n}$
  $q \leftarrow \mathbb{I}_{\text{Root}}$
  $f_t \leftarrow 0$
  $s_q \leftarrow -1$
  $\mathcal{I}_{\text{Root}} \leftarrow \mathcal{Z}_n \times \bar{\mathcal{Z}}_m$
  $\mathcal{J}_{\text{Root,Root}} \leftarrow \mathcal{Z}_n \times \bar{\mathcal{Z}}_m \times \bar{\mathcal{Z}}_m$
  Current Leaves$(f_t) := CL \leftarrow \{\text{Root}\}$
  while $|CL| \leq \max_{N_{\text{leaves}}}$ **do**
    for $l \in CL$ **do**
      $\mathcal{I}_l \leftarrow$ **Compute** according Equation (4.7)
      $\mathcal{J}_{l,l'} \leftarrow$ **Compute** according Equation (4.8)
      split, g $\leftarrow$ Find Best Split$(\mathcal{I}_l, \mathcal{J}_{l,l'}, \text{s}_q)$
    end for
    NodeToSplit, split $\leftarrow$ Node with higher g
    if $\text{s}_q < g$ **then**
      $\text{s}_q \leftarrow g$
      split NodeToSplit
      $CL \leftarrow CL +$ **Children**(NodeToSplit)
    end if
  end while
  return f_t.

4.4 Numerical Experiments and Analysis

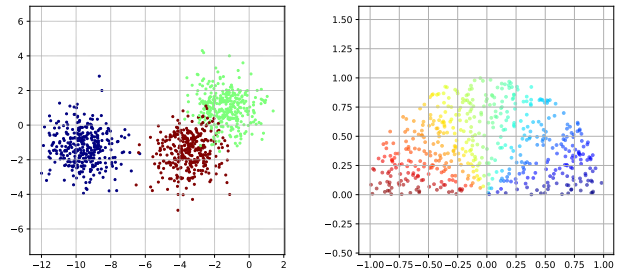
In this section we use synthetic and real data to compare our method (XGBoost Metric Learning, referred to as **XGBML**) against other methods referenced in the literature: a k-NN regressor (**k-NN**) using an Euclidean metric, a Nadaraya-Watson (**NW**) based on conventional kernels as in Hastie et al. (2001), and a Nadaraya-Watson with a learned local metric (**NW+LLM**, see Noh et al. (2017)).

4.4.1 Comparison with State-of-the-Art Metric Learning Methods

Results on synthetic data. In this section, we consider four synthetic datasets. First a classic swiss roll dataset **SR** in which we want to predict the radius. Then the two moons dataset **TM** in which we want to classify the point into two groups. The three blobs dataset **TB** used to classify point among each blob, and finally a half disk dataset **HD** for which the data are uniformly distributed over the unit half disk and we want to predict the angle. These four datasets are shown in Figure 4.2. For each dataset, the training sample size is $n = 500$, and the design sample size is $m = 50$. For the three experiments, our models achieves one of the lowest test errors. All results are provided in Table 4.1. Furthermore, the XGBoost Metric Learning provides interesting insights on the data. Figure 4.3 shows the similarities between a test point (indicated by a star) and the points of the design dataset. The width of the edges is increasing with the similarity between two points. We observe for instance that in Figure 4.3 that the highest similarities lie on the circle with radius of target-value. Furthermore, in Figure 4.4 we compute the Mean Squared Error (MSE) for different values of the ratio m/n. Our model achieves the lowest MSE for all ratios.

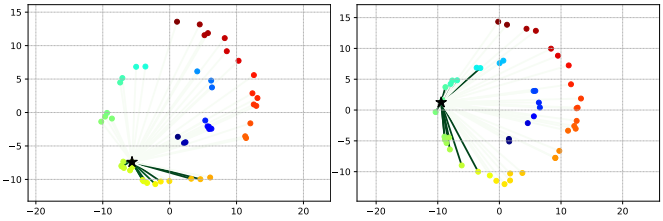


(a) Swiss Roll (b) Two Moons



(c) Blobs (d) Angles

Figure 4.2 – Scatter plots of the three synthetic datasets. The colormap stands for the target value y.



(a) Swiss Roll (b) Swiss Roll

Figure 4.3 – Graph representation (edges) of the similarities between the point of interest (star) and the points of the design dataset (dots).

	SR	TM	TB	HD
k-NN	2.339(±.05)	0.285(±.01)	0.243(±.01)	0.036(±.002)
NW	1.768(±.09)	0.358(±.02)	0.190(±.09)	0.043(±.002)
NW+LLM	1.720(±.12)	0.299(±.03)	**0.179**(±.03)	**0.033**(±.004)
XGBML	**1.613**(±.13)	**0.281**(±.03)	0.180(±.05)	**0.033**(±.004)

Table 4.1 – MSE of the compared methods for synthetic datasets.

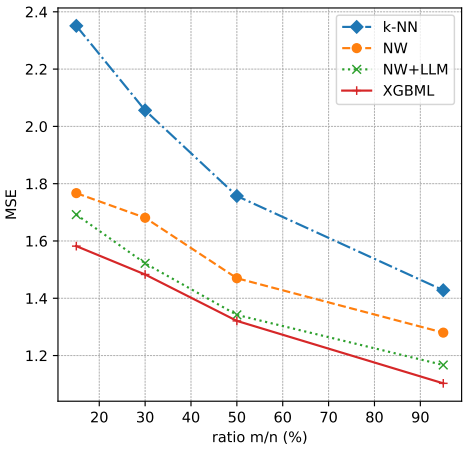


Figure 4.4 – MSE of the compared methods on the **SR** dataset, computed for different values of the ratio m/n and where $n = 500$.

	Boston	RSSI
k-NN	6.797	0.167
NW	7.516	0.219
NW+LLM	7.199	0.174
XGBML	**6.381**	**0.153**

Table 4.2 – RMSE of compared methods (± standard error).

Results on real data. In this section we consider two real world datasets. First, the classic housing dataset (**Boston**, available on *scikit-learn*[2]) in which we want to predict the house prices given 13 scalar features. Second, the Wireless Indoor Localization dataset (**RSSI**) which can be found on UCI website[3]. For both datasets, we proceed as follows. The dataset is split in 3: a design dataset $\bar{\mathcal{Z}}_m$ of size 50%, a training dataset $\mathcal{Z}_n$ of size 40%, and a test dataset of size 10% used to compute the MSE. The results are shown in Table 4.2. Despite the higher dimensionality of these data, known for deteriorating kernel regressors performances, our method again exhibits the lowest test error for both datasets.

[2] scikit-learn Boston Dataset
[3] UCI Wireless Dataset

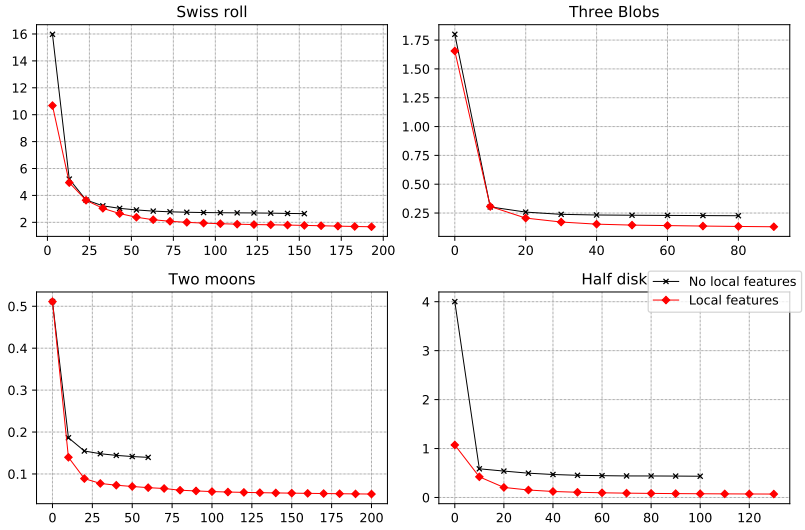


Figure 4.5 – MSE computed on test dataset with respect to the number of weak regressors. In black (resp. red) the model uses relative (resp. absolute and relative) information.

4.4.2 Benefits of Local Features on the Predictor Accuracy

We now highlight the benefits of the chosen pairwise position function Φ and in particular the improvements that the absolute position encoding have on the regression accuracy of our predictor. We build a predictor in which the features only encode for the relative information and compare it to our predictor. The models are compared The results are shown in Figure 4.5. Figure 4.5 shows that the addition of absolute position information drastically reduces the regression error of the estimator. Unsurprisingly, it also increases the impact of an additional tree on the error.

4.5 Conclusion and Perspectives

In this chapter, we proposed an original way to jointly learn a similarity function and the regressor based on this latter. Our approach consists in learning an similarity by directly minimizing the regression error. We further detailed at length an algorithm based on XGBoost to tackle this task. The proposed approach has been shown to be well adapted to minimize the objective function of interest and furthermore benefits from the well-known qualities of XGBoost. Experiments demonstrated both the competitiveness and the generality of our approach, as well as the relevance of our learned metric. This work leaves exciting avenues for future works such as the study of our estimator consistency as instance. Analyzing the Rademacher complexity of the induced class of functions by invoking only Talagrand's lemma Ledoux and Talagrand (2011) results in a bound depending linearly on m. A more in-depth study is thus needed to refine this bound and would be the purpose of future work.

5

Conditional independence testing via weighted partial copulas

Taille et vocabulaire ne sont pas indépendants ; mais ils le sont conditionnellement à l'âge.

As stated in Chapter 3, conditional independence is closely related to the statistical model that we proposed for the couple of variable $(\boldsymbol{X}, Y)$. Indeed, a direct consequence of the chosen model is that the coordinates of the variable $\boldsymbol{X}$ are independent given Y. In this chapter, we introduce the *weighted partial copula* function for testing conditional independence. The proposed test procedure results from the following ingredients: (i) the test statistic is an explicit Cramér-von Mises transformation of the *weighted partial copula*, (ii) the regions of rejection are computed using a bootstrap procedure which mimics conditional independence by generating samples from the product measure of the estimated conditional marginals. Under conditional independence, the weak convergence of the *weighted partial copula process* is established when the marginals are estimated using a smoothed local linear estimator. Finally, an experimental section demonstrates that the proposed test has competitive power compared to recent state-of-the-art methods such as kernel-based test.
This chapter covers the following publications:

- Elgui K., Bianchi P. & Portier F. Conditional independence testing via weighted partial copulas.arXiv e-prints, art. arXiv:2006.12839, June 2020.

5.1 Introduction

Let (Y_1, Y_2, X) be a triplet of real random variables. We say that Y_1 and Y_2 are conditionally independent given X if $\forall (y_1, y_2, x) \in \mathbb{R}^3$:

$$\Pr(Y_1 \leq y_1,\, Y_2 \leq y_2 \mid X = x) = \Pr(Y_1 \leq y_1 \mid X = x) \Pr(Y_2 \leq y_2 \mid X = x). \tag{5.1}$$

This is denoted by $Y_1 \perp\!\!\!\perp Y_2 \mid X$ and roughly speaking, it means that for a given value of X, the knowledge of Y_1 does not provide any further information on Y_2 (and vice versa). Determining conditional independence has become in the recent years a fundamental question in statistics and machine learning. For instance, it plays a key role in defining *graphical models* (Koller and Friedman, 2009; Bach and Jordan, 2003); see also Markowetz and Spang (2007) for a study dedicated to cellular networks. Moreover the concept of conditional independence lies at the core of *sufficient dimension reduction* methods (Li, 2018) and is useful to conduct variable selection in regression (Lee et al., 2016). Finally, conditional independence is relevant in many application fields such as economy (Huber and Melly, 2015) or psychometry (Bell et al., 1988). In this chapter, we propose a new statistical tests to assess conditional independence.

The approach taken is related to the well-studied problem of (unconditional) independence testing, in which the most intuitive way to proceed is to compute a distance between the estimated joint distribution and the product of the estimated marginals (Hoeffding, 1948). Inspired by Kendall (1948), rank-based statistics have been extensively used in independence testing (Ruymgaart, 1974; Ruschendorf, 1976; Ruymgaart and van Zuijlen, 1978). Because rank-based statistics do not depend on the marginals, they have appeared as a key tool for modelling the joint distribution of random variables without being affected by their margins. This has led to the introduction of the *copula function* (Deheuvels, 1981), defined as the cumulative distribution function associated to the ranks. We refer to Fermanian et al. (2004); Segers (2012) for recent studies on the estimation of the copula function. The copula function, which in principle measures the dependency between random variables, has been used with success in independence testing (Genest and Rémillard, 2004; Genest et al., 2006). Because the asymptotic distribution of the copula function is difficult to estimate, the related bootstrap estimate properties are of prime interest for inference (Fermanian et al., 2004; Rémillard and Scaillet, 2009; Bücher and Dette, 2010).

The conditional copula of Y_1 and Y_2 given X is defined in the same way as the copula of Y_1 and Y_2 but uses the conditional distribution of Y_1 and Y_2 given X instead of the joint distribution of Y_1 and Y_2. Compared to the copula, the conditional copula captures the conditional dependency between random variables and is thus useful to build conditional dependency measures (Gijbels et al., 2011). Therefore, as in the case of independence testing, the conditional copula appears to be a relevant tool for building statistical test of conditional independence. This has been pointed out as a an "interesting open issue" in (Veraverbeke et al., 2011, Section 4).

In this work, a new statistical test procedure called the "weighted partial copula test" is investigated to assess conditional independence. The proposed approach follows from the use of an integrated criterion, the *weighted partial copula*, a function that equals 0 if and only if conditional independence holds. Given estimators of the conditional marginals of Y_1 and Y_2 given X, the *empirical weighted partial copula* is introduced to estimate the weighted partial copula and the test statistic results from a Cramér-von Mises transformation.

From a theoretical standpoint, the use of an "integrated" criterion enables to establish, in a general non-parametric framework, a convergence rate of order $n^{-1/2}$ for the empirical weighted partial copula. More precisely, by using a smoothed local linear estimator for the conditional marginals, we obtain the weak convergence of the empirical weighted partial copula rescaled by $n^{1/2}$. The rate of convergence $n^{-1/2}$, which is the same as the one derived in the (unconditional) independence test, is notable because conditional copula estimates are known (Veraverbeke et al., 2011) to converge at a slower rate, $(nh^d)^{-1/2}$ where h is a smoothing parameter going to 0. Note finally that integrated criterion for testing has been frequently used in the *conditional moment restrictions* literature (see Lavergne and Patilea (2013) and the reference therein).

Inspired by the independence testing literature (Beran et al., 2007; Kojadinovic and Holmes, 2009), the computation of the quantiles is made using a bootstrap procedure which generates bootstrap samples from the product of the marginal estimators to mimic the null hypothesis. Thanks to this bootstrap procedure, one is allowed to perform the weighted partial copula test using any marginal estimates as soon as one can generate from them.

Related literature. Testing for conditional independence has been considered only recently in the literature. Some of the existing approaches are based on comparing the (estimated) conditional distributions involved in the definition of conditional independence. The distributions can be compared using their conditional characteristic functions as in Su and White (2007), their conditional densities as proposed in Su and White (2008), or their conditional copulas as studied in Bouezmarni et al. (2012). Unfortunately, the estimation of these conditional quantities are subjected to the well-known curse of dimensionality, i.e., the convergence rates are badly affected by the dimension of the conditioning variable. As a consequence, the power of the previous tests rapidly deteriorates if the conditioning variable has a large dimension. Note also that Bergsma (2010) uses partial copulas to derive the test statistic. Unfortunately, partial copulas fail to capture the whole conditional distribution and lead to detect a null hypothesis much larger than conditional independence.

Other approaches rely on the characterization of conditional independence using cross-covariance operators defined on reproducing kernel Hilbert spaces (Fukumizu et al., 2004). Extending the Hilbert-Schmidt independence criterion proposed in Gretton et al. (2008), Zhang et al. (2012) defines a kernel-based conditional independence test (KCI-test) by estimating the cross-covariance operator. A surge of recent research (Doran et al., 2014; Runge, 2017; Sen et al., 2017) has focused on testing conditional independence using permutation-based tests. The seminal work of Candès et al. (2018) had led to many conditional independence tests depending on the availability of an approximation to the distribution of $Y_1|X$, such as the conditional permutation test (CPT) proposed in **?**. In Sen et al. (2017), the authors propose to train a classifier (e.g., XGBoost) to distinguish between two samples, one is the original sample, another one is a bootstrap sample generated in a way that reflects conditional independence. According to the accuracy of the trained classifier the test rejects, or not, conditional independence. This is further referred to as the classifier based conditional independence test (CCI-test).

Outline. In Section 5.2, we introduce the weighted partial copula test and provide implementation details including the mentioned bootstrap procedure. In Section 5.3, we state the main theorems (weak convergence results). In Section 5.4, the theory is illustrated by numerical experiments. Our approach is compared to the ones described in Zhang et al. (2012) when facing simulated datasets. Proofs are given in a supplementary material file.

5.2 The Weighted Partial Copula Test

5.2.1 Set-up and Definitions

Let $f_{X,\mathbf{Y}}$ be the density function (with respect to the Lebesgue measure) of the random triplet $(X, \boldsymbol{Y}) = (X, Y_1, Y_2) \in \mathbb{R}^d \times \mathbb{R}^2$. Let f_X and $S_X = \{x \in \mathbb{R} : f_X(x) > 0\}$ denote the density and the support of X, respectively. The conditional cumulative distribution function of $\mathbf{Y}$ given $X = x$ is given by $\boldsymbol{y} \mapsto H(\boldsymbol{y} \mid x)$ for $x \in S_X$. The generalized inverse of a univariate distribution function F is defined as $F^-(u) = \inf\{y \in \mathbb{R} : F(y) \geq u\}$, for all $u \in [0, 1]$, with the convention that $\inf \emptyset = +\infty$. Since $H(\,\cdot\,|x)$ is a continuous bivariate cumulative distribution function, its copula is given by the function

$$C(\boldsymbol{u} \mid x) = H\left(F_1^-(u_1|x),\ F_2^-(u_2|x) \mid x\right),$$

for $\boldsymbol{u} = (u_1, u_2) \in [0,1]^2$ and $x \in S_X$, where $F_1(\cdot\,|x)$ and $F_2(\cdot\,|x)$ are the margins of $H(\cdot\,|x)$. We are interested in testing the null hypothesis that Y_1 and Y_2 are conditionally independent given X, that is,

$$\mathcal{H}_0 : Y_1 \perp\!\!\!\perp Y_2 | X.$$

By definition (Dawid, 1979), $\mathcal{H}_0$ is equivalent to $H(\boldsymbol{y}|x) = F_1(y_1|x)F_2(y_2|x)$, for every $\boldsymbol{y} \in \mathbb{R}^2$ and almost every $x \in S_X$. Using the conditional copula introduced before, it follows that

$$\mathcal{H}_0 \;\Leftrightarrow\; C(\boldsymbol{u} \mid x) = u_1 u_2, \qquad \text{for every } \boldsymbol{u} \in [0,1]^2, \text{and almost every } x \in S_X.$$

Let $w : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. The *weighted partial copula* is given by, for every $\boldsymbol{u} \in [0,1]^2$ and almost every $t \in \mathbb{R}$,

$$W(\boldsymbol{u}, t) = E\left[(C(\boldsymbol{u} \mid X) - u_1 u_2) w(t - X)\right].$$

The proposed test follows from the observation, that $\mathcal{H}_0$ is satisfied if and only if the function W is identically equal to 0 under a mild condition on w. This is presented in the following lemma whose proof is given in the supplementary material.

Lemma 5.1. *Suppose that $w : \mathbb{R}^d \to \mathbb{R}$ is integrable with respect to the Lebesgue measure and with a Fourier transform being non-zero almost everywhere, then $\mathcal{H}_0$ is equivalent to $W(\boldsymbol{u}, t) = 0$, for every $\boldsymbol{u} \in [0,1]^2$ and almost every $t \in \mathbb{R}$.*

5.2.2 The Test Statistic

In the following, we define a general estimator of W relying on some empirical copula construction that works for any estimate of the marginals F_1 and F_2 (see Section 5.2.5 for a typical example). That is, we first compute sample based observations of $F_j(Y_j|X)$, $j = 1, 2$, by estimating each marginal F_j. Those are usually called pseudo-observations. Second we define an estimate of W based on the ranks of the pseudo-observation. For the sake of generality, the estimator used for the conditional marginals is left unspecified in the subsequent development.

Let (X_i, Y_{i1}, Y_{i2}), for $i \in \{1, \ldots, n\}$, be independent and identically distributed random vectors, with common distribution equal to that of (X, Y_1, Y_2). We estimate the conditional margins in some way, producing random functions $\hat{F}_{n,j}(\cdot\,|x)$, $j = 1, 2$, and then we proceed with the pseudo-observations $\hat{U}_{ij} = \hat{F}_{n,j}(Y_{ij}|X_i)$. Let $\hat{G}_{nj}$, for $j \in \{1, 2\}$, be the empirical distribution function of the pseudo-observations $(\hat{U}_{1j}, \ldots, \hat{U}_{nj})$, i.e. $\hat{G}_{n,j}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{U}_{ij} \le u\}}$, for $u \in [0,1]$. From a conditioning argument, the weighted partial copula is given by

$$W(\boldsymbol{u}, t) = \mathbb{E}[(\mathbb{1}_{\{F_1(Y_1|X) \le u_1\}} \mathbb{1}_{\{F_2(Y_2|X) \le u_2\}} - u_1 u_2) w(t - X)]. \tag{5.2}$$

The previous expression suggests the introduction of the following so-called *empirical partial copula process*, given by

$$\hat{W}_n(\boldsymbol{u}, t) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{\hat{U}_{i1} \le \hat{G}^{-}_{n,1}(u_1)\}} \mathbb{1}_{\{\hat{U}_{i2} \le \hat{G}^{-}_{n,2}(u_2)\}} - u_1 u_2 \right) w(t - X_i). \tag{5.3}$$

The use of the transform $G_{n,1}^-$ and $G_{n,2}^-$ implies that $\hat{W}_n$ depends on $\hat{U}_{i1}$ and $\hat{U}_{i2}$ only through their ranks. Indeed, because $\hat{G}_{n,j}^-$ is a càd-làg function with jumps $1/n$ at each $\hat{U}_{ij}$, it holds that $\hat{U}_{ij} \leq \hat{G}_{n,j}^-(u_j)$ is equivalent to $(\hat{R}_{ij}-1)/n < u_j$, where $\hat{R}_{ij} = n\hat{G}_{n,j}(U_{ij})$ is the rank of U_{ij} among the sample $(\hat{U}_{1j}, \dots, \hat{U}_{nj})$. Hence, we have

$$\hat{W}_n(\boldsymbol{u}, t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{(\hat{R}_{i1}-1)<nu_1\}}\mathbb{1}_{\{(\hat{R}_{i2}-1)<nu_2\}} - u_1u_2\right)w(t - X_i). \tag{5.4}$$

The test statistic is given by

$$\hat{T}_n = \int_{[0,1]^2\times\mathbb{R}^d} \hat{W}_n(\boldsymbol{u}, t)^2 \,\mathrm{d}\boldsymbol{u}\mathrm{d}t. \tag{5.5}$$

Remark 5.2. *The test statistics $\hat{T}_n$ is of Cramér-von Mises type, as opposed to the Kolmogorov-Smirnov type (which would be defined taking the sup instead of integrating). In Genest and Rémillard (2004) these two types of statistics are introduced in the context of (unconditional) independence testing. In our context, the Cramér-von Mises type is preferred over the Kolmogorov-Smirnov for practical reasons. Indeed, as shall be seen in the next section, a closed form exists for $\hat{T}_n$.*

Remark 5.3. *The proposed estimate is of the same type as the copula estimator studied in Fermanian et al. (2004) and Portier and Segers (2018) but another definition might have been used. For instance, following the approach in Genest and Rémillard (2004), the statistic would be defined differently with $\{\hat{R}_{ik} \leq (n+1)u_k\}$ in place of $\{(\hat{R}_{ik}-1) < nu_k\}$, $k = 1, 2$. Using one version or the other does not influence our results.*

5.2.3 Computation of the Statistic

The following lemma provides a closed-form formula for the test statistics $\hat{T}_n$.

Lemma 5.4. *If $w : \mathbb{R}^d \to \mathbb{R}$ is an integrable function, then*

$$\hat{T}_n = n^{-2}\sum_{1\leq i,j\leq n} M\left(\hat{\boldsymbol{G}}_i, \hat{\boldsymbol{G}}_j\right) w^\star(X_i - X_j)$$

where $\hat{\boldsymbol{G}}_i = (\hat{R}_{i1}-1, \hat{R}_{i2}-1)/n$, $w^\star = w \star w_s$ with $w_s(x) = w(-x)$, and

$$M(\mathbf{u}, \mathbf{v}) = \left(1 - u_1 \vee v_1\right)\left(1 - u_2 \vee v_2\right) - \frac{1}{4}\left(1-u_1^2\right)\left(1-u_2^2\right) + \frac{1}{4}\left(1-v_1^2\right)\left(1-v_2^2\right) + \frac{1}{9}.$$

Remark 5.5. *The function w is left unspecified for the sake of generality. Examples include $w(t) = \exp(-t^2)$, $w(t) = \mathbb{1}_{|t|\leq 1}$ and other popular kernel functions such as the Epanechnikov kernel. In the simulations, we consider the Gaussian kernel as in this case, $w^\star$ remains Gaussian. In line with the result stated in Proposition 5.1, empirical evidences suggest that it does not have a leading role in the performance of the test. Another approach would have been to consider the (non-integrable) function $w(t) = \mathbb{1}_{\{X\leq t\}}$ as in Stute (1997). The same conclusion as in Lemma 5.1 remains valid in virtue of Proposition 16.10 in Billingsley (1995).*

Remark 5.6. *Computing $\hat{T}_n$ requires n^2 operations. A sampling strategy might be to rather compute*

$$\frac{1}{|I\times J|}\sum_{(i,j)\in(I\times J)} M(\hat{\boldsymbol{R}}_i, \hat{\boldsymbol{R}}_j) w^\star(X_i - X_j),$$

with $|I| = |J|$ denote random samples uniformly drawn in $\{1, \dots, n\}$.

5.2.4 Bootstrap Approximation

To compute the rejection regions of the test, we propose a bootstrap approach to generate new samples in a way that reflects the null hypothesis even when $\mathcal{H}_0$ is not realized in the original sample. This has been notified as a guideline for bootstrap hypothesis testing in (Hall and Wilson, 1991) and it enables, in practice, to control for the level of the test and to obtain a sufficiently large power.

The proposed bootstrap follows from the estimated conditional marginals of $Y_1|X$ and $Y_2|X$, respectively $\hat{F}_{n,1}$ and $\hat{F}_{n,2}$, and from the estimated distribution of X, denoted by $\hat{F}_n$. First we choose X^* uniformly over the $(X_i)_{i=1,\dots,n}$, that is, $X^* \sim \hat{F}_n$. Then we generate

$$Y_1{}^* \sim \hat{F}_{n,1}(\cdot|X^*), \qquad \text{and} \qquad Y_2^* \sim \hat{F}_{n,2}(\cdot|X^*),$$

and execute the previous steps n times until obtaining an independent and identically distributed bootstrap sample of size n. We denote by $(X_i^*, Y_{i1}^*, Y_{i2}^*)_{i=1,\dots,n}$ the obtained sample. We finally compute the test statistic based on this sample. We repeat this B times and obtain B realizations of the statistic under $\mathcal{H}_0$, denoted by $(T_{n,1}^*, \dots, T_{n,B}^*)$. Now define the cumulative distribution function of the bootstrap statistics $t \mapsto (1/B)\sum_{b=1}^{B} \mathbb{1}_{\{T_{n,1}^* \le t\}}$, and denote by $\xi_n(\alpha)$ its quantile of level $\alpha \in (0,1)$. The weighted partial copula test statistic with level α rejects $\mathcal{H}_0$ as soon as $\hat{T}_n > \xi_n(\alpha)$.

5.2.5 A Generic Example using Nadaraya-Watson Estimator

In this section, the aim is to illustrate the proposed test procedure when using the classical Nadaraya-Watson estimator for the margins F_j, $j \in \{1,2\}$ when $d = 1$.

Nadaraya-Watson estimator. Let $K : \mathbb{R} \to [0,\infty)$ be the standard Gaussian density function on $\mathbb{R}$. For $x \in \mathbb{R}$ and $h > 0$, put $K_h(x) = h^{-1} K(h^{-1}x)$. For $j \in \{1,2\}$, the Nadaraya-Watson estimator of $F_j(\cdot|x)$ is given by

$$\hat{F}_{n,j}(y|x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Y_{ij} \le y\}} K_{b_j}(x - X_i)}{\sum_{i=1}^{n} K_{b_j}(x - X_i)}, \qquad (y \in \mathbb{R}). \tag{5.6}$$

The choice of the bandwidths b_1 and b_2 is discussed below.

Post-nonlinear noise model. We consider $Y_1 = \cos(X+\epsilon_1)$, $Y_2 = \cos(X + aY_1 + \epsilon_2)$, where $a \ge 0$ and $X, \epsilon_1, \epsilon_2$ are independent standard Gaussian random variables with respective variances $\sigma_X^2, \sigma_1^2, \sigma_2^2$. When $a = 0$, $\mathcal{H}_0$ is true, i.e., $Y_1 \perp\!\!\!\perp Y_2 \mid X$. On the other hand, $\mathcal{H}_0$ is false when $a > 0$. The magnitude of a measures the distance to the null hypothesis. In the sequel we set $\sigma_1^2 = \sigma_2^2 = 0.2$ and $\sigma_X^2 = 1$. Two illustrative samples are shown in Figure 5.1, one is drawn from the previous model with $a = 0$ and one other with $a = 0.5$.

Cross-validation selection of the bandwidth. The bandwidths b_1 and b_2 have a critical effect on the shape of the resulting estimates, and thus on the performance of our test. Indeed, these estimates of the margins $\hat{F}_j(y_j|x)$ for $j \in \{1,2\}$ are used in the computation of the test statistic $\hat{T}_n$ as well as in the bootstrap procedure to simulate under the null (see Section 5.2.4). The idea is to assess the performance of each regression model $Y_1|X$ and $Y_2|X$ and to choose each bandwidth b_1 and b_2

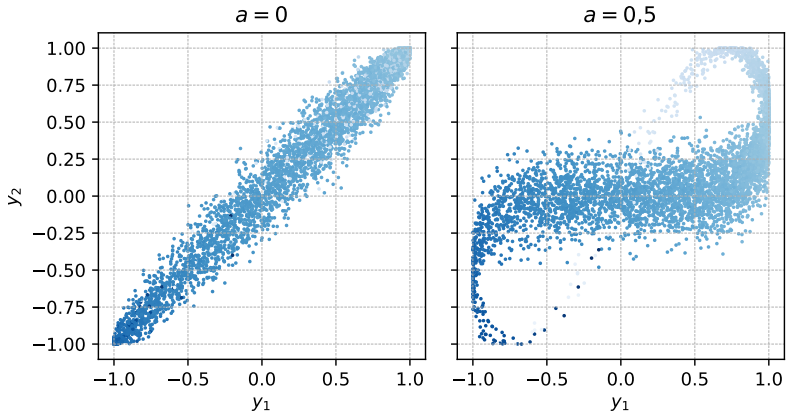


Figure 5.1 – On the left (resp. right) 2500 realizations of (X, Y_1, Y_2) from the post-nonlinear noise model when $a = 0$ (resp. $a = 0{,}5$). The shade of blue denotes for the value of X.

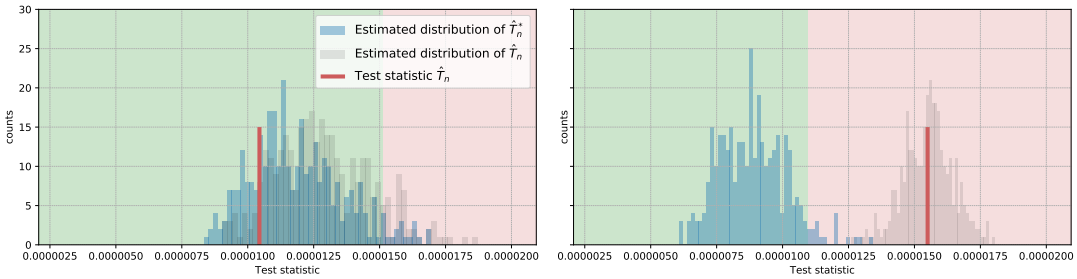


Figure 5.2 – The left (resp. right) figure corresponds to $a = 0$, i.e., $\mathcal{H}_0$ is valid (resp. $a = 0.5$, i.e., $\mathcal{H}_0$ is false). Based on a dataset of size $n = 1000$, $\hat{T}_n$ is computed (red line). Bootstrap statistics $(T^*_{n,b})_{b=1,\dots,B}$ $(B = 300)$, are used to obtain the distribution of $\hat{T}^*_n$ (blue). The red area is the rejection region at 5%. Independently, $\hat{T}_n$ is computed 300 times to obtain the distribution of in gray.

accordingly. We randomly divide the set of observations into K groups of nearly equal size. These groups are denoted by $\{I_k\}_{k=1,\dots,K}$. Define $\mathrm{MSE}_{j,k}(b) = (1/|I_k|)\sum_{i\in I_k}(Y_{ij} - \hat{g}^{(-k)}_{j,b}(X_i))^2$, where $\hat{g}^{(-k)}_{j,b}$ stands for the Nadaraya-Watson estimate of the regression $Y_j|X$ computed on $\{1,\dots,n\}\backslash I_k$ with bandwidth b. We choose $b_{n,j}$ as the minimizer of $(1/K)\sum_{k=1}^{K}\mathrm{MSE}_{j,k}(b)$ over b.

The success of the approach in distinguishing $\mathcal{H}_0$ from its contrary is illustrated on Figure 5.2 considering the generic post-nonlinear noise model Section 5.4.3.

Remark 5.7. *Though this example has been carried out using the Nadaraya-Watson estimate of the marginal distributions, other approaches to estimate the marginals can be used to conduct the weighted partial copula test. The only restriction on the employed marginal estimates comes from the bootstrap procedure in which the ability to generate*

according to the margins is required. For instance a k-nearest neighbours approach shall be considered in Section 5.4.

5.3 Weak Convergence

5.3.1 Smooth Estimator of the Margins

The theoretical results are provided in a general non-parametric setting, using a smoothed version of the *local linear estimator* (Fan and Gijbels, 1996) of the conditional marginals when $d = 1$ (see Remark 5.10 below). This estimate has been introduced in Portier and Segers (2018) and is a natural extension of the Nadaraya-Watson estimator (defined in the previous section). Such a non-parametric approach results in mild assumptions on the distribution of (X, Y_1, Y_2). Let $K : \mathbb{R} \to [0, \infty)$ and $L : \mathbb{R} \to [0, \infty)$ be two kernel functions, i.e., non-negative, symmetric functions integrating to unity. Let $(b_{n,j})_{n\geq 1}$ and $(h_{n,j})_{n\geq 1}$, for $j = 1, 2$, be four bandwidth sequences that tend to 0 as $n \to \infty$. For $(y, Y) \in \mathbb{R}^2$ and $h > 0$, put

$$\varphi_h(y, Y) = \int_{-\infty}^{y} L_h(t - Y)\,\mathrm{d}t. \tag{5.7}$$

with $L_h(y) = h^{-1} L(h^{-1}y)$. For $j \in \{1, 2\}$, we introduce the smoothed local linear estimator of $F_j(y_j|x)$ defined by

$$\hat{F}_{n,j}(y_j|x) = \hat{a}_{n,j}, \tag{5.8}$$

where $\hat{a}_{n,j}$ is the first component of the random pair

$$(\hat{a}_{n,j}, \hat{b}_{n,j}) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \Big\{ \varphi_{h_{n,j}}(y_j, Y_{ij}) - a - b(X_i - x) \Big\}^2 K\left(\frac{x - X_i}{b_{n,j}}\right), \tag{5.9}$$

where φ_h in (5.7) serves to smooth the indicator function $y \mapsto \mathbb{1}_{\{Y\leq y\}}$. Kernels K and L do not have the same role: L is concerned with "smoothing" over Y_1 and Y_2 whereas K "localises" the variable X at $x \in S_X$. For this reason, we purposefully use two different bandwidth sequences $(b_{n,j})_{n\geq 1}$ and $(h_{n,j})_{n\geq 1}$. We shall see that the conditions on the bandwidth $h_{n,j}$ for the y-directions are weaker than the ones for the bandwidth $b_{n,j}$ for the x-direction. The assumptions related to the two kernels and bandwidth sequences are stated in ((Gc)) and ((Gd)) below. Note that if the previous optimization would be carried out only over a we would recover the Nadaraya-Watson estimator with a smoothed indicator function.

5.3.2 Weak Convergence of the Weigh Partial Copula

We rely on the following Hölder regularity class. Let $\delta \in (0, 1)$, $k \in \mathbb{N}$, and $M > 0$ be scalars and let $S \subset \mathbb{R}$ be non-empty, open and convex. Let $\mathcal{C}_{k+\delta,M}(S)$ be the space of functions $S \to \mathbb{R}$ that are k times differentiable and whose derivatives (including the zero-th derivative, that is, the function itself) are uniformly bounded by M and such that every mixed partial derivative of order $l \leq k$, say $f^{(l)}$, satisfies the Hölder condition

$$\sup_{z\neq\tilde{z}} \frac{\left|f^{(l)}(z) - f^{(l)}(\tilde{z})\right|}{\left|z - \tilde{z}\right|^{\delta}} \leq M, \tag{5.10}$$

where $|\cdot|$ in the denominator denotes the Euclidean norm. In particular, $\mathcal{C}_{1,M}(\mathbb{R})$ is the space of Lipschitz functions $\mathbb{R} \to \mathbb{R}$ bounded by M and with Lipschitz constant bounded by M.

We need the following four assumptions:

(Ga) The law P admits a density $f_{X,\boldsymbol{Y}}$ on $S_X \times \mathbb{R}^2$ such that S_X is a nonempty, bounded, open interval. For some $M > 0$ and $\delta > 0$, the functions $F_1(\cdot|\cdot)$ and $F_2(\cdot|\cdot)$ belong to $\mathcal{C}_{3+\delta,M}(\mathbb{R} \times S_X)$ and the function f_X belongs to $\mathcal{C}_{2,M}(S_X)$. There exists $b > 0$ such that $f_X(x) \geq b$ for every $x \in S_X$. For any $j \in \{1, 2\}$ and any $\gamma \in (0, 1/2)$, there exists $b_\gamma > 0$ such that, for every $y_j \in [F_j^-(\gamma|x), F_j^-(1-\gamma|x)]$ and every $x \in S_X$, we have $f_j(y_j|x) \geq b_\gamma$.

(Gb) The function $w : \mathbb{R} \to \mathbb{R}$ is of bounded variation.

(Gc) The kernels K and L are bounded, non-negative, symmetric functions on $\mathbb{R}$, supported on $(-1, 1)$, and such that $\int L(u)\,\mathrm{d}u = \int K(u)\,\mathrm{d}u = 1$. The function L is continuously differentiable on $\mathbb{R}$ and its derivative is a bounded real function of bounded variation. The function K is twice continuously differentiable on $\mathbb{R}$ and its second-order derivative is a bounded real function of bounded variation.

(Gd) There exists $\alpha > 0$ such that for any $j = 1, 2$, the bandwidth sequences $b_{n,j} > 0$ and $h_{n,j} > 0$ satisfy, as $n \to \infty$,

$$nb_{n,j}^8 \to 0, \qquad nh_{n,j}^8 \to 0, \qquad b_{n,j}^{-1-\alpha/2} h_{n,j}^2 \to 0,$$
$$\frac{nb_{n,j}^{3+\alpha}}{\left|\log b_{n,j}\right|} \to \infty, \qquad \frac{nb_{n,j}^{1+\alpha} h_{n,j}}{\left|\log b_{n,j} h_{n,j}\right|} \to \infty.$$

Let $\mathbb{P}$ denote the probability measure on the underlying probability space associated to the whole sequence $(X_i, \boldsymbol{Y}_i)_{i=1,2,\ldots}$. Let $\ell^\infty(T)$ denote the space of bounded real functions on the set T, the space being equipped with the supremum distance. Define $U_{i1} = F_1(Y_{i1}|X_i)$, $U_{2i} = F_2(Y_{i2}|X_i)$, for any $i = 1, \ldots, n$, and

$$\tilde{W}_n(\boldsymbol{u}, t) = \hat{Z}_n(\boldsymbol{u}, t) - (f_X \star w)(t)(u_1 \hat{Z}_{n,2}(u_2) + u_2 \hat{Z}_{n,1}(u_1)), \tag{5.11}$$

for any $\boldsymbol{u} \in [0, 1]^2$, $t \in \mathbb{R}$, with

$$\hat{Z}_n(\boldsymbol{u}, t) = \quad n^{-1} \sum_{i=1}^n \left\{ w(t - X_i)(\mathbb{1}_{\{U_{i1} \leq u_1,\, U_{i2} \leq u_2\}} - u_1 u_2) \right\}, \tag{5.12}$$

and $\hat{Z}_{n,j}(u_j) = n^{-1} \sum_{i=1}^n \left\{ \mathbb{1}_{\{U_{ij} \leq u_j\}} - u_j \right\}$. We now state our main result. Its proof is provided in the supplementary material.

Theorem 5.8. *Assume that ((Ga)), ((Gb)), ((Gc)) and ((Gd)) hold. If $\mathcal{H}_0$ holds, then for any $\gamma \in (0, 1/2)$, we have when $n \to \infty$*

$$\sup_{\boldsymbol{u} \in [\gamma, 1-\gamma]^2,\, t \in \mathbb{R}} \left| \hat{W}_n(\boldsymbol{u}, t) - \tilde{W}_n(\boldsymbol{u}, t) \right| = o_{\mathbb{P}}(n^{-1/2}).$$

In addition, the process $\left\{ n^{1/2} \hat{W}_n(\boldsymbol{u}, w) \right\}_{\boldsymbol{u} \in [\gamma, 1-\gamma]^2, t \in \mathbb{R}}$ converges weakly in $\ell^\infty([\gamma, 1 - \gamma]^2 \times \mathbb{R})$ to a certain Gaussian process.

Remark 5.9. *Theorem 5.8 is a nontrivial extension of Theorem 2 in Portier and Segers (2018). By taking $w_t = 1$ we would recover their result.*

Remark 5.10. *The approach employed follows from an approximation of the process W_n by an oracle version of W_n where the estimated marginals are replaced by the true ones. In doing this, a crucial step consists in an embedding of some functions class involving estimated conditional quantiles into a Donsker class (van der Vaart and Wellner, 1996). First, because the estimated quantiles are difficult to control near the boundary of $[0, 1]$, we need to restrict the proof to the interval $[\gamma, 1-\gamma]^2$. We believe that the extension to the whole interval is an interesting avenue for further research. Second, the regularity properties of the local linear estimate defined in (5.8) are essential to obtain that the resulting quantile functions are sufficiently smooth to be contained in a Donsker class. Third, as noticed in Portier and Segers (2018), the extension to higher dimensions, though feasible, is not straightforward and represents an interesting topic for future work. In the case $d = 1$, covered by Theorem 5.8, the rate of convergence, $n^{-1/2}$, is not affected by the size of the different bandwidths. We conjecture that this remains true in multiple dimensions with the same rate of order $n^{-1/2}$.*

As a corollary of the previous weak convergence result, we obtain (invoking the continuous mapping theorem) the weak convergence, under $\mathcal{H}_0$, of a sightly modified version of $\hat{T}_n$.

Corollary 5.11. *Assume that ((Ga)), ((Gb)), ((Gc)) and ((Gd)) hold. If $\mathcal{H}_0$ holds, then for any $\gamma \in (0, 1/2)$ and any finite measure μ on $\mathbb{R}$, we have that*

$$n \int_{[\gamma,1-\gamma]^2 \times \mathbb{R}} \hat{W}_n(\boldsymbol{u}, t)^2 \mathrm{d}\boldsymbol{u}\mathrm{d}\mu(t)$$

converges weakly to a tight non-negative random variable as $n \to \infty$.

5.4 Numerical Experiments

In this section, we apply the proposed copula test on synthetic data to evaluate its performance based on the nominal level and the power of the test. We compare it with the KCI-test Zhang et al. (2012) presented in the related literature section. Since the level α is hard to set for the CCI-test of Sen et al. (2017), this approach will only be considered when the proportions of correct decision will be computed (see Figure 5.6a).

In all the experiments, the function w is a Gaussian kernel given by $w(t) = \exp(-t^2)$ and the estimate of the marginals is the k-nearest neighbors version of (5.6) using the cross-validation approach of Section 5.2.5 to tune k.

We use four datasets, the linear model and the post-nonlinear noise model are considered. We also test our method on a probabilistic graphical model for testing causality detection. Finally, we apply our test in a practical setting, using the movie watching based brain development dataset Richardson et al. (2018). In all of our simulations we set $\alpha = 5\%$ as the desired type-I error rate. All results are averaged over 300 trials, and we used $B = 200$ bootstrap realizations. The average CPU time taken by the tests in competition for $n = 1500, d = 1$ **copulas**: 18.3 s , **KCI-test**: 17.7 s, **CCI-test**: 19.7 s.

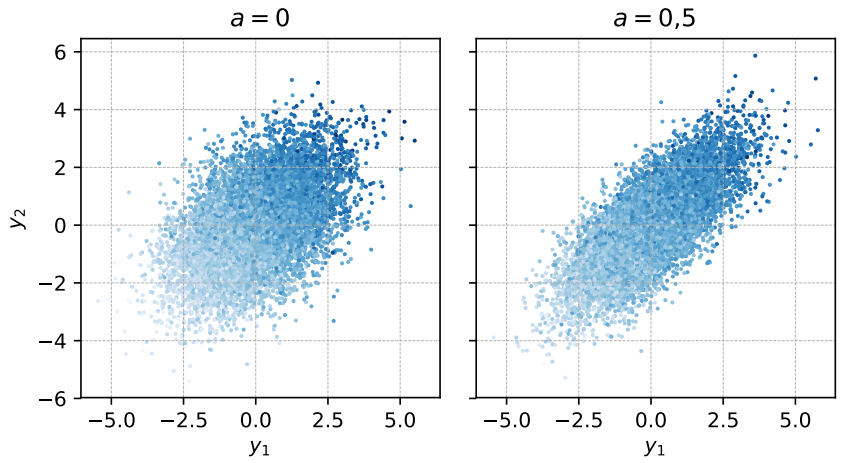


Figure 5.3 – On the left (resp. right) 10000 realizations of (X, Y_1, Y_2) drawn from the model in 5.4.1 when $a = 0$ (resp. $a = 0.5$). Here, $X \in \mathbb{R}^2$, and we set $\beta_1 = \beta_2 = (1/\sqrt{2}, 1\sqrt{2})^T \in \mathbb{R}^2$. The shade of blue denote for the value of $X^T\beta$.

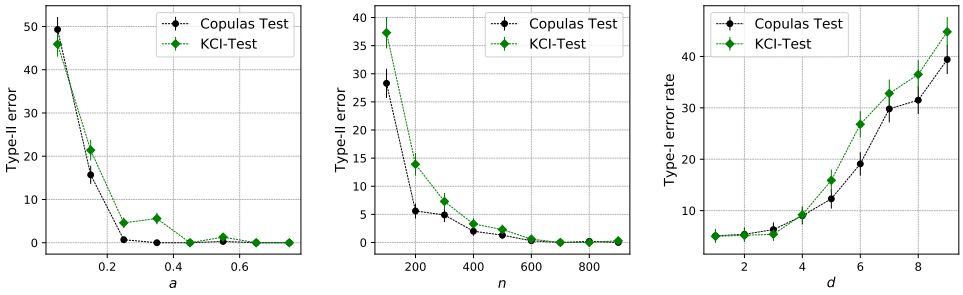
5.4.1 Linear Model

Consider the joint distribution given by $Y_1 = X^T\beta_1 + \epsilon_1, Y_2 = X^T\beta_2 + \epsilon_2$, where $X \sim \mathcal{N}(0, \mathsf{I}_d), \beta_1$and β_2 are two constant vectors of $[0, 1]^d$, and ϵ_1, ϵ_2 are two standard Gaussian variables with $\mathrm{Cov}(\epsilon_1, \epsilon_2) = a$. When $a = 0$, $\mathcal{H}_0$ is true. It is false when $a > 0$. We examine the effect of the constant $a > 0$, and the size of the dataset n on the type-II error rate. We also examine the type-I errors when the dimension of the variable X increases, in a setting where the null hypothesis $\mathcal{H}_0$ holds. Two illustrative samples are shown in Figure 5.3, one is drawn from the previous model with $a = 0$ and one other with $a = 0.5$. We set $\beta_1 = \beta_2 = (1/\sqrt{2}, 1\sqrt{2})^T \in \mathbb{R}^2$. Figure 5.4 shows the attractive performance of our test compared to the KCI-test. Notably, we can see that in high dimensions, our test is more accurate with respect to the level set α than the KCI-test.

5.4.2 Causality Discovery

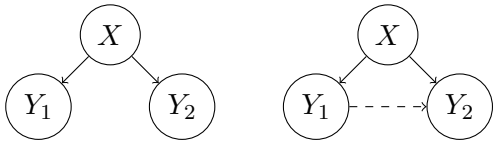
Hereinafter we consider a particular type of DAG called "Latent cause" model.

To draw samples from the alternative hypothesis, we break the conditional independence by adding an edge between the nodes Y_1 and Y_2. The resulting graphs are shown in Figure 5.5 in dashed lines. For the "Latent cause" model of interest we have $X \sim \mathcal{N}(0, 1)$, $Y_1|X \sim \mathcal{N}(X, 1)$, and $Y_2|X, Y_1 \sim \mathcal{N}(X + aY_1, 1)$. It is easy to verify that $\mathcal{H}_0$ is true when $a = 0$, and false otherwise. It can be seen in Figure 5.6 that for large sample size n, our test outperforms the ones in competition. Furthermore, our test is slightly more powerful than the KCI-test across a range of values of a but overall shows fairly similar performance.



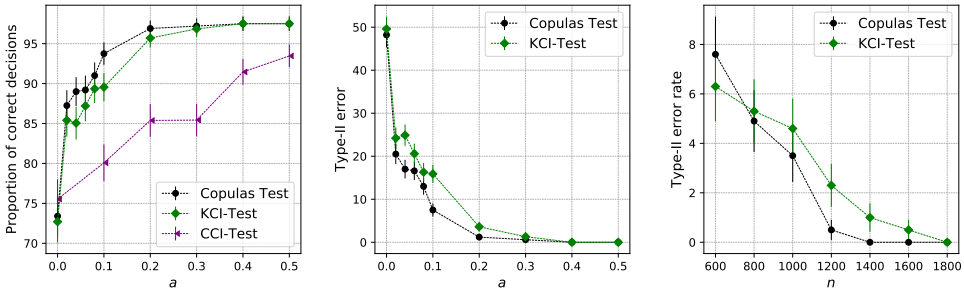
(a) $n = 1000$, various a, $d = 1$ (b) various n, $a = 0.3$, $d = 1$ (c) $a = 0$, $n = 1000$, various d

Figure 5.4 – Simulation results for the linear model. Figures 5.4a, 5.4b show the probability of acceptance (i.e. the type II error rate), plotted against the constant a and n. Figure 5.4c shows the probability of rejection (type I error) against d. The plots show the average probabilities with standard error bars.



(a) CI holds (b) CI fails

Figure 5.5 – Latent cause model. The CI holds when the dashed edge does not exist and fails otherwise.

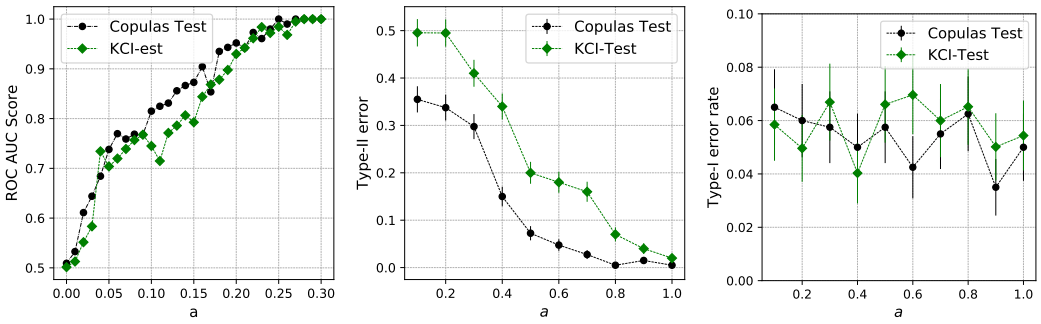


(a) $n = 1500$, various a (b) $n = 1500$, various a (c) $a = 0.3$, various n

Figure 5.6 – Simulation results for the causal inference datasets.

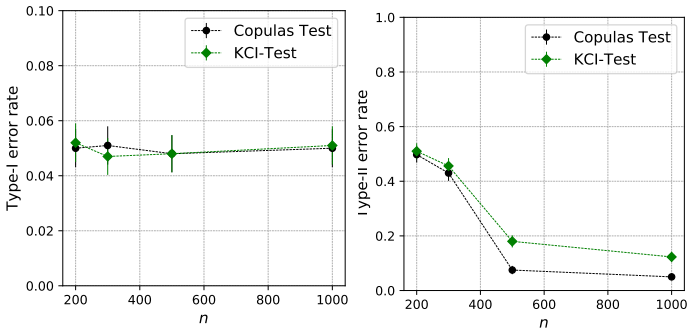
5.4.3 Post-Nonlinear Noise

We apply the proposed test on the post-nonlinear noise model described in Section 5.2.5. We first examine the effect of the constant $a > 0$ on the probability of type-I and type-II error of our test. The results are shown in Figure 5.7. As expected, larger values of a yields lower type-II error probabilities. For every value a, we observe that the type-I error probability is closed to α. The performances of the tests are also compared when the sample size n changes. The role of n is critical and the results are shown in Figure 5.8. We note that the type-I error probability is again closed to α and that the type-II quickly vanishes when n increases. In this experiment, the proposed procedure



(a) $n = 500$, various a and $d = 1$(b) $n = 500$, various a and $d = 1$(c) $n = 500$, various a and $d = 1$

Figure 5.7 – Simulation results for the post-nonlinear noise model. Figures 5.7a, 5.7b and 5.7c, show respectively the ROC AUC score, the probability of acceptance (i.e. the type II error rate), and the type I error rate plotted against the constant a with standard error bars.



(a) $a = 0.5$, various n, $d = 1$ (b) $a = 0.5$, various n, $d = 1$

Figure 5.8 – The probability of type-I (5.8b) and type-II (5.8b).

outperforms the KCI-test.

5.4.4 Classification of Age Groups using Functional Connectivity

In this paragraph, we apply our test in a practical setting, using the movie watching based brain development dataset Richardson et al. (2018) obtained from the OpenNeuro database[1]. The dataset consists in 50 patients (10 adults and 40 children). The fMRI data consists in measuring the brain activity in 39 Region of Interest (ROI). For every patient, 168 measurements are provided for each ROI. We denote for $j \in \{1, \dots, 39\}$ by X_j the variable that represents the j^{th} region signal value. Given two ROI j and j', we seek to test the null hypothesis that X_j and $X_{j'}$ are conditionally independent given $\mathbf{X}_{\setminus\{j,j'\}}$. The decisions given by our test allow us to obtain a map of connections between all the ROI, called *connectome*, given in Figure 5.9. In this figure, a line is drawn between two ROI whenever our test rejects the null for these two ROI. Here, due to the high dimension of the conditional variable, the margins are no longer estimated using a Gaussian kernel as in Section 5.2.5, but using a k-nearest neighbours approach.

[1] Accession number ds000228.

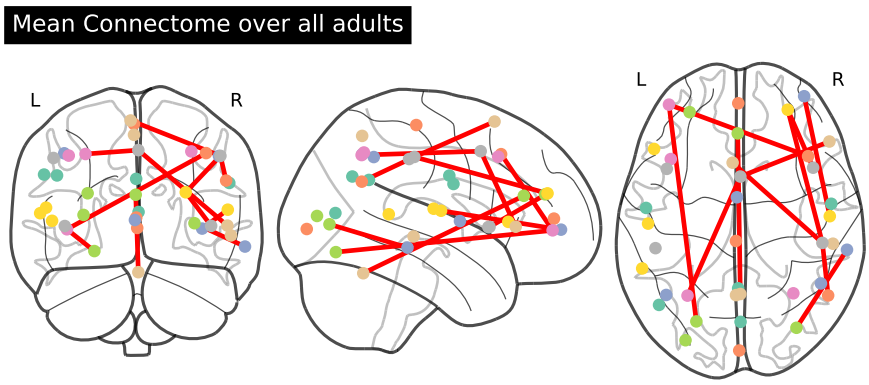


Figure 5.9 – Mean connectome provided by our test over all adults.

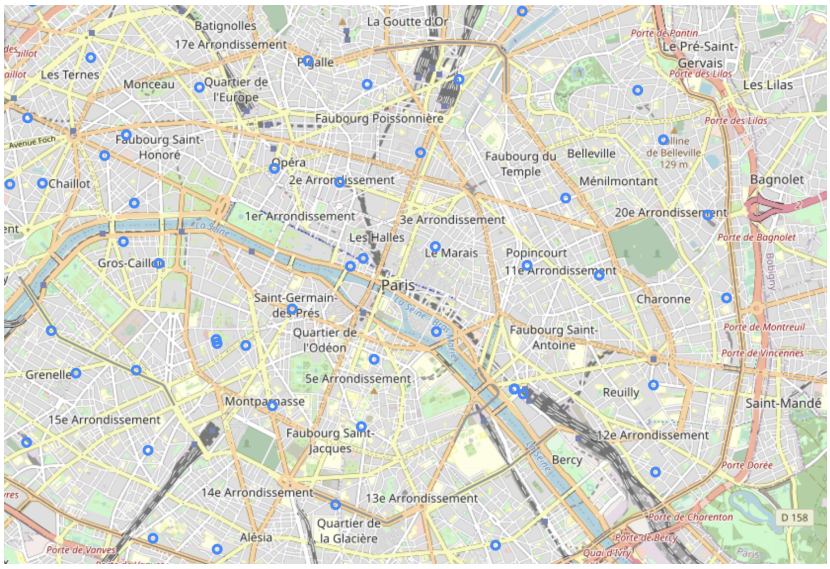


Figure 5.10 – The blue circles represent the GPS coordinates of the BS that are used in this numerical experiment.

For a given x, the mapping $\hat{F}_{n,j}(y|x)$ is estimated for every $y \in \mathbb{R}$ as the proportion of samples i amongst the k-nearest nearest neighbours of x which satisfy $Y_{ij} \leq y$. The integer k is select by cross-validation.

As a sanity check, our connectome is used as an input feature of a classifier (Linear Support Vector Classifier (SVC)) in order to distinguish children from adults. We estimate the classification accuracy of our classifier using k-fold. The obtained accuracy is 97.4%. This result outperforms the standard correlation method (91.3%) and is close to the so-called tangent method (98.9%) which is known to be fitted for this task Dadi et al. (2019).

5.4.5 Copulas CI Test to Corroborate the Model of Equation (3.1)

The last numerical experiment is dedicated to corroborate the statistical model of Equation (3.1). Under this statistical model, the components $X_1, \ldots, X_d$ of the random vector $\boldsymbol{X}$ are independent conditionally to the position Y. In this section, we want to test that, for any couple of BS $(i, j) \in \{1, \ldots, d\}^2, i \neq j, X_i \perp\!\!\!\perp X_j|Y$. The datasets consists

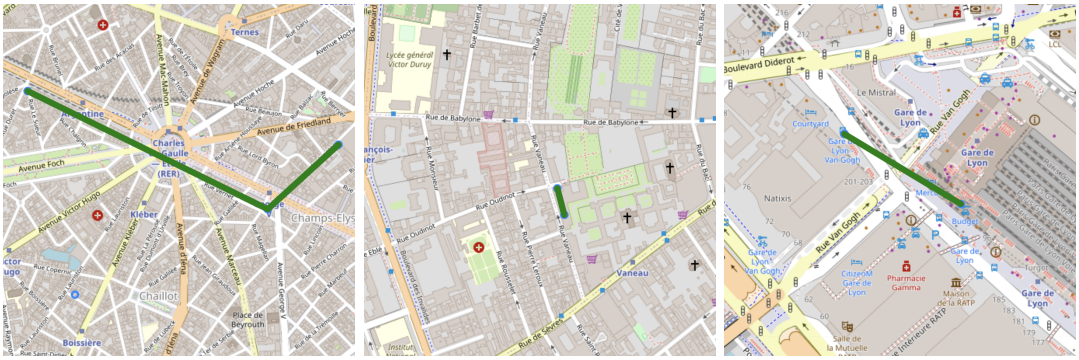


Figure 5.11 – four couples of BS for which the conditional independence given the position Y does not hold.

in $n = 2000000$ couples $(\boldsymbol{x}_i, y_i)$ collected for $d = 45$ BS displayed in Figure 5.10. For each pairs of BS, we perform our test. The outcome allow us to obtain couples of BS for which the conditional independence given the position Y does not hold. These 13 pairs are shown in Figure 5.11 and correspond to the pairs of nodes connected by an edge. The connected edges seem sound as they correspond to geographically closed BS, or BS that are placed on the same road. Sometimes, they also correspond to BS with the same coordinates.

5.5 Conclusion

In this chapter, we have developed test of conditional independence between Y_1 and Y_2 given X based on weighted partial copulas. First, under conditional independence, the weak convergence of the weighted partial copula process has been established under certain regularity conditions the marginals. We have shown that, empirically, our proposed test shows better performance, in terms of power, than recent state-of-the-art conditional independence tests such as the one based on a kernel embedding (Zhang et al., 2012). Furthermore, this test has been performed to corroborate one of the most commonly used assumption in RSSI-based geolocation literature, or more specifically, one consequence of this assumption. Numerical experiments have shown that, not only is this assumption not true for certain pairs of BS, but also that couples for whom the hypothesis was rejected share particular geographic dispositions with each others (geographic proximity or the fact that there are placed on a same road). An interesting topic would be to assess the improvement of prediction performance when these pairwise conditional independence are no longer assumed. Finally, the generality of the proposed approach makes it useful in many practical situations, and its soundness has been endorsed by conclusive numerical experiments.

Conclusion

In this thesis, we have proposed contributions to the RSSI-based geolocation problem in LPWAN.

First, we identified the limits of time-based approaches when applied to geolocation in the singular context of LPWAN; indeed, a preliminary investigation showed that the narrow band signals that are a requisite for IoT networks greatly jeopardize the performance of such geolocation methods. Then, we demonstrated that the well-known and widely used path-loss model is also not suited to address the task of geolocation because it fails to correctly represent the relation between the range to the emitter and the RSSI.

Then, we addressed this problem by means of recent techniques of machine learning, and presented most popular methods found in the literature. We furthermore proposed two new techniques. The first one relies on a semi-parametric Nadaraya-Watson (NW) estimator of the likelihood, followed by a MAP estimator of the object's position. The second one consists in learning a proper metric on the dataset, constructed by means of a Gradient boosting regressor. Contrary to prior works, the proposed methods take into consideration both the non-isotropy and the information of non-reception. This work emphasized how choosing a relevant distance on the RSSI space can greatly improve the quality of fingerprint methods estimation.

The quality of the prediction is indeed, highly related to the chosen distance on the RSSI space. The metric learning problem is therefore a fundamental issue to improve RSSI-based geolocation techniques. However, there are, as far as we know, few studies on the fingerprint problem from the perspective of metric learning. We therefore proposed an original way to jointly learn a similarity function and the regressor based on the latter. Our approach consists in learning a similarity by directly minimizing the regression error of the final predictor. We further detailed at length an algorithm based on XGBoost to tackle this task. The proposed approach has been shown to be perfectly adapted to minimize the objective function of interest. We then applied this method to popular regression problems such as the swiss roll dataset, and demonstrate that its range of applications go far beyond the field of geolocation.

Finally, independence conditional testing problem was tackled introducing the weighted partial copula function for testing conditional independence. The proposed test procedure is as follows: first, the test statistic is an explicit Cramér-von Mises transformation of the weighted partial copula. Second, the regions of rejection are computed using a boot-strap procedure which mimics conditional independence by generating samples from the product measure of the estimated conditional marginals. Furthermore, the weak convergence of the weighted partial copula process is established under the CI. Numerical experiments have shown that, not only is this assumption not true for certain pairs of BS, but also that couples for whom the hypothesis was rejected share particular geographic dispositions with each others (geographic proximity or the fact that there are placed on a same road). An interesting topic would be to assess the improvement of prediction performance when pair-wise conditional independence is no longer assumed.

Appendices

A Proofs of Chapter 6

A.1 Proofs of the basic lemmas of Section 5.2

A.1.1 Proof of Lemma 5.1

The "only if" part is obvious. Suppose that the function $W = 0$ and let $\boldsymbol{u} \in [0,1]^2$. Define $g(x) = C(\boldsymbol{u} \mid x) - u_1 u_2) f_X(x)$. We have $(g \star w) = 0$, a.e. on $\mathbb{R}^d$, where $\star$ stands for the standard convolution product with respect to the Lebesgue measure. Applying the Fourier transform gives that $\mathcal{F}(g)\mathcal{F}(w) = 0$ which, by assumption, yields $\mathcal{F}(g) = 0$. By the Fourier inversion theorem we obtain that $g = 0$ a.e. on $\mathbb{R}^d$. That is for any $\boldsymbol{u} \in [0,1]^2$ and any $x \in S_X$, $C(\boldsymbol{u} \mid x) = u_1 u_2$.

□

A.1.2 Proof of Lemma 5.4

Write

$$\hat{T}_n = n^{-2} \sum_{1 \le i,j \le n} \int_{[0,1]^2} \xi_i(\boldsymbol{u}) \xi_j(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \int_{\mathbb{R}^d} \omega(t - X_i) \omega(t - X_j) \, \mathrm{d}t,$$

where $\xi_k(\boldsymbol{u}) = \mathbb{1}_{\{\hat{G}_{k1} < u_1\}} \mathbb{1}_{\{\hat{G}_{k2} < u_2\}} - u_1 u_2$. It remains to compute the function M. Using the notation $\hat{\boldsymbol{G}}_i$, we have

$$\begin{aligned}
\int_{[0,1]^2} \xi_i(\boldsymbol{u}) \xi_j(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = & \int_{[0,1]^2} \mathbb{1}_{\{\hat{G}_{i1} < u_1\}} \mathbb{1}_{\{\hat{G}_{i2} < u_2\}} \mathbb{1}_{\{\hat{G}_{j1} < u_1\}} \mathbb{1}_{\{\hat{G}_{j2} < u_2\}} \mathrm{d}\boldsymbol{u} \\
& - \int_{[0,1]^2} \mathbb{1}_{\{\hat{G}_{i1} < u_1\}} \mathbb{1}_{\{\hat{G}_{i2} < u_2\}} u_1 u_2 \mathrm{d}\boldsymbol{u} \\
& - \int_{[0,1]^2} \mathbb{1}_{\{\hat{G}_{j1} < u_1\}} \mathbb{1}_{\{\hat{G}_{j2} < u_2\}} u_1 u_2 \mathrm{d}\boldsymbol{u} + \int_{[0,1]^2} (u_1 u_2)^2 \mathrm{d}\boldsymbol{u}.
\end{aligned}$$

First, let compute the first term of the right hand side. Let notice that the value of the integrand is 1 if $u_1 > \hat{G}_{i1} \vee \hat{G}_{j1}$ and $u_2 > \hat{G}_{u2} \vee \hat{G}_{j2}$ and 0 otherwise. Thus we obtain for this term:

$$\int_{[0,1]^2} \mathbb{1}_{u_1 > \hat{G}_{i1} \vee \hat{G}_{j1}} \mathbb{1}_{u_2 > \hat{G}_{i2} \vee \hat{G}_{j2}} \mathrm{d}\boldsymbol{u} = \left(1 - \hat{G}_{i1} \vee \hat{G}_{j1}\right) \left(1 - \hat{G}_{i2} \vee \hat{G}_{j2}\right). \tag{A.1}$$

Now let derive the second integral term of the right hand side, the third term will follow directly.

$$\begin{aligned}
\int_{[0,1]^2} \mathbb{1}_{\{\hat{G}_{i1} < u_1\}} \mathbb{1}_{\{\hat{G}_{i2} < u_2\}} u_1 u_2 \mathrm{d}\boldsymbol{u} &= \int_{[0,1]} \mathbb{1}_{\{\hat{G}_{i1} < u_1\}} u_1 \mathrm{d}u_1 \int_{[0,1]} \mathbb{1}_{\{\hat{G}_{i2} < u_2\}} u_2 \mathrm{d}u_2 \\
&= \frac{1}{4} \left(1 - \hat{G}_{i1}^2\right) \left(1 - \hat{G}_{i2}^2\right).
\end{aligned} \tag{A.2}$$

By combining (A.1) and (A.2) we obtain the desired result. □

A.2 Proof of Theorem 5.8

We use notation from empirical process theory. Let $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,\boldsymbol{Y}_i)}$ denote the empirical measure. For a function f and a probability measure Q, write $Qf = \int f \,\mathrm{d}Q$. The empirical process is

$$\mathbb{G}_n = n^{1/2}(P_n - P).$$

For any pair of cumulative distribution functions F_1 and F_2 on $\mathbb{R}$, put $\boldsymbol{F}(\boldsymbol{y}) = (F_1(y_1), F_2(y_2))$ for $\boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2$ and $\boldsymbol{F}^-(\boldsymbol{u}) = (F_1^-(u_1), F_2^-(u_2))$ for $\boldsymbol{u} = (u_1, u_2) \in [0,1]^2$.

A.2.1 Sketch of the proof

We introduce an oracle copula estimator, defined as the empirical copula based on the unobservable random pairs $(F_1(Y_{i1}|X_i), F_1(Y_{i2}|X_i))$, $i \in \{1,\dots,n\}$. Let $\hat{G}_{n,j}^{(or)}$ be the empirical distribution function of the uniform random variables $F_j(Y_{ij}|X_i)$, $i \in \{1,\dots,n\}$, i.e.,

$$\hat{G}_{n,j}^{(or)}(u_j) = P_n\{\mathbb{1}_{\{F_j \le u_j\}}\}, \qquad u_j \in [0,1].$$

Let $\hat{G}_{n,j}^{(or)-}$ be its generalized inverse. The oracle estimator of W is then

$$\hat{W}_n^{(or)}(\boldsymbol{u}, t) = P_n\left\{ w_t(\mathbb{1}_{\{\boldsymbol{F} \le \hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}} - u_1 u_2)\right\},$$

with $w_t(x) = w(t-x)$. A crucial result is that the processes $\hat{W}_n$ and $\hat{W}_n^{(or)}$ are asymptotically equivalent as stated in the following lemma.

Lemma A.1. *Assume that (G(Ga)), (G(Gb)), (G(Gc)) and (G(Gd)) hold. If $\mathcal{H}_0$ holds, then for any $\gamma \in (0, 1/2)$, we have when $n \to \infty$,*

$$\sup_{\boldsymbol{u}\in[\gamma,1-\gamma]^2,\, t\in\mathbb{R}} \left| n^{1/2}\left\{ \hat{W}_n(\boldsymbol{u}, t) - \hat{W}_n^{(or)}(\boldsymbol{u}, t)\right\}\right| = o_{\mathbb{P}}(1). \tag{A.3}$$

Using the notation from empirical process theory, introduced below, we have

$$\begin{aligned}
&\hat{Z}_n(\boldsymbol{u}, t) = P_n\left\{ w_t\left(\mathbb{1}_{\{\boldsymbol{F}\le\boldsymbol{u}\}} - u_1 u_2\right)\right\}, \qquad \boldsymbol{u} \in [0,1]^2,\, t \in \mathbb{R},\\
&\hat{Z}_{n,j}(u_j) = \hat{G}_{n,j}^{(or)}(u_j) - u_j, \qquad u_j \in [0,1],\\
&\tilde{W}_n(\boldsymbol{u}, t) = \hat{Z}_n(\boldsymbol{u}, t) - P\{w_t\}\left(u_1 \hat{Z}_{n,2}(u_2) + u_2 \hat{Z}_{n,1}(u_1)\right), \qquad \boldsymbol{u} \in [0,1]^2,\, t \in \mathbb{R}.
\end{aligned}$$

A second crucial result is the following one, where it is shown that $\hat{W}_n^{(or)}$ is asymptotically equivalent to $\tilde{W}_n$.

Lemma A.2. *Assume that (G(Ga)), (G(Gb)), (G(Gc)) and (G(Gd)) hold. If $\mathcal{H}_0$ holds, we have when $n \to \infty$,*

$$\sup_{\boldsymbol{u}\in[0,1]^2,\, t\in\mathbb{R}} \left| \hat{W}_n^{(or)}(\boldsymbol{u}, t) - \tilde{W}_n(\boldsymbol{u}, t)\right| = o_{\mathbb{P}}(n^{-1/2})$$

Based on Lemma A.1 and A.2, we deduce that

$$\sup_{\boldsymbol{u}\in[\gamma,1-\gamma]^2,\, t\in\mathbb{R}} \left|\hat{W}_n(\boldsymbol{u},t) - \tilde{W}_n(\boldsymbol{u},t)\right| = o_{\mathbb{P}}(n^{-1/2}).$$

Invoking Slutsky's Lemma, the process $\{\hat{W}_n(\boldsymbol{u},t)\}_{\boldsymbol{u}\in[\gamma,1-\gamma]^2,\, t\in\mathbb{R}}$ and $\{\tilde{W}_n(\boldsymbol{u},t)\}_{\boldsymbol{u}\in[\gamma,1-\gamma]^2,\, t\in\mathbb{R}}$ have the same weak limit in $\ell^\infty([\gamma,1-\gamma]^2\times\mathbb{R})$. Now note that $\{x\mapsto w_t(x) \,:\, t\in\mathbb{R}\}$ is a Euclidean or VC class with constant envelop $C_w = \sup_{x\in\mathbb{R}}|w(x)|$ (Nolan and Pollard, 1987, Lemma 22, (ii)), i.e., the covering numbers are polynomials. Moreover, the class of indicator functions is also Euclidean (van der Vaart and Wellner, 1996, Example 2.5.4). This implies that both classes have finite entropy integrals and therefore are Donsker (van der Vaart and Wellner, 1996, Chapter 2.1, equation (2.1.7)). Using the preservation of the Donsker property through products and sums (van der Vaart and Wellner, 1996, Example 2.10.7 and 2.10.8), the class $\{(\boldsymbol{y},x)\mapsto w_t(x)\mathbb{1}_{\{\boldsymbol{F}(\boldsymbol{y}|x)\le\boldsymbol{u}\}} \,:\, t\in\mathbb{R},\, \boldsymbol{u}\in[0,1]^2\}$ is Donsker. As a result, the process $\{\tilde{W}_n(\boldsymbol{u},t)\}_{\boldsymbol{u}\in[\gamma,1-\gamma]^2,\, t\in\mathbb{R}}$ converges weakly to a tight Gaussian process in $\ell^\infty([\gamma,1-\gamma]^2\times\mathbb{R})$.

A.2.2 Proof of Lemma A.1

Our proof is adapted from the proof of Theorem 1 in Portier and Segers (2018). For the sake of clarity, we start by recalling some of the results established in Portier and Segers (2018) that will be used further in our proof. Apart from this, the proof is self-consistent.

Fact 1. *On a sequence of events whose probabilities tend to one, it holds that for every $u_j\in[\gamma,1-\gamma]$ and every $(y_j,x)\in\mathbb{R}\times S_X$,*

$$\hat{F}_{n,j}(y_j|x) \le u_j \;\Leftrightarrow\; y_j \le \hat{F}^-_{n,j}(u_j|x).$$

This is shown page 170 in Portier and Segers (2018).

For $u_j\in[\gamma,1-\gamma]$, $x\in S_X$, and $j\in\{1,2\}$, define

$$\hat{\Delta}_{n,j}(u_j|x) = F_j\Big(\hat{F}^-_{n,j}\Big(\hat{G}^-_{n,j}(u_j)|\,x\Big)|\,x\Big) - \hat{G}^{(or)-}_{n,j}(u_j). \tag{A.4}$$

Fact 2. *We have for any $j=1,2$,*

$$\sup_{u_j\in[\gamma,1-\gamma]} \left| n^{1/2}\int \hat{\Delta}_{n,j}(u_j|x)\, f_X(x)\,\mathrm{d}x\right| = o_{\mathbb{P}}(1). \tag{A.5}$$

This is shown page 171 in Portier and Segers (2018).

Fact 3. *As established in (Portier and Segers, 2018, page 172), for each $j=1,2$,*

$$\left\{x\mapsto \hat{F}^-_{n,j}\Big(\hat{G}^-_{n,j}(u_j)|x\Big) \;:\; u_j\in[\gamma,1-\gamma]\right\} \subset \mathcal{C}_{1+\delta_1,M_1}(S_X),$$
$$\left\{x\mapsto F^-_j\Big(\hat{G}^{(or)-}_{n,j}(u_j)|x\Big) \;:\; u_j\in[\gamma,1-\gamma]\right\} \subset \mathcal{C}_{1+\delta,M_2}(S_X),$$

with probability going to 1.

We are based on Theorem 2.1 stated in (van der Vaart and Wellner, 2007) and reported below; for a proof see for instance van der Vaart and Wellner (1996, Lemma 3.3.5). Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed random elements of a measurable space $(\mathcal{X}, \mathcal{A})$ and with common distribution equal to P. Let $\mathbb{P}$ denote the probability measure on the probability space on which the sequence $\xi_1, \xi_2, \ldots$ is defined. Let $\mathbb{G}_{\xi,n}$ be the empirical process associated to the sample $\xi_1, \ldots, \xi_n$. Let $\mathcal{E}$ and $\mathcal{V}$ be sets and let $\{m_{v,\eta} \ : \ v \in \mathcal{V}, \, \eta \in \mathcal{E}\}$ be a collection of real-valued, measurable functions on $\mathcal{X}$.

Theorem A.3 (Theorem 2.1 in (van der Vaart and Wellner, 2007))**.** *Let $\hat{\eta}_n$ be random elements in $\mathcal{E}$. Suppose there exist $\eta_0 \in \mathcal{E}$ and $\mathcal{E}_0 \subset \mathcal{E}$ such that the following three conditions hold:*

1. $\sup_{v \in \mathcal{V}} P\Big(m_{v,\hat{\eta}_n} - m_{v,\eta_0}\Big)^2 = o_{\mathbb{P}}(1)$ *as* $n \to \infty$*;*
2. $\mathbb{P}(\hat{\eta}_n \in \mathcal{E}_0) \to 1$ *as* $n \to \infty$*;*
3. $\{m_{v,\eta} - m_{v,\eta_0} \ : \ v \in \mathcal{V}, \ \eta \in \mathcal{E}_0\}$ *is P-Donsker.*

Then it holds that

$$\sup_{v \in \mathcal{V}} \left| \mathbb{G}_{\xi,n} \left(m_{v,\hat{\eta}_n} - m_{v,\eta_0} \right) \right| = o_{\mathbb{P}}(1), \qquad n \to \infty.$$

The empirical process notation allows us to write

$$\hat{W}_n(\boldsymbol{u}, t) = P_n \left\{ w_t(\mathbb{1}_{\{\hat{\boldsymbol{F}}_n \leq \hat{\boldsymbol{G}}_n^-(\boldsymbol{u})\}} - u_1 u_2) \right\}, \quad \hat{W}_n^{(or)}(\boldsymbol{u}, t) = P_n \left\{ w_t(\mathbb{1}_{\{\boldsymbol{F} \leq \hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}} - u_1 u_2) \right\}.$$

where $w_t(x) = w(t - x)$. To establish (A.3), we rely on the decomposition

$$\begin{aligned}
&n^{1/2} \Big\{ \hat{W}_n(\boldsymbol{u}, t) - \hat{W}_n^{(or)}(\boldsymbol{u}, t) \Big\} \\
&= \mathbb{G}_n \left\{ w_t(\mathbb{1}_{\{\hat{\boldsymbol{F}}_n \leq \hat{\boldsymbol{G}}_n^-(\boldsymbol{u})\}} - \mathbb{1}_{\{\boldsymbol{F} \leq \hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}}) \right\} + n^{1/2} P \left\{ w_t(\mathbb{1}_{\{\hat{\boldsymbol{F}}_n \leq \hat{\boldsymbol{G}}_n^-(\boldsymbol{u})\}} - \mathbb{1}_{\{\boldsymbol{F} \leq \hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}}) \right\} \\
&= \hat{A}_{n,1}(\boldsymbol{u}, t) + \hat{A}_{n,2}(\boldsymbol{u}, t).
\end{aligned}$$

Let $\gamma \in (0, 1/2)$. The proof consists in showing that the empirical process term $\hat{A}_{n,1}(\boldsymbol{u}, t)$ goes to zero, uniformly over $(\boldsymbol{u}, t) \in [\gamma, 1-\gamma]^2 \times \mathbb{R}$, in probability (first step) and that the bias term $\hat{A}_{n,2}(\boldsymbol{u}, t)$ goes to zero, uniformly over $(\boldsymbol{u}, t) \in [\gamma, 1-\gamma]^2 \times \mathbb{R}$, in probability (second step). Assumption $\mathcal{H}_0$, will be crucial for treating the bias term in the second step.

First step: We show that

$$\sup_{\boldsymbol{u} \in [\gamma, 1-\gamma]^2, t \in \mathbb{R}} \left| \hat{A}_{n,1}(\boldsymbol{u}, t) \right| = o_{\mathbb{P}}(1), \qquad n \to \infty.$$

By Result 1, it holds that (with a slight abuse of notation)

$$\hat{A}_{n,1}(\boldsymbol{u}, t) = \mathbb{G}_n \left\{ w_t(\mathbb{1}_{\{\boldsymbol{Y} \leq \hat{\boldsymbol{F}}_n^-(\hat{\boldsymbol{G}}_n^-(\boldsymbol{u})|X)\}} - \mathbb{1}_{\{\boldsymbol{Y} \leq \boldsymbol{F}^-(\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})|X)\}}) \right\}.$$

Therefore we apply Theorem A.3 with $\xi_i = (X_i, \boldsymbol{Y}_i)$, $\mathcal{X} = S_X \times \mathbb{R}^2$, $\mathcal{V} = [\gamma, 1-\gamma]^2 \times \mathbb{R}$ and $\mathcal{E}$ the space of measurable functions valued in $\mathbb{R}^4$ and defined on $S_X \times [\gamma, 1-\gamma]^2$. Moreover, the quantities η_0 and $\hat{\eta}_n$ are given by, for every $\boldsymbol{u} \in [\gamma, 1-\gamma]^2$ and $x \in S_X$,

$$\eta_0(\boldsymbol{u}, x) = \left(\boldsymbol{F}^-(\boldsymbol{u}|x), \boldsymbol{F}^-(\boldsymbol{u}|x)\right),$$
$$\hat{\eta}_n(\boldsymbol{u}, x) = \left(\hat{\boldsymbol{F}}_n^-\left(\hat{\boldsymbol{G}}_n^-(\boldsymbol{u})|x\right), \boldsymbol{F}^-\left(\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})|x\right)\right).$$

Identifying $v \in \mathcal{V}$ with $(\boldsymbol{u}, t) \in [\gamma, 1-\gamma]^2 \times \mathbb{R}$ and $\eta \in \mathcal{E}$ with $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, where $\boldsymbol{\eta}_j$, $j \in \{1, 2\}$, are valued in $\mathbb{R}^2$, the map $m_{v,\eta} : \mathbb{R}^2 \times S_X \to \mathbb{R}$ is given by

$$m_{v,\eta}(\boldsymbol{y}, x) = w_t(x)(\mathbb{1}_{\{\boldsymbol{y} \leq \boldsymbol{\eta}_1(\boldsymbol{u}, x)\}} - \mathbb{1}_{\{\boldsymbol{y} \leq \boldsymbol{\eta}_2(\boldsymbol{u}, x)\}}),$$

Finally, the space $\mathcal{E}_0$ is the collection of those elements $\eta = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ in $\mathcal{E}$ such that

$$\{x \mapsto \boldsymbol{\eta}_1(\boldsymbol{u}, x) \,:\, \boldsymbol{u} \in [\gamma, 1-\gamma]^2\} \subset \left(\mathcal{C}_{1+\delta_1, M_1}(S_X)\right)^2,$$
$$\{x \mapsto \boldsymbol{\eta}_2(\boldsymbol{u}, x) \,:\, \boldsymbol{u} \in [\gamma, 1-\gamma]^2\} \subset \left(\mathcal{C}_{1+\delta, M_2}(S_X)\right)^2,$$

where M_2 depends only on b_γ and M. In the following we check each condition of Theorem A.3.

Verification of Condition (1) in Theorem A.3. Because the indicator function and w_t are bounded, we have

$$\int |w_t(x)| \left| \mathbb{1}_{\{\boldsymbol{y} \leq \hat{\boldsymbol{F}}_n^-(\hat{\boldsymbol{G}}_n^-(\boldsymbol{u})|x)\}} - \mathbb{1}_{\{\boldsymbol{y} \leq \boldsymbol{F}^-(\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})|x)\}} \right|^2 f_{X,\boldsymbol{Y}}(x, \boldsymbol{y}) \, \mathrm{d}(x, \boldsymbol{y})$$
$$\leq C_w \sum_{j=1}^{2} \sup_{u_j \in [\gamma, 1-\gamma]} \int \left| \mathbb{1}_{\{y_j \leq \hat{F}_{n,j}^-(\hat{G}_{n,j}^-(u_j)|x)\}} - \mathbb{1}_{\{y_j \leq F_j^-(\hat{G}_{n,j}^{(or)-}(u_j)|x)\}} \right|^2 f_{X,Y_j}(x, y_j) \, \mathrm{d}(x, y_j),$$

so that we can focus on each margin separately. Recall that if the random variable U is uniformly distributed on $(0, 1)$, then $\mathrm{E}(\mathbb{1}_{\{U \leq u_1\}} - \mathbb{1}_{\{U \leq u_2\}})^2 = |u_1 - u_2|$. Writing $\hat{a}_{n,x}(u_j) = \hat{F}_{n,j}^-\left(\hat{G}_{n,j}^-(u_j)|x\right)$, we have

$$\begin{aligned}
&\int \left| \mathbb{1}_{\{y_j \leq \hat{a}_{n,x}(u_j)\}} - \mathbb{1}_{\{y_j \leq F_j^-(\hat{G}_{n,j}^{(or)-}(u_j)|x)\}} \right|^2 f_{X,Y_j}(x, y_j) \, \mathrm{d}(x, y_j) \\
&= \int \left| \mathbb{1}_{\{F_j(y_j|x) \leq F_j(\hat{a}_{n,x}(u_j)|x)\}} - \mathbb{1}_{\{F_j(y_j|x) \leq \hat{G}_{n,j}^{(or)-}(u_j)\}} \right|^2 f_{X,Y_j}(x, y_j) \, \mathrm{d}(x, y_j) \\
&= \int_{S_X} \left| F_j(\hat{a}_{n,x}(u_j)|x) - \hat{G}_{n,j}^{(or)-}(u_j) \right| f_X(x) \, \mathrm{d}x \\
&= \int_{S_X} \left| \hat{\Delta}_{n,j}(u_j|x) \right| f_X(x) \, \mathrm{d}x,
\end{aligned}$$

where $\hat{\Delta}_{n,j}$ has been defined in (A.4). Result 2 permits to conclude.

Verification of Condition (2) in Theorem A.3. This is given by Result 3.

Verification of Condition (3) in Theorem A.3. It is enough to show that the class of functions

$$\left\{ (\boldsymbol{y}, x) \mapsto w_t(x)(\mathbb{1}_{\{\boldsymbol{y} \leq \boldsymbol{g_1}(x)\}} - \mathbb{1}_{\{\boldsymbol{y} \leq \boldsymbol{g_2}(x)\}}) \quad : \quad t \in \mathbb{R}, (\boldsymbol{g_1}, \boldsymbol{g_2}) \in \left(\mathcal{C}_{1+\delta_1, M_1}(S_X)\right)^2 \times \left(\mathcal{C}_{1+\delta, M_2}(S_X)\right)^2 \right.$$

is P-Donsker. Since the sum and the product of two bounded Donsker classes is Donsker (van der Vaart and Wellner, 1996, Example 2.10.8), it suffices to show that both classes

$$\left\{\mathbb{1}_{\{y\le g(x)\}} \;:\; g\in\mathcal{C}_{1+\delta,M}(S_X)\right\} \qquad \text{and} \qquad \{w_t\,:\,t\in\mathbb{R}\}$$

are Donsker. For any $\delta>0$ and $M>0$, the first one is Donsker since the class of subgraphs of $\mathcal{C}_{1+\delta,M}(S_X)$, under (G(Ga)), has a sufficiently small entropy (van der Vaart and Wellner, 1996, Corollary 2.7.3). The second one has been shown to be Donsker in Section A.2.1.

Second step: We show that

$$\sup_{\boldsymbol{u}\in[\gamma,1-\gamma]^2,\,t\in\mathbb{R}}\left|\hat{A}_{n,2}(\boldsymbol{u},t)\right| = o_{\mathbb{P}}(1), \qquad n\to\infty.$$

Under $\mathcal{H}_0$, we have, for every $\boldsymbol{u}\in[0,1]^2$, $t\in\mathbb{R}$,

$$\begin{aligned}
P\left\{w_t(\mathbb{1}_{\{\hat{\boldsymbol{F}}_n\le\hat{\boldsymbol{G}}_n^-(\boldsymbol{u})\}})\right\} &= \int w_t(x)\mathbb{1}_{\{\boldsymbol{y}\le\hat{\boldsymbol{F}}_n^-(\hat{\boldsymbol{G}}_n^-(\boldsymbol{u})|x)\}}\,f_{X,\boldsymbol{Y}}(x,\boldsymbol{y})\,\mathrm{d}(x,\boldsymbol{y})\\
&= \int w_t(x)F_1\Big(\hat{F}_{n,1}^-\Big(\hat{G}_{n,1}^-(u_1)|x\Big)\mid x\Big)F_2\Big(\hat{F}_{n,2}^-\Big(\hat{G}_{n,2}^-(u_2)|x\Big)\mid x\Big)\,f_X(x)\,\mathrm{d}x.
\end{aligned}$$

Consequently, using the bound $F_2\le 1$ and $\hat{G}_{n,1}^{(or)-}\le 1$,

$$\begin{aligned}
\left|\hat{A}_{n,2}(\boldsymbol{u},t)\right| &= \left|\int w_t(x)(F_1\Big(\hat{F}_{n,1}^-\Big(\hat{G}_{n,1}^-(\boldsymbol{u})|x\Big)\mid x\Big)F_2\Big(\hat{F}_{n,2}^-\Big(\hat{G}_{n,2}^-(\boldsymbol{u})|x\Big)\mid x\Big) - \hat{G}_{n,1}^{(or)-}(\boldsymbol{u})\hat{G}_{n,2}^{(or)-}(\boldsymbol{u}))\,f_X(x)\,\mathrm{d}x\right|\\
&\le C_w\int\left|F_1\Big(\hat{F}_{n,1}^-\Big(\hat{G}_{n,1}^-(\boldsymbol{u})|x\Big)\mid x\Big)F_2\Big(\hat{F}_{n,2}^-\Big(\hat{G}_{n,2}^-(\boldsymbol{u})|x\Big)\mid x\Big) - \hat{G}_{n,1}^{(or)-}(u_1)\hat{G}_{n,2}^{(or)-}(u_2)\right|\,f_X(x)\,\mathrm{d}x\\
&= C_w\int\left|\hat{\Delta}_{n,1}(u_1|x)F_2\Big(\hat{F}_{n,2}^-\Big(\hat{G}_{n,2}^-(u_1)|x\Big)\mid x\Big) + \hat{G}_{n,1}^{(or)-}(u_1)\hat{\Delta}_{n,2}(u_2|x)\right|\,f_X(x)\,\mathrm{d}x\\
&= 2C_w\max_{j=1,2}\sup_{u\in[\gamma,1-\gamma]}\int\left|\hat{\Delta}_{n,j}(u|x)\right|\,f_X(x)\,\mathrm{d}x
\end{aligned}$$

It remains to use Result 2 to obtain the conclusion.

A.2.3 Proof of Lemma A.2

Recall the definition of $\hat{W}_n^{(or)}(\boldsymbol{u},t)$ and $\hat{Z}_n(\boldsymbol{u},t)$ that are given in Section A.2.1 and that under $\mathcal{H}_0$, in virtue of Lemma 5.1, $P\{w_t(\mathbb{1}_{\{\boldsymbol{F}\le\boldsymbol{u}\}}-u_1u_2)\}=0$. Notice that

$$\begin{aligned}
&\hat{W}_n^{(or)}(\boldsymbol{u},t)-\hat{Z}_n(\boldsymbol{u},t)\\
&= P_n\{w_t(\mathbb{1}_{\{\boldsymbol{F}\le\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}}-\mathbb{1}_{\{\boldsymbol{F}\le\boldsymbol{u}\}})\}\\
&= n^{-1/2}\mathbb{G}_n\{w_t(\mathbb{1}_{\{\boldsymbol{F}\le\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}}-\mathbb{1}_{\{\boldsymbol{F}\le\boldsymbol{u}\}})\}+P\{w_t(\mathbb{1}_{\{\boldsymbol{F}\le\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u})\}}-\mathbb{1}_{\{\boldsymbol{F}\le\boldsymbol{u}\}})\}\\
&= R_{n,1}(\boldsymbol{u},t)+\left(\hat{G}_{n,1}^{(or)-}(u_1)\hat{G}_{n,2}^{(or)-}(u_2)-u_1u_2\right)P\left\{w_t\right\}\\
&= R_{n,1}(\boldsymbol{u},t)+\left(R_{n,2}(\boldsymbol{u})+u_1(\hat{G}_{n,2}^{(or)-}(u_2)-u_2)+u_2(\hat{G}_{n,1}^{(or)-}(u_1)-u_1)\right)P\left\{w_t\right\}
\end{aligned}$$

with

$$R_{n,1}(\boldsymbol{u},t) = n^{-1/2}(\hat{Z}_n(\hat{\boldsymbol{G}}_n^{(or)-}(\boldsymbol{u}),t) - \hat{Z}_n(\boldsymbol{u},t)),$$
$$R_{n,2}(\boldsymbol{u}) = (\hat{G}_{n,1}^{(or)-}(u_1) - u_1)(\hat{G}_{n,2}^{(or)-}(u_2) - u_2).$$

Now just recall the definition of $\tilde{W}_n$ to obtain that

$$\begin{aligned}&\hat{W}_n^{(or)}(\boldsymbol{u},t) - \tilde{W}_n(\boldsymbol{u},t)\\ &= R_{n,1}(\boldsymbol{u},t) + P\{w_t\}R_{n,2}(\boldsymbol{u}) + P\{w_t\}\left(u_1\rho_{n,2}(u_2) + u_2\rho_{n,1}(u_1)\right),\end{aligned}$$

with

$$\rho_{n,j}(u_j) = (\hat{G}_{n,j}^{(or)-}(u_j) - u_j) + (\hat{G}_{n,j}^{(or)}(u_j) - u_j).$$

From Vervaat's Lemma (Segers, 2015, Lemma 4.3), we have that

$$\sup_{u_j\in[0,1]} |\rho_{n,j}(u_j)| = o_{\mathbb{P}}(n^{-1/2}),$$
$$\sup_{\boldsymbol{u}\in[0,1]^2} |R_{n,2}(\boldsymbol{u})| = O_{\mathbb{P}}(n^{-1}).$$

Because the class of functions $\{(\boldsymbol{y},x) \mapsto w_t(x)\mathbb{1}_{\{\boldsymbol{F}(\boldsymbol{y}|x)\leq\boldsymbol{u}\}} : t\in\mathbb{R}, \boldsymbol{u}\in[0,1]^2\}$ is Donsker (as demonstrated in Section A.2.1), the process $\hat{Z}_n$ is asymptotically equicontinuous. This implies that

$$\sup_{\boldsymbol{u}\in[0,1]^2} |R_{n,1}(\boldsymbol{u})| = o_{\mathbb{P}}(n^{-1/2}).$$

Consequently, each quantity in the above decomposition of $\hat{W}_n^{(or)}(\boldsymbol{u},t) - \tilde{W}_n(\boldsymbol{u},t)$ is $o_{\mathbb{P}}(n^{-1/2})$, uniformly over $\boldsymbol{u}\in[0,1]^2$ and $t\in\mathbb{R}$, and so comes the conclusion.

B
Résumé en français

B.1 Contexte de la thèse

Cette thèse est le résultat d'une convention CIFRE (Convention Industrielle de Formation et de Recherche) entre Télécom Paris Saclay et Sigfox, un opérateur de télécommunications français créé en 2009 par Christophe Fourtet et Ludovic Le Moan. Sigfox est spécialisé dans le Machine to Machine (M2M) via des réseaux bas débit. Il contribue à l'Internet des objets (IoT) en permettant l'interconnexion via une passerelle. Sa technologie radio UNB ("Ultra narrow band") lui permet de construire un réseau cellulaire à bas débit et économe en énergie. Ce type de réseau est déployé dans les bandes de fréquences radio dites industrielles, scientifiques et médicales (ISM), disponibles dans le monde entier sans aucune licence.

Ces dernières années, l'Internet des objets (IoT) a suscité une grande attention dans des domaines très divers tels que l'agriculture ou les soins de santé. Les experts s'accordent à dire que 30 milliards d'objets feront partie de l'IdO d'ici 2023 et que 40% de ces objets devront être géolocalisés, par exemple pour le transport de marchandises. L'un des défis les plus importants pour ce domaine est le besoin de localisation. En effet, de nombreuses applications des réseaux de capteurs doivent suivre des objets mobiles, tels que des personnes, des animaux, des voitures, etc. Pour que ces applications soient viables, le coût des dispositifs devra être faible (de quelques dollars à quelques centimes selon l'application) et les dispositifs devront durer des années, voire des décennies, sans remplacement de la batterie. En outre, le réseau devra s'organiser sans modération humaine importante. En outre, afin de permettre la connectivité de milliards de dispositifs, la plupart des réseaux dédiés à l'IdO utilisent des communications à longue portée et à faible puissance. L'IoT a suscité beaucoup d'intérêt ces dernières années. Il désigne les réseaux de dispositifs physiques dotés de capacités de communication. L'attente que tout soit connecté est à l'origine de cette tendance. On prévoit que, d'ici 2020, il y aura plus de 20 milliards d'objets communicants dans le monde (Hatton). Ces objets sont capables à la fois de collecter et de transférer des informations. Comme la maintenance fréquente des batteries doit être évitée en raison du nombre élevé de dispositifs prévus, une faible consommation d'énergie est également une exigence forte pour l'IdO. Par conséquent, les défis des réseaux IoT sont d'atteindre une grande évolutivité pour gérer un nombre massif de dispositifs, d'atteindre un faible coût et d'avoir une large couverture tout en gardant une faible consommation d'énergie. Les appareils qui répondent à ces exigences sont difficiles à intégrer dans les réseaux cellulaires traditionnels. C'est pourquoi une technologie dédiée LPWAN (Low Power Wide Area Network), telle que l'UNB (Ultra Narrow Band), développée et brevetée par Sigfox, est apparue.

La connaissance de la géolocalisation de chaque appareil est une ressource très précieuse. En effet, elle permet à Sigfox de fournir cette information aux utilisateurs du réseau, ce qui débouche sur de nombreuses applications telles que la logistique ou le transport de marchandises, la surveillance et le suivi dans les bâtiments intelligents ou encore

le marketing et la publicité de proximité dans les centres commerciaux. Cependant, les techniques de localisation traditionnelles telles que le système de positionnement global (GPS) ne sont donc pas bien adaptées aux besoins particuliers de l'industrie IoT. L'installation d'un GPS sur chaque appareil représente un coût et une consommation d'énergie prohibitifs pour de nombreuses applications, et n'est en outre pas adaptée aux applications intérieures. Toutes ces exigences compliquent grandement la localisation de ces objets.

Les méthodes alternatives basées sur la distance utilisent des mesures telles que les deux présentées ici, le temps d'arrivée (TOA) ou la différence de temps d'arrivée (TDoA) (Ho and Chan, 1993; Cong and Zhuang, 2002), et l'indicateur de force du signal reçu (RSSI) pour estimer la distance entre un dispositif émetteur et une antenne réceptrice. D'autres méthodes de télémétrie sont couramment utilisées dans la littérature, comme l'angle d'arrivée (AOA) (Niculescu and Nath, 2003), ou la différence de fréquence d'arrivée (FDoA) (Amar and Weiss, 2008) mais leur étude dépasse le cadre de cette thèse.

Nous sommes, comme indiqué dans l'introduction, intéressés par des scénarios où l'objet suivi est équipé d'un dispositif de communication, mais pas nécessairement d'un dispositif de positionnement tel que le GPS.

Ce chapitre nous permet de mettre en perspective les enjeux de cette thèse. Il a pour but de permettre de mieux comprendre la singularité du problème de géolocalisation dans un réseau de capteurs tel que le réseau Sigfox.

Généralement, les mesures de distance et d'angle utilisées pour la localisation sont affectées par des erreurs variant dans le temps et des erreurs statiques, dépendantes de l'environnement. Les erreurs temporelles (dues, par exemple, au bruit additif et aux interférences) peuvent être réduites en faisant la moyenne de plusieurs mesures dans le temps. Les erreurs liées à l'environnement sont le résultat de la disposition physique des objets (par exemple, les bâtiments, les arbres et les meubles) dans l'environnement particulier dans lequel le réseau de capteurs fonctionne. Comme l'environnement est imprévisible, ces erreurs sont imprévisibles et doivent être modélisées comme aléatoires. Toutefois, dans un environnement particulier, les objets sont essentiellement stationnaires et, par conséquent, pour un réseau de capteurs essentiellement stationnaires, les erreurs liées à l'environnement seront largement constantes dans le temps. La majorité des applications des réseaux de capteurs sans fil impliquent des capteurs essentiellement stationnaires. Comme un certain délai est acceptable dans ces applications, chaque paire de capteurs effectuera plusieurs mesures dans le temps et fera la moyenne des résultats afin de réduire l'impact des erreurs variant dans le temps.

Tout d'abord, nous abordons le problème général de la géolocalisation dans Appendix B.2. Pour prédire la localisation d'un émetteur, les approches de pointe basées sur le canal consistent soit à estimer le délai entre l'émetteur et le récepteur, à partir duquel il est possible de déduire la distance entre l'émetteur et le récepteur; soit à prédire directement cette dernière distance à partir de la décroissance de puissance observée entre la puissance reçue et la puissance d'émission. Une introduction au canal sans fil, en particulier les paramètres clés pour le modéliser, et une présentation des prédicteurs discutés sont proposées dans la Section 1.3.

B.2 L'estimation de la localisation géographique

B.2.1 Principe de géolocalisation

Les termes géolocalisation et positionnement sont utilisés pour désigner l'estimation géographique dans le monde réel de l'emplacement d'un objet. Cette problématique est entrée dans notre société avec une diffusion massive. Il est utilisé dans de nombreuses applications telles que la navigation, la communication, les véhicules autopilotés, les objets connectés et les villes communicantes, ou plus récemment avec des problématiques telles que le contrôle de la contamination d'une population. Elle touche des domaines scientifiques très variés comme la Géolocalisation et Navigation par un Système de Satellites (GNSS), dont dépend de plus en plus notre économie. Les experts s'accordent à dire que 30% du produit intérieur brut dépendra en partie des GNSS d'ici 2030, contre 10% aujourd'hui.

La géolocalisation d'un objet fait référence aux coordonnées (latitude, longitude), c'est-à-dire à la position de l'objet sur la surface de la Terre. Parfois, le mot positionnement est plutôt employé lorsqu'il s'agit d'identifier l'emplacement d'un objet dans un espace particulier tel qu'un téléphone portable dans un centre commercial, ou un robot dans un bâtiment. Dans cette thèse, nous parlerons de géolocalisation lorsqu'il s'agit d'une localisation à l'échelle globale, alors que le terme positionnement est généralement utilisé dans des espaces intérieurs et/ou confinés.

Les méthodes radiofréquences sont utilisées pour la plupart des systèmes de géolocalisation. Cette famille de méthodes, également appelée méthodes de radiolocalisation, utilise les caractéristiques des ondes radio reçues pour prédire la localisation d'un objet émetteur. Les exemples sont nombreux. Le très répandu *Global Positioning System* (GPS) (voir Figure 1.1) est basé sur l'estimation du temps d'arrivée (TOA) d'un signal à un satellite. Lorsque l'heure de transmission, la vitesse de propagation et la position du satellite sont connues, le TOA conduit à un très bon estimateur de la distance entre l'objet émetteur et le satellite. La combinaison de plusieurs TOA conduit à l'estimation de la position de l'émetteur. L'utilisation de plusieurs récepteurs pour localiser un émetteur est connue sous le nom de *multilatération*. (illustré dans Figure 1.2). En téléphonie cellulaire, la radiolocalisation est effectuée directement par les stations de base (BS) du réseau cellulaire au moyen d'une ou plusieurs des caractéristiques suivantes :

- Le TOA (ou TDoA). Contrairement au GPS, ces quantités sont estimées par rapport à la station de base du réseau cellulaire.

- L'angle d'arrivée (AOA) correspond à la direction depuis laquelle le signal est reçu. Une façon pratique de déterminer l'AOA est de considérer que cette direction est celle où l'intensité du signal est maximale pendant une rotation complète de la station de base. En combinant plusieurs AOA, on obtient l'estimation de la position souhaitée.

- Le RSSI présente un grand intérêt dans cette thèse. Il correspond à la puissance du signal reçu moins la puissance du signal émis. Il donne un rendement à un système de télémétrie au moyen du modèle d'affaiblissement de chemin Log-distance décrit dans Section 1.3.2 ou par des méthodes basées sur les empreintes digitales (lorsque les différents lieux d'émission sont connus pour présenter des puissances très différentes signatures"). Ces méthodes sont étudiées en détail dans Chapter 4.

B.2.2 Géolocalisation basée sur le réseau

Cette thèse se concentre sur la géolocalisation basée sur le réseau. Ces méthodes utilisent uniquement l'infrastructure du réseau. Parmi toutes les méthodes de géolocalisation décrites, ces dernières sont les moins chères et nécessitent le moins d'énergie. Les méthodes de ce type ont rencontré un énorme succès avec l'apparition de l'Internet des objets (IoT) à la fin des années 1990. Essentiellement, le concept de l'IdO consiste à fournir à tout objet la capacité de transférer des données sur un réseau sans nécessiter d'interaction entre humains ou entre humains et ordinateurs. Aujourd'hui, l'ensemble des applications des dispositifs IoT est spectaculaire : maisons intelligentes (Samuel, 2016), applications médicales et de santé (Catarinucci et al., 2015), agriculture (Mekala and Viswanathan, 2017) ou même systèmes de transport (Zhou et al., 2012).

Nous passons ensuite à deux exemples concrets de géolocalisation rencontrés dans la pratique.

Logistique Le suivi des cargaisons et des biens via le réseau peut également présenter de grands avantages pour les systèmes de transport. On peut envoyer des alertes spécifiques lorsque des événements remarquables se produisent, comme l'arrivée dans un entrepôt.

Ecologie Il a ensuite été affirmé que l'IdO allait révolutionner le domaine de l'écologie. Tout d'abord, en termes de gestion de l'énergie : la connectivité d'un nombre important d'appareils consommateurs d'énergie (lampes, moteurs, pompes, etc.) peut leur permettre de communiquer avec les services publics non seulement pour équilibrer la production d'électricité mais aussi pour optimiser la consommation d'énergie dans son ensemble. Deuxièmement, les applications de surveillance environnementale de l'IdO utilisent des capteurs pour contribuer à la protection de l'environnement en surveillant, par exemple, la qualité de l'air ou de l'eau. D'autres applications, comme les systèmes de prévision des tremblements de terre ou des tsunamis, peuvent également être utilisées pour fournir une aide plus efficace.

Une idée naturelle est d'appliquer les méthodes de radiolocalisation décrites ci-dessus au réseau Sigfox. Comme indiqué précédemment, les performances de ces méthodes dépendent fortement de l'infrastructure du réseau. Sans entrer dans les détails ici, nous proposons de donner quelques éléments qui compromettent leur utilisation et motivent donc les travaux ultérieurs. Tout d'abord, les stations de base Sigfox manquent de directivité et ne peuvent donc pas discriminer les directions des ondes incidentes. Par conséquent, les méthodes basées sur l'AOA ne sont pas pertinentes ici. De plus, Sigfox base ses communications sur la technologie dite Ultra-Narrow-Band (UNB) qui permet d'atteindre à la fois une longue portée et une autonomie prolongée. Ainsi, chaque signal possède une bande de fréquence d'une largeur de 100 Hz dans la bande de fréquence sans licence (allant de 868.0 à 868.6 MHz). Cette bande est populaire car elle présente un bon équilibre entre la portée, la pénétration des bâtiments et la possibilité d'utiliser de petites antennes. Néanmoins, il est bien connu que les performances des méthodes temporelles utilisant l'estimation du TOA ou du TDoA dépendent fortement de la largeur de bande du signal. La limite inférieure de Cramér-Rao (CRLB) sur les estimateurs TOA est donc souvent employée pour quantifier cet effet. Elle stipule que la variance minimale de tout estimateur sans biais du TOA est inversement proportionnelle à $B^3 \times \text{SNR}$, où B représente la largeur de bande et Signal to Noise Ratio (SNR) est le rapport signal/bruit bien connu. Cette limite est donc défavorable à l'utilisation de telles approches et conduit par exemple à un écart-type des estimateurs de portée au

moins égal à 20 km avec l'infrastructure Sigfox US. Néanmoins, elle constitue également une première valeur de référence à laquelle doivent être comparées les méthodes de géolocalisation que nous proposons dans la suite. On montre (Boucher and Hassab, 1981) qu'un autre élément défavorable est qu'elles sont particulièrement gourmandes en mémoire puisque leurs performances sont directement liées au choix de l'intervalle d'échantillonnage.

B.2.3 Géolocalisation basée sur le RSSI

Le RSSI est une caractéristique pertinente pour la géolocalisation. Il est suffisamment explicatif sans constituer une charge mémoire. La littérature est extrêmement vaste et les méthodes sont souvent classée sous deux catégories: les méthodes "range-based" et les méthodes "free range". La première catégorie de méthodes utilise des mesures pour prédire la distance entre émetteurs et récepteurs (les coordonnées de la station de base sont connues dans ce cas). La combinaison de plusieurs portées estimées permet d'estimer la position de l'émetteur au moyen de certaines méthodes telles que la trilatération : (Thomas and Ros, 2005). Il y a donc toujours deux phases : une phase de télémétrie et une phase de localisation. En revanche, la deuxième famille de méthodes ne base pas ses prédictions sur l'estimation de la distance. Un exemple simple d'une telle méthode serait d'estimer la position comme le barycentre de la station de base réceptrice. Dans la mesure où le RSSI est la seule mesure dont nous disposons, les méthodes basées sur la distance montrent rapidement leurs limites (Chandrasekaran et al., 2009). En effet, elles reposent essentiellement sur le modèle log-distance path loss dont le manque de réalisme est mis en évidence dans Section 1.3.2.

B.3 Plan de la thèse et contributions

Dans cette thèse, nous proposons une autre classification des méthodes de géolocalisation basées sur le RSSI : les méthodes basées sur la vraisemblance et les méthodes basées sur les fingerprints. Les premières, qui englobent les méthodes "range based", consistent à apprendre (sur un jeu de données) un modèle pour le RSSI d'une station de base (notée X dans la suite) étant donné la position (notée Y dans la suite). Cette phase d'apprentissage est également appelée phase de calibrage dans la littérature. Une fois que ce modèle a été appris, on peut prédire la position de l'émetteur comme étant celle qui correspond au mieux au RSSI mesuré. Les secondes sont les méthodes basées sur les "fingerprints". Ces derniers font directement correspondre le RSSI à la position au moyen d'une fonction qui a été préalablement apprise sur un ensemble de données.

Les récentes avancées de l'apprentissage automatique et ses succès dans un large éventail de domaines ont poussé la communauté IoT à appliquer ces méthodes à la géolocalisation basée sur le RSSI. Présentons maintenant les contributions que nous avons développées pour répondre à cette problématique.

- Nous proposons d'abord des améliorations des méthodes utilisées pour le problème de géolocalisation basé sur le RSSI. La première technique proposée repose sur un estimateur semi-paramétrique de Nadaraya-Watson de la vraisemblance, suivi d'un estimateur du maximum a posteriori de la position de l'objet. La seconde technique consiste à apprendre une distance, construite au moyen d'un régresseur de type Gradient boosting : un algorithme de k-plus proches voisins est alors utilisé pour estimer la position. Les méthodes proposées sont comparées sur deux

jeux de données provenant du réseau Sigfox, et sur un jeu de données intérieur provenant d'un bâtiment de trois étages. Les expériences démontrent l'intérêt des méthodes proposées, tant en termes de performance d'estimation de la position, que de capacité à établir des régions de confiance sur nos estimées. Les résultats montrent également que la qualité de la prédiction est fortement liée à la distance choisie sur l'espace RSSI. Le problème de l'apprentissage métrique est donc une question fondamentale pour améliorer la technique de géolocalisation basée sur le RSSI.

- Deuxièmement, nous introduisons un objectif original pour apprendre une similarité entre des paires de points de données. Dans ce manuscrit, nous proposons de construire la similarité en minimisant directement l'erreur de régression d'un estimateur. Nous obtenons ainsi un objectif d'apprentissage orienté vers la tâche. Pour le minimiser, la similarité est choisie comme une somme d'arbres de régression et est apprise séquentiellement au moyen d'une version modifiée de XGBoost détaillée dans ce document. Cette méthode bénéficie des qualités bien connues de XGBoost, telles que son efficacité et ses capacités de mise à l'échelle. De plus, notre similarité, bien que non-paramétrique, ne nécessite pas un stockage de la taille du jeu de données. Enfin, les expériences montrent que notre modèle surpasse les autres modèles de régression à noyau sur plusieurs jeux de données de référence.

- L'indépendance conditionnelle a été largement utilisée dans la littérature sur la géolocalisation basée sur le RSSI afin de réduire la complexité des modèles statistiques tels que ceux présentés dans ce manuscrit. Tester l'IC est donc essentiel pour la performance de tels estimateurs. Nous introduisons la fonction de copule partielle pondérée pour tester l'indépendance conditionnelle. La procédure de test proposée résulte des ingrédients suivants : (i) la statistique de test est une transformation explicite de Cramér-von Mises de la copule partielle pondérée, (ii) les régions de rejet sont calculées à l'aide d'une procédure bootstrap qui imite l'indépendance conditionnelle en générant des échantillons. Sous CI, la faible convergence du processus de la copule partielle pondérée est établie et confirme la solidité de notre approche. Les expériences démontrent enfin la compétitivité de notre approche par rapport aux méthodes récentes de l'état de l'art.

Bibliography

Uzair Ahmad, Andrey Gavrilov, Sungyoung Lee, and Young-Koo Lee. Modular multilayer perceptron for wlan based localization. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pages 3465–3471. IEEE, 2006. page 48

H. Ahmadi and R. Bouallegue. Comparative study of learning-based localization algorithms for wireless sensor networks: Support vector regression, neural network and naïve bayes. In *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 1554–1558, Aug 2015. doi: 10.1109/IWCMC.2015.7289314. page 44

Alon Amar and Anthony J Weiss. Localization of narrowband radio emitters based on doppler frequency shifts. *IEEE Transactions on Signal Processing*, 56(11):5500–5508, 2008. pages 14, 106

Cédric Artigue. Method for searching for a useful signal in a multiplexing band, September 19 2017. US Patent 9,768,897. page 25

Francis R Bach and Michael I Jordan. Learning graphical models with mercer kernels. In *Advances in Neural Information Processing Systems*, pages 1033–1040, 2003. page 77

Paramvir Bahl, Venkata N Padmanabhan, Victor Bahl, and Venkat Padmanabhan. Radar: An in-building rf-based user location and tracking system. 2000. page 44

Paolo Barsocchi, Stefano Lenzi, Stefano Chessa, and Gaetano Giunta. A novel approach to indoor rssi localization by automatic calibration of the wireless propagation model. In *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pages 1–5. IEEE, 2009. page 47

Richard C Bell, Philippa E Pattison, and Graeme P Withers. Conditional independence in a clustered item test. *Applied Psychological Measurement*, 12(1):15–26, 1988. page 77

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013. pages 45, 64

Rudolf Beran, Martin Bilodeau, and P Lafaye de Micheaux. Nonparametric tests of independence between random vectors. *Journal of Multivariate Analysis*, 98(9):1805–1824, 2007. page 78

Wicher Bergsma. Nonparametric testing of conditional independence by means of the partial copula. *Available at SSRN 1702981*, 2010. page 79

Patrick Billingsley. Probability and measure. 1995. *John Wiley&Sons, New York*, 1995. page 81

Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007. page 56

Igor Bisio, Fabio Lavagetto, Andrea Sciarrone, and Simon Yiu. A smart 2 gaussian process approach for indoor localization with rssi fingerprints. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017. page 45

Cyril Botteron. *A statistical analysis of the performance of radio location techniques*. Graduate Studies, 2003. page 45

R Boucher and J Hassab. Analysis of discrete implementation of generalized cross correlator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3): 609–611, 1981. pages 17, 21, 109

Taoufik Bouezmarni, Jeroen VK Rombouts, and Abderrahim Taamouti. Nonparametric copula-based test for conditional independence with applications to granger causality. *Journal of Business & Economic Statistics*, 30(2):275–287, 2012. page 79

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996. page 38

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. page 48

Mauro Brunato and Roberto Battiti. Statistical learning theory for location fingerprinting in wireless lans. *Computer Networks*, 47(6):825–845, 2005. page 47

Mussa Bshara, Umut Orguner, Fredrik Gustafsson, and Leo Van Biesen. Fingerprinting localization in wireless networks based on received-signal-strength measurements: A case study on wimax networks. *IEEE Transactions on Vehicular Technology*, 59(1): 283–294, 2010. pages 46, 47

Axel Bücher and Holger Dette. A note on bootstrap approximations for the empirical copula process. *Statistics & probability letters*, 80(23-24):1925–1932, 2010. page 78

S. Campbell, N. O'Mahony, L. Krpalcova, D. Riordan, J. Walsh, A. Murphy, and C. Ryan. Sensor technology in autonomous vehicles : A review. In *2018 29th Irish Signals and Systems Conference (ISSC)*, pages 1–4, June 2018. doi: 10.1109/ISSC.2018.8585340. page 44

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection series b statistical methodology. 2018. page 79

G Carter and C Knapp. Time delay estimation. In *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 357–360. IEEE, 1976. page 20

G Clifford Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75 (2):236–255, 1987. page 20

G Clifford Carter, Albert H Nuttall, and Peter G Cable. The smoothed coherence transform. *Proceedings of the IEEE*, 61(10):1497–1498, 1973. page 20

Luca Catarinucci, Danilo De Donno, Luca Mainetti, Luca Palano, Luigi Patrono, Maria Laura Stefanizzi, and Luciano Tarricone. An iot-aware architecture for smart healthcare systems. *IEEE internet of things journal*, 2(6):515–526, 2015. pages 15, 108

Gayathri Chandrasekaran, Mesut Ali Ergin, Jie Yang, Song Liu, Yingying Chen, Marco Gruteser, and Richard P Martin. Empirical evaluation of the limits on localization using signal strength. In *2009 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pages 1–9. IEEE, 2009. pages 18, 109

Bharat S Chaudhari, Marco Zennaro, and Suresh Borkar. Lpwan technologies: Emerging application characteristics, requirements, and design considerations. *Future Internet*, 12(3):46, 2020. page 25

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016. pages 37, 39

Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005. pages 63, 65

Siu-Kay Chow and P Schultheiss. Delay estimation using narrow-band processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):478–484, 1981. page 21

Li Cong and Weihua Zhuang. Hybrid tdoa/aoa mobile user location for wideband cdma cellular systems. *IEEE Transactions on Wireless Communications*, 1(3):439–447, 2002. pages 14, 106

Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, Alzheimer's Disease Neuroimaging Initiative, et al. Benchmarking functional connectome-based predictive models for resting-state fmri. *Neuroimage*, 192:115–134, 2019. page 90

Huan Dai, Wen-hao Ying, and Jiang Xu. Multi-layer neural network for received signal strength-based indoor localisation. *IET Communications*, 10(6):717–723, 2016. page 48

Marzieh Dashti, Simon Yiu, Siamak Yousefi, Fernando Perez-Cruz, and Holger Claussen. Rssi localization with gaussian processes and tracking. In *2015 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2015. page 45

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007. page 64

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979. page 80

Paul Deheuvels. An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11(1):102–113, 1981. page 78

Amir Dembo, Thomas M Cover, and Joy A Thomas. Information theoretic inequalities. *IEEE Transactions on Information theory*, 37(6):1501–1518, 1991. page 33

Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000. page 39

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014. page 79

JE Ehrenberg, TE Ewart, and RD Morris. Signal-processing techniques for resolving individual pulses in a multipath signal. *The Journal of the Acoustical Society of America*, 63(6):1861–1865, 1978. page 19

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996. page 84

Zahid Farid, Rosdiadee Nordin, and Mahamod Ismail. Recent advances in wireless indoor localization techniques and system. *Journal of Computer Networks and Communications*, 2013, 2013. page 43

Wissam Farjow, Abdellah Chehri, Mouftah Hussein, and Xavier Fernando. Support vector machines for indoor sensor localization. In *2011 IEEE Wireless Communications and Networking Conference*, pages 779–783. IEEE, 2011. page 43

Jean-David Fermanian, Dragan Radulovic, Marten Wegkamp, et al. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860, 2004. pages 78, 81

Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925. page 31

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996. page 39

Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999. page 39

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000. page 41

Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002. pages 37, 40, 48

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004. page 79

Christian Genest and Bruno Rémillard. Test of independence and randomness based on the empirical copula process. *Test*, 13(2):335–369, 2004. pages 78, 81

Christian Genest, Jean-François Quessy, and Bruno Rémillard. Local efficiency of a cramér–von mises test of independence. *Journal of Multivariate Analysis*, 97(1):274–294, 2006. page 78

Irène Gijbels, Noël Veraverbeke, and Marel Omelka. Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, 55(5): 1919–1932, 2011. page 78

Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan R Salakhutdinov. Neighbourhood components analysis. In *Advances in neural information processing systems*, pages 513–520, 2005. page 65

Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8(4), 2007. page 65

Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008. page 79

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *2009 IEEE 12th international conference on computer vision*, pages 498–505. IEEE, 2009. page 65

Brian Ferris Dirk Hähnel and Dieter Fox. Gaussian processes for signal strength-based location estimation. In *Proceeding of Robotics: Science and Systems*. Citeseer, 2006. pages 45, 50

Peter Hall and Susan R Wilson. Two guidelines for bootstrap hypothesis testing. *Biometrics*, pages 757–762, 1991. page 82

Peter Hall, Byeong U Park, and Richard J Samworth. Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, pages 2135–2152, 2008. page 48

J Hassab and R Boucher. A quantitative study of optimum and sub-optimum filters in the generalized correlator. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 124–127. IEEE, 1979. page 21

J Hassab and R Boucher. Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):549–555, 1981. page 20

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. page 72

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009. page 27

Matt Hatton. The iot in 2030. URL https://iotbusinessnews.com/2020/05/20/03177-the-iot-in-2030-24-billion-connected-things-generating-1-5-trillion/. pages 13, 105

KC Ho and YT Chan. Solution and performance analysis of geolocation by tdoa. *IEEE Transactions on Aerospace and Electronic Systems*, 29(4):1311–1322, 1993. pages 14, 106

Wassily Hoeffding. A non-parametric test of independence. *The annals of mathematical statistics*, pages 546–557, 1948. page 78

Ville Honkavirta, Tommi Perala, Simo Ali-Loytty, and Robert Piché. A comparative survey of wlan location fingerprinting methods. In *Positioning, Navigation and Communication, 2009. WPNC 2009. 6th Workshop on*, pages 243–251. IEEE, 2009. pages 47, 52

Martin Huber and Blaise Melly. A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics*, 30(7):1144–1168, 2015. page 77

Zohaib Iqbal, Da Luo, Peter Henry, Samaneh Kazemifar, Timothy Rozario, Yulong Yan, Kenneth Westover, Weiguo Lu, Dan Nguyen, Troy Long, et al. Accurate real time localization tracking in a clinical environment using bluetooth low energy and deep learning. *PloS one*, 13(10):e0205392, 2018. page 44

Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988. page 64

Giovanni Jacovitti and Gaetano Scarano. Discrete time techniques for time delay estimation. *IEEE Transactions on signal processing*, 41(2):525–533, 1993. pages 21, 22

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013. pages 31, 39

Thomas Janssen, Rafael Berkvens, and Maarten Weyn. Benchmarking rss-based localization algorithms with lorawan. *Internet of Things*, page 100235, 2020. page 45

Esrafil Jedari, Zheng Wu, Rashid Rashidzadeh, and Mehrdad Saif. Wi-fi based indoor location positioning employing random forest classifier. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, pages 1–5. IEEE, 2015. pages 44, 48

Kamol Kaemarungsi and Prashant Krishnamurthy. Properties of indoor received signal strength for wlan location fingerprinting. In *Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004. The First Annual International Conference on*, pages 14–23. IEEE, 2004. page 46

Purushottam Kar and Prateek Jain. Similarity-based learning via data driven embeddings. In *Advances in neural information processing systems*, pages 1998–2006, 2011. page 64

Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993. page 33

Dor Kedem, Stephen Tyree, Fei Sha, Gert R Lanckriet, and Kilian Q Weinberger. Nonlinear metric learning. In *Advances in neural information processing systems*, pages 2582–2590. Citeseer, 2012. page 65

Philipp W Keller, Shie Mannor, and Doina Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 449–456, 2006. page 64

Maurice George Kendall. Rank correlation methods. 1948. page 78

Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4): 320–327, 1976. pages 19, 20

Ivan Kojadinovic and Mark Holmes. Tests of independence among continuous random vectors based on cramér–von mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154, 2009. page 78

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. page 77

Brian Kulis et al. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364, 2012. pages 63, 64, 65

Pascal Lavergne and Valentin Patilea. Smooth minimum distance estimation and testing with conditional estimating equations: uniform in bandwidth theory. *Journal of Econometrics*, 177(1):47–59, 2013. page 78

Michel Ledoux and Michel Talagrand. Probability in banach spaces. classics in mathematics, 2011. page 75

Kuang-Yao Lee, Bing Li, and Hongyu Zhao. Variable selection via additive conditional independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1037–1055, 2016. page 77

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018. page 77

Xinrong Li. Rss-based location estimation with unknown pathloss model. *IEEE Transactions on Wireless Communications*, 5(12), 2006. page 46

Zan Li, Torsten Braun, Xiaohui Zhao, Zhongliang Zhao, Fengye Hu, and Hui Liang. A narrow-band indoor positioning system by fusing time and received signal strength via ensemble learning. *IEEE Access*, 6:9936–9950, 2018. page 48

Somayya Madakam, Vihar Lake, Vihar Lake, Vihar Lake, et al. Internet of things (iot): A literature review. *Journal of Computer and Communications*, 3(05):164, 2015. page 24

Sandy Mahfouz, Farah Mourad-Chehade, Paul Honeine, Hichem Snoussi, and Joumana Farah. Kernel-based localization using fingerprinting in wireless sensor networks. In *2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 744–748. IEEE, 2013. page 45

Sandy Mahfouz, Farah Mourad-Chehade, Paul Honeine, Joumana Farah, and Hichem Snoussi. Non-parametric and semi-parametric rssi/distance modeling for target tracking in wireless sensor networks. *IEEE Sensors Journal*, 16(7):2115–2126, 2015. page 45

Oded Z Maimon and Lior Rokach. *Data mining with decision trees: theory and applications*, volume 81. World scientific, 2014. page 37

Guoqiang Mao, Brian DO Anderson, and Barış Fidan. Path loss exponent estimation for wireless sensor network localization. *Computer Networks*, 51(10):2467–2483, 2007. page 23

Florian Markowetz and Rainer Spang. Inferring cellular networks–a review. *BMC bioinformatics*, 8(6):S5, 2007. page 77

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Celer: a fast solver for the lasso with dual extrapolation. *arXiv preprint arXiv:1802.07481*, 2018. page 70

Santiago Mazuelas, Alfonso Bahillo, Ruben M Lorenzo, Patricia Fernandez, Francisco A Lago, Eduardo Garcia, Juan Blas, and Evaristo J Abril. Robust indoor positioning provided by real-time rssi values in unmodified wlan networks. *IEEE Journal of selected topics in signal processing*, 3(5):821–831, 2009. page 46

Mahammad Shareef Mekala and P Viswanathan. A survey: Smart agriculture iot with cloud computing. In *2017 international conference on microelectronic devices, circuits and systems (ICMDCS)*, pages 1–7. IEEE, 2017. pages 15, 108

Piotr Mirowski, Philip Whiting, Harald Steck, Ravishankar Palaniappan, Michael MacDonald, Detlef Hartmann, and Tin Kam Ho. Probability kernel regression for wifi localisation. *Journal of Location Based Services*, 6(2):81–100, 2012. page 45

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. page 35

Amir Navot, Lavi Shpigelman, Naftali Tishby, and Eilon Vaadia. Nearest neighbor based feature selection for regression and its application to neural activity. In *Advances in neural information processing systems*, pages 996–1002, 2006. page 65

Nam Nguyen and Yunsong Guo. Metric learning: A support vector approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 125–136. Springer, 2008. page 64

Dragos Niculescu and Badri Nath. Ad hoc positioning system (aps) using aoa. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, volume 3, pages 1734–1743. Ieee, 2003. pages 14, 106

Yung-Kyun Noh, Masashi Sugiyama, Kee-Eung Kim, Frank Park, and Daniel D Lee. Generative local metric learning for kernel regression. In *Advances in Neural Information Processing Systems*, pages 2452–2462, 2017. pages 65, 72

D. Nolan and D. Pollard. U-processes: rates of convergence. *The Annals of Statistics*, 15(2):780–799, 1987. page 99

Dieter Oosterlinck, Dries F Benoit, Philippe Baecke, and Nico Van de Weghe. Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits. *Applied geography*, 78:55–65, 2017. page 44

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962. page 34

Neal Patwari. *Location estimation in sensor networks*. PhD thesis, University of Michigan, 2005. page 47

Michał Piórkowski and Matthias Grossglauser. Constrained tracking on a road network. In *European Workshop on Wireless Sensor Networks*, pages 148–163. Springer, 2006. page 45

François Portier and Johan Segers. On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, 166:160–181, 2018. pages 81, 84, 86, 99

Theodore S Rappaport et al. *Wireless communications: principles and practice*, volume 2. prentice hall PTR New Jersey, 1996. pages 22, 23

Usman Raza, Parag Kulkarni, and Mahesh Sooriyabandara. Low power wide area networks: An overview. *IEEE Communications Surveys & Tutorials*, 19(2):855–873, 2017. page 24

Bruno Rémillard and Olivier Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100(3):377–386, 2009. page 78

Hilary Richardson, Grace Lisandrelli, Alexa Riobueno-Naylor, and Rebecca Saxe. Development of the social brain from age three to twelve years. *Nature communications*, 9(1):1–12, 2018. pages 86, 89

Greg Ridgeway. Generalized boosted models: A guide to the gbm package. *Update*, 1 (1):2007, 2007. page 39

Margaret Rouse. Internet of things (iot). iot agenda, 2020. page 15

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. *arXiv preprint arXiv:1709.01447*, 2017. page 79

Ludger Ruschendorf. Asymptotic distributions of multivariate rank order statistics. *The Annals of Statistics*, pages 912–923, 1976. page 78

Frederik Hendrik Ruymgaart. Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics*, pages 892–910, 1974. page 78

Frits H Ruymgaart and MCA van Zuijlen. Asymptotic normality of multivariate linear rank statistics in the non-iid case. *The Annals of Statistics*, pages 588–602, 1978. page 78

Hazem Sallouha, Alessandro Chiumento, and Sofie Pollin. Localization in long-range ultra narrow band iot networks using rssi. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017. page 21

S Sujin Issac Samuel. A review of connectivity challenges in iot-smart home. In *2016 3rd MEC International conference on big data and smart city (ICBDSC)*, pages 1–4. IEEE, 2016. pages 15, 108

Kent Scarbrough, Nasir Ahmed, and GC Carter. On the simulation of a class of time delay estimation algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):534–540, 1981. page 21

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*, pages 41–48, 2004. page 64

Anton Schwaighofer, Marian Grigoras, Volker Tresp, and Clemens Hoffmann. Gpps: A gaussian process positioning system for cellular networks. In *Advances in Neural Information Processing Systems*, pages 579–586, 2004. page 45

Johan Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18(3):764–782, 2012. page 78

Johan Segers. Hybrid copula estimators. *Journal of Statistical Planning and Inference*, 160:23–34, 2015. page 103

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2951–2961, 2017. pages 79, 86

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. page 42

Jun Shao. *Mathematical statistics: exercises and solutions*. Springer Science & Business Media, 2006. page 33

Galen R Shorack and GR Shorack. *Probability for statisticians*. Number 04; QA273, S4. Springer, 2000. page 32

Winfried Stute. Nonparametric model checks for regression. *The Annals of Statistics*, pages 613–641, 1997. page 81

Liangjun Su and Halbert White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007. page 79

Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008. page 79

Federico Thomas and Lluís Ros. Revisiting trilateration for robot localization. *IEEE Transactions on robotics*, 21(1):93–101, 2005. pages 17, 109

Joaquín Torres-Sospedra, Raúl Montoliu, Sergio Trilles, Óscar Belmonte, and Joaquín Huerta. Comprehensive analysis of distance and similarity measures for wi-fi fingerprinting indoor positioning systems. *Expert Systems with Applications*, 42(23): 9263–9278, 2015. pages 47, 52

David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005. page 18

Alexandre Tsybakov. Apprentissage statistique et estimation non-paramétrique. Course, 2013. page 51

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, 2009. pages 34, 35

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996. pages 86, 99, 100, 102

Aad W. van der Vaart and Jon A. Wellner. Empirical processes indexed by estimated functions. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 234–252. Inst. Math. Statist., Beachwood, OH, 2007. page 100

Noël Veraverbeke, Marek Omelka, and Irène Gijbels. Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, 38:766–780, 2011. page 78

Thaddeus Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93, 1975. page 47

Gang Wang, H Chen, Youming Li, and Ming Jin. On received-signal-strength based localization with unknown transmit power and path loss exponent. *IEEE Wireless Communications Letters*, 1(5):536–539, 2012. page 45

Yanzhao Wang, Chundi Xiu, Xuanli Zhang, and Dongkai Yang. Wifi indoor localization with csi fingerprinting-based random forest. *Sensors*, 18(9):2869, 2018. page 43

Kilian Q Weinberger and Gerald Tesauro. Metric learning for kernel regression. In *Artificial Intelligence and Statistics*, pages 612–619, 2007. pages 64, 65

Yaqin Xie, Yan Wang, Arumugam Nallanathan, and Lina Wang. An improved k-nearest-neighbor indoor localization method based on spearman distance. *IEEE Signal Process. Lett.*, 23(3):351–355, 2016. pages 44, 52

Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003. pages 63, 64

Caiming Xiong, David Johnson, Ran Xu, and Jason J Corso. Random forests for metric learning with implicit pairwise position dependence. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 958–966. ACM, 2012. pages 64, 65

Yaming Xu, Jianguo Zhou, and Peng Zhang. Rss-based source localization when path-loss model parameters are unknown. *IEEE communications letters*, 18(6):1055–1058, 2014. page 45

Simon Yiu, Marzieh Dashti, Holger Claussen, and Fernando Perez-Cruz. Wireless rssi fingerprinting localization. *Signal Processing*, 131:235–244, 2017. pages 44, 45, 47, 48, 50, 55

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012. pages 79, 86, 91

Wei Zhang, Kan Liu, Weidong Zhang, Youmei Zhang, and Jason Gu. Deep neural networks for wireless localization in indoor and outdoor environments. *Neurocomputing*, 194:279–287, 2016. page 44

Hong Zhou, Bingwu Liu, and Donghan Wang. Design and research of urban intelligent transportation system based on the internet of things. In *Internet of Things*, pages 572–580. Springer, 2012. pages 15, 108

Mu Zhou, Yanmeng Wang, Zengshan Tian, Yinghui Lian, Yong Wang, and Bang Wang. Calibrated data simplification for energy-efficient location sensing in internet of things. *IEEE Internet of Things Journal*, 6(4):6125–6133, 2018. page 45

Judit Tamás Zsolt Tóth. Miskolc iis hybrid ips: Dataset for hybrid indoor positioning. In *26st International Conference on Radioelektronika*, pages 408–412. IEEE, 2016. pages 46, 53

ECOLE DOCTORALE

Titre : Apprentissage statistique pour la géolocalisation d'objets connectés

Mots clés : Apprentissage de Similarité, Géolocalisation par le réseau, Indépendance Conditionnelle, Méthodes de Boosting

Résumé : La géolocalisation par le réseau a suscité beaucoup d'attention ces dernières années. Dans un contexte où les signaux sont à bandes étroites, par exemple pour l'Internet des Objets, les techniques de géolocalisation basées sur le RSSI se distinguent
Nous proposons tout d'abord quelques méthodes pour le problème de la géolocalisation basée sur le RSSI. En particulier, nous introduisons un estimateur semi-paramétrique de Nadaraya-Watson de la vraisemblance, suivi d'un estimateur de maximum à postériori de la position de l'objet. Les expériences démontrent l'intérêt de la méthode proposée en termes de performance d'estimation. Une approche alternative est donnée par une régression de type k-NN qui utilise une métrique appropriée entre les vecteurs de RSSI. Nous nous intéressons donc au problème de l'apprentissage de similarité et nous introduisons un objectif spécifiquemet choisi pour améliorer la géolocalisation. La fonction de similarité est choisie comme une somme d'arbres de régression et est apprise séquentiellement au moyen d'une version modifiée de l'algorithme eXtreme Gradient Boosting (XGBoost).
La dernière partie de la thèse est consacrée à l'introduction d'un test d'hypothèse d'indépendance conditionnelle (IC). En effet, pour de nombreux estimateurs, les composantes des vecteurs RSSI sont supposées indépendantes sachant la position. La contribution est cependant fournie dans un cadre statistique général. Nous introduisons la fonction de copule partielle pondérée pour tester l'indépendance conditionnelle. La procédure de test proposée résulte des éléments suivants : (i) la statistique de test est une transformation de Cramér-von Mises de la copule partielle pondérée, (ii) les régions de rejet sont calculées à l'aide d'une procédure de "boot-strap" qui imite l'indépendance conditionnelle en générant des échantillons. Sous l'hypothèse nulle, la faible convergence du processus de la copule partielle pondérée est établie et confirme le bien-fondé de notre approche.

Title : Contributions to RSSI-based Geolocation

Keywords : Metric Learning, Network-Based Geolocation, Conditional Independence, Boosting Methods

Abstract : The Network-Based Geolocation has raised a great deal of attention in the context of the Internet of Things. In many situations, connected objects with low-consumption should be geolocated without the use of GPS or GSM. Geolocation techniques based on the Received Signal Strength Indicator (RSSI) stands out, because other location techniques may fail in the context of urban environments and/or narrow band signals.
First, we propose some methods for the RSSI-based geolocation problem. The observation is a vector of RSSI received at the various base stations. In particular, we introduce a semi-parametric Nadaraya-Watson estimator of the likelihood, followed by a maximum a posteriori estimator of the object's position. Experiments demonstrate the interest of the proposed method, both in terms of location estimation performance, and ability to build radio maps. An alternative approach is given by a k-nearest neighbors regressor which uses a suitable metric between RSSI vectors. Results also show that the quality of the prediction is highly related to the chosen metric. Therefore, we turn our attention to the metric learning problem. We introduce an original task-driven objective for learning a similarity between pairs of data points. The similarity is chosen as a sum of regression trees and is sequentially learned by means of a modified version of the so-called eXtreme Gradient Boosting algorithm (XGBoost).
The last part of the thesis is devoted to the introduction of a Conditional Independence (CI) hypothesis test. The motivation is related to the fact that for many estimators, the components of the RSSI vectors are assumed independent given the position. The contribution is however provided in a general statistical framework. We introduce the weighted partial copula function for testing conditional independence. The proposed test procedure results from the following ingredients : (i) the test statistic is an explicit Cramér-von Mises transformation of the weighted partial copula, (ii) the regions of rejection are computed using a boot-strap procedure which mimics conditional independence by generating samples. Under the null hypothesis, the weak convergence of the weighted partial copula process is established and endorses the soundness of our approach.

Institut Polytechnique de Paris
91120 Palaiseau, France