université
PARIS-SACLAY

High dimensional pattern learning applied to symbolic time-series

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale de Mathématique Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques appliquées
Unité de recherche : Centre Borelli (ENS Paris-Saclay), UMR 9010 CNRS
Référent : Ecole normale supérieure de Paris-Saclay

Thèse présentée et soutenue le 11 octobre 2021, par

Amir DIB

Au vu des rapports de :

Stephane Gaiffas
Professeur, Unversité Paris-Diderot (Laboratoire de Probabilités, Statistique et Modélisation)
Jean-Michel Loubès
Professeur, Université Toulouse Paul Sabatier (Institut de mathématiques de Toulouse)

Composition du jury :

Laurent Oudre Professeur, ENS Paris-Saclay (Centre Borelli)	Président
Stephane Gaiffas Professeur, Unversité Paris-Diderot (Laboratoire de Probabilités, Statistique et Modélisation)	Rapporteur
Jean-Michel Loubès Professeur, Université Toulouse Paul Sabatier (Institut de mathématiques de Toulouse)	Rapporteur
Eli Upfal Professeur, Brown University	Examaniteur
Mathilde Mougeot Professeur, ENS Paris-Saclay (Centre Borelli)	Co-directrice
Nicolas Vayatis Professeur, ENS Paris-Saclay (Centre Borelli)	Directeur

Thèse de doctorat

NNT: 2021UPASM038

Acknowledgments

My first thanks go to my supervisors Nicolas Vayatis, Mathilde mougeot and Héloise Nonne for the opportunity to work on this most interesting problem. I also want to express gratitude towards Jean-Michel Loubès and Stéphane Gaiffas for doing me the honor of reviewing this thesis. The same gratitude goes for Eli Upfal to have accepted to be a jury member of my thesis committee and Laurent Oudre for presiding it.

To all the member and staff of the Borelli Center, thank you for these years of - sometimes heated - discussions and good moments. Thanks to Véronique Almodovar, Alina Muller, Sandra Doucet, Batiste Le Bars, Myrto Limnios, Etienne Boursier, Alice Nicolaï, Firas Jarboui, Antoine Mazarguil, Pierre-Yves Masse, Brian Tervil, Mona Michaud, Ludovic Minvielle, Quentin Laborde, Mounir Atiq, Ioannis Bargiotas, Miguel Colom, Mathilde Fekom, Antoine de Mathelin, Mathieu Jedor, Guillaume Richard, Albane Moreau, Vianney Perchet, Argyris Kalogeratos, Julien Audiren, Théo Saillant. Working with you was a pleasure, I hope that we'll meet again.

A special thank to Cyrus Cousins for our collaboration and the inspiring discussions about math, space and things of life.

Finally, I'm grateful to the teams of the HIA Percy and IRBA research center to have allow my colleagues Alice Nicolai, Antoine Mazarguil, Brian Terville, Pierre-Yves Masse, Albane Moreau and I to work on the wonderful ONADAP and Covid-related projects. In these special times, thank you for giving us the opportunity to do our part.

Abstract

While the adoption of machine learning in many applied contexts has been growing rapidly in the last decade, there remain challenges to use it in certain industrial settings. The main reason is the clash between established historical procedures with the uncertainty and lack of transparency of a machine learning pipeline's decision process. Another reason is that the input needed to feed a traditional machine learning model does not fit the available type or quality of available data. Most industrial databases have not been developed for statistical analysis but to comply with the regulatory requirements and to perform administrative tasks. In particular, non-numerical or symbolic features are common as it is a versatile way of recording events of interest. Examples of such data are textual documents, sequence of log-events or DNA sequences. The exponential number of possible patterns typically dominates the complexity associated with learning relevant information from symbols.

This thesis's applicative framework and primary motivation is to design efficient, human-readable and computationally tractable methods for predictive maintenance on the french train fleet. To that end, we propose to go beyond standard approaches by using a combination of traditional machine learning algorithms with pattern mining techniques to allow human experts to understand and interact with the algorithmic layer of the predictive maintenance pipeline. This thesis's main objective is to tackle these issues by proposing approaches that can be generally applied to a symbolic sequence of data with a human-readable output and trained at a reasonable computational cost. To that end, we begin by constructing a complete machine learning pipeline solution for predictive maintenance on a large fleet of rail vehicles that can be computed at a reasonable cost and provides valuable insight on the underlying symbol dynamic of the degradation process. As a second contribution, we propose a new method for symbolic data set based on a Bayesian generative model for patterns that can increases score accuracy in an interpretable fashion for any symbolic data set. As a third contribution, we introduce a new progressive mining method based on local complexities to obtain sharper statistical bounds on the pattern frequency. Finally, a new and general stochastic optimization method based on alternative sampling is proposed. This

method can be applied to the specific use case of Bayesian learning through the Variational Inference setting. In this instance, we provide theoretical and empirical proof of the superiority of this approach compared to the most advanced methods.

Contents

Part I

Introduction

Chapter 1

Scope and motivation of thesis

1.1 Context of thesis

General context. While the adoption of machine learning in many applied contexts has been growing rapidly in the last decade, there remain challenges to use it in certain industrial settings. The main reason is the clash between established historical procedures with the uncertainty and lack of transparency of a machine learning pipeline's decision process. Another reason is that the data standards needed to feed a traditional machine learning model do not fit the available type or quality of available data. Most industrial databases have not been developed for statistical analysis but to comply with the regulatory requirements and to perform administrative tasks. In particular, non-numerical or symbolic features are common as it is a versatile way of recording events of interest. Examples of such data are textual documents, sequence of log-events or DNA sequences. This thesis's main objective is to tackle these issues by proposing approaches that can be generally applied to a symbolic sequence of data with a human-readable output and trained at a reasonable computational cost.

Predictive maintenance for the French Fleet of Trains. This thesis is sponsored by the Société Nationale des Chemins de fer Français (SNCF), literally *French National Railway Company*, the state-owned railway company which operates all French railway traffic in France. Each day in France, 15000 trains operate. Paris urban area alone counts 3.2 million travelers a day and 60000 stops at train stations. SNCF have to deal with a context of increasing mass transit: in the last ten years, the number of travels in Paris increased by 30%. This context puts increasing pressure on the railway network and calls for a more automated approach toward maintenance. In the last few years, SNCF developed an alerting system based on carefully constructed rules from experts. Even though successful,

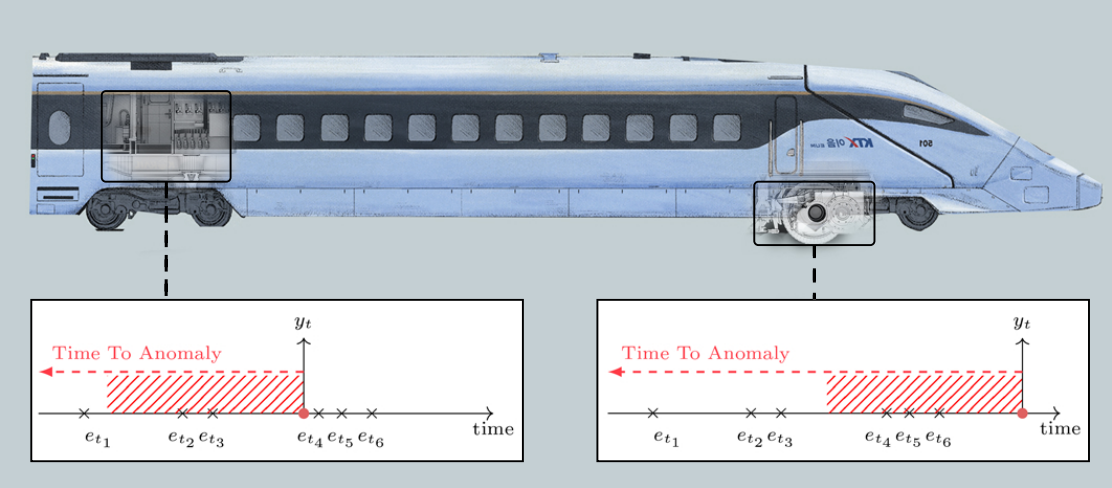


Figure 1.1: A rail vehicle is a complex electromechanical system composed of several subsystems. The figure shows the transformer (left) and engine block (right) subsystems. Each of them is composed of many components that emits time-stamped log events or *error codes* (e_t) at different times t. A breakdown or anomaly Y_t^S at time t can be linked to a specific subsystem S.

this approach is time-consuming and does not allow for automatic discovery of new rules that are not already known. Moreover, a set of rules designed by this method is specific to a class of vehicle and cannot be applied to new equipment.

1.2 Motivations

The task of *predictive maintenance* aims to anticipate critical failures of a large industrial system to plan early and cost-effective interventions. The method for preventing the critical failure of a component during exploitation was historically based on *preventive maintenance*. Knowing the average lifetime or law of deterioration of the component, repairs are planned to reduce the chance of unanticipated equipment failures. It is a step forward from *reactive maintenance*, which will only replace and maintain equipment in case of observed failure. Predictive maintenance is a broad term for performing equipment maintenance based on observed or recorded signs of deterioration. More precisely, it is a maintenance strategy that monitors the health condition of machinery in real-time and makes an optimal maintenance decision. Even though predictive maintenance leads to greater availability and reduced costs, it needs much more time, effort and resources to be performed. A high level of skills is required to collect, model, and interpret the data and reorganize the maintenance process.

Predictive maintenance for rolling stocks. This thesis's primary goal is to construct an end-to-end rolling stocks predictive maintenance solution from data collection to prediction. Trains are complex electromechanical systems that use many interconnected components to offer short and secured travel for passengers and ought to be energy-efficient. In France, a good covering of the territory implies exposition to the possibly harsh environment (regarding the topology of the tracks, weather) and is thus exposed to high failure rates. The case for predictive maintenance is particularly crucial in this context since the impact of a rolling stock failure generally has global consequences on the entire railway system. Because the train operates on a highly interconnected network, any malfunction leads to the complete immobilization of the train and propagates delays to a large portion of the transportation network. In that regard, the railway system makes a particularly relevant case for the added value of a predictive maintenance system.

In the context of SNCF, one of the challenges was to identify a set of relevant features that can inform about the deterioration state of the train. The sequence of a particular set of events, the sequences of *error codes*, was identified to be especially informative. Error codes are time-stamped strings of text emitted at regular or irregular intervals by a train's specific system. The emission of a particular type of code corresponds to a (sometime arcane) manufacturer's rule. For instance, on the train door's system, a code emission can correspond to the crossing of a threshold for the door's DC motor voltage response. Note that there is a slight abuse of language in the use of the term *error code* since an error code does not necessarily inform on a malfunction but can indicate the nominal functioning of a system. One of the main advantages of this pattern is that expert uses it for *a posteriori* diagnostics of a failure. When a specific train breaks down, it is sent to the maintenance factory for inspection. To determine the cause of the breakdown, the logs are pulled down from the systems and analyzed by the maintainer. The expert search for specific *patterns* and known recurrences of error codes in these codes to track the malfunction's root cause. We underline that this procedure is a widely used in practice for predictive maintenance in industrial context beyond the railway domain such as the automobile industry (Sung et al., 2020), manufacturing processes (Gutschi et al., 2019) or anomaly detection on various IT systems (Wang et al., 2017a; Wang, Vo, and Ni, 2015; Zhang et al., 2016).

Machine learning for symbolic data. Most of our daily tasks such as speech, reading, or episodic memory usage, rely on symbolic rather than numerical data. What fundamentally differentiates symbolic from numerical data is the *ordering* property. For instance, there is a natural way to compare two physical measurements of an electrical signal but none to compare two symbols. This type of data is ubiquitous in broad range of domains such biology with DNA and RNA tran-

scription (Schölkopf, Tsuda, and Vert, 2004; Aubin-Frankowski and Vert, 2020), chemistry for molecular structure prediction and classification (Elton et al., 2018), graph analysis (Mansha et al., 2016; Shang et al., 2017; Zheng et al., 2013), and in music theory to extract patterns that have the same harmonic function (Rompré, Biskri, and Meunier, 2017).

In general, symbolic data are not suitable for most of the machine learning algorithms as a common hypothesis made in machine learning theory is that the d-dimensional feature vector is a random variable valued in $\mathbb{R}^d$. A first approach is to consider kernel methods (Kung, 2014) which extends the use of common machine learning techniques to non-numerical data. More precisely, it relies on choosing a kernel function that maps the symbols data in a structured space. The principal drawbacks of kernel methods are the difficulty to interpret the results, which is a requirement for a predictive solution to be used in an industrial context. A second approach consists of transforming the process to a numerical one by aggregating (by counting or considering some statistics) the events one a chosen time windows and has been widely used for Anomaly Detection (He et al., 2016; Bogojeski et al., 2020; Aggarwal et al., 2018; Laredo et al., 2019). Even though popular (Basora, Olive, and Dubot, 2019), classification based solely on this construction is often unable to capture critical patterns of events that can be highly relevant in predictive maintenance. More crucially, it does not directly provide explainable output in terms of sets of log event or *patterns*.

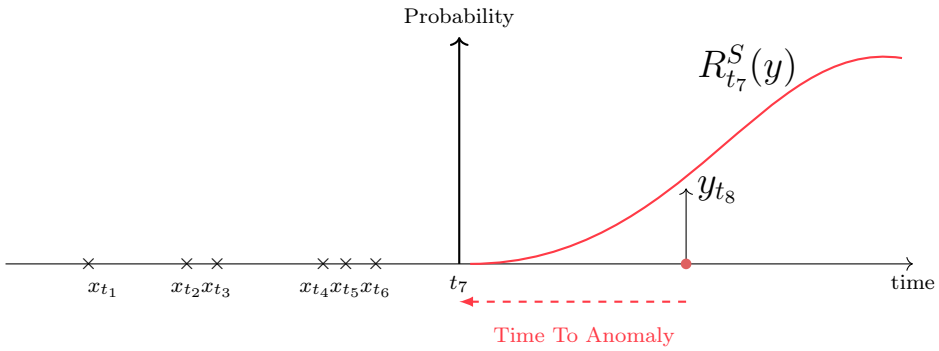


Figure 1.2: Regression function $R^S_{t_7}$ at time t_7 (red line) given the events $(x_{t_1}, \ldots, x_{t_6})$ on the subsystem S. At t_7, the past events are used to produce the density probability function of a breakdown appearing in the future. This density is compared with the true occurrence of an anomaly y_{t_8}.

1.3 Background

This section formally introduces predictive maintenance as a statistical regression based on symbolic data. The pattern mining task is then presented and reformulated as a Bayesian inference problem. Finally, the *stochastic optimization* procedure is described with a highlight on variance reduction methods.

1.3.1 Symbolic time series for predictive maintenance

As mentioned, symbolic data play a crucial role in predictive maintenance thanks to their versatility and historical use by the maintainers. Formally, the errors codes are a finite dictionary or set $E = (c_1, \ldots, c_d)$ of size d. At a time $t \in \mathbb{R}_+$, an event can be emitted by the subsystem S_i. The subsystem identifier define the block of component involved (such as the Engine Block) as well as the train and vehicle identifier. Finally, the feature space must be enriched by information that are correlated with the underlying degradation process (see chapter 2). In our application and in general it will generally consists of a real vector in $\mathbb{R}^K$. Example of such context is, for instance, the number of kilometers since last maintenance, weather data or additional context information at the time of the error code emission. We denote $X^S_t = \mathcal{E} \times \mathbb{R}_K$ the description space the subsystem S at time t with $\mathcal{E} = \mathcal{P}(E)$ being the set of all subsets of E. At time $t \in \mathbb{R}_+$ we can observe the occurrence of a breakdown $Y^S_t \in \{0, 1\}$ on the subsystem S. The goal of any predictive maintenance algorithm is to compute the *regression function* at each time t defined as

$$R^S_t(y) = \mathbb{P}[Y^S_t = y | (X^S_{t_0})_{t_0 \leq t}], \tag{1.1}$$

where $y \in \{0, 1\}$ denotes a set of malfunction. Figure C.2 illustrates the construction of such function. At time $(t_1, \ldots, t_6)$ the error codes $(e_{t_1}, \ldots, e_{t_6})$ are emitted and enriched to produce $(x_{t_1}, \ldots, x_{t_6})$. At t_7, the regression function estimates the probability of occurrence of a breakdown on the subsystem S for each time in the future. A broad range of techniques based on stochastic process model (Guan, Tang, and Xu, 2016; Chen et al., 2016; Cha and Pulcini, 2016), kernel methods (Kung, 2014) or deep learning approaches (Guo et al., 2017; Liu et al., 2018; Karpat et al., 2020) can be used to model such regression function. As mentioned, all these methods suffer for poor explainability and are incompatible with established maintenance processes which are based on pattern of codes. The goal is thus to constructed model based on small sets of codes that occur shortly and specifically before failures, which is a challenging task. Finding these combinations of codes is typically intractable due to the exponential number of possible patterns. It is thus necessary to resort the class of *pattern mining* techniques (Agrawal, Imielinski, and Swami, 1993).

1.3.2 Background on pattern mining

The Data Mining domain stems from the need for computational tools to extract useful information from large databases collected by administrations and industries. These databases are typically large records of numerous variables or *features* primarily constructed for administrative tasks such as accounting and regulatory compliance.

Deterministic approaches. The seminal work of (Agrawal, Imielinski, and Swami, 1993) on Frequent Itemset Mining (FIM) for basket analysis sparked interest as it offers a tractable procedure to tackle a real-world problem with vast commercial application. The problem posed was to find with a given level of precision, the association or *patterns* of common products that were bought together based on a database of purchases. Given a number d of possible articles to purchase, and a database of receipts, the complexity associated with querying the database to find the number of time each pattern of products were bought together, or *support*, is in $\mathcal{O}(2^d)$. The computation of such patterns is thus intractable even for a moderate-sized *dictionary* of itemsets. The proposed solution was to exploit the *antimonoticity* of the set of patterns $\mathcal{E}$: for two patterns $x, y \in \mathcal{E}$, if x derives from y in the sense that $x \subseteq y$ then the support of y is no greater than the support of x. Setting a *minimum support threshold* $\mu \in [0, 1]$, an algorithm can mine the support of the itemsets in a breadth-first search fashion (Zuse, 1972; Moore, 1959) by generating new pattern *candidates* at each step and halt the tree exploration whenever it encounters a pattern with support less to μ. This procedure constitutes the APRIORI algorithm (Agrawal and Srikant, 1994) and has been a

significant milestone for Data Mining related tasks. Even though APRIORI is an efficient algorithm when the average size of patterns present in the database is not too large (Hegland, 2007), it has several drawbacks. First, it requires multiple scans of the database for each evaluated pattern, and the need for computing a new set of patterns to test during the procedure leads to an exponential memory complexity of $\mathcal{O}(2^d)$. Improvements over the APRIORI algorithm such as ECLAT (Zaki, May-June/2000) proposes depth-search algorithm with a vertical data format which alleviate the need for multiple queries of the database. A different strategy for FIM has been taken by Han et al. (2004) called FP-TREE. The authors use a tree structure to encode the sorted set of transactions which allow for only two scans of the database. Crucially, the tree structure avoids generating unnecessary itemsets, leading to a much more memory-efficient procedure compared to APRIORI (Fournier-Viger et al., 2017). The CP-TREE (Tanbeer et al., 2008) algorithm extends FP-TREE by only requiring one scan of the database, reducing by a factor N the computational requirements.

We stress that FIM is the starting point of various techniques related to data mining tasks. For instance, Association Rule Mining (ARM) (Agrawal and Srikant, 1994; Zaki and Hsiao, 2005) considers the problem of finding rules between itemsets at a given confidence level. For two patterns $x, y \in \mathcal{E}$, the goal is to mine rules $x \to y$ such that the support $s(x \vee y)$ and confidence measure $c(x, y) = \frac{s(x \vee y)}{s(x)}$ are no greater than two threshold $\mu, \nu \in [0, 1]$. The confidence measure inform about the co-occurrence of two patterns while taking into account their frequency in the database. Episode Rule Mining (Mannila, Toivonen, and Verkamo, 1997; Zimmermann, 2014) consider the problem of finding the rules of the form $x \to y$ that appears regularly in an user-defined window. There has been numerous applications in anomaly and fraud detection (Qin and Hwang, 2004; Su, 2010; Wang et al., 2017b), sensor analysis (Li et al., 2017a), traffic data (Fournier-Viger et al., 2017) and in the medical field (Patnaik, Sastry, and Unnikrishnan, 2008). The Periodic Pattern Mining problem's goal is to extract patterns that repeat themselves over the transactions of the database (Venkatesh et al., 2016) and is commonly used for biomedical application (Zhang et al., 2007) and temporal sequence analysis (Sirisha, Shashi, and Raju, 2014). An original approach was taken by Vreeken, van Leeuwen, and Siebes (2010) by searching for the set of patterns that compresses best the database without loss. The resulting KRIMP algorithm first perform a FIM before using the Minimum Description Length principle to summarize the database. Finally, the Progressive Pattern Mining tasks consists of performing FIM on a properly sized subset of the database to approximate the support uniformly at a given level of confidence (Riondato and Upfal, 2015). We mention other methods that derive from FIM such as Sub-graph mining (Santhi and Padmaja, 2015), Discriminative Pattern Mining (Hämäläinen and Webb,

2019) and Sequential Pattern Mining (Fournier-Viger et al., 2017).

Bayesian approaches. The methods mentioned can successfully extract the pattern from a large database with efficient memory usage but still have a time exponential computational complexity for low support threshold μ since the problem has been shown to be NP-hard (Yang, 2004). Additionally, these models do not assume any stochasticity on the underlying process generating the database. In contrast, in the vast majority of cases, the transactions can be viewed as the result of an underlying but unknown generative process. As a result, no probabilistic confidence interval can be derived to assess the results' statistical significance.

Generative models have been proposed to perform various FIM tasks to address these fundamental issues. The multivariate tree distribution model (Hegland, 2007) fit a probability distribution on the $\binom{d}{2}$ pairwise items and a tree structure on the attributes. Fowkes and Sutton (2016) use a Bayesian Network Model to model the transaction database. Since the inference requires solving for the intractable weight covering problem (Korte and Vygen, 2006), the authors used a greedy approximation to infer interesting itemsets. Pavlov, Mannila, and Smyth (2003) empirically compare several generative models such as the independence model (Hegland, 2007), the multivariate tree distribution model (Chow and Liu, 1968), and the Mixture model in the equivalent framework of sparse binary dataset querying. Notably, the use of these probabilistic approaches goes beyond FIM and can serve as a tool to derive convergence bound for APRIORI-like algorithms (Hegland, 2007). Note that these approaches are closely related to the MDL principle (Vreeken, van Leeuwen, and Siebes, 2010) since the entropy of a probability model define the maximum lossless compression achievable by any compression algorithm (we refer the interested reader to (Friedman, Geiger, and Goldszmidt, 1997; Lam and Bacchus, 1994)).

Mode inference. Contrary to the deterministic approach, the probabilistic methods rely on the assumption that the database $\mathcal{D}$ is the outcome of a stochastic process. This assumption opens up the possibility for applying common statistical tools to infer the set of frequent itemsets. The common goal of all these methods is to find for every pattern x the probability distribution of the support $p(x|z, \mathcal{D})$. Given the generative model, finding a closed formula to compute $s(x)$ can be difficult and often involves intractable enumeration of all possible patterns (Fowkes and Sutton, 2016). Considering simpler models such as *mixture models* (Hegland, 2007) solve this issue and permit to control the complexity of this computation by the choice of the number of components in the mixture distribution. Under this representation the task of extracting the most frequent pattern becomes a *Bayesian optimization* task. The next sections formally describe the technical framework of

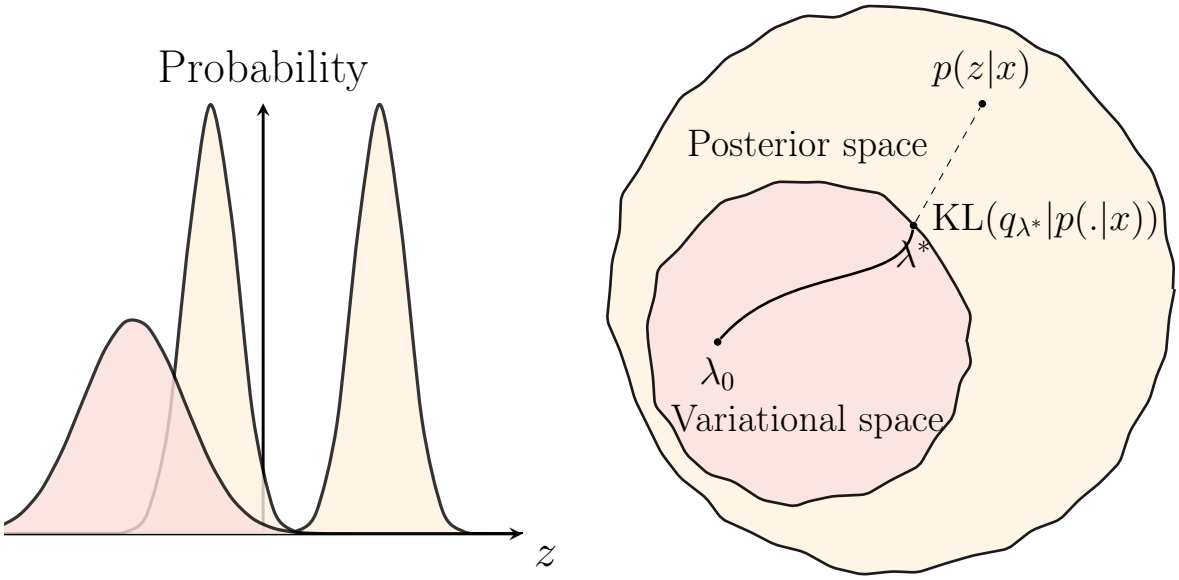


Figure 1.3: **Variational inference**. Left: The variational distribution q_λ (orange) parametrized by λ and the true posterior distribution $p(z|x)$ (green). Right: The Variational Inference procedure consists of finding the optimal λ^*, starting from λ_0 to minimize the Kullback–Leibler divergence between the true posterior and the variational distribution (represented as dashed line).

such inference and strategies to speed up the procedure.

1.3.3 Background on Bayesian statistics

In this section, we introduce the Bayesian statistics framework and basic notations. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(E, \|.\|)$ be a vector space equipped with the distance d and the induced norm $\|.\|$ and consider a random variable $X\colon (\Omega, \mathcal{A}) \longrightarrow (E, \mathcal{B}(E))$. In the Bayesian setting, the parameter space $\mathcal{Z}$ is equipped with a measure Π on $\mathcal{T}$ such that $(\mathcal{Z}, \mathcal{T}, \Pi)$ is a probability space and X is distributed according to a parametric model $\mathcal{P}_z$ from the parametric family of distribution $\mathcal{P} = \{P_z\colon z \in \mathcal{Z}\}$. In most cases, $\mathcal{Z}$ is a subset of an Euclidean space and applications often consider the d-dimensional real case $\mathcal{Z} \subset \mathbb{R}^d$. In addition, assume that for every z in $\mathcal{Z}$, the measures P_z and Π admit a density function such that

$$\begin{aligned} dP_z &= p(.|z)d\mu \\ d\Pi &= \pi d\nu, \end{aligned} \tag{1.2}$$

where μ, ν are σ-finite measures on respectively $\mathcal{B}(E)$ and $\mathcal{T}$. Then, the likelihood function $z \mapsto p(z|x)$ such that $p(z|x) = p(x|z)\pi(z)$ is a density with respect to

the product measure $\mu \otimes \nu$. The difference with the *frequentist* point of view is that the parameter z is itself a random variable distributed according to the *prior* distribution π and, conditionnaly on the data x, has the following distribution

$$p(z|x) = \frac{p(x|z)\pi(z)}{\int p(x|z)\pi(z)d\nu(z)}. \tag{1.3}$$

The Bayesian inference setting thus depends on the ability to simulate z from Equation C.3. Computing $p(z|x)$ requires evaluating the prior predictive distribution and thus integrating over all latent variables which lead to intractable computation (except in the prior conjugate case) even for simple models (Gelman et al., 2013). A common approach is to use methods such as Gibbs Sampling, Monte Carlo Markov Chain or Hamilton Monte Carlo (Betancourt, 2018; Homan and Gelman, 2014; Brooks et al., 2011) which rely solely on the unnormalized posterior distribution (freeing us from the need to compute $p(y)$) and the ability to sample from the posterior. These methods are consistent but associated with heavy computation, high sensitivity to hyperparameters and potential slow to converge to the true target distribution.

Variational inference

The posterior distribution in Equation C.3 can be exactly computed under some condition on the prior distribution when closed-form is available (Gelman et al., 2013). For most of the applications, such condition is not fulfilled, and one needs to resort to either asymptotically exact procedure or rely on approximation. One approximation approach that became the prominent framework for approximate Bayesian computation is Variational Inference (VI). It relies on building a proxy for the posterior distribution parametrized by a *variational family* distribution $Q = \{\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \Lambda\}$. In this method, a metric is chosen so that the distance between the true target distribution p and the variational distribution q is minimized. A common choice is the Kullback–Leibler (KL) divergence. Denoting x the data, z the latent variable space and $p(z|x)$ the likelihood, and $q_{\boldsymbol{\lambda}}$ the variational distribution parametrized by $\boldsymbol{\lambda}$, variational inference consists of a minimization problem (Saul, Jaakkola, and Jordan, 1996)

$$q_{\lambda^*} = \underset{q_{\boldsymbol{\lambda}} \in \mathcal{Q}}{\operatorname{argmin}} \operatorname{KL}\left(q_{\boldsymbol{\lambda}}(z) \| p(z|x)\right), \tag{1.4}$$

with $\operatorname{KL}\left(q_{\boldsymbol{\lambda}}(z) \| p(z|x)\right) = E_q[\log q_{\boldsymbol{\lambda}}(x) - \log p(z|x)]$ the Kullback–Leibler divergence. Even though KL remain the most used metric, other measures on the distribution space have been investigated (Ambrogioni et al., 2018). The reason for the popularity of such techniques is the fact that KL divergence can be linked

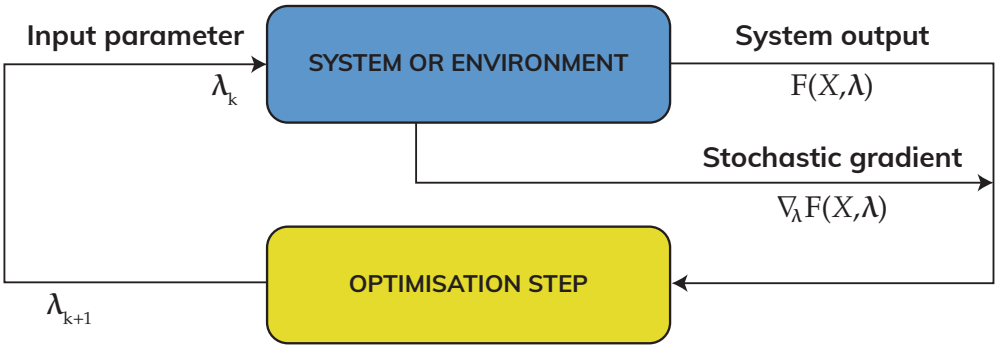


Figure 1.4: A typical stochastic optimization process composed of two steps; simulation (yellow) and optimization (green). The simulation phase produces a simulation of the stochastic system or interaction withthe environment, as well as unbiased estimators of the gradient (adapted from (Mohamed et al., 2020)).

to the Evidence Lower Bound (ELBO) that does not depend on the posterior distribution (Saul, Jaakkola, and Jordan, 1996)

$$\log p(y) = \mathrm{ELBO}(\boldsymbol{\lambda}) + \mathrm{KL}\left(q_{\boldsymbol{\lambda}}(z) \| p(z|x)\right), \tag{1.5}$$

where the ELBO is defined as

$$\mathrm{ELBO}(\boldsymbol{\lambda}) = \mathbb{E}_{z\sim q_{\boldsymbol{\lambda}}}\left[\log p(z,x) - \log q_{\boldsymbol{\lambda}}(z)\right]. \tag{1.6}$$

Since the marginal likelihood $p(y)$ does not depend on the parameters z, it follows that maximizing the ELBO with respect to $q_{\boldsymbol{\lambda}}$ leads to find the best approximation of $p(z|x)$ for the Kullback–Leibler (KL) divergence. Intuitively, this procedure minimizes the information loss subsequent to the replacement of the likelihood by $q_{\boldsymbol{\lambda}}$ but other distances can be used (Ambrogioni et al., 2018).

In practice, the distribution class $\mathcal{Q}$ is chosen in a distribution family that can be easily sampled from. A common choice is to pick from the normal distribution family $\mathcal{Q} = \{\mathcal{N}(\mu, \Sigma) \,|\, (\mu, \sigma) \in \mathbb{R}^K \times M_{K\times K}\}$ with $M_{K\times K}$ the space of symmetric positive-definite matrix on $\mathbb{R}^{K\times K}$. In this instance, performing VI consists of finding the optimal set of parameters (μ^*, Σ^*) such that equation C.6 is minimized.

Again, there is generally no closed formula for computing the ELBO or its gradient and one must rely on a Stochastic Optimization (Bottou, Curtis, and Nocedal, 2018) to perform this task. With this method, the minimization is carried by performing a Stochastic Gradient Descent (SGD) procedure on the ELBO objective function.

1.3.4 Stochastic optimization

One of the most prominent optimization problem in modern statistics consists of finding the root of an *objective function* which is an expectation of a random variable (Bottou, Curtis, and Nocedal, 2018). This problem has vast and known applications in Machine Learning (Bottou, Curtis, and Nocedal, 2018; Sutton and Barto, 2018; Gelman et al., 2013; Simsekli et al., 2019) but also in Finance for sensitivity analysis (Pagès, 2018; Glasserman, 2013), transport network management CITE and Supply Chain. Given a μ-distributed random variable $X\colon \Omega \longrightarrow E$ on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the general Stochastic Optimization problem reads as minimizing the following objective function

$$\begin{aligned} f(\boldsymbol{\lambda}) &= \mathbb{E}\left[F(X, \boldsymbol{\lambda})\right] \\ &= \int_E F(x, \boldsymbol{\lambda})\mu(dx), \end{aligned} \tag{1.7}$$

with respect to $\boldsymbol{\lambda} \in \mathbb{R}^K$ where $F\colon E \times \mathbb{R}^K \longrightarrow \mathbb{R}$ is a real function in $L_1(\Omega, \mathcal{A}, \mathbb{P})$. Under the regularity condition that f is continuous differentiable (or a least that a *sub-gradient* can be computed), this problem can be solved by finding the points where the gradient $\boldsymbol{g} = \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda})$ vanishes since $\boldsymbol{\lambda}^* \in \operatorname{argmin}\{\boldsymbol{g}_\lambda = 0\}$.
This problem can interpreted as optimizing a *cost* or *loss function* F with respect to $\boldsymbol{\lambda}$ with a noisy interference distributed according to μ. In Machine Learning applications (such as training a neural network), F represents the expected *loss* of a model parameterized by $\boldsymbol{\lambda}$ for a training set distributed according to μ. In this case, it has been shown that finding the optimal set of parameters $\boldsymbol{\lambda}^*$ is NP-hard even for a simple binary classification model (Feldman et al., 2012). More generally, the main difficulty in finding a solution to C.7 is that it involves computing a potentially high dimensional expectation which is prohibitively expensive. Even when the distribution is known, there is typically no closed-form available for computing the gradient. Nowadays, quadrature methods (Leader, 2004) to compute integral at given accuracy is feasible only for dimension up to ten or twenty which make it unusable for most modern application. Additionally, in most frameworks such as statistical learning, the distribution μ is unknown and only samples from the distribution μ are available.

Alternative sampling for the mean estimator.

Alternative sampling has been introdcuded to accelerate stochastic optimization procedures. Finding an approximation for the optimization problem in C.7 crucially depends on the ability to compute efficiently a sample dependent approximation of the expectation.

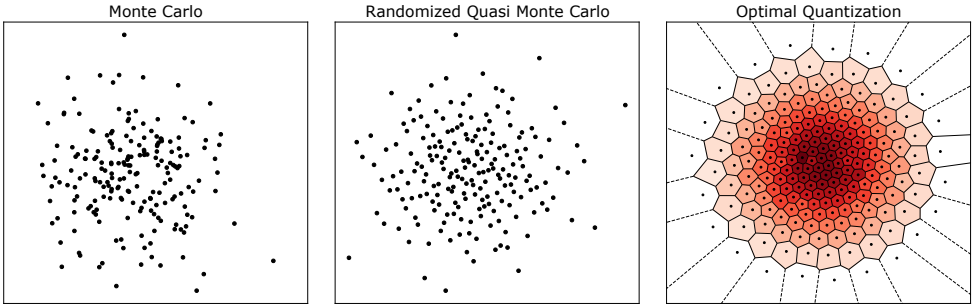


Figure 1.5: Monte Carlo (left), Randomized Monte Carlo (center) and Optimal Quantization with the associated Voronoi Cells (right), for a sampling size $N = 200$ of the bivariate normal distribution $\mathcal{N}(0, I_2)$. (Dib, 2020)

Monte Carlo. The most commonly used numerical procedure to approximate the expectation in C.7 is based on the Law of Large Number. It relies on replacing the expectation with an empirical mean estimator. Let $(X_1, \dots, X_n)$ be an *i.i.d.* sequence of μ_X-distributed random variable, F any measurable real valued function, and consider the following *Monte-Carlo* estimator

$$I_n^{MC} = \frac{1}{n} \sum_{i=1}^{n} F(X_i). \tag{1.8}$$

By the Strong Law of Large Numbers, I_n^{MC} converges towards $\mathbb{E}\big[F(X)\big]$ μ-almost surely and, provided that $F(X) \in L_2(\Omega, \mathcal{A}, \mathbb{P})$, at a rate of $\mathcal{O}(n^{-\frac{1}{2}})$ with quadratic error

$$\| I_n^{MC} - \mathbb{E}\big[F(X)\big] \|_{L_2(\Omega,\mathcal{A},\mathbb{P})} = \frac{\mathbb{V}F(X)}{\sqrt{n}}. \tag{1.9}$$

The Monte-Carlo method relies only on the ability to draw from distribution μ at reasonable cost. In addition, the Central Limit Theorem can be used to produce asymptotic confidence interval.

Quasi Monte-Carlo. Methods have been designed to improve on the convergence rate, mostly by considering alternative sampling methods for generating $(X_1, \dots, X_n)$. The most widely used are the Quasi Monte Carlo methods (Dick, Kuo, and Sloan, 2013). Theese methods are based on generating sequences of *pseudo-random numbers* that mimic the statistical properties of a target *i.i.d.* sequence of samples. More precisely, let X be a random variable which admits a density ψ with respect to the d-dimensional Lebesgue measure and consider an uniformly distributed random variable $U \sim \mathcal{U}([0,1]^d)$. Then, the random variable $\psi^{-1}(U)$ is distributed according to X and for every measurable function H we have

that $\mathbb{E}\left[H(X)\right] = \mathbb{E}\left[H \circ \psi^{-1}(U)\right]$. A *low-discrepancy* sequence $\boldsymbol{u} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$, with $(\boldsymbol{u}_i)_{i=1}^n$ valued in the d-dimensional hypercube $[0,1]^d$, is produced and evaluated through the inverse density probability distribution function ψ^{-1} (Pagès, 2018). Since $\boldsymbol{u}$ converges weakly towards the Lebesgue measure on $[0,1]^d$, the following holds for the QMC estimator

$$\begin{aligned} I_n^{QMC} &= \frac{1}{n}\sum_{i=1}^{n} F \circ \psi^{-1}(\boldsymbol{u}_i) \\ &\xrightarrow[n\to\infty]{} \mathbb{E}\big[F(X)\big]. \end{aligned} \tag{1.10}$$

Intuitively, if $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ is similar to the realization of an *i.i.d.* sequence of an uniformly distributed random variable, the sequence $(\psi^{-1}(\boldsymbol{u}_1), \dots, \psi^{-1}(\boldsymbol{u}_n))$ will be similar to the target *i.i.d.* set of samples $(X_1, \dots, X_n)$. The quality of such approximation is controled by the *star discrepancy* measure which is defined as the ℓ_∞ distance between the cumulative distribution of the empirical and Lebesgue measure

$$D_n^*(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n) = \sup_{\boldsymbol{b}\in[0,1]^d} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{\boldsymbol{u}\in[0,\boldsymbol{b}]\right\}} - \lambda_d([0,\boldsymbol{b}]) \right|. \tag{1.11}$$

For the a sequence $\boldsymbol{u}$ whose, the Hlawka-Koksma inequality (Koksma, 1942; Hlawka, 1961) states that the approximation error of C.10 is upper bounded by its discrepancy measure for h with finite variation. Since there are several sequences $\boldsymbol{u}$ which exhibit a discrepancy measure such that

$$D_n^*(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n) \leq c_d \frac{(\log n)^{d-1}}{n}, \tag{1.12}$$

the QMC estimator I_n^{QMC} can thus achieve much better convergence rate than the MC estimator I_n^{MC} of equation C.9. There exists several methods to compute such low-discrepancy sequences such as Halton, Faure or Sobol sequences. We also mention that a stochastic version of the QMC method exists, the Randomized Quasi Monte Carlo (RQMC), which is obtained by carefully introducing randomness in the sequence $\boldsymbol{u}$ (Owen, 2008; Gerber, 2015). The RQMC estimator is obtained as previously by mean averaging the produced sequence. Contrary to I_n^{QMC}, the produced estimator is unbiased and has recently been shown to achieve a $\mathcal{O}(n^{-1})$ rate of integration under square integrability hypothesis (Gerber, 2015).

Stochastic Gradient Descent

The SGD methods introduced by (Robbins and Monro, 1951) was specifically designed as a first-order stochastic zero-search procedure for a noisy objective

function. This class of algorithms and its variants (Polyak and Juditsky, 1992; Kingma and Ba, 2015; Duchi, Hazan, and Singer, 2011a; McMahan and Streeter, 2010) gained rapid attention due to its simplicity and broad range of applications. In modern problems, it relates to numerous application in statistics and machine learning ("Stochastic Approximation Approach to Stochastic Programming"; Bottou and Le Cun, 2005). The original Gradient Descent method (Cauchy, 1847; Hadamard, 1908; Rumelhart, Hinton, and Williams, 1985) uses a gradient estimate to recursively updates λ at time t as following

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \alpha_t \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_t). \tag{1.13}$$

In the setting described in C.7 we do not have access to the total expectation $f(\boldsymbol{\lambda})$ but only a noisy estimator. The gist of the *stochastic gradient descent* method is to replace the true gradient with its estimator, resulting in

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \alpha_t \boldsymbol{g}(\boldsymbol{\lambda}_t). \tag{1.14}$$

The choice of the *learning rate* α_t is crucial as it control how large the updates can be. A set of sufficient conditions known as *Robbins-Monro* conditions ensure that the procedure C.14 convergences if the decreasing update schedule is such that $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$. The choice of the learning rate is challenging by iteself and influences greatly the rate of convergence (Bottou, Curtis, and Nocedal, 2018). A simple choice consists of taking $\alpha_t = ct^a$ for a real power a and c some real constant. Modern methods uses *adaptive learning rates* for tuning the learning rate such as AdaDelta (Zeiler, 2012), AdaGrad (Duchi, Hazan, and Singer, 2011b) or Adam (Kingma and Ba, 2015). Theoretical guarantee on the rate of convergence can be obtained giving some regularity assumption on f. For instance, assuming smoothness and strong-convexity, Bottou, Curtis, and Nocedal (2018) show that the error $f(\boldsymbol{\lambda}_t) - f(\boldsymbol{\lambda}^*) = \epsilon$ is in $\mathcal{O}(t^{-1})$.

Gradient variance. If two gradient estimators are available at the same computational cost, the one with lower variance should generally be preferred since the convergence of Stochastic Optimization methods crucially depend on the variance. Most of these optimization procedures rely on gradient descent optimization over the parameters associated with the variational family and subsequently depending heavily on the $\ell_2(\mathbb{R}^K)$ (with K the number of variational parameters) norm of the expected gradient (Bottou, Curtis, and Nocedal, 2018; Domke, 2019). Low variance of the gradient estimators allows for taking larger steps in the parameter space and result in faster convergence if the induced bias can be satisfyingly controlled. Several methods have been used to reduce gradient variance such as filtering (Miller et al., 2017; Roeder, Wu, and Duvenaud, 2017) control variate (Geffner and Domke, 2018) or alternative sampling (Tran, Nott, and Kohn, 2017;

Ruiz, Titsias, and Blei, 2016; Buchholz, Wenzel, and Mandt, 2018). These methods generally suffer from several drawbacks. First, it commonly requires restraining assumptions on the variational distribution. For instance, QMCVI is only valid for distribution with invertible density function. Second, most of the time, the theoretical guarantee on the solution's goodness is not properly established. Finally is that it often involves a complex computation framework and can be challenging to implement.

1.4 Contributions

Pattern based learning applied to predictive maintenance. We propose an extensive overview of the field of predictive maintenance with a highlight over predictive maintenance recent advances in the context of the railway industry. This use case has is particularly challenging; the industrial system of railway spans across a vast territory with various environments and involves complex heterogeneous and interconnected systems. The second contribution consists of designing an industrial prediction pipeline to tackle the predictive maintenance problem in an industrial context. To overcome computational complexity that comes with a high number of possible hyperparameters, we design a two-sample test based method to prune the tree of operations to perform. Various algorithms and sets of hyperparameters are tested and compared on the two classes of french train fleet over a two year period.

Bayesian generative model for pattern mining. We develop methods using a Bayesian Generative Model for Pattern Mining and show superiority over the traditional deterministic methods on various tasks. First, we show that the set of frequent itemsets can be efficiently mined using Stochastic Approximation methods. We propose a Bayesian approach with a variational inference scheme to obtain the space of frequent itemsets with high accuracy.

Second, we use a Bayesian Mixture Model to infer with a low computational cost the discriminative itemsets (Hämäläinen and Webb, 2019) with empirical proof of the general use of such discriminative patterns by considering them as features for the classification task. This results in a method that can extract an interpretable set of attributes and significantly improve any classifier. Moreover, the Bayesian generative model allows for computing the posterior distribution and estimating the confidence intervals. Finally, additional expert-knowledge can be naturally introduced in the model *via* the choice of prior (Gelman et al., 2013). This method is applied to the predictive maintenance task and significantly improves the classification score in an interpretable fashion.

Part of this work corresponds to the paper (Dib et al., 2021) published in *29th IEEE European Signal Processing Conference (EUSIPCO) proceedings*.

Local rademacher complexity for infrequent pattern mining. The progressive sampling task consists of computing the size of the subset of the database n needed to obtain an estimation of any frequency at precision $\varepsilon \in [0, 1]$ with probability at least $1-\delta$. It thus relates to bounding an empirical process generated by an unknown distribution indexed on a finite functional space (Boucheron, Lugosi, and Massart, 2013).

Existing methods use (global) Rademacher averages to mine *frequent* or *top-k* itemsets, which is appropriate, as we do not require sharp bounds on low-frequency itemsets. Notably, Riondato and Upfal, 2015 uses an analytical counting argument to get a loose bound on the global empirical Rademacher average. In the same fashion, Pellegrina et al., 2020 followed this path by using a Monte-Carlo approximation strategy to get sharper bound at the cost of additional computation.

This work marks the first use of localized Rademacher complexity to the low-frequency pattern mining problem. We show that localized Rademacher averages are sufficient to obtain relative confidence interval estimates on pattern frequencies, as well as other interestingness measures, such as the *lift*, *confidence*, or *odds ratio*, whereas previous techniques fail to do so for low-frequency patterns.

Our methods rely on standard tools in the pattern mining domain, such as closed pattern families, antimonotonicity, and Monte-Carlo Rademacher averages, as well as new techniques we introduce to address the problem-specific computational challenges arising from evaluating the localized Rademacher average. The performance of our approach is empirically demonstrated on real-world datasets, wherein exhibit fast convergence rates for the considered subclass of patterns, sharply contrasting existing work.

This work corresponds to the preprint (Cousins* and Dib*, 2021)[1] submitted to the *IEEE International Conference on Data Mining (ICDM 2021).*

Alternative sampling for stochastic optimization. We develop an new approach for Stochastic Optimization technique based on Optimal Quantizer (OQ) (Graf and Luschgy, 2000; Pagès, 2018). We show that using OQ produces an optimal gradient-free gradient estimate at the cost of introducing asymptotically decaying bias with a theoretical guarantee. The method is applied to the Bayesian Learning setting for Evidence Lower Bound (ELBO) maximization and show that using the Quantized Variational Inference framework leads to fast convergence for both score function and the reparametrized gradient estimator at a comparable computational cost than traditional Monte Carlo Variational Inference. Subsequently, we propose a Richardson extrapolation type method (Richardson and

[1] equal contributions.

Glazebrook, 1911; Pagès, 2007) to improve the asymptotic bound and reduce the produced bias. Two new algorithms, QVI and RQVI, are evaluated on several large scale experiments and exhibit superior performance compared state-of-the-art methods (Miller et al., 2017; Buchholz, Wenzel, and Mandt, 2018).

Part of this work corresponds to the paper (Dib, 2020) published in *Advances in Neural Information Processing Systems 33 Proceedings (NeurIPS 2020).*

1.5 Outline of the thesis

- Part II: Anomaly detection for rolling stock maintenance.
  - Chapter 2: A systematic review of predictive maintenance.
  - Chapter 3: Pattern extraction for anomaly detection for rolling stock maintenance. This chapter describes the approach taken to tackle the complex issue of predictive maintenance on the French fleet of high-speed train.
- Part III: Pattern Mining.
  - Chapter 4: Probabilistic view for pattern extraction problem. A Bayesian approach to the famous itemset mining problem is described with various experiments.
  - Chapter 5: Localized complexity for progressive sampling. This section describes the use of localized Rademacher Averages to tackle the progressive mining problem. We show how this method can lead to faster pattern mining with theoretical guarantee.
- Part IV: Optimal Quantization for stochastic optimization.
  - Chapter 6: Background on optimal quantization. We give theoretical background on Voronoi Tesselation and propose to use this alternative sampling for stochastic optimization. Theoretical results on the quality of the approximation are developed.
  - Chapter 7: Quantized Variational Inference. We introduce a new algorithm for ELBO maximization. We show that thanks to the variance free gradient, this method outperforms the state-of-the-art on various real-world experiments, including the case of the bayesian pattern extraction problem.

Part II

Anomaly detection for rolling stock maintenance

Chapter 2

Predictive maintenance: a selective review

Once an operator has ensured his system's nominal functioning to perform a task, he has to assure the continuity of the exploitation. Most of our industrial tools have to operate during extended periods and often in a non-ideal environment, exposing them to aging or external damaging. In that sense, maintenance is defined as the procedure by which a degraded system is regenerated to a satisfying level of functioning. Hence, the first step towards completing a maintenance task is the measure of the state system and Predictive Maintenance (PM) is the domain of statistics which aim to detect this degraded state.
This chapter aims to make the reader familiar with the basic element of Predictive Maintenance. Additionally, this chapter positions this thesis's work in the broad field of anomaly detection for industrial systems. Section 2.1 presents the main strategies deployed for maintenance depending. Section 2.2 gives a broad view of the data used for PM, the associated physical measures, and some literature examples. The rest of the chapter is devoted to present the different classes of models in section 2.4, the possible output or target of the pipeline in section 2.3 and some commonly used metrics in section 2.5.

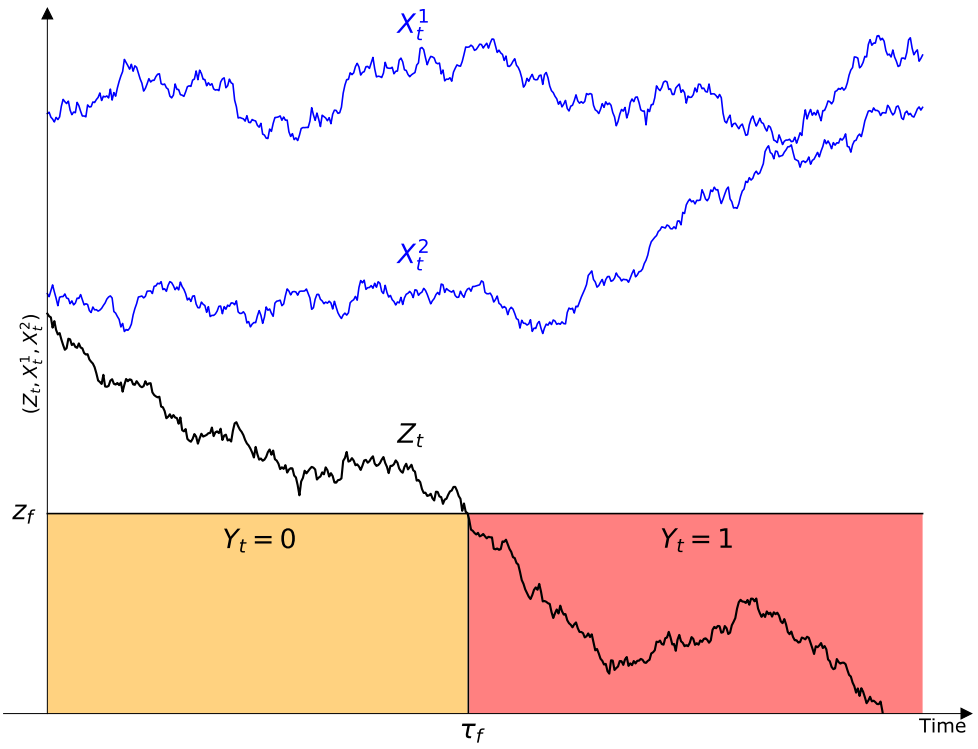


Figure 2.1: Evolution of a degradation process Z_t (black line), the feature vectors X_t^1, X_t^2 (blue) and the binary health status Y_t. The system is considered in its functioning state (orange) if the health process remain beyond the threshold z_f. At the time τ_f, the degradation process cross the threshold and the system enters a deteriorated state (red).

2.1 Typology of Predictive Maintenance

Every industrial system undergoes a process of degradation that leads, if not addressed, to the asset's failure and unavailability. Thus, it is crucial to control, adjust, repair, maintain, and upgrade the system throughout its lifetime. There several possible strategies to maintain a certain level of availability; each requires its own skill sets and has different associated costs.

The status of a system can be modeled by considering degradation process Z_t that reflect the true state of the component at each time t. It can be directly observed in some instances as, for example, when considering the wearing growth of a wheel but is actually hidden in the vast majority of use cases. Typically, the observed quantities are a set of covariates linked to the degradation process through a complex mapping. These covariates can be any measurement identified as informative on the asset's health status, such as physical measurement (temperature,

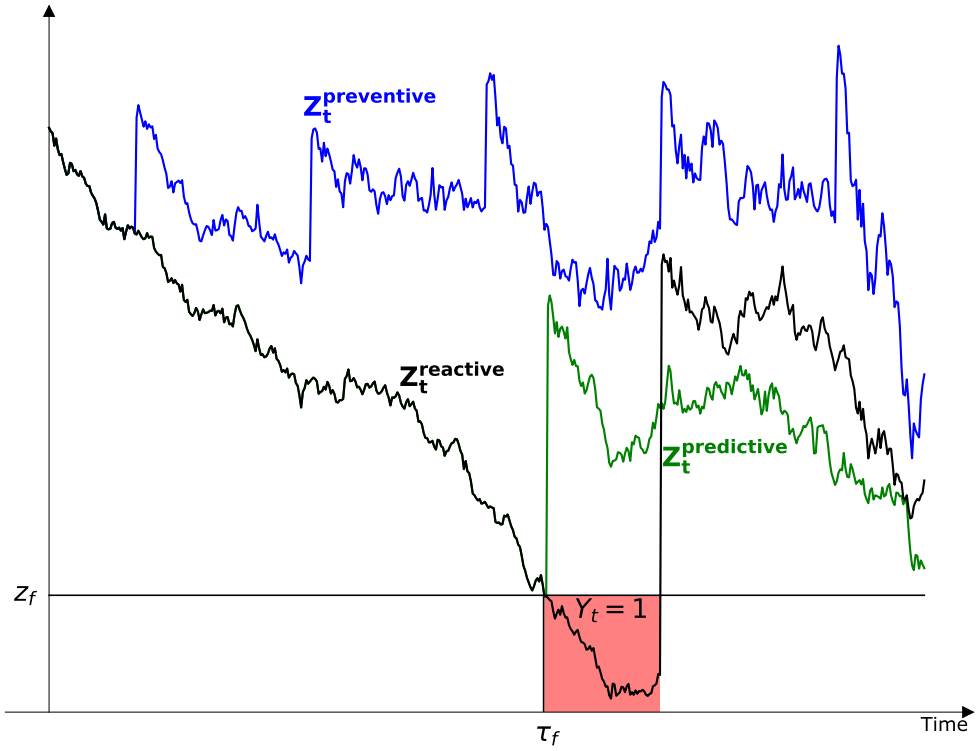


Figure 2.2: Evolution of a degradation processes Z_t for different maintenance strategies; reactive (black), preventive (blue) and predictive (green) maintenance. The reactive maintenance strategy lead to an extended period of unavailability of the asset (red area) while preventive maintenance avoid breakdown at the cost of unneeded regeneration. The preventive maintenance strategy regenerate the system at the right time, before breakdown, and constitutes the ideal scenario.

electrical current, particle count), environment variables such as weather condition, or historical events as the time since the last maintenance operation. The health state Z_t degrades during operation and will eventually reach a threshold z_f at time τ_f. From this point, the asset is in a failure state and needs intervention to be regenerated. Figure 2.1 shows an example of evolution of covariates X_t and degradation process Z_t. The goal of any predictive maintenance strategy is to minimize the unavailable time of an asset (red area).

The goal of predictive maintenance is to minimize the costs associated with this degradation phenomenon. To that end, predictive maintenance policies can be categorized into three main strategies. In the following, we formally introduce these maintenance strategies and give concrete examples of usage.

Reactive Maintenance The simplest alerting event that can be taken to trigger maintenance is the breakdown itself. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and a filtration $\mathcal{A} = \cup_{t=0}^{\infty} \mathcal{A}_t$ such that the application $Z_t \colon \Omega \longrightarrow \mathbb{R}$ is a measurable real random variable representing the degradation process of a system indexed by continuous-time $t \in \mathbb{R}_+$ and z_f a threshold indicating if the system is considered malfunctioning. Additionally, we denote $\boldsymbol{X_t} \colon \Omega \longrightarrow \mathbb{R}^d$ a set of random covariates forming the feature space.

Reactive Maintenance (RM) consists of carrying maintenance at the time of failure

$$\tau_f = \inf \left\{ t \in \mathbb{R}_+ \mid Z_t < z_f \right\}.$$

Since $\mathbb{1}_{\{\tau_f \leq t\}}$ is $\mathcal{A}_t$-measureable, the random time τ_f is thus a stopping time. This strategy has the advantage of requiring only the knowledge of the breakdown set of events $\left\{ \omega \in \Omega \mid Z_t(\omega) < z_f \right\}$ and not the complete degradation process or feature space. However, this approach is usually economically nonviable since it leads to an unprogrammed shutdown of the system and induces costs associated with its unavailability. To illustrate this drawback, take the example of rolling stock in the railway transport sector. An unplanned breakdown means to block an entire portion of the railway network leading to the delay of a large number of other vehicles. This time delay can propagate to numerous vehicles on the network, even if they are far away from the initial malfunctioning train (Corman and Kecman, 2018).

Figure 2.2 shows that this strategy (black line) leads to unavailability of the equipment that is typically associated with high costs.

Preventive Maintenance Another maintenance strategy consists of periodically regenerating some key component of the system. The period T_p of that procedure is typically based on some statistics computed on historical data such as the mean (or better, the median) of the historical time before breakdown. Using previous notation, this quantity corresponds to the expectation of the stopping $T_p = \mathbb{E}\left[\tau_f\right]$. It is a very common strategy since the average lifetime of a component is usually provided by the technical specifications. This average lifetime can be biased by the fact this quantity is often estimated in a controlled environment designed by the constructor to reflect the real condition of use.
An important aspect (and sometimes overlooked) is that Preventive Maintenance can be the best strategy in the case of degradation without aging. Considering such system is equivalent to making the assumption that the degradation process at time t is independent of the past, or memory-less. Formally, the degradation process Z_t is independent of $\mathcal{A}_{t_0}$ for $t_0 < t$ and so does the stopping time. In that case, the information about the past is irrelevant to the future state of the

system. This case is not hypothetical even tough memory-less processes are uncommon; some part of the railroad such as the railroad switches are composed of Hadfield steel (which is extremely robust) and can be considered to obey to a degradation without aging process (Gertsbakh, 2013; González-González, Praga-Alejo, and Cantú-Sifuentes, 2016).
This approach is actually efficient (depending on the concentration of the time of failure around its mean) but can lead to regenerate the system a long time before it becomes necessary (Calixto, 2016).

Figure 2.2 illustrates the preventive maintenance strategy (blue line) which regenerate the equipment several time during exploitation even though it is not necessary.

Predictive Maintenance Predictive maintenance comes naturally as the strategy that uses all the information available before a time of prediction t to accurately predict τ_f. Giving an event $\omega \in \Omega$, providing that there is enough time between the prediction and $\tau_f(\omega)$ to intervene, it is possible to optimally repair before the breakdown. The goal of the learning procedure is to find a measurable function f^* such that for any time t and a *loss function* $L \colon \mathbb{R}^2 \longrightarrow \mathbb{R}$

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}\left[L(f(\boldsymbol{X}_{0:t}), Z_{0:t}), \tau_f)\right], \tag{2.1}$$

with the functional $\mathcal{F}$ is often in some parametric function class of model (we defer to section 2.4 for details). If reactive maintenance is too late and preventive maintenance too soon, then predictive maintenance is optimal in the sense that it would lead to the best availability of the system at a minimal cost. Even though exhaustive, this approach presents several challenges. One has to construct the feature space $\boldsymbol{X_t} \in \mathbb{R}^d$ and collect the data. This is actually the most time-demanding step as most of the data available needs to be engineered so that it can be linked to the system functioning. Moreover, the model has to be chosen to take into account the dynamic of the past data which poses a computational complexity problem. Additionally, that degradation process Z_t is often only observed through the set of events generated by τ_f which is (hopefully) a rare event target (Lei et al., 2018). Finally, establishing a metric that reflects the economic utility of a maintenance plan is challenging and must be carefully designed and coded in the pipeline's model (Si et al., 2011).

Figure 2.2 gives an example the predictive maintenance approach (green line) which regenerates the equipment just before crossing the failure threshold z_f. At fixed maintenance costs, this can be considered as the optimal time to intervene.

2.2 Typology of data used in Predictive Maintenance

This section is devoted to giving an extensive (but not exhaustive) overview of the data used for Predictive Maintenance. Given the (potentially unobserved) degradation's process Z_t at time t, the first step of any statistical learning is to identify a set of relevant random variables $\boldsymbol{X_t} \in \mathbb{R}^d$ that are covariates of the health state of the system at time t. Some of these variables can be internal as, for instance, an electrical current, or external such as the weather (Jalili Hassankiadeh, 2011; Li et al., 2019). An ideal feature space is both complete and pairwise-independent in the sense of giving information about any relevant physical measurement linked to the degradation process and nonredundant. Careful consideration of any relevant source of data is a crucial part of establishing any machine learning solution (see Table 2.2 for examples of the physical properties of a system that can be measured).

Vibration The vibration of a physical system is its movement around an equilibrium position (or rest position). It is associated with a periodic signal with characteristics linked to the state of the system. For instance, the frequency of this signal is related to the fundamental modes of the object. In a controlled environment (excluding external perturbations), any degradation of this signal from its normal behavior can be interpreted as a degradation of the system.
Vibration analysis may be the most common data collection techniques due to the availability of the sensor and the historical use of such data (Renwick and Babson, 1985). This can be used for assessing all king of physical variation in a component. However, interpreting this signal typically requires some additional knowledge about the system to identify relevant features (Wu et al., 2007; Ugechi et al., 2009).
Vibration analysis is commonly used for assessment of roller bearing defauts (Al-Ghamd and Mba, 2006; Khadersab and Shivakumar, 2018; Malla and Panigrahi, 2019; Dyer and Stewart, 1978) but has been applied to larger components such industrial pumps (Amihai et al., 2018), automobile gearboxes (Praveenkumar et al., 2014) or computer numerical control (Luo et al., 2019).
Vibrations can be collected through accelerometer data to be used for fault detection. Nunez, Jamshidi, and Wang (2019) use a Pareto model to map the health state of railway tracks based on an evaluation of the acceleration profile of the train wheel. Ma et al. (2019) proposes a track detection system based on train recorded vibrations and a CNN-LSTM architecture ((Ma et al., 2019)[Table 1]) to assess the condition of the railway tracks.

Acoustic Analysis The acoustic analysis consists of directly observe and measure the sound wave of a system. If air sound pattern is taken as reference, deterioration can be signaled through the change in the property of this wave. Everyday life offers numerous accounts of fault diagnostic based on sound waves. For instance, a small crack in a plate will produce a distinctive sound when gently tapped. The advantage over direct vibration analysis of the vibration is that a sensor can be installed to monitor and detect any deviation from the normal aggregated pattern of all systems, even when there are not physically connected. Acoustic analysis is used for any degradation that will induce degradation in the sound wave signal, such as hidden flaws in metallic structure (Liggan and Lyons, 2011), lathe system (Garg et al., 2015), robot swarms (Tarapore, Christensen, and Timmis, 2017), Heating Ventilation and Air conditioning systems (Srinivasan et al., 2017) or electrical motors (Grandhi and Krishna Prakash, 2021).

Another approach is to use an acoustic source tuned on a specific frequency (corresponding to the fundamental mode of the studied system) and analyze the reflected signal. This method is used in (Kocbek and Gabrys, 2019; Jiang et al., 2019) to spot structural deterioration of rail tracks.

Imaging The use of images is especially developed in the field of predictive maintenance due to the low cost of high-quality camera sensors and the versatility of its usage. In the railway industry, it has been historically used in maintenance centers to quickly evaluate any structural damage on some parts of the rail vehicles. Combining simple feature extraction methods (such as edge detection) and pattern recognition algorithm can lead to interpretable and well-performing anomaly detection pipelines (Lu, Liu, and Shen, 2018). Thermal Imaging, a method that measures the temperature of a body remotely and provides the thermal image of the entire component or machinery, has also been widely used for a long time to detect mechanical or electrical problems that cause temperature anomalies (Pathirathna et al., 2018; Meola, 2007). Thermal Imaging has been successfully utilized for several condition monitoring applications such as civil structures [(Grinzato et al., 2002; Clark, McCann, and Forde, 2003; Meola, 2007), inspection of electrical equipment (Jadin and Taib, 2012), monitoring of plastic deformations (Badulescu et al., 2011), evaluation of fatigue damages in materials (Luong, 1998; Pastor et al., 2008; Bagavathiappan et al., 2013) and inspection of machineries (Bagavathiappan et al., 2013) like rotating machinery (Janssens et al., 2015). In the railway sector, Karakose and Yaman (2020) use two thermal cameras to monitor the pantograph and rail health state on each trip. The same Fuzzy-logic based model is then used to evaluate the health system of each system and perform fault diagnosis. We mention that the use of data from Optical sensors is in active development, and many other devices are considered in the railway sector such as optical fibers (Tam

et al., 2018).

Symbolic data A very common source of data consists of a process of symbolic data. Examples of such processes are a database of documents or log sequences produced by some components that are common in predictive maintenance since logging is still considered the most versatile way of recording events of interest (He et al., 2016; Bogojeski et al., 2020). These types of processes are particularly challenging since the common binary operations are not defined and that there is no natural order on E, thus making most machine learning models unusable. Several solutions exist to give a structure to these types of sets. A first approach is to transform the process to a numerical one by aggregating (e.g. counting or considering some statistics) the events on a chosen time windows (see Figure 3.3). Another one is to consider an embedding by a kernel in a well-structured feature space allowing to apply a broad range of classical machine learning techniques (Kung, 2014). Finally, one can consider the set $\mathcal{E} = \mathcal{P}(E)$. This set is the set of patterns on E on which it is possible to apply *pattern mining* techniques (Agrawal, Imielinski, and Swami, 1993).
In the predictive maintenance domain, most of the studies focus on aggregation and finding a satisfactory embedding. Heidarysafa et al. (2018) uses a classical word embedding layer with a Multi Layer Perceptron and Long Short Term Memory network to detect railway tracks failures. Another study (Kauschke, Fürnkranz, and Janssen, 2016) takes the log events as input, combined with a simple Decision Tree, to predict railway car breakdown. Seep recognition techniques have recently been applied to automatically analyze incident report *a posteriori* in order to enrich the feature space or give an interpretation on a particular prediction (Heidarysafa et al., 2018). The kernel approach is considered in (Li et al., 2019), where the authors used a Multiple Kernel Learning model on maintenance log historical data of the Sydney Trains database.

Industrial data sets We stress that most of the data used in the literature are not produced by real functioning industrial systems but are rather collected in a laboratory environment in which a component is tested in isolation or during stress tests (Amruthnath and Gupta, 2018). Mostly for confidentiality issues and difficulties associated with communicating specific expert knowledge, it is difficult to acquire public real-world data for academic purposes. Table 2.3 lists the most publicly available data used in literature.

2.3 Model Output

We recall that, at time t, the degradation process Z_t informs of the system's health status. This degradation process may be directly observed. For example, in (Xu et al., 2018) the degradation process corresponds to the wearing of train wheels. However, in most situations, the true degradation process is modeled or completely unobserved. In this case, the only data available is the failure date $\tau_f(\omega)$ (or the date at which the system does not meet specific performance criteria).

Remaining Useful Life The Remaining Useful Life (RUL) (Si et al., 2011) (or Time To Failure (TTF) in the broad domain of anomaly detection) is defined as the time remaining to the failure of a critical component. Giving $\boldsymbol{X_t}$ be the feature vector, let $Y_t =$ the binary health status of the system at time $t \in \mathbb{R}_+$ indicating if the system is failling and define the failure time as $\tau_f = \min_t \left\{t \in \mathbb{R}_+ \mid Y_t = 1\right\}$. For any time t_0, the Remaining Useful Life is defined as

$$R(t_0) = \tau_f - t_0. \tag{2.2}$$

If we denote $\mathcal{A}_t$ the filtration $\sigma(\{Z_{t_0}\}_{t_0 \leq t})$, then the failure time τ_f and the RUL are $\mathcal{A}_t$-measurable random variables. The process of failure is often an equivocal concept since one needs to define the state of failure of the system. In the case where the operator have access to a measure of health of the system Z_t at any time $t \in \mathbb{R}_+$ he can rely on an expert based threshold z_f to determine if the system is malfunctioning. In this case, the binary process of failure is the random variable $Y_t = \mathbb{1}_{\{Z_t \leq z_f\}}$.

The RUL is commonly used since it can be easily interpreted but one should keep in mind that the expected value of the RUL is most of the time insufficient in industrial context (Saxena et al., 2010). In general, the RUL will be given with a level of confidence corresponding to percentiles of the probability distribution of τ_f. Several use cases are presented in (Daniyan et al., 2020) and we refer to Table 2.1 for references in the context of railway transport.

Failure indicator Apart from the degradation process itself, a commonly used output is the binary failure Y_t at time $t \in \mathbb{R}_+$. We remind that this binary variable is the indicator function of the usability of the system. Its corresponds to the *binary classification* framework where, given covariates $(X_1, \ldots, X_n)$ and a test set $(Y_1, \ldots, Y_n)$, one wants to find a $\{0, 1\}$-valued function that minimizes a loss with respect to true binary failure indicator.

2.4 Models

Once the space of features has been carefully described and the system monitored, one needs to choose a function class that modelizes the dependence between the degradation process and the feature space.

In this section, we describe the most commonly used model for Predictive Maintenance. Our goal is to exhibit the links and differences between these types of models, state their advantages and limitations, and give literature use cases.

A common distinction is made between a physical model of deterioration (also referred as the model-based approach) and a statistical approach (or data-based inference) (Si et al., 2011). Model-based failure methods rely on the physics of the underlying degradation process to predict the onset of failures, while data-driven approaches attempt to derive models directly from the collected data.

The area of Model-driven prognostics deals with predicting the degradation process of critical components by explicitly choosing a physical model of the degradation phenomenon, usually a set of differential equations. Experts typically choose this approach as it requires some layer of understanding of the phenomena involved in the degradation process. For large interconnected systems, this approach may be infeasible but will generally more explicitly exhibit a variable's influence over the deterioration process. For instance, Nappi et al. (2020) design a purely model-based approach by deriving a set of differential equations (Nappi et al., 2020)[Equation 1] to be solved to estimate the normal behavior of a railway wheel system. Fu et al. (2019) construct a model-based solution to simulate the degradation process of a bogie component. The system is modeled by a Complex Network Model in which the probability of failure is evaluated through a walk upon the induced graph structure ((Fu et al., 2019)[Figure 3].

The data-driven approach constructs a model primarily based on the collected training data with modeling tools commonly used by the artificial intelligence community: temporal prediction series, trend analysis techniques, Artificial Neural Network, Neuro-fuzzy Systems, Hidden Markov Model or commonly used machine learning algorithm (Zhang, Yang, and Wang, 2019). Contrary to the first, this approach would not typically require in-depth knowledge of the phenomena, but intensive work on the data collection and preparation needs to be performed.

Some approaches try to take the best of both worlds by proposing hybrid models. As an illustration, we refer to the work of Wang, Bu, and He, 2020 in which the authors design a model-based sample generator to simulate the degradation process of a power equipment plant and use the generated data to train a Long Short Term Memory based neural network for failure detection (for more details,

see (Wang, Bu, and He, 2020)[Figure 1]).

Typically, the model-based approach requires to perform numerous study on test bench to establish to establish a set of physical equations that describe the behavior of the system. It can be particularly interesting for a component that is widely used and for which a physical response model has been established, but is very costly to develop for a new asset or complex systems that are composed of many interdependent parts. Moreover, there is little use for production databases that are typically available in large volume in a industrial system. Thus, for its versatility and use of already available data, the data-driven approach is the proeminent framework of predictive maintenance (Mosallam, Medjaher, and Zerhouni, 2015). For these reasons, we only focus on the data-driven methods.

2.4.1 Data-driven model based on machine learning methods

In this section, we detail the data-driven approaches for the RUL and Binary Failure estimation, focusing on standard machine learning models. Since most of them are well-known and have been extensively documented, we restrict ourselves to reviewing interesting applications.

Statistical models Tree-based algorithms such as Support Vector Machine (SVM), Decision Tree (DT) or Gradient Boosting Model (GBM) are commonly used in practice thanks to the low computational resources needed to train them and their overall relatively good performance on a broad range of real use case. Moreover, they can be easily interpreted by analyzing the tree structure (Zien et al., 2009) or through indirect methods such as the Shapley value (Lipovetsky and Conklin, 2001; Štrumbelj and Kononenko, 2014). Allah Bukhsh et al. (2019) build a model based on three Tree-based models to evaluate the failure probability of railway switches based on regular visual inspection data and maintenance logs. When labeled data is unavailable, a common approach is to use traditional unsupervised learning techniques such as PCA, Hierarchical clustering, K-Means. A simple use case is given by the work of Amruthnath and Gupta, 2018 where K-means clustering is used to evaluate an industrial fan's state. Table 2.1 gives a view of the recent application of such techniques to predictive maintenance in the railway sector.

Artificial Neural Network In (Ugechi et al., 2009), the authors use vibration data as an input to a Multi Layer Perceptron to classy vibration's data to predict faulty centrifugal pumps. Wu et al. (2007) construct an ANN based decision system for the rotating equipment to infer the distribution of the time before breakdown. Liu et al. (2018) used a Generative Adversarial Network to perform

unsupervised classification on rotatory machines under several types of regimes. Guo et al. (2017) proposed a recurrent neural network to estimate the Remaining Useful Life on rolling bearings but only compared it with a self-organizing map-based method. As mentionned in section 2.2, image data has been long used for predictive maintenance. Convolutional Neural Network (Bengio, Goodfellow, and Courville, 2017) are thus suitable architectures to extract interesting patterns. For instance, Chen et al. (2017) proposes a CNN-based architecture to detect surface default of catenary fasteners. CNN need not to be used on image data and it became common to see one dimensional CNN applied to numerical temporal signals (Karpat et al., 2020)

Wavelet transform Wavelet transform is a way of decomposing a signal and finding a typical component related to a specific behavior. It is based on a wavelet series used to define the orthonormal basis for a Hilbert space. (Chiementin, Bolaers, and Dron, 2007) proposed the early detection technique based on adaptive wavelet for fatigue damage measurement on the inner and outer race of ball bearings. A procedure was designed for analyzing signals using this wavelet. The method seemed to improve fault detection in the presence of noise also. Jiang et al. (2019) used wavelet transformation of a vibration signal delivered by a dedicated measurement tool for railway track in an ablation study. The authors showed that transformation applied to the source data improved prediction compared to the raw signal.

Yang et al. (2008) proposed and validated a new wavelet-based adaptive filter for CM of wind turbines. Conventionally vibration measurement and lubrication oil analysis were used as CM systems in wind turbines. However, both these methods suffered from some drawbacks as the former method required high hardware costs, and the latter could not detect electrical abnormalities in the turbine generator and electrical system. The wind turbine's power energy was used as an indicator of wind turbine condition as wavelet-based adaptive filter extracts the power energy at prescribed fault-related frequencies with both varying and constant rotational speeds.

2.4.2 Data-driven model based on processes

Stochastic processes

The degradation process is inherently the result of a complex stochastic dynamical system due to the great variability and uncertainty associated with the measurement of such a state. A natural idea is to use simplistic instances of the class of stochastic processes to capture the dynamic of the Remaining Useful Life. One of

the most simple to consider is the Wiener process

$$dZ_t = \mu dt + \sigma dW_t, \tag{2.3}$$

where W_t is the Brownian motion at time t and μ, σ the *drift* and *diffusion* coefficients. The main advantage of such a model is that he is very well studied, particularly in the finance sector (Rolski et al., 2009), and that the time to failure can be analytically computed or closely approximated. In the context of RUL estimation, it was considered in several studies recent studies (Guan, Tang, and Xu, 2016; Lorton, Fouladirad, and Grall, 2013; Nicolai, Dekker, and Van Noortwijk, 2007). Even though easy to use, this model has some fundamental drawbacks. For one, the degradation process is monotonically decreasing where Winner processes can only be parametrized to drift in expectation. At time t and $s < t$, the value of the Winner process W_t will be independent on the event in the filtration $\mathcal{F}_t = \sigma(W_1, \ldots, W_s)$ as it is a markovian process. Hence it can only model degradation without aging. Finally, Winner processes are path continuous and cannot account for sudden jumps that often occur when considering a system's physical degradation.
Gamma processes fix most of the issues raised by Wiener processes since they are monotonic and can be inhomogeneous, taking into account past temporal evolution. Gamma processes are popular models to describe a monotonic degradation process as in wear processes or fatigue crack propagation. We refer to (Cha and Pulcini, 2016; Crowder and Lawless, 2007; Huynh et al., 2012; Liao and Elsayed, 2006) for use in the context of Predictive Maintenance and to (van Noortwijk, 2009) for a complete review of the Gamma Process and its properties.

Hidden Markov model

A Hidden Markov Model is composed of two discrete-time stochastic processes, a hidden Markov chain $(Z_n)_{n\geq 0}$, which is unobservable and represents the real state of the deterioration, and an observable process $(Y_n)_{n\geq 0}$, which is the observed condition information from monitoring and tests (we confer to section B for a detailed and technical introduction).
Hidden Markov Model based approaches are particularly suitable for predictive maintenance in an industrial context since they can model the latent state, which represents the machine's health condition. Zhao et al. (2019) use a constrained left to right Hidden Markov Model to estimate the Remaining Useful Life of the NASA engine degradation data. HMM has been used to model industrial process in a discrete event system using different structure of Markov automate for the hidden state (Robles et al., 2013), Diesel Engine (Simões et al., 2017), lubricating oil for engine with Remaining Useful Life prediction (Du, Wu, and Makis, 2017), vehicle maintenance (Kamlu and Laxmi, 2019), bearing condition evaluation (Cartella

et al., 2012) and online condition assesment (Lee, Li, and Ni, 2010). In the more broad domain of anomaly detection, Chen et al. (2016) use discrete observations of known sequences of online behavior to detect intrusions and Song et al. (2009) propose a nonparametric HMM that extends traditional HMM to structured and non-Gaussian continuous distributions (*via* kernel embedding) and derives a kernel spectral algorithm for learning.

In all of the above examples, the hidden state is chosen to be an unknown degradation state. However, basic Hidden Markov Model models suffer from the fact that the state duration of any hidden non absorbing state is geometric. To overcome this, the semi-Markov processes were developed independently by (D'Amico, Janssen, and Manca, 2009; Serfozo, 1972). HSMM is traditionally defined by allowing the unobserved state process to be a semi-Markov chain (Wang et al., 2014) and used, for instance, in speech recognition or equipment health diagnostics and prognostics. In the context of predictive maintenance, it has been shown (Tobon-Mejia et al., 2011; Medjaher, Tobon-Mejia, and Zerhouni, 2012) to effectively estimate the Remaining Useful Life by using the the duration time random variable. An extended summarize of different duration model classes has been made by Yu (2010) . Most state duration models used in the literature are nonparametric discrete distributions (Yu and Kobayashi, 2006; Yu and Kobayashi, 2003; Wang et al., 2014) Hidden semi-Markov models enjoy the versatility of the HMM and the possibility to represent temporal structures in the signal but are the difficulty to infer the optimal parameters (Wang et al., 2014).

2.5 Metrics

This section presents several used metrics emphasizing the framework in which each one should be considered. The evaluation of a Predictive Maintenance solution raises several challenges. Most of them are associated with the industrial context, which imposes additional constraints on exploitation. For instance, in the case of railway transport, a prediction made just a few hours before the breakdown will usually not be useful since this information cannot be used to reduce maintenance cost since there is not enough time to prevent maintenance shutdown of the system. Another aspect is that an online algorithm providing risk estimation for failure would rarely predict on real-time data but rather with a time delay. A metric for PM should also accommodate practical aspects such as safety, cost efficiency, and mission priority (Saxena et al., 2010).

Regression

Let us define the the point-wise error $\mathcal{E}(t)$ at each time $t \in \mathbb{R}_+$

$$\mathcal{E}(t) = R(t) - \hat{R}(t). \tag{2.4}$$

The error $\mathcal{E}(t)$ is a random variable representing, for a specific set of test data, the difference between the true and estimated RUL. Given a time horizon $T \in \mathbb{R}_+$, it is more interesting to consider the scaled version of this error on the total duration since large absolute true Remaining Useful Life can produce a large error. Thus the introduction of a Relative Error (Medjaher, Tobon-Mejia, and Zerhouni, 2012)

$$\text{RE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\mathcal{E}(t)}{R(t)}\right|. \tag{2.5}$$

Medjaher, Tobon-Mejia, and Zerhouni, 2012 develop an accuracy metric for RUL estimation, first proposed by Vachtsevanos and Vachtsevanos, 2006, is discussed. It writes

$$\text{Accuracy } = \frac{1}{T}\sum_{t=1}^{T} e^{-\frac{\left|R(t)-\hat{R}(t)\right|^{\alpha}}{R(t)}}, \tag{2.6}$$

for $\alpha \in [0, 1]$. This measure is similar to the Relative Error but is easy to interpret: a prediction is considered acceptable if its accuracy is close to one and not if the accuracy is close to zero.

Classification

The metrics used in the classification framework introduced in section 2.3 are commonly used in machine learning where we consider a predictor $\hat{f}$. The True Positive (TP) and True Negative (TN) measures correspond respectively to the number of correctly classified samples and the number of sequences incorrectly classified when compared to the true Failure indicator function. These two types of error are not very informative by themselves since we could design a simple algorithm that would return 1 for each test example and reach a TP measure of 1 (similarly, we could reach a minimal TN measure). It is more fruitful to introduce the following quantities.

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{TNR} &= \frac{\text{TN}}{\text{TP} + \text{FN}}. \end{aligned} \tag{2.7}$$

In some case, the predictor $\hat{f}$ will output a probability vector instead of a binary vector. Let us consider a probability threshold p such that the predictor will output one if $\mathbb{P}(\hat{f}(X) = 1) > p$. In this case, at any level p, we have that

$$\begin{aligned} \mathrm{TPR}(p) &= \int_p^\infty \hat{f}_1(x), \\ \mathrm{TNR}(p) &= \int_p^\infty \hat{f}_0(x), \end{aligned} \tag{2.8}$$

with $\hat{f}_1$ and $\hat{f}_0$ being respectively the density function associated with output one and zero. The Receiver Operating Characteristic (ROC) curve is defined as the graph of the following function

$$\begin{array}{rccc} \mathrm{ROC}\colon & [0,1] & \longrightarrow & [0,1]^2 \\ & p & \longmapsto & (\mathrm{TPR}(p), \mathrm{TPN}(p)). \end{array} \tag{2.9}$$

A good classifier will be above the bisector of the identity function representing the random classifier (properly weighted if the class is imbalanced). We refer to (Bradley, 1997) for details and proofs.

Study	System	Data	Model	Output
(Kocbek and Gabrys, 2019)	Railway Tracks	Acoustic; Visual Inspection	RF;LR;SVM;DT	Accuracy
(Jiang et al., 2019)	Railway Tracks	Acoustic	SVM	Accuracy
(Hu and Liu, 2016)	Railway Tracks	Numerical; Track physical measurements	SVM	Accuracy
(de Bruin, Verbert, and Babuska, 2017)	Railway Tracks; Circuit	Electrical measure	RNN;LSTM	Accuracy
(Zhao, Xu, and Hai-feng, 2014)	Rail Vehicle	Symbol; Logs	Pattern Mining	Accuracy
(Li et al., 2019)	Railway Tracks	Symbol; Numerical	KL	AUC;ROC
(Braghin et al., 2006)	Rail Vehicle; Wheels		model-based	Degradation Process
(Karakose and Yaman, 2020)	Pantograph; Railway Tracks	Thermal Imaging	Fuzzy-logic	Degradation Process
(Xu et al., 2018)	Rail Vehicle	Vibration	PCA;LR	Degradation Process
(Ma et al., 2019)	Railway Tracks	Vibration	Deep Learning	Degradation Process
(Fu et al., 2019)	Rail Vehicle; Bogy	Degradation Process	Model-based	Degradation Process

(Tam et al., 2018)	Rail Vehicle; Railway Tracks	Optical Signal		Non quantitative
(Pathirathna et al., 2018)	Various component	Thermal Imaging		Non quantitative
(Karpat et al., 2020)	Rail Vehicle; Gearboxes	Vibration	CNN	None
(Nappi et al., 2020)	Rail Vehicle	see Table I	model-based	None
(Wang, Bu, and He, 2020)	switchgears	Degradation process	LSTM	RUL
(Daniyan et al., 2020)	Railway Vehicle; Wheel Bearing	Temperature	LR	RUL
(Fumeo, Oneto, and Anguita, 2015)	Axel Bearings	Vibration	SVM	RUL
(Li and He, 2015)	Rail Vehicle	Numerical; Mechanical misalignment	RF	RUL
(Heidarysafa et al., 2018)	Railway Tracks	Documents	Deep Learning	TP;TN;FN;Accuracy
(Kauschke, Fürnkranz, and Janssen, 2016)		Symbols	DT	TP;TN;FN;Accuracy
(Allah Bukhsh et al., 2019)	Railway Switches	Visual Inspection	RF;DT;GB	TP;TN;FN;Accuracy
(Consilvio, Febbraro, and Sacco, 2020)	Railway Tracks	Model-based	Degradation Process; model-based	Degradation Process; model-based
(Chen et al., 2017)	Catenary	Images	CNN	Accuracy;Precision

(Lu, Liu, and Shen, 2018)	Rail Vehicle	Images	Data-based	Accuracy
(Mercier, Meier-Hirmer, and Roussignol, 2012)	Railway Tracks	Numerical; Track physical measure-ments	Gamma Process	Degradation Process
(Jalili Hassankiadeh, 2011)	Railway Switches			TP;TN;FN;FP
(Nunez, Jamshidi, and Wang, 2019)	Railway Tracks	Vibration		Degradation Process

Table 2.1: Survey of Predictive Maintenance applied to Railway.

Table 2.2: Physical measurement of the system state for each category of source data.

Category	Parameter
Vibrations	Imbalance, Eccentricity, Misalignment of couplings and bearings, Resonance problems, Mechanical looseness/weakness, Rubbing, Bent shafts, Shaft cracks, Worn or damaged gears and bearings, Defective/misadjusted drive belts and chains, Sleeve-bearing problems, Turbulence, Turbine/fan blade defects.
Thermal	Temperature, heat flux, heat dissipation
Electrical	Voltage, current, resistance, inductance, capacitance, charge, polarization, electric field, frequency power, noise level, impedance, Mechanical looseness, corroded electrical connection.
Mechanical	Length, area, volume, velocity or acceleration, mass flow, force, torque, stress, strain, density, stiffness, strength, acoustic intensity, power, acoustic spectral distribution, angular, direction, pressure.
Chemical	Species concentration, gradient, re-activity, mess, molecular weight
Humidity	Relative humidity, absolute humidity
Optical	Intensity, phase, wavelength, polarization, reflection, transmittance, refraction index, distance, vibration, amplitude and frequency
Magnetic	Magnetic field, flux density, magnetic moment, permeability, direction, distance, position, flow
Acoustic	Bearing inspection, Steam trap inspection, Integrity of seals, pipe systems and large walk-in boxes, Pump cavitations, Compressor valve analysis, Electrical arcing
Oil Analysis	Particle count, Water content, Viscosity, Additive content, Acid or base number, Flashpoint

Name	Monitoring object	Data type	Classification	Prediction	Transfer diagnosis
CWRU	Motor bearing	Multi-vibration	√	×	√
PHM 09	Gearbox, bearing, shaft	Multi-vibration	√	×	√
Paderborn	Bearing	Current-vibration	√	×	√
IMS	Bearing	Vibration	√	√	×
C-MAPSS	Turbofan engine	21 sensor data	×	√	√
PHM 10	Milling cutter	Current-vibration-AE	×	√	×
FEMTO	Bearing	Temperature-vibration	×	√	√

Table 2.3: Popular public datasets for Anomaly Detection.

Chapter 3

Predictive Maintenance: the case of the French train fleet

This chapter presents a practical approach to implement predictive maintenance in the French train fleet's industrial context. We mainly focus on tackling the issues that bar from in-production use of these techniques. We build a complete and ready to be industrialized system to signal probable breakdown on rail vehicle to improve the availability and safety of the rail transportation network. To provide the best possible result and allow for easy use on other transportation systems, we implement a computational pipeline that includes state of the art preprocessing and prediction model techniques. Subsequently, we propose a method to reduce the computational cost associated with executing this pipeline. The method is applied on two large fleets of train and showed superior results to best-known expert techniques for early anomaly detection.

This chapter is organized as follows. Section 3.1 presents the general problem and the most relevant related work. Section 3.3 describes the data used and the feature and target space construction along with the complete computational pipeline used for prediction. In Section 3.4 we introduce a method to prune the computational tree of calculation quickly. Finally, Section 3.5.1 presents the study results along with pattern mining extraction for interpretability.

3.1 Introduction

The use of machine learning techniques for predictive maintenance in industrial contexts has proven to be a fertile approach in various application areas (Carvalho et al., 2019). Studies, as well as several proofs-of-concept, have demonstrated the potential of this approach in improving the safety, reliability and efficiency of a transportation system (Kocbek and Gabrys, 2019; Allah Bukhsh et al., 2019;

Nappi et al., 2020). As of today, this potential remains largely untapped due to the youthfulness of new modes of data governance. Data from different sources were siloed in the past, with each source flowing into its own database in isolation from the others. This led to analyses based mainly on technical expertise and the implementation of specific rules and maintenance procedures. Conversely, the power of modern predictive maintenance systems relies on the cross-referencing of numerous data sources to enrich the space of variables and increase their predictive power.

The equipment engineering team at SNCF recently set up a tool to detect upstream failures based on the fault codes transmitted by the rolling stock. Although promising, this new form of maintenance, which we could describe as predictive, is still far from replacing the historical model that consists of mixed operation of corrective and planned maintenance (see chapter 2). And for good reason, understanding and then translating abnormal behavior into an algorithm requires considerable expert time. In addition, a project called "zero LGV faults" aimed at limiting the risk of line failures on high speed rolling stock (line stops, speed restrictions, incidents) have been launched. In this context, both entities see the emergence of automatic learning techniques as an opportunity to accelerate the construction of these rules capable of anticipating breakdowns.

The development of an automated pipeline for predictive maintenance comes with several difficulties when they are considered for industrial production. One of the main drawbacks is the scarcity of the target signal (Ran et al., 2019). Indeed, the average number of failures per day is (fortunately) low, so the number of relevant signals (in the sense of helpful in predicting a failure) is rare with regard to the total number of signals produced. Hence, the generalization error may be heavily impacted. Moreover, this target is the result of different types of readings; some failures will be automatically transmitted while others depend on an operator's report. Thus, the time of occurrence of a failure and its type are potentially imprecise, and the quality of the target data is inhomogeneous. Second, the output of the predictive model needs to be compatible with the previous maintenance approach. At SNCF, the maintenance procedure is a well-established process with precise guidelines that need to be followed to ensure passenger security and train availability. Hence, an operator needs to understand the output of any algorithm to link any predicted anomaly to a subsystem of the vehicle. Finally, most traditional methods cannot be efficiently used and tested on a large-scale system (Carvalho et al., 2019). Constructing the feature and target space, testing and maintaining the machine learning solution, requires introducing many hyper-parameters (Canizo et al., 2017). Finding a good set of these hyper-parameters comes with a heavy computational price.

The high impact and high costs potentially associated with system failure led

to extended research in system monitoring and fault detection (Carvalho et al., 2019). Research on model-based failure detection relying on the underlying physical degradation process recently trimmed back in favor of a data-driven approach which aims to apply generic statistical models from condition monitoring (Ran et al., 2019). However, authentic physics-based models are difficult to build for some components and systems because equipment dynamic response and damage propagation processes are very complex.

In the data-driven area, variety of statistical approaches have been developed including regressive models (Medjaher, Tobon-Mejia, and Zerhouni, 2012; Kocbek and Gabrys, 2019; Jiang et al., 2019) and state space (Bayesian Network, Hidden Markov Models, etc.) models (Mercier, Meier-Hirmer, and Roussignol, 2012; Zhao et al., 2019) or Deep Neural Networks (Chen et al., 2017; Heidarysafa et al., 2018). Yan, Koç, and Lee (2004) performed a logistic regression to establish performance model of an elevator's door, and the remaining life is estimated using ARMA model based on historical data. ANN are used in two fashion: in classification to predict trends and system failure and in regression, commonly with feedback connection, to model dynamical processes and give an expectation of Remaining Useful Life (Laredo et al., 2019). A prediction of the health of a roller bearing by modelisation of the vibration root mean square value was developed by Laredo et al. (2019). On the same topic, Gebraeel and Lawley (2008) performed ball bearing remaining life prediction by using the output of the ANN as a condition monitoring measurement. A great variety of applications continue to emerge: generalized ANN that can deal with multiple measurement inputs (Tian, 2012), integration of failure and suspension data to improve accuracy ANN-based time series prediction to deal with insufficient data Tian, Wong, and Safaei (2010). Adding a feedback loop to an ANN allows accounting for the past input to influence the new network output opening up the use of Recurrent Neural Networks (Lipton, Berkowitz, and Elkan, 2015). In the railway domain, de Bruin, Verbert, and Babuska (2017) used a recurrent neural network to predict the condition of railway tracks. Tian and Zuo (2010) also developed a recurrent ANN-based time series prediction method to deal with situations where sufficient faulty events are not available. Aggarwal et al. (2018) utilized Long Short Term Memory to forecast the damage propagation trend of rotating machinery to both predict failure and the RUL.

Fitting expert knowledge to machine learning modern problems is a well-known issue in practical application. The term *black box* means that it is very difficult or even impossible to have physical explanations of the networks' outputs. Besides, as models grow in size, training can be challenging. For example, how many hidden layers should be included and the number of processing nodes used for each layer are difficult questions for model developers. Garga et al. (2001) proposed a hybrid reasoning method for prognostics. A feed-forward neural network was

trained using explicit domain knowledge to get a parsimonious representation in this approach. A Dynamic Bayesian network (Murphy, 2012), also called belief network, is a directed graphical model of stochastic processes that enables users to monitor and update the system as time proceeds. A Bayesian network is a field-proven tool for modelisation in domains with uncertainty (Sakib and Wuest, 2018). Their graphical representation, showing the conditional independencies between variables, is easy to understand for humans experts. As historical example, Sheppard and Kaufman (2005) used Bayesian networks for prognosis systems. The authors construct a Bayesian Belief network incorporating information on instrument uncertainty, knowledge about false indication and failure probability. Then, to acknowledge change over time, the prognosis is performed by using a dynamic bayesian network. In the prognostic model developed by Gebraeel et al. (2005), the Bayesian approach is employed to update the prior distributions for estimation of subsequent failure times. Dong and Yang (2008) investigated a DBN-based model to predict remaining life for drill-bits. The authors built a DBN-based model and corresponding inference algorithms. A prognosis procedure based on particle filtering algorithms is used to predict RUL of the drill-bits of vertical drilling machine.

A industrial machine learning pipeline is a complex sequence of computationally expansive operations and typically involves many steps. Each of theses steps is associated with some user-defined variables that control how the data are cleaned and filtered, the feature and target space constructed or how the final prediction model learns. We call *parameters* all the variables of the computational pipeline that are directly used by the final model to make a prediction (such as the regression coefficients in a Linear Regression model) and *hyperparameters* the ones that are involved in the prediction only through the training (such as a filter parameter or the learning rate of a gradient descent procedure) (Murphy, 2012). Sometimes overlooked, the choice of hyperparameters is crucial and can greatly affect the model performance (Van Rijn and Hutter, 2017).
Traditional approaches aims at inferring an approximation of the best set of hyperparameters with respect to the model score on a validation data set. Naive approaches purely based on grid-search are computationally intractable since the number of evaluation grows exponentially with the number of hyperparameters (Bergstra et al., 2011). Bergstra and Bengio (2012) show that selecting at random the hyperparameters to evaluate outperform the extensive grid-search approach. A more refined approach proposed by Li et al. (2017b) allows for setting a given level of computational resource to be allocated to search an approximation of the best hyperparameters by iteratively dropping some portion of the search space. Bayesian optimization approaches build a bayesian probabilistic model to map the output of the pipeline to the hyperparameter space (Pe'er, 2005; Shahriari et al.,

2015). Derived from this approach, efficient algorithms such as FABOLAS (Klein et al., 2017) can efficiently and at given computational cost identifies the most promising hyperparameters by careful balancing of evaluation and exploration of the promising hyperparameters. This method has the main drawback of requiring surrogate model to map the hyperparameters to the output of the prediction pipeline which is often arbitrarily chosen and itself costly to estimate (Mendoza et al., 2016).

One must keep in mind that most of the mentioned above and within Chapter 2 studies are not performed on real in-production systems but rather on a public and conveniently preprocessed dataset. The consequence is that computational costs and limitations are rarely addressed. As mentioned, the industrial process of implementing such a machine learning approach involves a large number of parameters to adjust a optimize on (Thomas et al., 2016). Moreover, most of these approaches do not question the compatibility between the model's output and the well-established maintenance process.

In this work, we propose a complete pipeline solution for predictive maintenance from data identification and selection, preprocessing and modeling in this work. Our method is used on real-world systems during exploitation at a reasonable computational cost. To prune the computational tree and identify a set of hyperparameter, we use a non parametric *two-sample* test approach based on the Maximum Mean Discrepancy measure on the labeled training data set to automatically prune the hyperparameter space. This approach is efficient (Gretton et al., 2008), does not involve the choice of a mapping between the space of hyperparameters and the output and straightforward to implement. The constructed prediction pipeline has a superior predictive power than expert-knowledge rules and can be easily exported on different classes of components (such as railway tracks, overhead lines, etc).

3.2 Problem Description

Each day, the SNCF operates 15,000 trains over 35,000 kilometers of rail network. As a result, the rail stocks evolve in very diverse operation's environments. Additionally, the fleet of vehicles comes from various class of trains that have been in function for various amount of time. This great variability poses challenges to construct a predictive maintenance solution. For this reason, the perimeter of the study has been reduced to two types of trains, namely *NAT* and *High-speed trains*, during a controlled period over which special events (such as extreme weather condition or infrequent breakdown events) have been carefully analyzed and taken into account. As mentioned in chapter 2, designing a machine learning pipeline for predictive maintenance tasks is by itself a challenge. For this study, we follow the design process described in section 2.2. The first step is to identify and collect variables that are linked to the degradation process. Numerous data sources were considered to tackle this problem. Joint work with technical experts and operational maintainers allowed the construction of a feature space as a function of data considered to be sufficiently linked with the degradation process. The following summarizes this step of feature and target construction.

3.2.1 NAT

The *Class Z 50000* railway vehicles also known as Nouvelle Automotrice Transilien (NAT) (which stands for *New Self-Propelled Transilien*) is a multiple unit electric regional trains built by Bombardier that operates in Paris and its suburbs since 2009.

Error Codes

Error codes consist of time-stamped signals provided by TrainTracer, a software for collecting and processing data on onboard equipment. According to rules to which the end-user does not have access, these codes are produced during events deemed relevant by the manufacturer (exceeding the threshold of an electrical signal, malfunction) according to rules to which the end-user does not have access.

The collected database contains 6069329 code events distributed in a dictionary of code Σ of size $|\Sigma| = 754$ over a period spanning from 2014-01-01 to 2015-05-27. Each error code is associated with a particular event and identifiable through a correspondence table. For example, table 3.1 shows the meaning of the ten most frequent codes. The codes have simple regularities that reflect the normal operation of the system. Figure 3.2a shows the number of codes issued as a function of time over several aggregation periods. It shows a sustained activity around peak hours (7h-9h and 18h-19h) due to more intensive operation. Similarly, the time

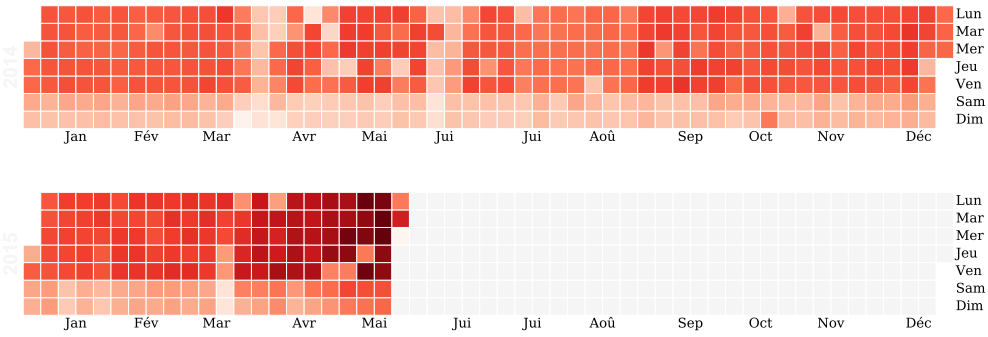
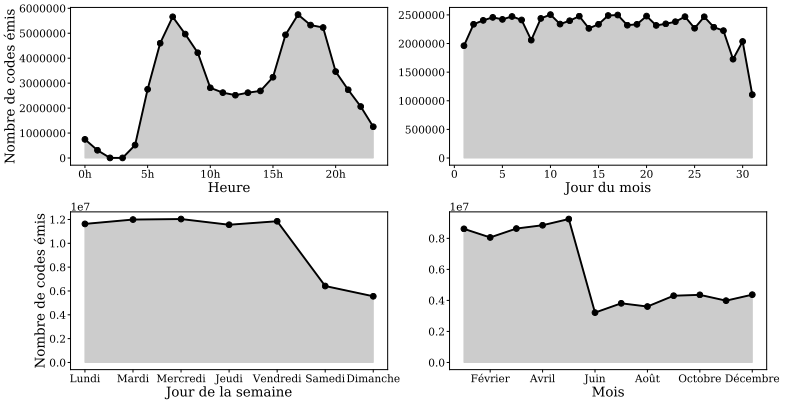


Figure 3.1: Calendar of error codes emissions of NAT train class over the studied period. The number of events is lower during the weekend since it is typically a period of reduced service (except during holidays). The number of codes tend to increase during winter as the weather condition is a known factor of deterioration. Moreover, the number of error codes (and anomalies) tends to increase over the year due to the aging of the train fleet.

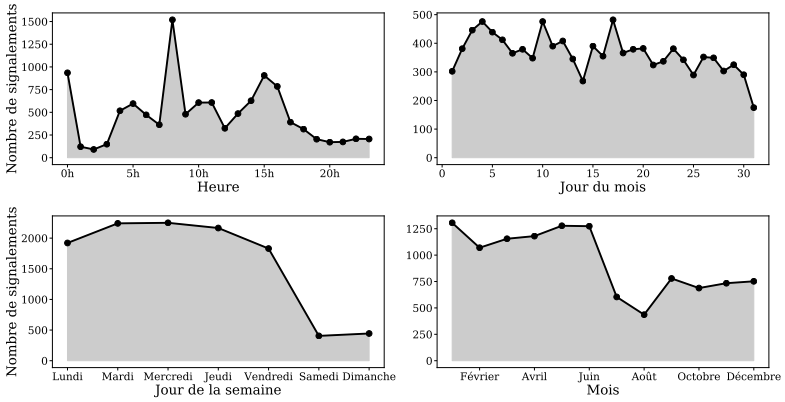
distribution by day of the week shows stable activity on weekdays and reduced activity on Saturdays and Sundays. The monthly time distribution is not more surprising: our entry set contains the first six months of 2015, so we find more events over this period. The number of codes seems to increase in winter. This may be due to a higher number of particular events and the use of components that are inactive during the rest of the year (heating) or less active (anti-skid/anti-scratch). Since we only have a single winter history, it is not possible to draw any further conclusions. Several updates to the onboard system have resulted in changes in the type and number of codes issued.

Breakdowns

The target variable for the NAT class was constructed upon the history of signaling. A signal is an event logged either by the train operator, maintenance technicians or automatically produced by onboard system for critical detected anomalies. Since a signal is not necessarily linked to a breakdown, these signals need to be linked to maintenance reports. Each time a train is inspected, whereas it is a planned or an unplanned operation, a report is produced. A report consists of free text and structured information and summarizes the state of the equipment and the possible



(a) Error Codes.



(b) Breakdowns events.

Figure 3.2: Temporal distribution for error codes (top) and constructed breakdown events (bottom) within a day (upper left), month (upper right), week (bottom left) and year (bottom right).

operation performed. To make this information suitable for machine learning, each maintenance operation is associated with specific equipment that corresponds to a specific system or function of the rail vehicle. Moreover, this classification of the report into a class of equipment allows for targeting critical failure. For instance, an engine failure leads to complete immobilization of the train, whereas air conditioning only impacts the passenger's comfort. Table 3.2 report the breakdown classes identified that constitutes the multi-class target event variable. Finally, these signal events are enriched by various table to precisely identify the localization of the signaling (not only the position on the network but the direction of traffic), the specific vehicle of the train impacted and the history of the equipment (age of the component, last intervention, etc).

Table 3.1: The top-10 most common error codes with meaning.

Error Code	Wording
8025	RightDoor closing(GDIR E99)
8425	RightDoor closing(GDIR E99)
16111	Def. camera 2
20071	Zone balise 2
20070	Zone balise 1
20052	LT Autorisation open RD
20058	LT Autorisation open LD
20053	LT Autorisation dev right
16110	Def. camera 1
20059	LT Autorisation dev left

Table 3.2: Type of breakdowns.

Class	System
B	Cashiering / Boiler Room
C	Body Lining
D	Interior Fittings
E	Running Gear
F	Power Device / Drive Train
G	Drive Train Control / Brakes
H	Auxiliary Equipment
J	Safety and Monitoring Equipment
K	Lighting
L	Air Conditioning
M	Other Equipment
N	Door
P	Passenger information and operating assistance systems
Q	Hydraulic and pneumatic equipment
R	Brakes (brake system / components)
S	Interconnections

3.2.2 High Speed Train fleet

The French High Speed Train fleet is a class of vehicles that travels at speeds around 300 km/h on a dedicated rail system that allows short inter-city travel time. As for the NAT fleet, we based our analysis on the onboard log system and the history of recorded breakdown events over a period of two years that spans from 2017-10-01 to 2019-10-01.

Error Codes

MyTrainData (MTD) collects events recorded by rail vehicles during operations managed by the Equipment Department. Each event is emitted by a train system (door, engine block, onboard computer, etc.) by a command control unit. An event contains an error code that as well as a *context code* that provides geographical and technical information about the state of the train at the time of code emission. In addition, various databases are needed to identify the type of train, position it on the railway network. The extracted data covers 18 months from 2017-10-01 to 2019-10-01. The collected data consists of various files, tables, and reports from several databases for a total volume of approximately one terabyte of data. To analyse the degradation process and the influence of the covariate, it is needed to gather all the information available about a system at a specific time. As mentioned, these data are scattered through various departments, are of heterogeneous quality and have not been designed to be crossed. For instance, it is important to exclude from the analysis the period of maintenance. During maintenance, the onboard system can fire events that are not useful for predictive maintenance in operation. The solution found was to list and geolocalise all maintenance centers and filter out any events that have been emitted in these areas. Another example is the need to identify the exact type of train that operates. Indeed, a train is composed of a heterogeneous class of railroad car, which can have specific behavior. Analysis should thus include a reference to the class of railroad cars that have emitted an error code. To that end, specific databases has been used to extract the rail plan of each trip and recover the composition of each train. Numerous database crossing similar to that was performed to enrich the state features. These enriched events can then be used to look at different use cases, focus on particular series, and find the maximum number of explanatory elements for failures. At the end of this process, all the information available at each time of the lifetime of the subsystem are gathered to create the feature space. It may factor environment variables such as railway characteristics on which the rail vehicle operates at the time considered, or weather conditions. It also includes system information such as the age (in terms of exploitation time) or a particular state variable called *context code*. For instance, some context code gives the speed

of the train measured as the radial speed of the vehicle wheel. This information can be decisive to identify a malfunction; if a code of the door subsystem indicates an open position while the context code indicates that the train is in motion, it is considered as a critical safety risk and a malfunction of the door system.

Breakdowns

All breakdowns that occur on high-speed trains are subject to an intervention by a maintenance operator who establishes an Intervention Report (CRI). Apart from breakdown events, regular planned maintenance operations are carried out to track for early signs of deterioration of the rail vehicle and similarly logged in an IR. During a preventive operation, not all the subsystem is inspected. For instance, our database accounts for almost 2500 preventive operations per train set over two years. Among those interventions are only about twenty operation where the doors are looked at (approximately every month) and about ten where the Engine Block are looked at (approximately every six months). Over the studied period and for the class of train selected, we collected 400 thousands intervention reports which amounts to approximately fifty per month and per train set. Among them, there are about two IR per month and per train that concern the doors (including one breakdown) and five that relate to the Engine Block (including between three and four breakdowns per month per train set). The first step was to determine which IR can be considered as indicative of failures on a subsystem with experts. The second was to reconstruct the life of a specific subsystem over the period. Since 2011, about thirty train sets have been broken down and then reassembled to create new ones. Thus, the subsystems can have been on several train over a period of time. Finally, each subsystem was associated to multiples time series indicative of various environmental and historical information about the component, such as journeys made and the number of kilometers traveled (accumulated) each day or the time since the last maintenance operation. Once all these databases are crossed, we finally obtain a consistent historical view of each subsystem.

3.3 Construction of a production machine learning pipeline for predictive maintenance

In this section, we describe the process of construction a complete algorithmic pipeline for predictive maintenance. As described in chapter 2 and the introduction, several challenges are associated with such construction at each step of the process. For an industrial application, one has to test multiple possibilities in order to find the best pipeline to put in production. Moreover, for this work to be used in slightly different context (such as predictive maintenance on railtrack), the process must be sufficiently general to be adapted. The design of this pipeline is thus complex and involved several exchanges and collaboration with expert-knowledge to be suitable. In the following we brush out the main steps of the produced algorithmic solution. The construction of the pipeline brokes down into three phases; the import and cleaning, the windowing phase and the prediction step.

Data preparation. The data preparation step includes, app art from the importing and cleaning steps, the preprocessing phase. The preprocessing consists of constructing the feature space by adding or removing variables and filtering data.

Most error codes are issued at regular intervals, without signaling a noteworthy event. This can be a signal to open a door, start, stop or the inconsequential activation of a command. Therefor periodic signals in the signal do not indicate the occurrence of a fault. Figure 3.8 shows the median time between two transmissions for the twenty most frequent codes in a box diagram. For instance, half of the 20064 codes are transmitted every five minutes. Several filter have been designed to take into account the fact that some code are not necessarily linked to a malfunction. For instance, the Gaussian filter apply a rule to delete every emission that is outside a time bandwidth of $t \times \sigma$ for $t \in [0, 1]$. Simpler filter such as quantile filtering or by the top-k most relevant codes are considered. All these case are parameter of the production pipeline with the goal of finding the most suitable

Window aggregation. The prediction procedure uses event logs, which are time-stamped error codes e_t taken from a dictionary E of d distinct codes. These events are collected and processed by onboard equipment according to dedicated rules to which the end-user does not have access. These codes are produced during events deemed relevant by the manufacturer (for instance exceeding the threshold of an electrical signal or a malfunction).

Procedures that make use of log events are particularly challenging since there is no natural order or distance on the space of symbols, thus making most machine learning models unsuitable. This issue can be overcome by kernel methods (Kung, 2014) but these approaches are difficult to interpret, which is a requirement for a

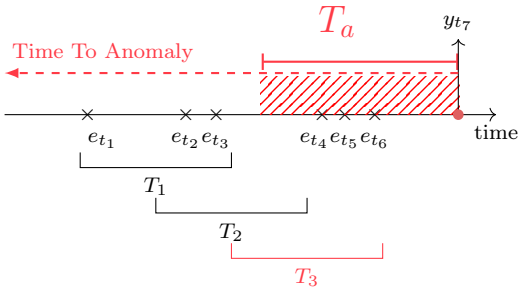


Figure 3.3: Temporal aggregation of log-events (e_{t_1}, ..., e_{t_6}) over sliding windows (T_1, T_2, T_3). In red, events that occur in the period T_a before y_{t_7} are considered anomalous and labeled $l = 1$. The aggregation produces the itemsets $x_1 = \{e_{t_1}, e_{t_2}, e_{t_3}\}, x_2 = \{e_{t_2}, e_{t_3}\}, x_3 = \{e_{t_4}, e_{t_5}, e_{t_6}\}$ and the labels $l_1 = 0$, $l_2 = 0$ and $l_3 = 1$. The goal is to correctly predict the labels l_i from the itemsets x_i.

predictive solution to be used in an industrial context. Another common strategy consists in transforming the prediction task into a binary classification task. In a nutshell, the signal is aggregated over sliding temporal windows (possibly overlapping) of fixed size. Features are simply the set of collected events within the window (called itemsets). For a given user-defined threshold period $T_a > 0$, a window is considered as anomalous (label "1") if it contains codes emitted in the period T_a before a failure, and normal (label "0") otherwise. This aggregation procedure is schematically illustrated on Figure 3.3. Even though popular (Basora, Olive, and Dubot, 2019), classification based solely on this construction is often unable to capture critical patterns of events that can be highly relevant in PM.

There is several variation of the process illustrated in Figure 3.3. A window aggregation procedure is associated with multiple parameter that will lead to different feature construction; mainly, a window size, an overlap parameter (can be set to zero for purely consecutive windows) and a time range anomaly T_a. Moreover, the choice of the aggregation function greatly influence the model performance. The simplest aggregation function consists of a binary indication indicating the presence or absence of a certain code type. A more involved solution considers a statistic on the count of error code (count, distance to a mean count, mean inter emission time for specific code in a window, etc).

Prediction. The definition of the training and test set is itself a challenge in the framework of predictive maintenance. Classical train-test split will consider random sub sampling of the training data set. However, the sequentially of the signal is lost by such split. To take this effect in account a time split is considered; the training data consists of the signal over an uninterrupted period reflecting the normal condition of exploitation. Finally, we also consider a train-test split by rail

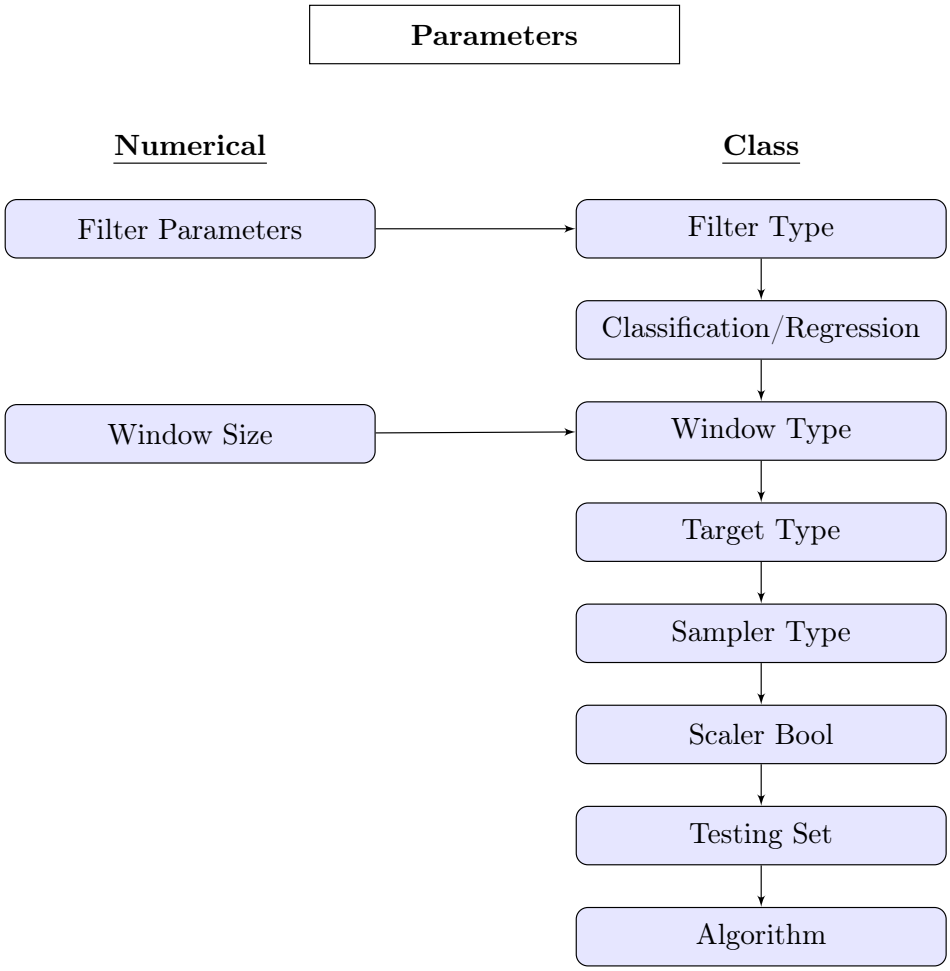


Figure 3.4: Simplified computational pipeline used for prediction. Each step, from preprocessing to end prediction, involves multiple hyper parameters to fit.

vehicle; the algorithm is trained on a subset of the available rail vehicle and tested on data from never seen vehicles. A typical experiment will output the result for each split strategy and the performance difference is valuable information to evaluate the true performance of the model. To tackle the imbalance problem, we consider various under sampling and over sampling methods, including implementations of both traditional new methods such as random sampling or the cluster centroid method (Santoso et al., 2017) but also SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002) and the Instance Hardness Threshold approach (Smith, Martinez, and Giraud-Carrier, 2014). Finally, the statistical algorithm is set and a grid search of the model hyperparameter is performed.

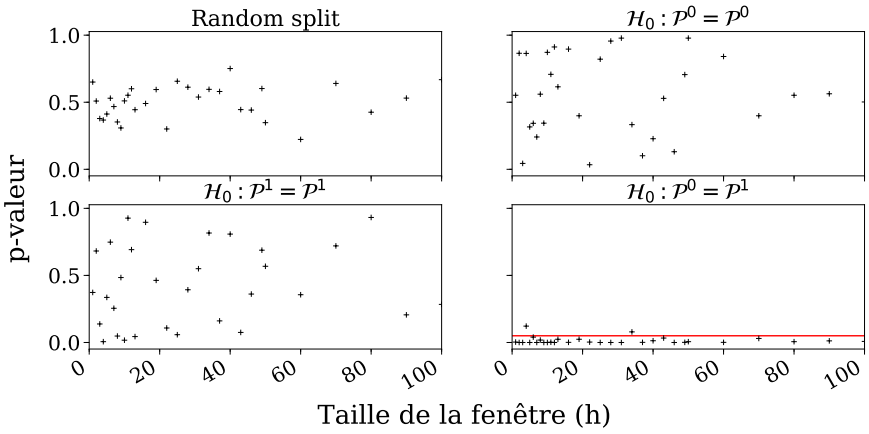


Figure 3.5: p-value of the statistical test as a function of the window size. Upper left: *Random Split* is the random assignment of the binary target. Upper right: comparison performed on two negatively labeled subsets. Lower left: comparison performed on two positively labeled subsets. Lower right: comparing the set of samples labeled positively and the set labeled negatively (the threshold in red corresponds to the level of rejection of the statistical test at 5 %)

.

At each step, the choice of parameters affects the overall performance of the model. Figure 3.4 shows the set of adjustable parameters of an experiment.

3.4 A two sample test for pipeline pruning

The tree of computation described in the previous section spans a large number of hyperparameters. Exploring even a portion of this hyperparameter space through *grid search* approaches (Bergstra and Bengio, 2012) requires an great amount of computational power in term of parallel threads and memory requirements. In this section, we propose to exploit the binary nature of the target output to design a pruning criteria based on a measure of the statistical distance between the two classes of the target.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and consider X the set of input variables and Y the binary output, such as $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. We denote $\mathbb{P}$ the join probability distribution on $\sigma(\mathcal{X} \times \mathcal{Y})$. The goal is to determine whether the distributions associated with each target class are different. More precisely, if we call $\mathbb{P}^0_X$ and $\mathbb{P}^1_X$ the distributions associated respectively with the versions of the conditional expectation on $\mathcal{G}^0 = \sigma(Y = 0)$ and $\mathcal{G}^1 = \sigma(Y = 1)$ we wonder if we can reject the hypothesis $\mathbb{P}^0_X = \mathbb{P}^1_X$. Answering this question comes down to perform a statistical

test on the samples from each class and relates to the well studied two-sample test problem (Gretton et al., 2008).

In the following, we present the formal framework of the *Maximum Mean Discrepancy* method (Smola et al., 2007; Muandet et al., 2017).

3.4.1 Maximum mean discrepancy

Formalism Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions on $\mathcal{X}$. The laws are unknown, but we have the realizations of the law $\mathbb{P}$ and m realizations of the law $\mathbb{Q}$ such that $\mathbb{Q}$ tel que $\{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$ and $\{\mathbf{y}_j\}_{j=1}^m \sim \mathbb{Q}$ and the null hypothesis such as

$$\begin{aligned}\mathcal{H}_0 &: \mathbb{P} = \mathbb{Q},\\ \mathcal{H}_1 &: \mathbb{P} \neq \mathbb{Q}.\end{aligned}$$

Definition 1 (MMD)**.** Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions and $\mathcal{H}$ a functional element space defined on $\mathcal{X}$ and with real values. The MMD of $\mathbb{P}$ and $\mathbb{Q}$ is defined by

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) := \sup_{f\in F}\left[\mathbb{E}_{\mathbb{P}}\mathbf{f}(\mathbf{X}) - \mathbb{E}_{\mathbb{Q}}\mathbf{f}(\mathbf{Y})\right] \tag{3.1}$$

Under the null hypothesis we have indeed $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = 0$ for any $\mathcal{H}$). The question is to determine under which conditions on $\mathcal{H}$ the reciprocal proposition is true, *ie* the necessary condition such as $\mathrm{MMD}.(\mathbf{P},\mathbf{Q};F) = 0$ ssi $.\mathbf{P} = \mathbf{Q}$). We can demonstrate (Jitkrittum et al., 2016) that equivalence holds in the case where $\mathcal{H}$ is a Hilbert space with a reproducible kernel. It is shown that in this framework, maximization comes down, after choosing the kernel, to the computation of a standard.

Definition 2 (RHKS)**.** Let $\mathcal{H}$ be a Hilbert space of functions defined on $\mathcal{X}$ and with real values. A $kfunction : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a Reproducible Core of $\mathcal{H}$ and $\mathcal{H}$ a Hilbert to Reproducible Core space if k satisfies the following conditions

$$\begin{aligned}\forall x\in\mathcal{X}, \quad & k(\cdot,x)\in\mathcal{H},\\ \forall x\in\mathcal{X},\forall f\in\mathcal{H}, \quad & \langle f,k(\cdot,x)\rangle_{\mathcal{H}} = f(x).\end{aligned}$$

The kernel thus allows the evaluation of a function by the calculation of a scalar product. In particular,

$$k(x,y) = \langle k(\cdot,x),k(\cdot,y)\rangle_{\mathcal{H}} \tag{3.2}$$

And for any function $f,g\in\mathcal{H}$

$$|f(x)-g(x)| = \left|\delta_x(f-g)\right| \leq \lambda_x\|f-g\|_{\mathcal{H}} \quad \forall f,g\in\mathcal{H}. \tag{3.3}$$

If we note $\mathcal{H}'$ the topological dual of $\mathcal{H}$, the Reitz representation theorem applied to $\mathbb{E}_{\mathbb{P}}[. \in \mathcal{H}'$ allows to simplify the equation (1) and to demonstrate the existence and uniqueness of an element $\mu_{\mathbb{P}} \in \mathcal{H}$ such as

$$\begin{array}{rccc} \mu: & \mathcal{M}(X) & \longrightarrow & \mathcal{H} \\ & \mathbb{P} & \longmapsto & \mu_{\mathbb{P}} = \int k(\cdot, \mathbf{x}) \mathrm{d}\mathbb{P}(\mathbf{x}) \end{array} \tag{3.4}$$

$$\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle, \quad f \in \mathcal{H}.$$

The equation (1) is simplified by using (3.4.1) and allows to obtain the evaluation of the MMD by the standard defined on $\mathcal{H}$.

Theorem 1. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions and $\mathcal{H}$ a functional element space defined on $\mathcal{X}$ and with real values. The* MMD *of $\mathbb{P}$ and $\mathbb{Q}$ is*

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}, \mathcal{H}) = \left\| \mu_{\mathrm{P}} - \mu_{\mathbb{Q}} \right\|_{\mathcal{H}}^2 \tag{3.5}$$

The equation (3.5) is used to derive an empirical evaluation of the MMD.
Let $\{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$ and $\{\mathbf{y}_j\}_{j=1}^m \sim \mathbb{Q}$)

$$\widehat{\mathrm{MMD}}_u^2(P, Q) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k\left(x_i, x_j\right) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k\left(y_i, y_j\right) \tag{3.6}$$

$$- \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k\left(x_i, y_j\right) \tag{3.7}$$

3.4.2 Pruning algorithm

Each instanciation of the pipeline produces a label sequence of $\mathbb{P}$ distributed samples $\mathcal{D}_n = (x_i, y_i)$ at the preprocessing step. This dataset is used as input for the remaining part of the computational tree until a predictor $\hat{f}$ and a predicted output $\hat{f}(x_i) = \hat{y}_i$ for each sample i is produced. The gist of the method consists of discarding any computational branch that does not meet the requirement in term of statistical differences between $\mathbb{P}_0$ and $\mathbb{P}_1$. In other word, any hyperparameter set that does not produce a sufficiently dissimilar positively and negatively dataset is discarded. The discarded experiment correspond to feature and target space construction with low predictive power.

The first test to apply the kernel two-sample test is the choice of the functional space $\mathcal{H}$ that can be reduced to the choice of a kernel (which uniquely determines $\mathcal{H}$). This kernel only requires to induce a bijection of the operator μ in Equation 3.4 whose existence and uniqueness is guaranteed by the Riesz theorem (Sriperumbudur et al., 2010). In this case, the kernel is said to be characteristic. The

most commonly used characteristic kernel is the Gaussian kernel which as been empirically demonstrated to produce tests with high statistical power (Muandet et al., 2017). In all experiment, we chose the bandwidth parameter to fixed at the median of the euclidean distance of the sample and the p-value threshold of reject at $p_0 = 5\%$.

To validate the approach, the statistical test on the computational pipeline described in section 3.3 using the SNCF datasets (see Table 3.3) and a varying window size hyperparamter. We apply the kernel two-sample test with $\mathbb{P}_0 = \mathbb{P}$ and $\mathbb{P}_1 = \mathbb{Q}$. Two compute the p-value from the sample data, we use the permutation method described in (Gretton et al., 2009). The figure 3.5 shows the result of the statistical test with respect to the window size (see section 3.3) in four cases; when considering only the positive label, the negative label, when the dataset is randomly shuffled and with the true dataset. In all of the three first cases, the null hypothesis cannot be rejected at p_0. On the contrary, in the case where the statistical test is performed between the labeled samples, the null hypothesis can be rejected at 5%.

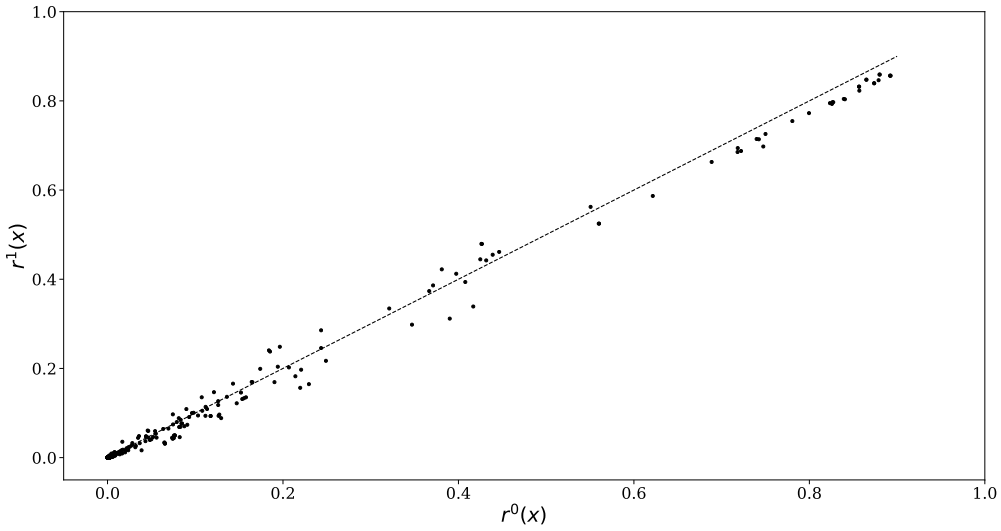


Figure 3.6: Regression function for unitary error codes in the framework of binary classification.

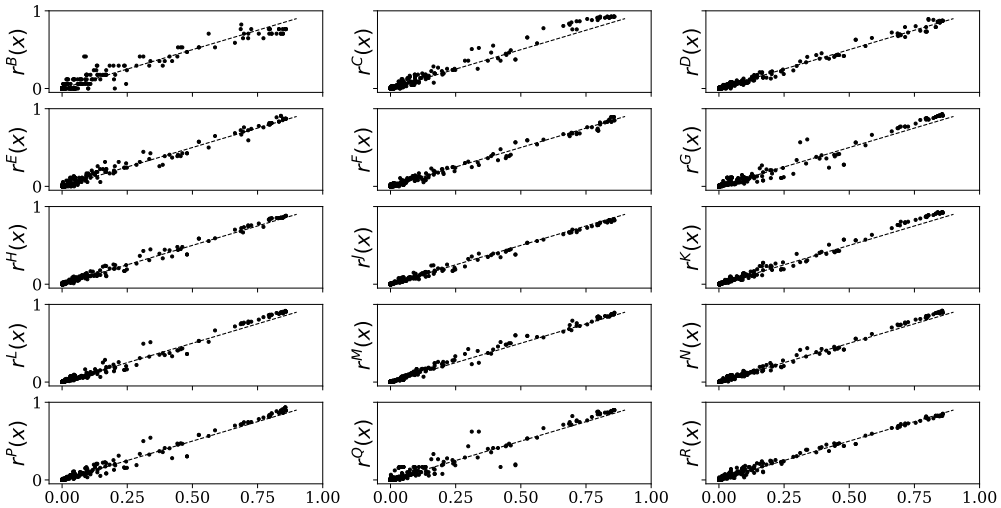


Figure 3.7: Error code regression function in the framework of multi-class classification. The failure classes are those described in the table 3.2. We observe that some anomalies have more dependency to individual occurrence of codes than other.

3.5 Experimental results

This section presents the result obtained by the predictive pipeline on two types of fleet and several subsystem for various algorithms with hyperparameter analysis.

	NAT	TGV BM	TGV Doors
Instances	6069329	233060	134352
Fleet	158	91	170
T (months)	14	24	24
$\|\Sigma\|$	555	198	73
K	0	10^2	10^2
$\|E\|$	12	1	1
Target Size	11256	6447	4341
Regression	✓	✓	✓
Binary Class.	✓	✓	✓
MultiLabel Class.	✓	×	×

Table 3.3: Characteristics of Rolling Stock SNCF data sets. Is reported for each use case the period of study T, the size of the dictionnary of error codes $|\Sigma|$, the number of external and internal contextual variable K, the number of target or type of breakdowns $|E|$ and the number of critical anomaly reported over the period.

3.5.1 Prediction pipeline results

Use cases. The study focuses on three use cases using the data from two fleet of trains described in section 3.2. For each case, we construct a state timeline by crossing several databases to include internal and external variables (see section 3.3) for each subsystem at every time. It can include contextual data at time t such as instantaneous speed and acceleration, binary state of some component of the train or internal data such as the historical use of the subsystem (in term of age and number of kilometers traveled), the history of maintenance at the time considered or instantaneous measures on the component considered. Additionally, we built a referential using GPS tracking data at wach time to include very precise exogenous information as, for instance, the characteristics of the rail tracks on which the train circulates or the weather condition (using Meteo France API service) at each time t. The characteristics of the three studied data sets is described in table 3.3.

Setup. Each instance of the prediction pipeline, also referred as *experience*, produces a dataset and a prediction. For each experiment, the models' hyperparameters are grid-search optimized. The main metric used to compare models is the AUC (area under the curve) of the ROC curve, which quantifies the ability of the model to detect if there is a signal in the data and to distinguish itself from a random model (this score is 0.5 for a random model, and 1 for a perfect ideal model). The higher the AUC on the test set, the more relevant the predictions. We also track the maximum precision, recall and F_1 scores calculated on the test

set. These scores are computed according to the failure classification threshold computed on the training set. The higher the F_1 score compared to the failure rates on the test set, the more valid the predictions are. This score is used to compare the models and to evaluate their business performance. As benchmark, we also use My Train Data Alerts (MTDA), which is an alerting system based on expert knowledge that has been developed over the years.

The procedure is parallelized across multiple threads and we implement a breadth-first schedule to run the pipeline tree described in section 3.3. By executing according to this strategy, duplicated operations are reduced and, most importantly, the memory requirement is lower. The space of computation is pruned using the two-sample test pruning method described in section 3.4; The p-value is set to 5% and every experiment that does not pass the threshold is halted and canceled.

The data processing, pipeline and scheduler are implemented using Python 3.7 and runs on AWS Instances with 8 core Xeon Platinum 8000 @ 2.5 Ghz. For the frequent itemset mining, we use the FP-GROWTH algorithm for the exact frequent itemset mining leveraging the extensive SPMF library (Fournier Viger et al., 2016) originally wrote in java.

Comparaison of multiple algorithms. The table 3.4 report, for all use case, the best result over all the hyperparameters of the pipeline described in section 3.3 for specific aggregated window sizes. We remind that the error codes and context logs are aggregate on the consecutive periods of sizes ω. We excluded the days where the car was out of service for maintenance or test. Hence, we consider only consecutive days of service. Besides, several temporal features are added. For instance, $\omega = [1]$ represents a single day of usage whereas $\omega = [1, 7]$ represents the concatenation of the two data set obtained by the windowing method described in section 3.3. This approach allow for capturing information that may only exist in longer sequence at the cost of doubling the size of the feature space.

For the NAT use case, table 3.4 shows the superiority of Random Forest over the other models for the default experiment for an area under the curve of 0.65 . The cross-validated AUC score shows a slightly higher value at 0.66 and is also the fastest to train. For the Engine Block and Doors use case the best model scores at an AUC of 0.74 and 0.75 respectively. These scores are similar or better than the expert based system of rules constructed of the year by the maintenance team. The F_1 score of each model, for the test sample, is always higher than the rate of CRIs among these data. The F_1 score is the harmonic mean of the rate of predicted CRIs and the accuracy of those predictions. Thus, an F_1 score greater than the CRI rates indicates that we will detect CRI-potential trains more efficiently compared to random detection. One way to exploit these predictions would be to prioritize

maintenance operations on trains for which a CRI is predicted. On the other hand, the prediction probabilities are also exploitable. For some anomalies in the Reliability database, the corresponding CRI probabilities are among the highest. For instance, the chance that, on 10-07-19, the engine block of train number 236 breaks down the next day is among the highest with 3.49% of one-day probabilities. Furthermore, the reliability database shows an anomaly for this EB on 11-07-19. In addition, there were no MTDA alerts raised for this EB between 08-07-19 and 10-07-19. Therefore, the prediction probabilities can be associated with the MTBA alerts to detect the anomalies of a TGV train set. Similarly, the study of the most incidentogenic explanatory variables in our models can be used to enrich the equations of the BAT alerts.

Using a sampling strategy to correct for the class imbalance problem improves AUC of every use case for almost all algorithm. For instance, the best AUC for the TGV Engine Block data set is 0.70 without correction and 0.75 with sampling correction. The method used is found to not impact signifantively the metric scores on all experiments; either over or under sampling with any method is equivalent in term of model performance. Though, it is not inn term of computational resource and number of parameter to adjust. Hence, the random under sampling method is used on all experiment for every use case.

Variable dependence In the following, we apply the 'Top-N first' filter to select the N most frequent codes and evaluate the effect of adding less frequent codes and increasing the window size. Precisely, for each N, we compute different 5x cross-validated metrics (precision, recall, macro, F_1) for a set of window sizes. The results are presented in the form of a heat map in Figure 3.10. Notably, it shows an improvement of the area under the curve when increasing the dimension of the variable space. The maximum is reached for a window of twelve hours over all dimensions. We note a significant and constant improvement of the prediction value as a function of the size of the sequences. Nevertheless, this effect is probably due to the mechanical effect of the number of positive sequences relative to the number of negative sequences on the test set.

3.5.2 Pattern regression analysis

We consider the framework of the supervised classification. Let $\mathcal{Y} \in \{0, 1\}$ and $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ the training dataset. In the following, the entries may represent the number of occurrences of a default code on a sequence of a given size, the sum of the times of appearance in the sequence or another form of aggregation along with contextual data. The target associated with each vector obtained will be, with few exceptions, the appearance of a failure in a given time interval around the sequence as described in section 3.3

A central random variable is the so-called *regression function* defined as follow.

Definition 3 (Regression function)**.** . Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F} = \sigma(\mathcal{X}\times\mathcal{Y}), \mathbb{P})$ be a measured space, $x \in \mathbb{R}^d$ we call regression function

$$r^i(x) = \mathbb{P}[Y = i|X = x]. \tag{3.8}$$

In the classification framework we will note $r^k(x)$ the regression function in $Y = k$. This function will be frequently used to assess the relevance of the extracted patterns. For instance, an important question is to evaluate whether codes alone had explanatory power. To that end, we only look at the presence or absence of a pattern of code in the sequences constructed by the methodology described in the section 3.3. Therefore, for a dictionary of error codes size $d = |\Sigma|$, there is 2^d possible pattern and $\mathcal{X} = \{0, 1\}^d$.

The figure 3.6 shows the relative frequency of isolated error codes. Formally, we compute $r(x_i)$ with $x_i = \{0, ..., 1, ..., 0\}$ (the vector everywhere null except on its i-th component) and each point represents the frequency of occurrence in the healthy sequences (labelled as $Y = 0$) versus the frequency of occurrence in the faulty sequences (labelled as $Y = 1$). A point on the bisector $y = x$ indicates that the pattern x is non discriminating between healthy and faulty sequences. It shows that codes taken alone have a limited explanatory power for NAT use case. Hence it is necessary to consider associations of pattern to hope to capture the signal linked to a breakdown. In the same fashion, figure 3.7 plots the relative frequencies for each type of breakdowns reported.

Pattern extraction for explainability We perform a pattern extraction by a type *a priori* algorithm which searches for frequent closed items of support greater than a threshold $\mu \geq 0$ (Agrawal and Srikant, 1994). The sequences are divided into two classes according to the method described in 3.3 then the patterns are extracted for different values of μ and a maximum length of the pattern. We then compare these patterns by calculating their support on each class. In all, several thousands of patterns have been extracted by this method. Figures 3.9 show the result of this search. Among all these patterns, we are only interested in those that best discriminate the two positive sequences from the negative ones. We extract the patterns farthest from the bisector $(f(x) = x)$ on $[0, 1]$ in order to submit them to the business expertise to evaluate their relevance.

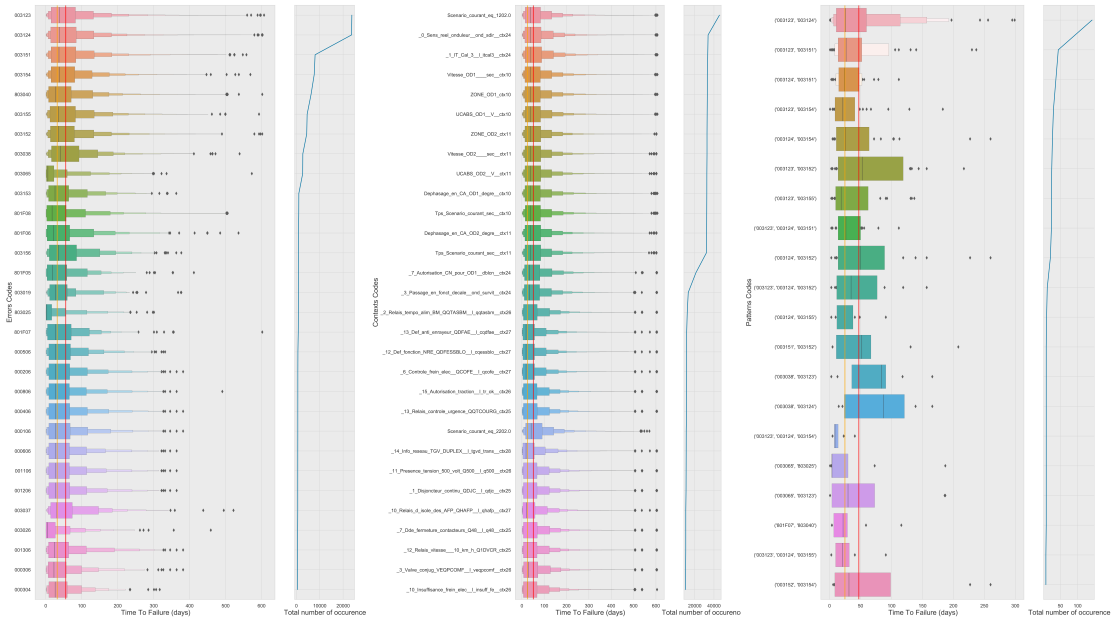


Figure 3.8: Boxen plot of the time to anomaly for the most frequent error codes (left), context codes (middle) and patterns (right) aggregated on one day time window (see section 3.3) for the high speed train use case. On each plot, the left figure gives the distribution of the time to anomaly with respect to the considered event and the right figure the frequency of occurrence in the whole data set.

Table 3.4: Test Accuracy, Recall and AUC 5× cross-validated on datasets reported in Table 3.3.

	X Gradient Boosting			Random Forest			Light Gradient-Boosting Machine			Categorical Boosting			Linear Regression			k-Nearest Neighbors		
	[1]	[1,7]	[1,7,14]	[1]	[1,7]	[1,7,14]	[1]	[1,7]	[1,7,14]	[1]	[1,7]	[1,7,14]	[1]	[1,7]	[1,7,14]	[1]	[1,7]	[1,7,14]
NAT																		
AUC	0.728	0.769	**0.927**	0.726	0.767	**0.913**	0.732	0.769	**0.926**	0.727	0.768	**0.927**	0.714	0.732	**0.899**	0.614	0.643	**0.841**
Accuracy	0.906	0.907	**0.929**	0.906	0.907	**0.928**	0.906	0.907	**0.929**	0.906	0.907	**0.93**	0.905	0.905	**0.918**	0.89	0.897	**0.922**
Recall	0.0398	0.0465	**0.403**	0.0411	0.0479	**0.416**	0.0238	0.0372	**0.401**	0.0413	0.0474	**0.407**	0	0.0002	**0.245**	0.106	0.105	**0.419**
F1	0.0742	0.0862	**0.519**	0.0762	0.0885	**0.523**	0.0455	0.0702	**0.516**	0.0765	0.0877	**0.523**	0	0.0003	**0.362**	0.154	0.16	**0.505**
TGV Engine Block																		
AUC	0.698	0.745	**0.769**	0.699	0.74	**0.766**	0.713	0.759	**0.774**	0.675	0.724	**0.729**	0.697	0.732	**0.738**	0.561	**0.568**	0.56
Accuracy	0.65	0.698	**0.707**	0.643	0.691	**0.704**	0.652	0.701	**0.714**	0.615	0.667	**0.675**	0.649	0.688	**0.692**	0.546	**0.556**	0.546
Recall	0.558	0.612	**0.655**	0.57	0.615	**0.656**	0.557	0.597	**0.653**	0.403	0.487	**0.577**	0.562	0.549	**0.596**	0.526	0.542	**0.543**
F1	0.614	0.669	**0.69**	0.615	0.665	**0.689**	0.616	0.666	**0.695**	0.511	0.594	**0.64**	0.615	0.637	**0.66**	0.537	**0.549**	0.545
TGV Doors																		
AUC	0.728	0.73	**0.758**	0.72	0.725	**0.749**	0.733	0.73	**0.756**	0.634	0.632	**0.659**	0.699	0.707	**0.725**	**0.582**	0.578	0.562
Accuracy	0.659	0.668	**0.692**	0.659	0.674	**0.683**	0.671	0.669	**0.692**	0.597	0.594	**0.608**	0.645	0.653	**0.667**	**0.567**	0.556	0.55
Recall	0.591	0.59	**0.609**	0.608	0.616	**0.625**	0.575	0.564	**0.597**	0.611	**0.645**	0.628	0.547	0.531	**0.561**	0.542	**0.552**	0.541
F1	0.634	0.64	**0.664**	0.641	0.654	**0.663**	0.636	0.63	**0.659**	0.602	0.613	**0.616**	0.606	0.605	**0.627**	**0.556**	0.554	0.546

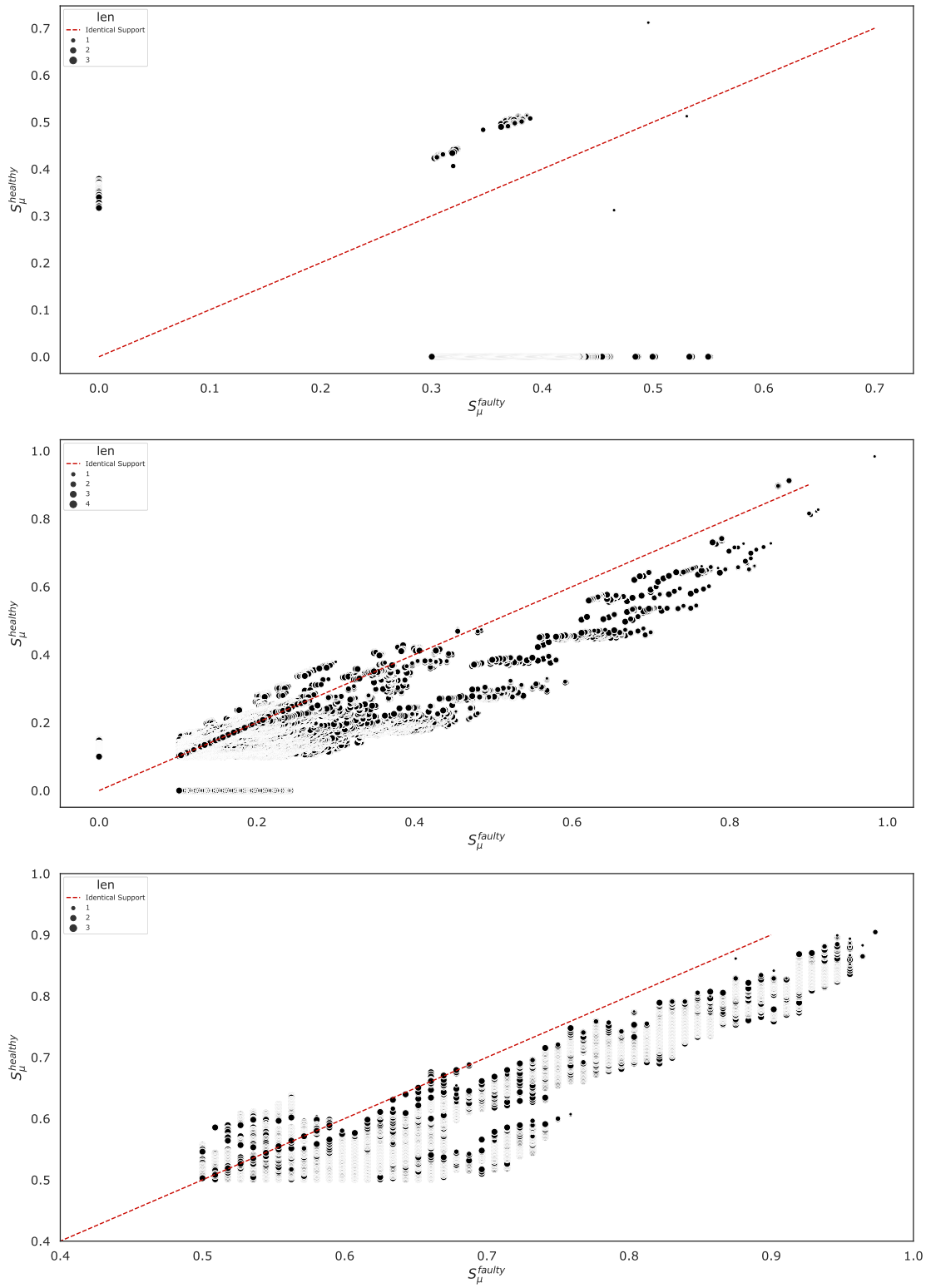


Figure 3.9: Support on each class of patterns extracted by algorithm *a priori* (Agrawal and Srikant, 1994) for $\mu = 1\%$ and $\mu = 4\%$ and patterns of different sizes for the Engine Block (up), TGV Doors (middle), NAT (bottom) dataset (see Table 3.4. Each black point is a pattern of codes with size representing the length of the pattern. Patterns that are in the upper half of the figure are the patterns that appears mostly near breakdowns events and pattern that are in the bottom half of the bisector (red dotted line) are the one appear in period without breakdowns..

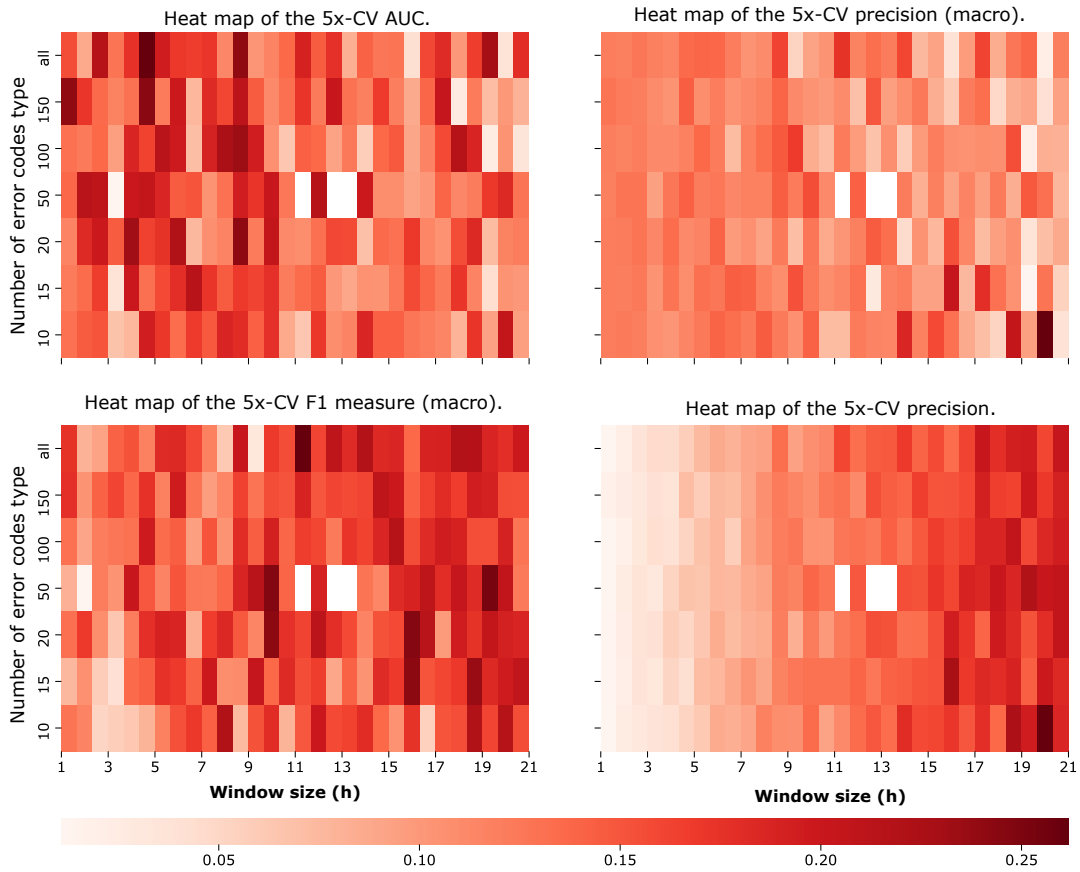


Figure 3.10: 5× cross-validated score for Random Forest on NAT dataset for various scores; AUC (upper left), precision macro (upper right), F_1 (bottom left) and precision (bottom right).

Part III

Pattern Mining

Chapter 4

Bayesian Feature Discovery for Predictive Maintenance

This chapter corresponds to the paper (Dib et al., 2021) published in *29th IEEE European Signal Processing Conference (EUSIPCO) proceedings*.

Abstract: This paper considers predictive maintenance, which is the task of predicting rare and anomalous events (typically, system failures) using event logs data, which are series of time-stamped symbolic codes emitted at regular or irregular intervals by a monitored system. Our objective is to find small sets of codes (called itemsets or patterns) that occur shortly before failures. Current prediction methods either produce patterns at a high computational cost or resort to kernel approaches which are often difficult to interpret. We introduce Bayesian Pattern Feature Discovery (BPFD), a new generic algorithm for pattern discovery. Our method, based on a pattern mining technique, produces informative and explainable features and is computationally efficient. The performance of BPFD is highlighted on real-world data sets, showing that enriching the feature space with the discovered patterns improves significantly the prediction power of a broad range of predictors and offers useful insight on the predictive maintenance task.

Key Words: Bayesian learning, pattern mining, predictive maintenance, variational inference.

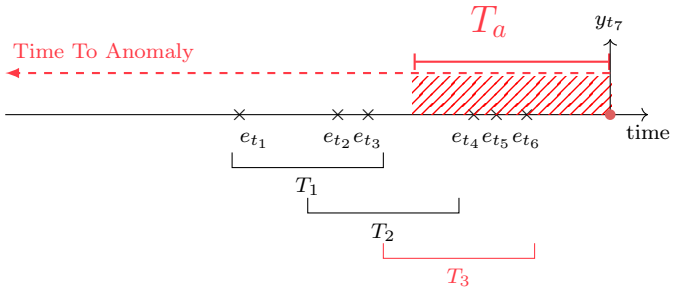


Figure 4.1: Temporal aggregation of log-events $(e_{t_1}, \ldots, e_{t_6})$ over sliding windows (T_1, T_2, T_3). In red, events that occur in the period T_a before y_{t_7} are considered anomalous and labeled $l = 1$. The aggregation produces the itemsets $x_1 = \{e_{t_1}, e_{t_2}, e_{t_3}\}, x_2 = \{e_{t_2}, e_{t_3}\}, x_3 = \{e_{t_4}, e_{t_5}, e_{t_6}\}$ and the labels $l_1 = 0$, $l_2 = 0$ and $l_3 = 1$. The goal is to correctly predict the labels l_i from the itemsets x_i.

4.1 Introduction

Predictive Maintenance (PM) aims to anticipate critical failures of large industrial systems to plan early and cost-effective interventions. Since maintenance can amount from 15% to 70% of the total operational cost (Bevilacqua and Braglia, 2000), PM is an important task to study, with far-reaching applications for the maintenance management of a number of industrial structures: transportation network (Ghofrani et al., 2018), power equipment (Koukoura et al., 2017), factory plant (Kolokas et al., 2018). Many fault-predicting procedures are based on event logs that provide information on the monitored system's health status. Event logs typically consist of event codes emitted at regular or irregular intervals. Formally, such data can be seen as temporal point processes of symbols taken from a finite dictionary. In that context, PM essentially amounts to identifying characteristic sequences (or patterns) of symbols that occur shortly before failures. The management of a railway fleet illustrates particularly well the importance of PM. SNCF, France's main railway company, uses event logs to predict failures of the train door system, one of the most critical equipments of its rolling stock. Any malfunction leads to the complete immobilization of the train and propagates delays to a large portion of the transportation network.

This work's main driver is to design an interpretable and efficient machine learning pipeline to detect potential occurrences of breakdowns of rolling stocks.

The prediction procedure uses event logs, which are time-stamped error codes e_t taken from a dictionary E of d distinct codes. These events are collected and processed by on-board equipment according to dedicated rules to which the end-user does not have access. These codes are produced during events deemed relevant by the manufacturer (for instance exceeding the threshold of an electrical signal

or a malfunction).

Procedures that make use of log events are particularly challenging since there is no natural order or distance on the space of symbols, thus making most machine learning models unsuitable. This issue can be overcome by kernel methods (Kung, 2014) but these approaches are difficult to interpret, which is a requirement for a predictive solution to be used in an industrial context. Another common strategy consists in transforming the prediction task into a binary classification task. In a nutshell, the signal is aggregated over sliding temporal windows (possibly overlapping) of fixed size. Features are simply the set of collected events within the window (called itemsets). For a given user-defined threshold period $T_a > 0$, a window is considered as anomalous (label "1") if it contains codes emitted in the period T_a before a failure, and normal (label "0") otherwise. This aggregation procedure is schematically illustrated on Figure 4.1. Even though popular (Basora, Olive, and Dubot, 2019), classification based solely on this construction is often unable to capture critical patterns of events that can be highly relevant in PM.

To tackle this issue, one can resort to methods from the related domains of Frequent Itemset Mining (FIM) and Discriminative Pattern Mining (DPM). FIM is the task of finding the most common patterns of a set in an exponentially large class of all possible combinations (Agrawal, Imielinski, and Swami, 1993). A famous application is the shopper recommendation problem, where the goal is to find the most common products that are bought together. DPM aims at searching for the set of patterns that best differentiate two subsets of a data set in the sense that a pattern occurs significantly more frequently in one of the classes. This framework has many applications such as consumer behavior analysis, RNA and DNA gene expression, subgraph mining, and anomaly detection. Generally, DPM algorithms start with a FIM step, where the most frequent itemsets are identified, then compute a statistical test for each itemset to determine if its presence is significantly different between two subsets (Hämäläinen and Webb, 2019). This often leads to an exponential number of statistical tests to perform and make many DPM methods computationally intensive.

In this work, we propose a Bayesian approach to explore the space of frequent itemsets in an efficient way. More precisely, we use a Bayesian Mixture Model to infer with a low computational cost the both frequent and discriminative itemsets. Also, we offer empirical proof of the general use of such discriminative patterns by considering them as features for the PM task. This results in a method that can extract an interpretable set of attributes and significantly improve any PM algorithm. Moreover, the Bayesian generative model allows for computing the posterior distribution and estimating the confidence intervals. Finally, additional expert-knowledge can be naturally introduced in the model *via* the choice of prior (Gelman et al., 2013).To the extent of our knowledge (and as pointed in (Hämäläi-

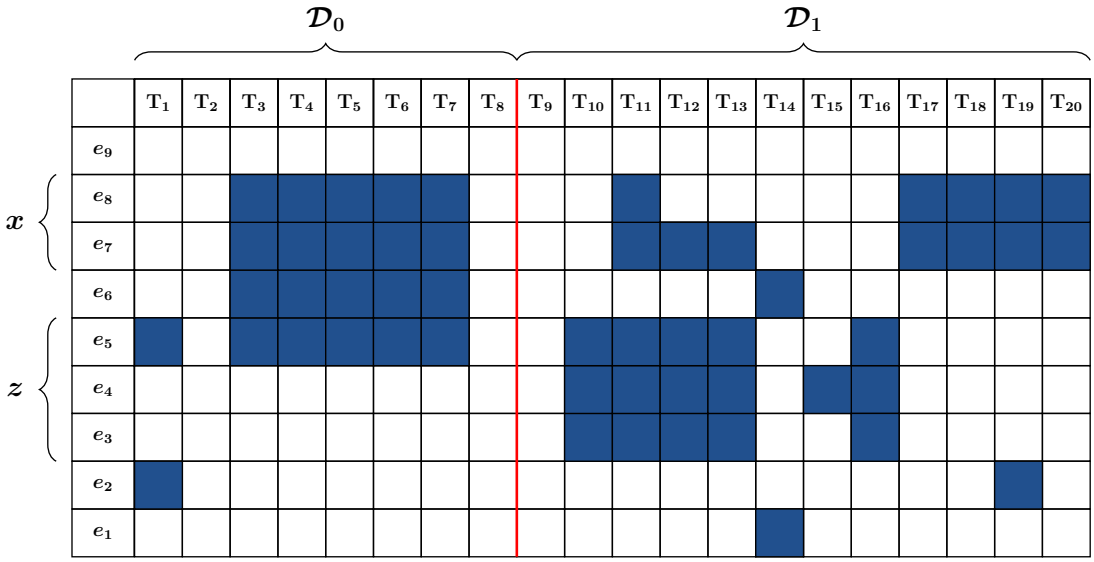


Figure 4.2: An example data set of events $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$. Row corresponds to items in $E = (e_1, \ldots, e_9)$ and columns to $n = 20$ samples. A blue colored area indicates that the item is present in the sample column considered. In this data set, the pattern $x = \{e_7, e_8\}$ in $\mathcal{E}$ seems to be nondiscriminative since $s_0(x) = s_1(x)$. On the contrary, the pattern $z = \{e_3, e_4, e_5\}$ appears to be specific to the positive class $l = 1$.

nen and Webb, 2019)), it is the first Bayesian approach towards DPM, and there has been no investigation of using pattern discovery methods based on discriminant pattern to the Predictive Maintenance task.

In Section 4.2, the basic concepts of FIM are introduced. Section 4.3 presents our approach to the DPM problem and application to signals of log events. The experiments are described and commented in Section 4.4.

4.2 Background

This Section introduces the concepts and main approaches of FIM and DPM.

4.2.1 Frequent Itemset Mining

Let $E = (e_1, \ldots, e_d)$ the base dictionary of events and $\mathcal{E} = \mathcal{P}(E)$ the collection of all 2^d possible patterns on E. The windowing procedure described in Fig. 4.1 transforms the sequence of log events into a database $\mathcal{D} = \{(x_i, l_i)_{i=1}^n\}$ of elements of $\mathcal{E} \times \{0, 1\}$ with the binary variable l indicating if a breakdown event occurred soon after the code emission. Note that the set $\mathcal{E}$ can be identified with the

Table 4.1: Contengency table for a pattern E and a database $\mathcal{D} = \mathcal{D}_0 \cup D_1$ to compute p_F.

	x	x^c	Size
$\mathcal{D}_1$	$s_1(x)$	$\lvert\mathcal{D}_1\rvert - s_1(x)$	$\lvert\mathcal{D}_1\rvert$
$\mathcal{D}_0$	$s_0(x)$	$\lvert\mathcal{D}_0\rvert - s_0(x)$	$\lvert\mathcal{D}_0\rvert$
Column totals	$s(x)$	$n - s(x)$	n

d-dimensional hypercube $\mathcal{X} = \{0, 1\}^d$, leading to the equivalence with the binary representation described in Fig. 4.2. We also denote $\mathcal{D}_0$ (respect $\mathcal{D}_1$) the samples in $\mathcal{D}$ associated with the target value $l = 0$ (respect $l = 1$) so that $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$.

The support of a pattern $x \in \mathcal{E}$ is defined as the number of samples of the database in which any pattern greater (with respect to $\subseteq$) than x appears. Formally,

$$s(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x \in \{z \in \mathcal{E} | x_i \subseteq z\}}. \tag{4.1}$$

In the same fashion, we denote $s_j(x)$ the support of the pattern $x \in \mathcal{E}$ in $\mathcal{D}_j$. In the context of predictive maintenance, $s_1(x)$ represents the number of times that a pattern of events appears close to a breakdown. Given a threshold $\mu \in [0, 1]$, the FIM task consists of finding the collection $\mathcal{TH}(\mathcal{E}, \mathcal{D}, \mu)$ of all *frequent patterns* in $\mathcal{E}$ defined has having support greater or equal than μ. The computation of such a collection is challenging since any algorithm has to explore a space size of $|\mathcal{E}| = 2^d$ elements and will exhibits exponential complexity $\mathcal{O}(n2^d)$. The key for pruning the set of possible patterns is the anti-monotonicity constraint which states that every sub-pattern of a frequent pattern is frequent. This approach spans a class of problems referred to as the Frequent Itemset Mining algorithms that can be used to extract $\mathcal{TH}(\mathcal{E}, \mathcal{D}, \mu)$ at reasonable computational cost(Agrawal, Imielinski, and Swami, 1993; Fournier Viger et al., 2016).

4.2.2 Discriminative Pattern

The classical DPM pattern procedure requires to perform a FIM procedure as described in Section 4.2.1 to obtain $\mathcal{TH}(\mathcal{E}, \mathcal{D}_0, \mu)$ and $\mathcal{TH}(\mathcal{E}, \mathcal{D}1, \mu)$ and compute the *contingency table* (Hämäläinen and Webb, 2019). Table 4.1 describes the complete contingency table for a pattern $x \in \mathcal{E}$ as the record of the support of x and x^c (which is the complementary pattern such that $x \cup x^c = E$) in $\mathcal{D}_0$ and $\mathcal{D}_1$. For instance, Fig. 4.2 displays the occurrence of each code in E in the sample i aggregated over the window T_i. The pattern $x = \{e_7, e_8\}$ produces a

contingency table with $s_0(x) = s_1(x)$. Since the data set $\mathcal{D}$ is the result of a stochastic process, one needs to design a statistical test to evaluate the statistical significance of the discrepancy between $s_0(x)$ and $s_1(x)$. The hypergeometric model with a Fisher test is the most commonly used framework for finding statistically significant pattern. Under the null hypothesis, the probability of observing the contingency table associated with x with $s_1(x) = a$ is

$$p_F(a) = \frac{\binom{|\mathcal{D}_1|}{a}\binom{|\mathcal{D}_0|}{s(x)-a}}{\binom{n}{s(x)}}. \tag{4.2}$$

The p-value is then obtained as the probability of observing a contingency table at least as extreme as the observed one. Since, in the worst case, a number of 2^d patterns must be considered, the probability of false discovery increases drastically and requires corrections. This is the goal of recent work on DPM algorithm such as LAMP and SPuManTe (Pellegrina, Riondato, and Vandin, 2019).

Nevertheless, all the above methods require the costly computation of $\mathcal{TH}(\mathcal{E}, \mathcal{D}_0, \mu)$ and $\mathcal{TH}(\mathcal{E}, \mathcal{D}_1, \mu)$ and can be challenging to interpret as the choice of the threshold for the p-value is a notoriously difficult problem that leads to misuses (Goodman, 2008).

4.3 Method

This Section introduces a new Bayesian approach for the DPM problem and its application to the signal of log events.

4.3.1 Bayesian interference for pattern discovery

Once the signal of error codes has been processed according to the procedure described in Fig. 4.1, we need to choose a generative model for the pattern database $\mathcal{D}$. We believe that a good trade-off is achieved between generality and complexity with a model assuming that the training data set is the result of a Bayesian Mixture Model (BMM) process with K mixture components (Pearson, 1894). This model assumes conditional independence given the mixture class and that the database is the result of sampling from multiple distributions p_k. The final number of parameters to evaluate for a K Bayesian Mixture Model is $K \times d$. We stress out that the choice of K controls the complexity of the model. Taking the number of components K to be large approximates the most exhaustive choice, which is the fully correlated Bernoulli model with 2^d parameters and is computationally intractable for even a moderate dimension d. The simple case of $K = 1$ is the independent and homogeneous Bernoulli model with *i.i.d.* samples. Simple combinatory

calculus gives a support function which only depends on the length of the pattern. Intuitively, it is similar to the experiment of throwing d identical coins with probability θ_0 and computing the probability of a given arrangement with given a number of heads. The too simple previous model assumes interchangeability on the elements e_i, complete independence between them and a similar distribution for all samples of the training data set. In the use case of DPM, this approach has the advantage of allowing computation of any quantity of interest; one computation is needed to infer the parameters and all conclusions can be drawn from it by sampling the posterior predictive distribution. The following gives a formal definition of the model.

Let $\mathrm{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be an i.i.d.sample of the pattern in the binary labeled database $\mathcal{D} = \{(\boldsymbol{x_i}, l_i)\}_{i=1}^n$ with $\boldsymbol{x_i} = (x_{ij})_{j=1}^d$ elements of $\{0,1\}^d$ and suppose the underlying model is a BMM with K components. For $k \in \{1, \ldots, K\}$, the k-ith sampling distribution $p_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})$ depends only on the parameter $\boldsymbol{\theta_k} = (\theta_{kj})_{j=1}^d$. Denoting λ_k the probability of sampling from the k-th component with $\sum_{k=1}^K \lambda_k = 1$, the global sampling distribution writes

$$p_(\boldsymbol{x_i}|\Theta, \boldsymbol{\lambda}) = \sum_{h=1}^{K} \lambda_k p_k(\boldsymbol{x_i}|\boldsymbol{\theta_k}), \tag{4.3}$$

where $\Theta = (\boldsymbol{\theta_k})_{k=1}^K$ and $\boldsymbol{\lambda} = (\lambda_k)_{k=1}^K)$. The conditional independence hypothesis for each Bernoulli component applied to the mixture distribution p_k leads to

$$p_k(\boldsymbol{x_i}|\boldsymbol{\theta_h}) = \prod_{j=1}^{d} \theta_{kj}^{x_{ij}} (1 - \theta_{kj}^{x_{ij}}).$$

Since it is unknown to which component $k \in \{1, \ldots, K\}$ a sample i belongs to, it is needed to introduce the unobserved indicator w_{ik} defined by

$$w_{ik} = \begin{cases} 1 & \text{if sample } i \text{ drawn from the } k\text{-th component,} \\ 0 & \text{otherwise.} \end{cases}$$

Knowing the mixture component parameter $\boldsymbol{\lambda}$, the component indicator $\boldsymbol{w_i} = (w_{i1}, \ldots, w_{iK})$ for the sample i is thus distributed as Multin($\boldsymbol{\lambda}$). Finally, the joint distribution is derived as

$$\begin{aligned} p(\mathrm{X}, \mathrm{W}|\Theta, \boldsymbol{\lambda}) &= p(\mathrm{W}|\boldsymbol{\lambda}) p(\mathrm{X}|\mathrm{W}, \Theta) \\ &= \sum_{k=1}^{K} \lambda_k \prod_{i=1}^{n} p_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})^{w_{ik}}. \end{aligned}$$

The last step is to choose a proper prior distribution on the parameters. The natural choice (Gelman et al., 2013) is to respectively set a Beta and Dirichlet distribution for the mixture probability of occurrence Θ and the mixture parameters vector $\boldsymbol{\lambda}$. For a set of parameter $\Gamma = (\Theta, \boldsymbol{\lambda}, K)$ associated with the Bayesian Mixture Model $\mathcal{M}$ is summarized as follow

$$\begin{aligned} \boldsymbol{\lambda}|\boldsymbol{\alpha} &\sim \text{Dirichlet}\,(\boldsymbol{\alpha})\,, \\ \boldsymbol{w_i}|\boldsymbol{\lambda} &\sim \text{Multin}(\boldsymbol{\lambda}), \\ \theta_{kj}|\boldsymbol{\beta}, \boldsymbol{\gamma} &\sim \text{Beta}(\boldsymbol{\beta}, \boldsymbol{\gamma}), \\ x_{ij}|\theta_{kj} &\sim \text{Bernoulli}(\theta_{kj}). \end{aligned} \tag{4.4}$$

4.3.2 The BPFD algorithm

The BPFD algorithm is based on choosing the model described in Section 4.3.1 as a generative model for the samples $\mathcal{D}$ and computing the *odd ratio support* to compare the patterns between classes. The steps are described in the following.

Preprocessing The first step is to transform the sequential data to a binary matrix as described in Fig. 4.1. Note that any continuous feature can be transformed into a multi-categorical feature.

Inference Set the hyperparameter $\boldsymbol{\alpha} = (\frac{1}{K}, \ldots, \frac{1}{K})$. An Expectation Minimization (Dempster, Laird, and Rubin, 1977) procedure is performed on $\mathcal{D}_0$ and D_1 to infer the set of parameters Γ_0 and Γ_1 associated with the models $\mathcal{M}_0$ and $\mathcal{M}_1$.

Discriminant Pattern computation The discriminative power of a pattern $x \in \mathcal{E}$ is evaluated through the odd ratio support

$$r(y) = \frac{p(\mathcal{M}_1 \mid x)}{p(\mathcal{M}_0 \mid x)} \tag{4.5}$$

$$= \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)} \times \frac{p(x \mid \Gamma_1)}{p(x \mid \Gamma_0)}. \tag{4.6}$$

Classification The best discriminative patterns are then added to the original training data set $\mathcal{D}$ and classification is performed.

The main advantage of this automatic feature extraction method is that it can be applied to any data and will return new features that will often be easy to interpret. The method does not require a threshold μ and can thus discover patterns that the traditional approach would not explore. Additionally, since the posterior sampling distribution can be simulated thanks to 4.4, the confidence

interval on the value of $r(y)$ can be directly obtained. Note that the potential imbalance between the two classes is naturally taken into account by the prior distribution effect (Gelman et al., 2013). Finally, the method is computationally efficient since the EM algorithm converges rapidly to a local minimum of the log posterior distribution.

4.3.3 Identifiability issue

The identifiability is a fundamental issue of mixture models of finite measures that's largely overlooked in the literature when statistical inference is performed on such class of probability distributions. There's two main identifiability problem. the first arise from the invariance of the log-likelihood under any permutation of the component also known as the Label Switching Phenomenon (LSP) and the second from the non-uniqueness of the optimal solution given a dataset of samples. We stress out that these issues are essential to tackle in order to obtain a relevant inference.

Let $\mathcal{T}_K$ be the set of permutation on the set $(1, \ldots, K)$. For $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K) \in \mathcal{T}_K$ we define the corresponding permutations $\boldsymbol{\tau\theta} = \left(\theta_{\tau_1}, \ldots, \theta_{\tau_K}\right)$, $\boldsymbol{\tau w} = \left(w_{\tau_1}, \ldots, w_{\tau_K}\right)$ and $\boldsymbol{\tau z} = (\tau_{z_1}, \ldots, \tau_{z_n})$. A symmetric prior distribution (as in the case of a non-informative prior) will be invariant under any permutation

$$\forall \boldsymbol{\tau} \in \mathcal{T}_K, \quad p(\tau\boldsymbol{\theta}, \tau\boldsymbol{w}) = p(\boldsymbol{\theta}, \boldsymbol{w}).$$

Given that $p(\boldsymbol{\theta}, \boldsymbol{w}|\boldsymbol{x}) \propto L(\boldsymbol{\theta}, \boldsymbol{w}; \boldsymbol{x}) p(\boldsymbol{\theta}, \boldsymbol{w})$, the posterior of the model is then himself invariant under any permutation of the component, thus any Monte Carlo Markov Chain (MCMC) method will switch between the different permutation of the parameters and exhibit $K!$ modes in his sample distribution. On multiple chain sampling, this will likely result to a poor $\hat{R}$ score (also known as Gelman-Rubin score (Kucukelbir et al., 2015)) since the different chains will explore different area of the parameter space.

4.4 Experiments

The BPFD was initially designed to tackle the problem of Discriminative Pattern Mining for Predictive Maintenance on rail stock. Nevertheless, this approach is general and can be applied to any supervised classification problem. To demonstrate the validity and effectiveness of our approach and ensure full reproducibility, we evaluate the BPFD algorithm on various widely used and publicly available[1] data sets as well as on the industrial **Doors** data set. In addition, the method is compared across multiple classifiers against the Base Classifier (BC) and the popular

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

Polynomial Feature (PF) approach (Kuhn and Johnson, 2019). The results are reported in Table 4.3.

4.4.1 Setup

The BPFD algorithm presented in Section 4.3.2 and the Expectation-Minimization procedure are implemented using the Tensorflow 2.4 and Python 3.8. The experiments run on a Quad-core Intel i7 10th Gen @ 2.5 GHz. The source code and complementary experiments, including additional classifiers and data sets, are available online[2] for reproducibility.

[2] https://github.com/amirdib/bpfd

Table 4.2: Test Accuracy, Recall and AUC 10× cross-validated for BPFD, PF and BC classifiers (with grid-search hyperparameter tuning) on datasets reported in Table 4.3.

	X Gradient Boosting			Random Forest			Light Gradient-Boosting Machine			Categorical Boosting			Linear Regression			k-Nearest Neighbors		
	BC	PF	BPFD	BC	PF	BPFD	BC	PF	BPFD	BC	PF	BPFD	BC	PF	BPFD	BC	PF	BPFD
ijcnn1																		
AUC	0.728	0.769	**0.927**	0.726	0.767	**0.913**	0.732	0.769	**0.926**	0.727	0.768	**0.927**	0.714	0.732	**0.899**	0.614	0.643	**0.841**
Accuracy	0.906	0.907	**0.929**	0.906	0.907	**0.928**	0.906	0.907	**0.929**	0.906	0.907	**0.93**	0.905	0.905	**0.918**	0.89	0.897	**0.922**
Recall	0.0398	0.0465	**0.403**	0.0411	0.0479	**0.416**	0.0238	0.0372	**0.401**	0.0413	0.0474	**0.407**	0	0.0002	**0.245**	0.106	0.105	**0.419**
F1	0.0742	0.0862	**0.519**	0.0762	0.0885	**0.523**	0.0455	0.0702	**0.516**	0.0765	0.0877	**0.523**	0	0.0003	**0.362**	0.154	0.16	**0.505**
cod-rna																		
AUC	0.776	0.496	**0.815**	0.776	0.496	**0.815**	0.776	0.496	**0.815**	0.776	0.496	**0.815**	0.765	0.495	**0.813**	0.706	0.5	**0.764**
Accuracy	0.718	0.667	**0.775**	0.718	0.667	**0.775**	0.717	0.667	**0.775**	0.718	0.667	**0.775**	0.713	0.667	**0.774**	0.688	0.591	**0.739**
Recall	**0.588**	0	0.383	**0.585**	0	0.386	**0.592**	0	0.384	**0.588**	0	0.384	**0.512**	0	0.364	0.483	0.231	**0.516**
F1	**0.581**	0	0.532	**0.58**	0	0.534	**0.583**	0	0.532	**0.581**	0	0.532	**0.544**	0	0.518	0.503	0.263	**0.568**
a9a																		
AUC	0.89	**0.896**	0.88	0.863	0.869	**0.875**	0.894	0.9	**0.903**	0.894	0.9	**0.904**	0.893	0.902	**0.902**	0.837	0.848	**0.85**
Accuracy	0.841	0.844	**0.846**	0.825	0.826	**0.829**	0.844	0.846	**0.849**	0.844	0.847	**0.848**	0.841	**0.849**	0.847	0.817	**0.826**	0.824
Recall	0.597	0.604	**0.615**	0.564	**0.582**	0.578	0.606	0.613	**0.626**	0.595	0.606	**0.611**	0.581	**0.611**	0.604	0.566	0.584	**0.589**
F1	0.643	0.649	**0.658**	0.607	0.616	**0.619**	0.651	0.656	**0.666**	0.646	0.654	**0.66**	0.637	**0.659**	0.655	0.597	0.616	**0.617**
Doors																		
AUC	0.707	0.691	**0.736**	0.713	0.707	**0.753**	0.706	0.697	**0.739**	0.722	0.715	**0.749**	0.635	0.629	**0.637**	0.557	**0.574**	0.574
Accuracy	0.643	0.629	**0.679**	0.655	0.645	**0.686**	0.647	0.637	**0.681**	0.663	0.657	**0.684**	**0.6**	0.592	0.597	0.546	**0.551**	0.551
Recall	0.614	0.608	**0.642**	0.594	0.585	**0.608**	0.595	0.577	**0.619**	0.569	0.56	**0.592**	0.652	**0.674**	0.648	**0.545**	0.526	0.526
F1	0.632	0.62	**0.667**	0.632	0.622	**0.659**	0.627	0.613	**0.66**	0.627	0.619	**0.652**	0.62	**0.623**	0.617	**0.545**	0.539	0.539

Table 4.3: Characteristic of the experimental datasets.

Name	n	d	$\frac{\lvert\mathcal{D}_0\rvert}{\lvert\mathcal{D}_1\rvert}$
ijcnn1	91701	35	0.10
cod-rna	271617	17	0.5
a9a	32561	124	0.31
Doors	6349513	153	0.03

4.4.2 Experiments

Data sets The BPFD algorithm is tested on three public data sets commonly used for benchmark: **ijcnn1** consists of binarized maintenance data, **cod-rna** is a table of labeled strains of RNA and **a9a** is a record of census data to predict income of a household. The **Doors** data set has been provided by the French National Railway Company and consists of a database of log-events emitted by 143 trains' doors collected over twenty-four months. For each data set, the number of samples n, the size of the base dictionary $d = |E|$ and the class imbalance $\frac{\mathcal{D}_0}{\mathcal{D}_1}$ is reported in Table 4.3.

Feature Discovery We consider the $10\times$ cross-validated F_1, Area Under the Curve (AUC), Recall and Accuracy metrics to evaluate the improvement over the classifiers reported in the result Table 4.2 with $70\% - 30\%$ train-test split. In particular, the proposed approach improves the overall AUC score for almost all data sets and classifiers considered. For instance, the **ijcnn1** experiment exhibits an AUC of 0.927 for the Extreme Gradient Boosting (XGB) classifier whereas the vanilla approach scores at 0.769. On all data sets, the gain seems particularly significant for the Recall metrics. It seems that the discriminating pattern mined allows the classifier to be more sensitive. This is particularly important in the Predictive Maintenance domain where false negatives are generally the most costly type error that can be made.

Discriminative Patterns BPFD is compared with state-of-the-art SPuManTe (Pellegrina, Riondato, and Vandin, 2019) test and retrieve most of the patterns with comparable significance level. These patterns revealed to be very informative about the link between a breakdown and pattern of code emission as well as explaining why a given algorithm would produce an incorrect prediction. As an example, in the case of **Doors** fault prediction, the Base Classifier would typically raise the probability of breakdowns after a manual blocking of a door by the

onboard personnel represented by the event $e_m = \{"Locking\ Door"\}$. Our approach shows that some patterns that indicate whether this blocking is intended or not. For instance, the pattern $x = \{"Locking\ Door", "Unlocking\ Door"\}$ is not interpreted as an alert with BPFD as it is likely to be a handling error. More complex events have been extracted and their relevance validated with maintenance experts.

4.5 Conclusion

In this work, we introduced a new algorithm for DPM and derived a Feature Discovery method to improve performance of any classifier in the supervised learning framework. This method is tested on various real-world and production data. In addition to the metric score improvement, our approach offers explainable insights on the classification task. Some extensions of this work could include using the bread-stick model to alleviate the need for a mixture parameter K. The present framework can easily be extended to multi-categorical classification. We plan to consider it in future work.

Chapter 5

Localized Pattern Mining

This chapter corresponds to the preprint (Cousins* and Dib*, 2021)[1] submitted to the *IEEE International Conference on Data Mining (ICDM 2021)*.

Abstract: This paper considers the problem of finding the best sampling strategy for pattern mining problems, which can be stated as computing the frequency at which a pattern, or a set of events, occurs in a database. This problem is ubiquitous in data mining, and is typically intractable due to the exponentially large number of possible patterns that must be evaluated. Recent approaches use traditional tools from statistical learning theory to obtain uniform additive bounds on these frequencies, which are effective for *frequent patterns*, but are generally unsatisfying for *infrequent patterns*, which are typically the hardest to mine exactly. In this work, we propose the first bound based on *localized Rademacher averages* (LRAs) in the context of pattern mining. We show that localized Rademacher averages are sufficient to obtain relative confidence interval estimates on pattern frequencies, as well as other interestingness measures, such as the *lift*, *confidence*, or *odds ratio*. In contrast, previous techniques fail to do so for low-frequency patterns. Our methods rely on standard tools in the pattern mining repertoire, such as closed pattern families, antimonotonicity, Monte-Carlo Rademacher averages, and new techniques we introduce to address the problem-specific computational challenges arising from evaluating the localized Rademacher average. The performance of our approach is empirically demonstrated on real-world datasets, wherein exhibit fast convergence rates for the considered subclass of patterns, sharply contrasting existing work.

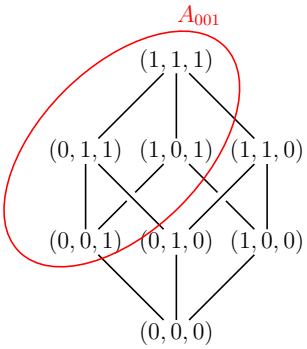
[1] equal contributions.

5.1 Introduction

Consider the independent observations of a stochastic process of events such as random graphs, sentences, or DNA sequences. A most natural question to ask is *how many* or *which patterns* of events occur with *at least* a given frequency. Unfortunately, answering such questions requires to enumerate the exponential number of possible association of events which is NP-HARD (Yang, 2004). This is known as the *pattern mining* task, and it is one of the most prominent problem in Data Mining, with various application in a broad range of domains that span from query database (Pavlov, Mannila, and Smyth, 2003), graph mining (Mansha et al., 2016; Shang et al., 2017; Zheng et al., 2013), sequence mining (Sirisha, Shashi, and Raju, 2014), anomaly detection (He et al., 2016; Bogojeski et al., 2020; Aggarwal et al., 2018; Laredo et al., 2019).

Since the seminal work of (Agrawal, Imielinski, and Swami, 1993), most of the proposed algorithms to tackle the frequency pattern mining task leverage the *anti-monoticity property*, which can be stated as follows: any pattern is at most as frequent as any pattern that contains it. Using a user-specified frequency threshold μ in $[0, 1]$, it is thus possible to prune the space of possible patterns in a breadth-first search fashion by generating new pattern candidates at each step, and stop the exploration whenever patterns with support less than μ is encountered. Although efficient, these methods suffer from several limitations. For one, the requirement for a frequency threshold bars from estimating the frequency of rare patterns. Second, time and memory complexity can still be prohibitive, even for a high-frequency threshold, when the number of samples in the database is large, as is often the case in modern applications. Finally, deterministic pattern mining techniques do not consider the fact that the database results from a random generative process. Hence, it is not possible to obtain a confidence interval for the resulting mined patterns.

Each of this limitation is overcome by sampling methods REF. In this setting, one only considers a subset of size n of the pattern database to mine from and compute a bound on the obtained frequencies. It leads to a relation between the size of the subset and precision of the estimated frequencies with respect to the true unknown frequency. Formally, the task consists of computing the size of the subset of the database n needed to obtain an estimation of any frequency at precision $\varepsilon \in [0, 1]$ with probability at least $1 - \delta$. It thus relates to bounding an empirical process generated by an unknown distribution indexed on a finite functional space (Boucheron, Lugosi, and Massart, 2013).



	e_1	e_2	e_3	e_4
t_1	1	1	0	0
t_2	1	1	0	1
t_3	1	1	1	1
t_4	1	0	1	0
t_5	0	1	1	1

Figure 5.1: Lattice representation of $\mathcal{X}$ (left) and table view of a database from a realisation ω (right). In the left figure, the red circle represents the subset associated with the itemset $(0, 0, 1) = \{e3\}$. The right table gives that support of this item is $\frac{3}{5}$.

5.1.1 Related work

This work uses tools and concepts from statistical learning theory to bound the precision of estimated frequencies computed on a sample.

Stemming from the work of Agrawal and Srikant (1994), the frequent pattern mining task consists of computing the frequency of a set of items, or patterns, that appear in a database more than a certain threshold $\mu \in [0, 1]$. The frequency $s(x)$, sometimes referred as *support*, of a pattern x is defined as the number of samples of the database on which a pattern greater (with respect to $\subseteq$) than x appears. We stress that FIM is the starting point of various techniques related to data mining tasks. For instance, Association Rule Mining (Agrawal and Srikant, 1994; Zaki and Hsiao, 2005) considers the problem of finding rules between itemsets at a given confidence level. For two patterns $x, y \in \mathcal{E}$, the goal is to mine rules $x \to y$ such that the support $s(x \vee y)$ and confidence measure $c(x, y) = \frac{s(x \vee y)}{s(x)}$ are no greater than two threshold $\mu, \nu \in [0, 1]$. The confidence measure inform about the co-occurrence of two patterns while taking into account their frequency in the database.

From the problem of FIM quickly emerged the computation time in the high dimensional case. One idea introduced and empirically tested by ("Efficient Algorithms for Discovering Association Rules") extracted FIS by sub-sampling the database. Toivonen (1996) proposes a method using the Hoeffding bound (Hoeffding, 1963), wherein a sample size in $bigO(\frac{d \ln \frac{1}{\delta}}{\varepsilon^2})$ is sufficient to estimate all itemset frequencies. (Zaki et al., 1998) showed that this method leads to a drastic reduction in computation time (by one order of magnitude) and memory requirement. However, even though quick and straightforward to compute, this method does not take account of the sparsity of the data, leading to very conservative estimation.

	Union Bound	Rade.	LRA
Analytic	(Toivonen, 1996)	AMIRA	Bartlett / Oneto
Monte-Carlo	N/A	MCRapper	This Work

Table 5.1: Summary of related literature and positioning of this work.

More refined bounds can be obtained by using distribution-dependent *complexity measures* from statistical learning theory. Recent work (Riondato, 2014) bounds the VC dimension by the *D-index*, which is an upper bound on the maximum transaction size, and shows that $n \in bigO(\frac{D \ln \frac{1}{\delta}}{\varepsilon^2})$ samples suffice.

With *data-dependent* bounds, the additive error ε is a function of the sample, thus the sufficient sample size n cannot be explicitly computed *a priori*. At a given level ε and for δ, the value of n cannot always be explicitly computed due to the data-dependence of epsilon with the sample. This framework corresponds to the *progressive sampling* method and roughly consists of drawing iteratively larger samples, until the desired ε threshold is met.

Existing methods use (global) Rademacher averages to mine *frequent* or *top-k* itemsets, which is appropriate, as we do not require sharp bounds on low-frequency itemsets. In particular, (Riondato and Upfal, 2015) uses an analytical counting argument to get a loose bound on the empirical Rademacher average, whereas (Pellegrina et al., 2020) use a Monte-Carlo approximation strategy to get a sharp bound, at the cost of additional computation. Moreover, in some settings like *k-mer* frequency estimation, all patterns are low-frequency, but some are significantly lower than others. This led to domain-specific methods to efficiently estimate low-frequencies with biased window-based domain-specific estimators (Pellegrina, Pizzi, and Vandin, 2020). In contrast, our method provides sharp frequency estimates for both low and high-frequency patterns in generic pattern mining settings.

The classical notion of Rademacher complexities, which considers the entire hypothesis functional space, only allow for establishing the slow rate $bigO(\frac{1}{\sqrt{n}})$ although empirical studies reported *fast rate* in $bigO(\frac{1}{n})$. This fact was one of the main driver for the localized Rademacher framework (Bartlett, Bousquet, and Mendelson, 2005).

5.1.2 Contributions

This work marks the first use of localized Rademacher complexity to the low-frequency pattern mining problem. We also introduce the Monte Carlo localized Rademacher average, improving over existing analytic methods to more sharply bound LRAs.

Computational challenges arise in both the Monte-Carlo (evaluating Monte-Carlo RAs) and localization (computing fixed points involving localized RAs) aspects of this work. In particular, we present Algorithm 1 to efficiently compute MC-LRAs, give applications to contrast pattern mining, and discuss and many other settings where such bounds prove invaluable. We then formalize a target task for low-frequency pattern mining, and present Algorithm 2, which uses progressive sampling to realize this objective efficiently, with applications to mining importance measures with *relative error guarantees*. After tackling these issues, we find that the natural combination of these ideas leads to sophisticated finite-sample guarantees, with strong performance for low-frequency patterns.

Our experimental evaluation shows that localized methods soundly improve over a Bennett Union bound approach (also appropriate for low-frequency estimation), and that Monte Carlo LRAs improve over looser analytic LRA bounds. This mirrors the well-known progress in the literature from Hoeffding-union bounds, to loose analytic global Rademacher average pounds, to Monte Carlo Rademacher averages; the difference is of course that the aforementioned techniques all produce uniform estimation guarantees, whereas our methods produce better guarantees for low-frequency estimation patterns, and are the suitable for relative pattern frequency estimation objectives.

5.2 Background

This section introduces the necessary tools and concepts from the Data Mining and Statistical Learning domain.

5.2.1 Pattern mining

Let $E = (e_1, \ldots, e_d)$ represents a list of items, $\mathcal{P}(E) = \mathcal{X}$ be the set of all itemsets or patterns, and $\mathcal{D} \in \mathcal{X}^{\otimes N}$ the transaction database of size N. The support or frequency $s(t)$ of an itemset $t \in \mathcal{X}$ is defined as the number of times it appears on the transaction database $\mathcal{D}$ divided by N.
Note that the set $\mathcal{X}$ is in bijection with the hypercube of dimension d, thus we can set $\mathcal{X} = \{0,1\}^d$. Additionally, consider the collection of set $\mathcal{A} = \{A_t : t \in \mathcal{X}\}$ with $A_t = \{z \in \mathcal{X} : z \supseteq t\}$ the set of all itemsets greater than $t \in \mathcal{X}$ with respect to $\supseteq$. We can define the functional space associated with this collection of events

as $\mathcal{F} = \{f_t : t \in \mathcal{X}\}$ with $f_t = \mathbb{1}_{A_t}$. Figure 5.1 illustrates the case in which $d = 3$ with A_{001}.
Given distribution P, the support $s(t)$ of any itemset $t \in \mathcal{X}$ is defined as the probability of the event A_t

$$\begin{aligned} s(t) &= P\left(A_t\right) \\ &= Pf_t. \end{aligned}$$

In other words, the support of a pattern $x \in \mathcal{E}$ is defined as the number of samples of the database in which any pattern greater (with respect to $\subseteq$ than x appears. The computation of such a collection is intractable since any algorithm will have to evaluate the $|\mathcal{E}| = 2^d$ elements and will thus exhibit exponential complexity $bigO(n2^d)$. Instead, the sampling approach consists of computing these supports only for a small subset of $n \ll N$ and providing deviation bound for the true but unknown support. This approach relates to the problem of estimating bound on the suprema of empirical process (Boucheron, Lugosi, and Massart, 2013).

5.2.2 Suprema of an empirical process

This section presents the key tools of statistical learning theory and associated results. The central quantity is the deviation process over the function class $\mathcal{F}$. For every function f_t in $\mathcal{F}$, the deviation from the true support can be expressed as

$$\mathcal{S}_n\mathcal{F} = \sup_{f \in \mathcal{F}} \left| \hat{P}_n f - Pf \right|, \tag{5.1}$$

where $\hat{P}_n$ is the empirical counterpart of Pf for every function in $\mathcal{F}$. The main difficulty arises from the fact that this quantity depends on the unknown underlying distribution P. Most modern approaches use Rademacher Complexities (Boucheron, Lugosi, and Massart, 2013) as it generally leads to sharper bounds and can be easily empirically evaluated.

Definition 4 (Rademacher Averages)**.** Let $\sigma_1, \ldots, \sigma_n$ be an i.i.d. sample of Rademacher variables independent of the samples $(x_1, \ldots, x_n)$ and valued in $\{-1, 1\}$ with equal probability. Then, the (global) *empirical Rademacher average* complexity of $\mathcal{F}$ is defined as

$$\hat{\mathcal{R}}_n(\mathcal{F}, \boldsymbol{x}) = \frac{1}{n} \left| \sum_{i=1}^{n} \sup_{f \in \mathcal{F}} \sigma_i f(x_i) \right|, \tag{5.2}$$

and the *global Rademacher average* as $\mathcal{R}_n\mathcal{F} = \mathbb{E}\left[\hat{\mathcal{R}}_n(\mathcal{F}, \boldsymbol{x})\right]$.

Intuitively, Rademacher averages measure the richness of the functional family by evaluating its ability to fit random noise $\boldsymbol{\sigma}$. Moreover, the use of Rademacher complexity to uniformly bound processes is a standard tool in statistical learning thanks to famous Bousquet's inequality (Boucheron, Lugosi, and Massart, 2013). As a *sub-Poisson*, bound, we must first introduce the associated functions, used throughout this paper to sharply bound probabilistic quantities.

Definition 5 (Poisson Fenchel-Legendre Dual and its Inverses)**.**

$$\begin{aligned}
&\phi(a) = (1+a)\log(1+a) - a, \;\; a > -1 \\
&\hat{\phi}(a) = 1 - \exp\left[1 + W_{-1}\left(\tfrac{a-1}{e}\right)\right], \;\; \forall a \in [0,1] : \; \phi[-\hat{\phi}(a)] = a \\
&\breve{\phi}(a) = \exp\left[1 + W_{0}\left(\tfrac{a-1}{e}\right)\right] - 1, \quad \forall a \geq 0 : \; \phi[\bar{\phi}(a)] = a
\end{aligned} \tag{5.3}$$

Note that these functions are generally evaluated around 0, where $\phi(a) \approx \frac{a^2}{2}$, thus $\hat{\phi}(a) \approx \breve{\phi}(a) \approx \sqrt{2a}$. Given any $x > 0$, the following inequality holds with probability $1 - e^{-x}$ (Boucheron, Lugosi, and Massart, 2013)

$$\sup_{f \in \mathcal{F}} \hat{P}_n f - Pf \leq 2\mathcal{R}_n(\mathcal{F}, \boldsymbol{x}) + \nu\phi^{-1}(\frac{x}{\nu}), \tag{5.4}$$

where $\nu = 2\mathcal{R}_n(\mathcal{F}, \boldsymbol{x}) + \sigma^2$. $\hat{\mathcal{R}}_n(\mathcal{F}, \boldsymbol{x})$ is typically in $\mathcal{O}(\frac{1}{\sqrt{n}})$, as is the bound of Equation 5.4. Note that The bound in equation 5.4 is considered over the whole functional family. By considering only a fraction of the class $\mathcal{F}_r = \{f \in \mathcal{F} : T(f) < r\}$ that depend on the *localization* parameter r and a function $T : \mathcal{F} \to \mathbb{R}$ it is possible to greatly improve this result and obtain fast rate in $bigO(\frac{1}{n})$ under some mild conditions (Bartlett, Bousquet, and Mendelson, 2005) that are straightforward to verify in the application context of pattern mining. There is two main challenge toward this approach. The first one is to select a suitable radius r^* that can be obtained through a subroot function analysis. The second is the constant factor that can make a bound obtained by such method vacuous. These two challenges are overcome by the use of the Oneto's (Oneto et al., 2015) approach that we adapt to obtain a very sharp bound for the subset pattern mining problem.

5.3 Localized Pattern Mining

This section begins by describing the theoretical results for localized pattern mining. We then present the LocalMiner algorithm to bound the expected support Pf in terms of the empirical support $\hat{P}f$ for all $f \in \mathcal{F}$, with applications to discriminative pattern mining. Finally, we introduce the RLPS algorithm to derive relative confidence intervals, which we use to bound *interestingness measures*.

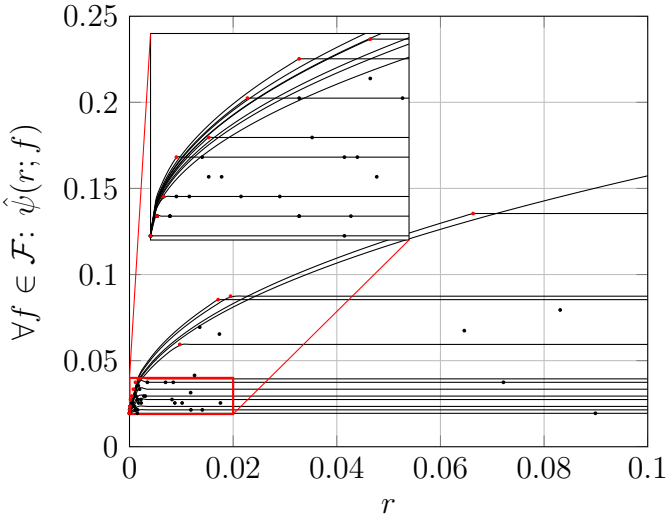


Figure 5.2: Illustration of piecewise $\hat{\psi}(r)$ for each function in $\mathcal{F}$, for a toy database generated by a homogeneous Bernoulli distribution with $d = 4$ and $n = 100$.

5.3.1 Localized empirical bound for pattern mining

At the heart of the localization method is the so-called *r-star-localized* family (Bartlett, Bousquet, and Mendelson, 2005), defined as

$$\mathcal{F}_r \doteq \left\{ \alpha f \middle| \alpha \in [0,1], f \in \mathcal{F}, P\left[(\alpha f)^2\right] \leq r \right\}. \tag{5.5}$$

Star-localization has the effect of scaling down high-variance functions, thus uniform convergence within the localized class $\mathcal{F}_r$ yields sharper guarantees for low-variance functions. More precisely, the Rademacher average of the localized class $\hat{\mathcal{R}}(\mathcal{F}_r)$ is decreasing in r, with $\mathcal{F}_1 = \mathcal{F}$. Unfortunately, since Pf is not known *a priori*, we must approximate $\mathcal{F}_r$ with an *empirically localized class* $\hat{\mathcal{F}}_r$. Following Oneto, 2020, Equation 5.102, we define

$$\hat{\mathcal{F}}_r \doteq \left\{ \alpha f \middle| \alpha \in [0,1], f \in \mathcal{F}, \hat{P}_n\left[(\alpha f)^2\right] \leq 3r + 5r\hat{\phi}\left(\frac{\ln \frac{4}{\delta}}{5nr}\right) \right\}. \tag{5.6}$$

Since $\hat{\mathcal{F}}_r \subseteq \mathcal{F}_r$ with high probability, considering the class $\hat{\mathcal{F}}_r$ allow to obtain variance-sensitive bounds across $\mathcal{F}$ (Oneto et al., 2015). To that end, we need to introduce the following quantities.

Definition 6. Let $\mathcal{F}$ be the functional family and $\hat{\mathcal{F}}_r$ the empirical star localized class (5.6). For Rademacher trial count m, sample size n, and any $\delta \in [0,1]$, define

the following

$$\hat{\psi}_n(r) \doteq 2\hat{\mathcal{R}}_n\left(\hat{\mathcal{F}}_r, \boldsymbol{x}\right) + r\hat{\phi}\left(\frac{2\ln\frac{3}{\delta}}{nr}\right), \tag{5.7}$$

with $\hat{r} \doteq 3r + 5r\hat{\phi}\left(\frac{\ln\frac{3}{\delta}}{5nr}\right)$ and consider the *fixed point* $\hat{r}_n^*$ such that $\hat{r}_n^* = \hat{\psi}_{n,m}(\hat{r}_n^*)$. For all $K > 0$, we set $r^{\mathrm{U}}(K)$ to be the fixed point w.r.t. r of the following equation

$$\sqrt{r\hat{r}_n^*} + \left[2\sqrt{r\hat{r}_n^*} + r\right]\breve{\phi}\left(\frac{\frac{1}{n}\ln\frac{3}{\delta}}{2\sqrt{r\hat{r}_n^*} + r}\right) = \frac{r}{K}. \tag{5.8}$$

Note that using $\breve{\phi}(u) \approx 2\sqrt{u}$, we can roughly understand the fixed point $r^{\mathrm{U}}(K)$ in two cases. The first one is when $\hat{r}_n^*$ is large (and $\breve{\phi}$ negligible) leading to $r^{\mathrm{U}}(K) \approx K^2\hat{r}_n^*$. The second case correspond to the situation where $\hat{r}_n^*$ is small ($\hat{r}_n^* \approx 0$) which result in $r^{\mathrm{U}}(K) \approx \frac{4K^2\ln\frac{4}{\delta}}{n}$. As these situations are mutually exclusive, we take $r^{\mathrm{U}}(K) \approx K^2(\hat{r}_n^* + \frac{4}{n}\ln\frac{4}{\delta})$ to be our initial guess in solving this fixed-point equation.

The fixed point $\hat{r}_n$ exists as it can be shown that $\hat{\psi}_{n,m}$ is sub-root (Oneto et al., 2015). The behavior of the $\hat{\psi}_{n,m}$ is displayed in figure 5.2. Note that the *Monte-Carlo* correction term $2\hat{r}\breve{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nm\hat{r}}\right)$ can be brought arbitrarily close to 0 by raising the number of Monte-Carlo trials m, though it is rapidly dominated by the Bousquet term $r\hat{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nr}\right)$.

Lemma 1 (Frequency-Constrained MCLRA)**.** *Suppose upper and lower ERA bounds*

$$\hat{\mathcal{R}}_n^{\downarrow}(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathcal{R}}_n(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathcal{R}}_n^{\uparrow}(\mathcal{F}, \boldsymbol{x}),$$

and consider

$$\hat{\psi}_{n,m}^{\downarrow\uparrow}(r) \doteq 2\hat{\mathcal{R}}_n^{\downarrow\uparrow}\left(\hat{\mathcal{F}}_r, \boldsymbol{x}, \boldsymbol{\sigma}\right) + 2\hat{r}\breve{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nm\hat{r}}\right) + r\hat{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nr}\right),$$

with fixed points

$$\hat{r}_{n,m}^{*\downarrow\uparrow},$$

which upper and lower bound the corresponding quantities for the MCLERA in expectation.

Let frequency threshold *and* truncated pattern family

$$\mu^{\downarrow} \doteq \hat{\mathcal{R}}_n(\hat{F}_{\hat{r}^{*\downarrow}_{n,m}}, \boldsymbol{x}), \quad \mathcal{F}' \doteq \{f \in \mathcal{F} : \hat{P}f \geq \mu^{\downarrow}\}.$$

For any sample and Rademacher variable $\boldsymbol{\sigma}$*, we bound the localized Monte-Carlo ERA as*

$$\hat{\mathcal{R}}_{n,m}(\hat{F}'_{\hat{r}^*_{n,m}}, \boldsymbol{x}, \boldsymbol{\sigma}) \leq \hat{\mathcal{R}}_{n,m}(\hat{\mathcal{F}}_{\hat{r}^*_{n,m}}, \boldsymbol{x}, \boldsymbol{\sigma}) \leq \\ \frac{1}{m}\sum_{i=1}^{m} \max\left(\hat{\mathcal{R}}^{\downarrow}_{n,m}(\hat{\mathcal{F}}_{\hat{r}^*_{n,m}}, \boldsymbol{x}), \hat{\mathcal{R}}_{n,1}(\hat{\mathcal{F}}'_{\hat{r}^*_{n,1}}, \boldsymbol{x}, \boldsymbol{\sigma}_i)\right).$$

Furthermore, in expectation we have

$$\hat{\mathcal{R}}_n(\hat{F}'_{\hat{r}^*_{n,m}}, \boldsymbol{x}) \leq \hat{\mathcal{R}}_n(\hat{\mathcal{F}}_{\hat{r}^*_{n,m}}, \boldsymbol{x}) \leq \\ \mathbb{E}_{\boldsymbol{\sigma}}\left[\max\left(\hat{\mathcal{R}}^{\downarrow}_{n}(\hat{\mathcal{F}}_{\hat{r}^*_{n,m}}, \boldsymbol{x}), \hat{\mathcal{R}}_{n,1}(\hat{\mathcal{F}}'_{\hat{r}^*_{n,1}}, \boldsymbol{x}, \boldsymbol{\sigma})\right)\right].$$

Furthermore, with high probability, fixed points of $\hat{\psi}^{\uparrow\downarrow}_{n,m}$ *approximately (precisely in expectation?) sandwich those of* $\hat{\psi}_{n,m}$ *as*

$$\hat{r}^{*\downarrow}_{n,m} \lessapprox \hat{r}^*_{n,m} \lessapprox \hat{r}^*_{n,m}.$$

We can now state the main theorem of this work that derives from (Oneto et al., 2015), which offers fully empirical bounds for the supremum deviation for mining a dataset of sample of size n.

Theorem 2 (Monte-Carlo Localization Bounds)**.** *Consider the fixed point* $r^{\mathrm{U}}(K)$ *of Definition 6. With probability at least* $1-\delta$ *and for function* $f \in \mathcal{F}$ *we have*

$$Pf \geq \sup_{K>0} \min\left\{\frac{K}{K+1}\hat{P}_n f, \hat{P}_n f - \frac{r^{\mathrm{U}}(K)}{K}\right\}, \\ Pf \leq \inf_{K>1} \max\left\{\frac{K}{K-1}\hat{P}_n f, \hat{P}_n f + \frac{r^{\mathrm{U}}(K)}{K}\right\}.$$

The bounds in Theorem 2 only contain empirical quantities although computing $\hat{\mathcal{R}}(\hat{\mathcal{F}}_r)$ is generally intractable. The next subsection resolves this issue and presents an efficient algorithm to compute the bounds.

Algorithm 1: Localized Pattern Miner

Function `LocalMiner(`$\mathcal{F}$, r_0, $\boldsymbol{x}$, m, δ`):`

Input: Pattern family $\mathcal{F} \subseteq \{0,1\}^d \to \{0,1\}$, mining frequency threshold r_0, transaction database $\boldsymbol{x} \in \{0,1\}^{d\times n}$, Monte-Carlo trial count m, confidence $(1-\delta)$

Output: Localized bound $r^{\mathrm{U}}(\cdot)$ (see Theorem 2 and lemma 2)

$\mathcal{F}^C \leftarrow \textsc{ClosedPatterns}(\mathcal{F}, \boldsymbol{x}, r_0)$

$\boldsymbol{\sigma} \leftarrow \mathcal{U}^{m\times n}(\pm 1)$ ▷ Draw Rademacher sequences

$\mathcal{F}^P(i) \leftarrow \textsc{Pareto}\left\{\left(-\sum f(\boldsymbol{x}), \left|\boldsymbol{\sigma}_i \cdot f(\boldsymbol{x})\right|\right) : f \in \mathcal{F}^C\right\}$

$\textsc{Let}\ \hat{r} \doteq 3r + 5r\hat{\phi}\left(\frac{\ln\frac{5}{\delta}}{5nr}\right) \forall r$

$\hat{\psi}_i^{r_0}(r) \leftarrow \sqrt{r_0 \cdot \min(r_0, \hat{r}\frac{m+|\mathbf{1}\cdot\boldsymbol{\sigma}_i|}{2m})}$

$\hat{\psi}_i(r) \leftarrow 2\max\left(\hat{\psi}_i^{r_0}(r),\ \sup_{f\in\mathcal{F}^P(i)} \min\left(1, \sqrt{\frac{\hat{r}}{\hat{P}_n f}}\right)\left|\frac{\boldsymbol{\sigma}\cdot f(\boldsymbol{x})}{n}\right|\right)$

$\hat{\psi}(r) \leftarrow \frac{1}{m}\sum_{i=1}^{m}\hat{\psi}_i(r) + r\hat{\phi}\left(\frac{2\ln\frac{5}{\delta}}{nr}\right) + 2\hat{r}\breve{\phi}\left(\frac{2\ln\frac{5}{\delta}}{nm\hat{r}}\right)$

$\hat{r}^* \leftarrow r : r = \hat{\psi}(r)$

$r^{\mathrm{U}}(K) \leftarrow r : \sqrt{r\hat{r}^*} + \left[2\sqrt{r\hat{r}^*} + r\right]\breve{\phi}\left(\frac{\ln\frac{5}{\delta}}{n\left(2\sqrt{r\hat{r}^*}+r\right)}\right) = \frac{r}{K}$

return $r^{\mathrm{U}}(\cdot)$ ▷ Function $r^{\mathrm{U}}(\cdot)$ fully specifies the bound

5.3.2 Estimating itemset frequencies with LocalMiner

This section is devoted to the computational aspects of deriving localized bounds for pattern mining. The procedure consists of a double fixed-point resolution. First, we compute the function $\hat{\psi}_{n,m}(\cdot)$ and determine its unique fixed point. Then, it is used to solve for the second fixed point $r^{\mathrm{U}}(K)$. We now describe LocalMiner (Algorithm 1), an efficient algorithm for the computation of $\hat{\psi}_{n,m}(\cdot)$ for pattern mining tasks.

Using the definitions of $\hat{\psi}_{n,m}(r)$, $\hat{\mathcal{F}}$, $\hat{\mathcal{F}}_r$, and, letting $\hat{r} \doteq 3r + 5r\hat{\phi}\left(\frac{\ln\frac{4}{\delta}}{5nr}\right)$ for all r and every function f in $\mathcal{F}$, we define

$$\hat{\psi}_{n,m}(r, f) \doteq \gamma + \hat{\mathcal{R}}_n(\{f\}, \boldsymbol{x}, \boldsymbol{\sigma}) \min\left(1, \sqrt{\frac{\hat{r}}{\hat{P}_n f}}\right). \tag{5.9}$$

We then re-express $\hat{\psi}_{n,m}(r)$ in terms of these quantities as

$$\hat{\psi}_{n,m}(r) = \sup_{f \in \mathcal{F}} \hat{\psi}_{n,m}(r, f) + 2\hat{r}\breve{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nm\hat{r}}\right) + r\hat{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nr}\right) \tag{5.10}$$

$$= \sup_{f \in \mathcal{F}} \min\left(1, \sqrt{\frac{\hat{r}}{\hat{P}_n f}}\right) \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right| \tag{5.11}$$

$$+ 2\hat{r}\breve{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nm\hat{r}}\right) + r\hat{\phi}\left(\frac{2\ln\frac{4}{\delta}}{nr}\right). \tag{5.12}$$

This function has a unique fixed-point as $\hat{\phi}$ and $\breve{\phi}$ amount to small corrections, and the remaining terms are piecewise curves consisting of parabolic and horizontal linear segments. The piecewise behavior of $\hat{\psi}_{n,m}(\cdot)$ that is described by Equation (5.10) is illustrated in Figure 5.2. It is easy to see that any f such that $(\hat{P}_n f, \boldsymbol{\sigma} \cdot f(\boldsymbol{x}))$ is not Pareto-optimal REF can not realize the supremum of $\hat{\psi}_{n,m}(\cdot)$, which further simplifies computing $\hat{\psi}_{n,m}(\cdot)$ and fixed points thereof.

Theorem 3 (Local Miner (Algorithm 1) Guarantees)**.** *Suppose* $\boldsymbol{x}$, $r^{\mathrm{U}}(\cdot) \leftarrow \textsc{LocalMiner}(\mathcal{F}, \boldsymbol{X}, m, \delta)$*, and take* $\hat{P}_n$ *to be the empirical measure on* $\boldsymbol{x}$*. The conclusions of Theorem 2 then holds with* P*,* $\hat{P}_n$*, and* $r^{\mathrm{U}}(\cdot)$*.*

A rich line of work explores the *discriminative pattern mining* task, where the goal is not to estimate a set of frequent itemsets but rather to determine whether two transaction databases are drawn from the same distribution, i.e., hypothesis testing for a statistically significant difference. Similar methods apply: e.g., union bounds to control for multiple testing or permutation testing for the maximum deviation between populations. Like the (global) Rademacher average, the permutation test is inherently insensitive to low-frequency patterns due to them being, in a sense, overpowered by high-frequency patterns. Algorithm 1 can be applied immediately to this task and makes an interesting tradeoff in test power, as it is inherently more sensitive to frequency differences on *low frequency* patterns than unlocalized methods. Due to space constraints, we do not further explore the topic here, but we note that Theorem 3, applied individually to 2 samples, allows us to reject the null hypothesis if the bounds for any itemset are disjoint.

Algorithm 2: Relative Localized Progressive Sampling

Function `Rlps`($\mathcal{F}$, $\boldsymbol{X}$, m, α, ε, δ)**:**

Input: Pattern family $\mathcal{F} \subseteq \{0,1\}^d \to \{0,1\}$, transaction database $\boldsymbol{X} \in \{0,1\}^{d \times m}$, Monte-Carlo trial count m, low-frequency threshold $\alpha \in (0,1]$, multiplicative error $\varepsilon > 0$, confidence $(1-\delta)$

Result: Sample $\boldsymbol{x}$, relative guarantee ε, localized bound $r^{\mathrm{U}}(\cdot)$

$\beta \leftarrow 2; I \leftarrow 1 \vee \lfloor \log_\beta \frac{d}{2\alpha\varepsilon} \rfloor; n_0 \leftarrow \frac{1}{\alpha\varepsilon} \ln \frac{2I}{\delta}$. ▷ Schedule

for $i \in 1, \ldots, I$ **do**

  $n_i \leftarrow \lceil n_0 \beta^i \rceil$

  $\boldsymbol{x} \leftarrow \mathcal{U}^{n_i}(\boldsymbol{X})$ ▷ Subsample n_i transactions

  $r^{\mathrm{U}}(\cdot) \leftarrow \textsc{LocalMiner}(\mathcal{F}, \boldsymbol{x}, n, \frac{\delta}{I})$ ▷ Get LRA bounds

  $K^* \leftarrow K : \frac{K+1}{K} r^{\mathrm{U}}(K) = \alpha$

  if $\frac{1}{K^*} \leq \varepsilon$ **then**

    return $\boldsymbol{x}$, $\frac{1}{K^*}$, $r^{\mathrm{U}}(\cdot)$

  end

end

5.3.3 Relative Frequency Estimation with Progressive Sampling

We define a novel objective for low-frequency pattern mining which we term as the α-ε-δ *relative frequency estimation* task. Frequency estimation is particularly relevant in the context of low-frequency pattern mining. Formally, in an ε-δ relative guarantee, the following bound holds with probability $1-\delta$

$$\tfrac{1}{1+\varepsilon}\hat{P}_n f \leq Pf \leq \tfrac{1}{1-\varepsilon}\hat{P}_n f, \tag{5.13}$$

or equivalently, $Pf \in [\frac{1}{1\mp\varepsilon}]\hat{P}_n f$. This equation captures the ideal of sharper bounds for low-frequency patterns. However, this target is too optimistic, as the sample complexity of ε-relative estimation grows unboundedly as $Pf \to 0$. To circumvent this limitation we introduce the frequency threshold parameter α and define a α-ε-δ relative estimator for a family of patterns $\mathcal{F}$ as any estimator that produces some $\hat{P}_n$. In particular, given a threshold α in $[0,1]$, we require

$$\mathbb{P}\left(\bigcap_{f \in \mathcal{F}: Pf \geq \alpha} Pf \in [\tfrac{1}{1\mp\varepsilon}]\hat{P}_n f \right) \geq 1-\delta. \tag{5.14}$$

Note that we consider the $\frac{1}{1\mp\varepsilon}$-relative error for convenience, but any other standard relative error concept can be alternatively used with small changes to the

algebra. As illustration, Figure 5.3 compares the different choice to be considered in Lemma 2 for bounds involving various relative error concepts.

Using Algorithm 1, we construct Algorithm 2 for the α-ε-δ relative frequency estimation task. The goal of any progressive sampling algorithm is to design a sampling schedule, ranging from optimistic minimal sample size to pessimistic maximal sample size, such that the desired probabilistic bounds hold at all steps of the schedule with probability at least $1 - \delta$. We first state a technical lemma relating LRA to relative error guarantees and then proceed to derive an appropriate sampling schedule.

Lemma 2 (Monte-Carlo Localization Bounds: Relative Error)**.** *Under the same assumptions than Theorem 2, let $\alpha \in \mathbb{R}_+$ and $f \in \mathcal{F}$ such that $Pf \geq \alpha$. Then*

1. $\frac{K}{K+1}\alpha \geq r^{\mathrm{U}}(K) \implies Pf \geq \frac{K}{K+1}\hat{P}_n f$,
2. $\frac{K}{K-1}\alpha \geq r^{\mathrm{U}}(K) \implies Pf \leq \frac{K}{K-1}\hat{P}_n f$.

Furthermore, suppose K such that $\alpha \geq \frac{K+1}{K} r^{\mathrm{U}}(K)$. Then

3. $\varepsilon \doteq \frac{1}{K^2} \implies Pf \in \frac{K^2}{K^2-1}\hat{P}_n f[1 \pm \varepsilon]$,
4. $\varepsilon \doteq \frac{1}{K} \implies Pf \in \hat{P}_n f \left[\frac{1}{1\pm\varepsilon}\right]$.

Due to the complexity of localized bounds, and the difficulty of bounding such terms in the absence of any *a priori* knowledge, we use the following generic lower bound on Bernoulli mean estimation

$$n^{\downarrow}(\alpha\varepsilon, \tfrac{\delta}{I}) \doteq \ln(\frac{I}{\delta})\frac{1}{\alpha\varepsilon} \tag{5.15}$$

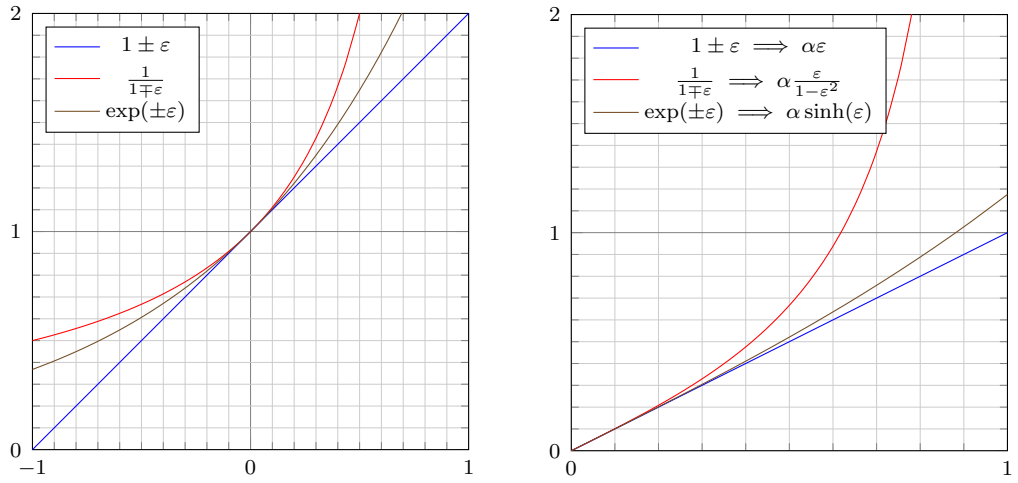
$$\leq \ln(\frac{2I}{\delta})\frac{1}{\alpha\varepsilon}. \tag{5.16}$$

This bound is tight within constant factors for datasets containing only near-empty transactions. For the upper bound, we use the standard Hoeffding-union bound (Boucheron, Lugosi, and Massart, 2013)

$$n^{\uparrow}(\alpha\varepsilon, \tfrac{\delta}{I}) = \ln(\frac{2^{d+1}I}{\delta})\frac{1}{2\alpha^2\varepsilon^2} \tag{5.17}$$

$$\leq \ln(\frac{2I}{\delta})\frac{d}{2\alpha^2\varepsilon^2}. \tag{5.18}$$

Note that the technique could be refined to instead use VC-theoretic bounds(Riondato and Upfal, 2015) with a possible improvement of $\mathcal{O}(\ln\ln d)$ to $\mathcal{O}(\ln\ln D)$ terms at the cost of having to evaluate the empirical VC dimension for application.



(a) Relative Error Concepts (b) Additive Bound Radii

Figure 5.3: Comparison of Various Relative Error Concepts.

The final step is to divine a sampling procedure to find the smallest n to consider at a given precision requirement. We use a doubling ($\beta = 2$ geometric) sampling schedule to interpolate between $n^{\downarrow}$ and $n^{\uparrow}$. As multiple rounds induce multiple comparisons error, we take a union bound over the I iterations which adds a $\ln \ln \frac{d}{\alpha\varepsilon}$ sample complexity factor.

Theorem 4 (Progressive Sampling (Algorithm 2) Guarantees)**.** *Suppose* $\boldsymbol{x}, \varepsilon,$ $r^{\mathrm{U}}(\cdot) \leftarrow \textsc{Lrpm}(\mathcal{F}, \boldsymbol{X}, m, \alpha, \varepsilon, \delta)$ *and take* $\hat{P}_n$ *to be the empirical measure on* $\boldsymbol{x}$*. For all* f *in* $\mathcal{F}$ *such that* $Pf \geq \alpha$ *the following holds with probability at least* $1 - \delta$*,*

$$\hat{P}_n f(1 - \varepsilon) \leq Pf \leq \hat{P}_n f(1 + \varepsilon). \tag{5.19}$$

In other words, $\hat{P}_n \mathcal{F}$ *is an* α*-*ε*-*δ relative frequency estimate *of* $P\mathcal{F}$*.*

Proof. The result follows via the union bound over the guarantee of Algorithm 1 (Theorem 3) applied to each of its (up to) I applications, and the relative frequency estimation guarantees of Lemma 2. □

Note that this algorithm can immediately be used to compute α-ε-δ approximations of many interestingness measures (c.f. Table 5.3) as long as all of the relevant probabilities exceed the threshold α and support values have been estimated with sufficient accuracy. The Table 5.3 offers a brief overview of several popular interesting measures. We highlight that relative estimation guarantees are easily composed via products, ratios, and square roots, unlike additive estimates, which can become large when divided.

Table 5.2: Characteristics of the experimental datasets.

Name	n	d
Accidents	340183	468
Chess	3196	75
Connect	67557	129
pumsb	49046	2108
Retail	88162	14089
Mushroom	8416	119

5.4 Experimental Evaluation

Setup. The implementation is in python 3.7 and runs on [redacted for anonymity]. the full source code of each experiment is available at online[2] for reproducibility. We used the FP-GROWTH algorithm for the exact frequent itemset mining, leveraging the extensive SPMF library (Fournier-Viger et al., 2017). We carry out experiments on synthetic, standard and real-world datasets to demonstrate the performance of the proposed method for progressive sampling. As a baseline, we use the Hoeffding-Union bound (CU) and MC-RAPPER, as the latter is considered to be state of the art.

Unless otherwise noted, all experiments use the *itemsets* pattern family $\mathcal{F} \doteq \{f_t : t \in \mathcal{X}\}$.

5.4.1 Comparative Analysis of statistical guarantees for support

In this first experiment (Figure 5.4), we compare Algorithm 1 to the Hoeffding-union bound, and the Monte-Carlo empirical Rademacher average bound of Pellegrina et al. (2020) on several real-world datasets whose main characteristics are displayed in Table 5.2. The competing approaches all yield uniform confidence intervals, but our method yields sharper balance for low-frequency itemsets, so we visualize all algorithm outputs as *upper and lower bounds* (y-axis) on true pattern frequencies at a given *empirical frequency* (x-axis).

For comparison, we also contrast our algorithm with an *empirical Bennett union bound* approach. Bennett's inequality consists of a sub-Poisson inequality for random variables of bounded range and variance. It performs similarly to Hoeffding's inequality when variance is maximal but yields much sharper confidence

[2] https://anonymous.4open.science/r/lra-pattern-mining-C5CF/

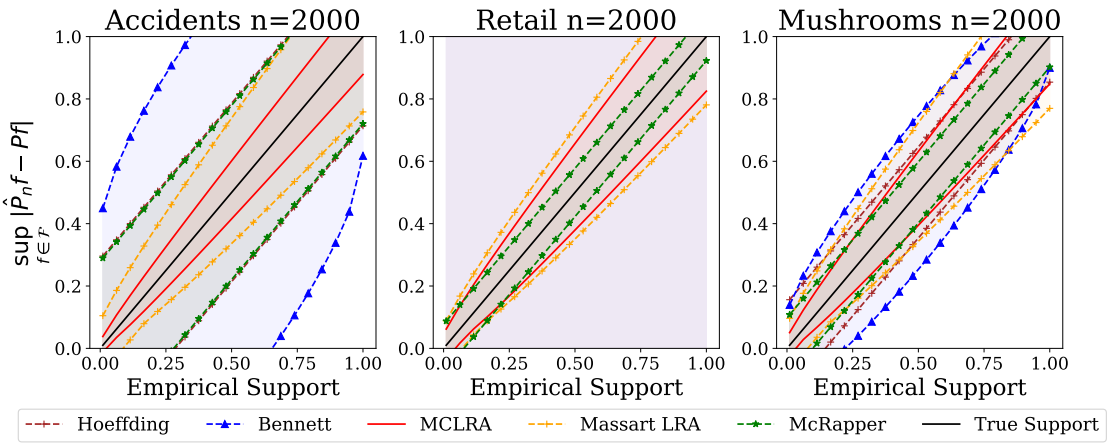


Figure 5.4: Experimental comparison of upper and lower bounds (y-axis) given empirical frequencies (x-axis), of our method (Algorithm 1) to existing work on real-world datasets.

intervals when variance is small.

Additionally, our work makes two major leaps; the first is to use localized Rademacher averages for the low-frequency pattern mining problems, and the second is in using Monte-Carlo estimation to sharply bound LRAs. To assess their impact in isolation, we contrast with a localized bound that uses a simple loose upper bound on the appropriate empirical Rademacher averages instead of our sharply estimated MC-LRA.

Setup In these experiments we take a conference parameter $1-\delta = 1-10^{-3}$. We plot estimated frequencies on the x-axis, and true frequencies on the y-axis, of all closed itemsets. Due to the one-in-a-thousand error guarantee, we unsurprisingly observe no bound violations in any experiment.

Results It is clear that the Hoeffding and Bennett bounds are always the worst methods. Among the uniform bounds, we see that is uniformly the worst method, and then we observe that MCRapper is at least as good, if not much better, in all datasets.

Our method makes an interesting tradeoff, with clear superiority for low frequency and medium frequency itemsets but weaker bounds after a dataset-dependent frequency threshold is met. This is inherently the price to pay for increased sensitivity to low-frequency itemsets, but we know that this deficit may be repaired by applying both methods (with a union bound correction).

In the (high dimensional) retail dataset, the Bennett and Hoeffding bounds are vacuous, and in the accident dataset, the Bennett bound is not vacuous, but it is

dominated by the other bounds. However, while Bennett has not performed particularly well in the mushrooms dataset, there are ranges of low-variance function for which it beats the uniform bounds.

The localized methods take this theme of improvement for low-functions and develop it further. By avoiding the costly union bound, they can show strong performance for low-frequency functions. In particular, our Monte-Carlo localized bounds uniformly dominate (are always superior to) the analytic (Massart) localized bounds for all datasets and all frequency ranges. Furthermore, for sufficiently small empirical frequencies, both localized bounds always beat all uniform bounds, including the state-of-the-art McRapper. Clearly, the frequency threshold below which localized bones become superior is higher (better) for our Monte-Carlo localized bounds than for the analytic bounds.

The contrast between the analytic and Monte-Carlo bounds mirrors the improvement from AMIRA to MCRapper in the uniform (non-localized) case.

5.4.2 Progressive Sampling

In our final experiment (Figure 5.5), we visualize the execution of Algorithm 2. In particular, we show how our localized bound evolves as a function of sample size, by plotting our bounds at multiple sample sizes, and contrasting with the target ε-δ relative guarantees. In particular, for each sample size, an ε-δ relative guarantee is reached for all frequencies exceeding some α', and the algorithm terminates once $\alpha' \leq \alpha$.

The termination condition is equivalent to checking when the upper and lower bounds are fully contained by the region between the α and $\frac{1}{1\mp\varepsilon}$ lines.

We observe that the bounds (in particular the lower bound) for low-frequency patterns improve faster than those for high-frequency patterns. Based on their variance, the sample complexity of estimating low-frequency patterns should indeed be lower than the one when estimating high-frequency patterns, and this behavior confirms that the proposed LRA-based method exhibits this behavior.

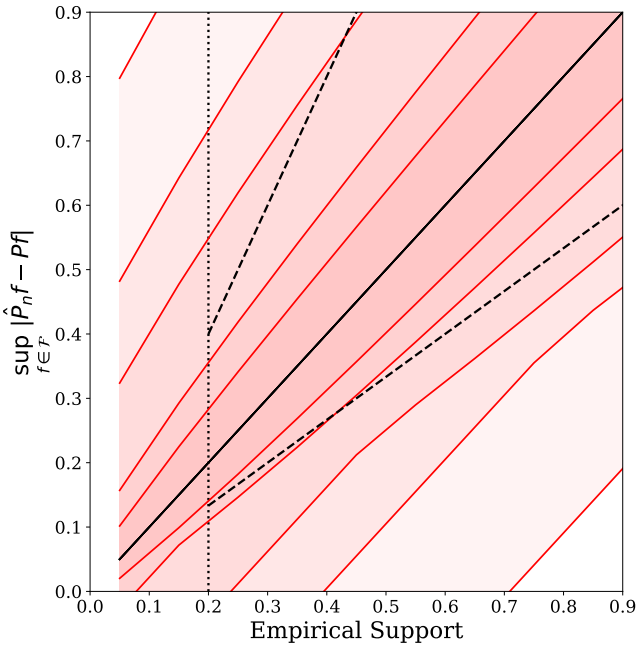


Figure 5.5: Progressively generated upper and lower bounds on true pattern frequencies with progressively doubling sampling sizes from Algorithm 2 on the Accidents dataset, with $\alpha = 0.1$, $\varepsilon = 0.25$, and $\delta = 0.1$. The low-frequency threshold α is visualized as a dotted vertical line and the $\frac{1}{1 \mp \varepsilon}$ relative guarantee as dashed diagonal lines around the true frequency line $y = x$. There is no ambiguity, as the bound becomes sharper with each doubling iteration.

5.5 Conclusion

We re-examine the pattern mining setting with a focus on the often-overlooked low-frequency patterns. We identify two key applications for low-frequency pattern mining; namely that of *contrast pattern mining* and bounding *interesting measures*. We introduce the α-ε-δ relative frequency estimation problem to formalize low-frequency pattern estimation and develop *sampling methods* to efficiently perform these tasks with small sample size.

We show that Local Rademacher Average (LRA) are an effective tool for low-frequency pattern estimation, whereas previous work, which relied on global ERAs, was only sufficient for ε-δ additive estimation. This work is of independent interest beyond the pattern of mining setting, as it is, to our knowledge, the first to sharply bound LRA This mirrors the improvement from analytic (Riondato and Upfal, 2015) to Monte-Carlo (Pellegrina et al., 2020) (global) ERAs in the pattern mining setting.

Support	$P(XY)$
Confidence	$P(Y \mid X)$
Lift/Interest	$\frac{P(Y\|X)}{P(Y)}$ or $\frac{P(XY)}{P(X)P(Y)}$
Jaccard	$\frac{P(XY)}{P(X)+P(Y)-P(XY)}$
Certainty Factor	$\frac{P(Y\|X)-P(Y)}{1-P(Y)}$
Odds Ratio	$\frac{P(XY)P(\bar{X}\bar{Y})}{P(\bar{X}Y)P(X\bar{Y})}$
Yule's Q	$\frac{P(XY)P(\bar{X}\bar{Y})-P(X\bar{Y})P(\bar{X}Y)}{P(XY)P(\bar{X}\bar{Y})+P(X\bar{Y})P(\bar{X}Y)}$
Yule's Y	$\frac{\sqrt{P(XY)P(\bar{X}\bar{Y})}-\sqrt{P(X\bar{Y})P(\bar{X}Y)}}{\sqrt{P(XY)P(\bar{X}\bar{Y})}+\sqrt{P(X\bar{Y})P(\bar{X}Y)}}$

Table 5.3: Several common interestingness measures.

We provide two algorithms: Algorithm 1 (local miner) shows how to compute LRAs efficiently for finite pattern families, and Algorithm 2 uses progressive sampling for the specific problem of α-ε-δ relative frequency estimation. These have immediate applications in contrast to pattern mining and in interestingness measure mining, and we think that future work could extend our statistical ideas to new application-specific quantities.

Our experiments confirm that localized Rademacher averages are an effective tool for low-frequency pattern mining. It can be used to obtain variable-width confidence intervals that are sharper for such low-frequency patterns than global methods, which are bottlenecked by the necessarily larger confidence intervals for high-frequency patterns. Furthermore, we see that our methods are competitive with the state-of-the-art global methods and vastly improve the more simplistic Bennett-union method for low-frequency pattern mining.

A straightforward extension of this work consists of applying our methods to related pattern concepts, wherein new computational routines may be required to bound LRAs (in particular the $\hat{\psi}_{n,m}(\cdot)$ function), i.e., with *infinite pattern families* or *utility pattern families*, and to tasks with other objectives. Algorithm 1 provides a generic recipe to obtain sharp bounds, and Algorithm 2, with minor changes (in particular to the *termination condition*), is a powerful tool to dynamically adapt sample consumption to the needs of a particular task.

While our methods soundly beat the state-of-the-art, we are under no delusions about their efficiency in an absolute sense. This work marks the first application of localized Rademacher averages in the pattern-mining setting, and we expect subsequent research into LRAs to yield further improvements to our bounds. Furthermore, the progressive sampling schedule of Algorithm 2 is likely suboptimal, i.e., considers a schedule of sample sizes that is *too optimistic* on the low-end and *too pessimistic* on the large end, but this is manifest only as suboptimal sample

complexity in $\log\log(\cdot)$ terms. Future work could optimize further the computations performed by our algorithms. In particular, while we use standard techniques in pattern mining to consider only the set of *closed patterns*, and perform most computation on *Pareto-optimal subsets*, we note that McRapper (Pellegrina et al., 2020) uses a branch-and-bound search to compute Monte-Carlo (global) ERA without necessarily enumerating all closed patterns. Computing the fixed point r^* is much more involved, as we effectively need to compute ERAs for many r-star-localized families, but similar search methods would likely be applicable to our framework to avoid enumerating closed patterns.

Part IV

Optimal Quantization for stochastic optimization

Chapter 6

Voronoi Tesselation for stochastic optimization

This section is devoted to the use of quantization in the context of stochastic optimization. The term himself takes his origin in the domain of signal processing. It relates to finding a satisfactory (in term of information loss) finite approximation of a continuous signal and has been the main tool for signal compression (Gersho and Gray, 1992). The intuition behind considering that the redundancy of information in a random object can be exploited through *codebooks* can be traced back to (Shannon, 1951). In the context of probability, the term quantization historically referred (Graf and Luschgy, 2000) to the best approximation in the measure space metricized by the Wasserstein distance of a $\mathbb{R}^d$ valued random variable X by a random variable $\hat{X}$ with finite support. This section presents the theoretical foundations and methods for the construction of such quantizer.

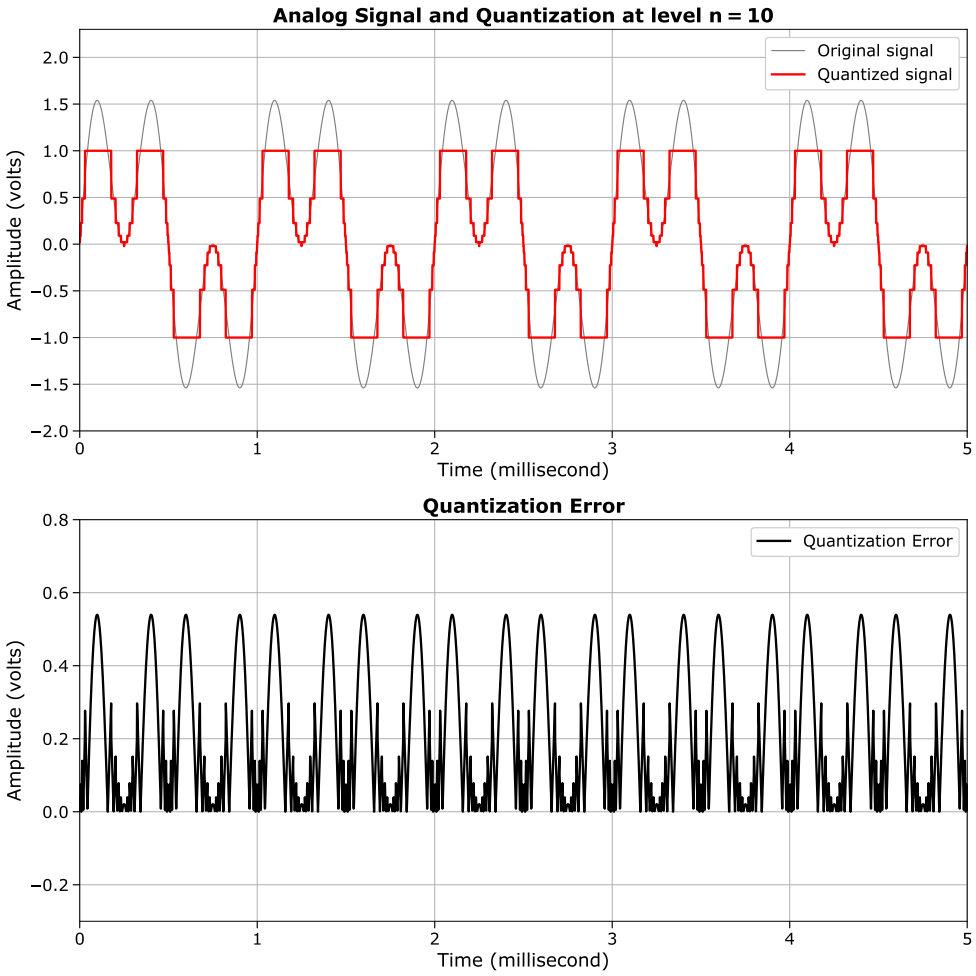


Figure 6.1: **Quantization of an analog signal**. A signal amplitude in volts, his non uniform quantization (top) and the absolute value of the quantization error produced (bottom).

6.1 The Voronoi partition

We first present the geometrical aspect of Optimal Quantizer as it relates to fundamental properties. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(E, \|.\|)$ be a vector space equipped with the distance d and the induced norm $\|.\|$. The space of p-integreable measure in $\mathcal{M}(E)$ is denoted $P_p(E)$ and $L_p(\Omega, \mathcal{A}, \mathbb{P})$ the quotient space (for the equivalence relation defined by the $\mathbb{P}$-a.s. equality) of probability distribution such as

$$L_p(\Omega, \mathcal{A}, \mathbb{P}) = \left\{ f \text{ measurable} \mid \int_E \|f(x)\|^p \mu(dx) < +\infty \right\}.$$

When considering a random variable $X\colon (\Omega, \mathcal{A}) \longrightarrow (E, \mathcal{B}(E))$, the measure is given by $\mu = \mathbb{P} \circ X^{-1}$. We say that a measure π in $\mathcal{M}\left(\mathcal{B}(E)^{\otimes 2}\right)$ has marginals μ and ν if for any borel set $A \subset E$, $\pi(A \times E) = \mu(A)$ and $\pi(E \times A) = \nu(A)$.

For x, y in E the *line segment* between x and y is defined as the set $[x, y]$ such as

$$[x, y] = \left\{(1-t)x + ty \mid 0 \leq t \leq 1\right\}.$$

For any scalar $\lambda \in \mathbb{R}$ and set $A \subseteq E$ we denote $\lambda A = \left\{\lambda a \mid a \in A\right\}$. In the same fashion, given any real form $L\colon E \longrightarrow \mathbb{R}$, $LA = \left\{Ta \mid a \in A\right\}$ is the image of A by the application L.

At the heart of OQ is the concept of Voronoi subset.

Definition 7. Let $\Gamma \subset E$ be a finite bounded set of size n.

Voronoi diagram. For every x_i in Γ, the *Voronoi cell* associated with x_i is defined as

$$V(x_i, \Gamma) = \left\{z \in E \mid \|z - x_i\| = \min_{x_j \in \Gamma} \|z - x_j\|\right\},$$

and the *Voronoi diagram* of E associated with Γ

$$\mathcal{V}(\Gamma) = \left\{V(x_i, \Gamma)\colon x_i \in \Gamma\right\}.$$

Voronoi partition. A *Voronoi partition* $\mathcal{V}(\Gamma) = \{W(x_i, \Gamma)\}_{i=1}^{n}$ of E associated with Γ is a borel set of E such as for all $x_i \in \Gamma$

$$W(x_i, \Gamma) \subseteq V(x_i, \Gamma).$$

The borel sets $V(x_i, \Gamma)$ consists of the point that are the closest of x_i for the distance induced by $\|.\|$. Note that the set of Voronoi partition associated with a set of point Γ is clearly non empty as it contains the *Voronoi diagram* of Γ. Moreover, it is not unique depending on the choice for the boundary points in $\partial W(x_i) = W(x_i) \cap \overline{(E \backslash W(x_i))}$.

The geometry of the Voronoi cells are determined by the set of hyperplanes separating a couple of *Voronoi cell* associated with x_i and x_j defined by

$$H(x_i, x_j) = \left\{z \in E \mid \|z - x_i\| \leq \min_{x_j \in \Gamma} \|z - x_j\|\right\}.$$

The Voronoi cells are then the intersection of all hyperplanes surrounding x_i in the sense that

$$V(x_i, \Gamma) = \bigcap_{x_j \in \Gamma} H(x_i, x_j).$$

We can deduce from that characterization that the geometrical properties of the cells will heavily depend on the regularity of the underlying norm. Figure 6.2 shows the Voronoi diagram generated by a set of points for different norm in the real case $E = \mathbb{R}^2$.

More restrictive conditions on the norm are needed to obtain cells that are sufficiently regular to have interesting property when measured. For instance, under the hypothesis that the underlying norm is strictly convex it can be shown (Lugosi and Wegkamp, 2004; Pagès, 2015) that the interior of the boundary $\partial W(x_i, \Gamma)$ is empty. The following proposition establishes general results similar to Luschgy and Pagès (2008) in the general case of vector space.

Proposition 1. *Let $\Gamma \subset E$ be a finite non empty bounded set of size n and for each $x_i \in \Gamma$ the associated Voronoi cells $V(x_i, \Gamma) \subset E$.*

1. *The Voronoi diagram $\mathcal{V}(\Gamma)$ is borel cover of E in the sense that*

$$\bigcup_{x_i \in \Gamma} V(x_i, \Gamma) = E.$$

2. *The Voronoi cells are star shaped relatively to their center* i.e. *for any element x in $V(x_i, \Gamma)$, the segment joinging x_i and x is in $V(x_i, \Gamma)$.*

3. *In the real case $E = \mathbb{R}^d$ and for the euclidean norm, the Voronoi cells are convex polytopes.*

We refer to Graf and Luschgy (2000)[Proposition 1.2] for the star-shaped property (stated in the real case but essentially the same for a normed space E) and to section 6.5.1 for the euclidean norm. In the former case, the separating plane between two points is the kernel of the continuous linear map $\psi(z) = \left\langle x_i - x_j, z - \frac{1}{2}(x_i + x_j) \right\rangle$ which the middle hyperplane between x_i and x_j.

The figure 6.2 shows that the Voronoi cells associated with each points are convex subsets of $E = \mathbb{R}^2$ for some $\ell_p\left(\mathbb{R}^2\right)$ norms.

A crucial property of Voronoi cells are their *invariance property* or the fact that they are preserved under a certain type of transformation.

Proposition 2. *Let $\Gamma \subset E$ be a non empty finite bounded set of size n and $V(x_i, \Gamma)$ the associated Voronoi cells for each $x_i \in \Gamma$. Let T be a* scaling function *with scaling factor c that is any application such as $\|\mathrm{T}\, x - \mathrm{T}\, y\| = c\|x - y\|$. Then, for all x_i in Γ we have*

$$V(\mathrm{T}\, x_i, \mathrm{T}\, \Gamma) = \mathrm{T}\, V(x_i, \Gamma). \tag{6.1}$$

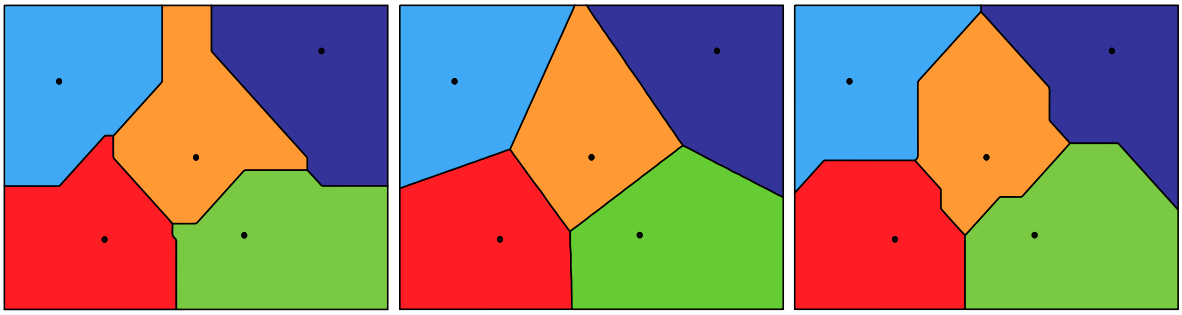


Figure 6.2: Voronoi diagram for a finite subset $\Gamma \subset R^2$ with size $n = 5$ for the $\ell_1\left(\mathbb{R}^2\right)$ norm (Manhattan distance, left), $\ell_2\left(\mathbb{R}^2\right)$ norm (Euclidean distance, center) and $\ell_\infty\left(\mathbb{R}^2\right)$ norm (Chebyshev distance, right). Each point x of $\mathbb{R}^2$ is colored by it's associated Voronoi cell. Notably, the Voronoi cells are star-shaped for all considered distances (see Proposition 1), are convex polytopes in the euclidian case and the separating sets are hyperplanes of $\mathbb{R}^2$.

Given a Voronoi diagram $\mathcal{V}(\Gamma)$, proposition 2 states that all the Voronoi diagram of any scaled transformation of Γ is known. This property is of particular interest when the Voronoi diagram is known up to a scaling function and is a key element of the *quantized variational inference* procedure described in Chapter 7.

Proposition 3 (Geometric regularity (Graf and Luschgy, 2000))**.** *Let $E = \mathbb{R}^d$ equiped with a norm $\|.\|$, a subset Γ of $\mathbb{R}^d$ and consider for each x_i in Γ the Voronoi cells $V(\Gamma, x_i)$ induced Γ as in definition 7. If the norm on $\mathbb{R}^d$ is strictly convex or $\mathbb{R}^d$ is equipped with the $\ell_p\left(\mathbb{R}^d\right)$ norm with $1 \leq p \leq \infty$, then for all x_i in Γ*

$$\lambda_d(V(\Gamma, x_i)) = 0, \tag{6.2}$$

where λ_d is the Lesbeague measure on $\mathcal{B}(\mathbb{R}^d)$.

Note that the conjecture made in (Graf and Luschgy, 2000, Conjecture 1.12) that states that proposition 3 holds for any norm and any dimension has proven to be false for any dimension d greater than three (Gao, 2005).

6.2 Optimal quantization for random variables

An *quantifier* is a Borel function E which is valued in a set with cardinal less or equal to n such as $\hat{X} = f(X)$. It is equivalent to say that a quantizer of X is a $\sigma(X)$-measureable random variable with finite support less or equal to n. Since Ω is an element of $\sigma(X)$, any constant is a quantizer of X. Of course, we want to find the quantizer that is a satisfactory approximation of X in a sense that will

be made clear. Such good finite approximation is called a *optimal quantizer*. This section is devoted to formalize this definition and explore the properties of this object.

6.2.1 Projection on quantization grid

Constructing the optimal quantizer requires to find a set of point $\Gamma \subset E$, also called *quantization grid* such as the minimum $L_p(\Omega, \mathcal{A}, \mathbb{P})$ distance is attained. The following shows how to explicitly construct such finite measure by the closest projection of $X \in L_p(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ onto a finite closed set of E. Let $\Gamma \subset E$ and $\mathcal{V}(\Gamma)$ an associated Voronoi diagram. Consider the closest projection onto the Voronoi cells defined by

$$\begin{array}{rccc} \Pi_\Gamma \colon & E & \longrightarrow & E \\ & z & \longmapsto & \sum_{i=1}^N x_i \mathbb{1}_{z \in V(\Gamma, x_i)}. \end{array} \tag{6.3}$$

The function Π_Γ is $(E, \mathcal{B}(E))$-$(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ measurable so that the following random projection of X onto the Voronoi cells is well defined

$$\begin{aligned} \widehat{X}_\Gamma &= \Pi_\Gamma X \\ &= \sum_{i=1}^N x_i \mathbb{1}_{X \in V(\Gamma, x_i)}. \end{aligned} \tag{6.4}$$

Note that $\widehat{X}_\Gamma$ is not unique in general as there is many Voronoi diagram associated with Γ. Since the Voronoi diagram depends only on the way that the points on the boundary $\partial V(\Gamma, x_i)$ are associated, two quantization are μ-a.s. equal whenever the $\mu(\partial V(\Gamma, x_i)) = 0$ for all x_i in Γ (see proposition 3 in the for the real case).

The criteria to evaluate the quality of a quantizer of X is the *distortion* of μ at level n.

Definition 8 (Optimal Quantizer)**.** Let $X \colon ((\Omega, \mathcal{F})) \longrightarrow (E, \mathcal{B}(E))$ be a random variable in $L_p(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ and consider a finite subset $\Gamma \subset E$ of size n. The $L_p(\Omega, \mathcal{A}, \mathbb{P})$ *distortion function* $\mathcal{D}_{p,\mu}$ of μ at level n is defined by

$$\begin{array}{rccc} \mathcal{D}_{p,\mu} \colon & E^n & \longrightarrow & \mathbb{R}_+ \\ & \Gamma & \longmapsto & \mathbb{E}\left[\min_{x_i \in \Gamma} \|X - x_i\|^p\right] \end{array}, \tag{6.5}$$

and the *quantization error* function by

$$e_{p,\mu} = \mathcal{D}_{p,\mu}^{\frac{1}{p}}. \tag{6.6}$$

The minimizer of $e_{n,\mu}(x)$ is called a $L_p(\Omega, \mathcal{A}, \mathbb{P})$ *optimal quantizer of* μ *at level* n.

The existence of an optimal quantizer is established in (Graf, Luschgy, and Pagès, 2007)[Proposition 1] for any Banach space E and E-valued radon measure μ. The optimal quantizer is not unique in general. For an optimal quantizer $\widehat{X}_{\Gamma^*}$ of X at level n, each of the 2^n permutation of elements of Γ^* will produce an optimal quantizer. If radial invariance of the probability density function is assumed will result in an infinite number of the optimal quantizer. A result of (Kieffer, 1982) proves uniqueness for log-concave probability density function in the one dimensional real case (Pagès, 2018)[Theorem 5.3].

Note that for the projected quantizer 6.4, the $L_p(\Omega, \mathcal{A}, \mathbb{P})$ norm of the pointwise error is

$$\|X - \widehat{X}_\Gamma\|_{L_p(\Omega,\mathcal{A},\mathbb{P})} = e_{p,\mu}(\Gamma). \tag{6.7}$$

Example 1 (Optimal Quantizer at level 1)**.** In the real case $E = \mathbb{R}^d$, the quadratic ($p = 2$) optimal quantization grid at level $n = 1$ is the element x of $\mathbb{R}^d$ that minimizes

$$\left\|X - x\right\|_{L_2(\Omega,\mathcal{A},\mathbb{P})},$$

which is a strictly convex function admitting an unique solution $x^* = \mathbb{E}[X]$, an associated optimal quantization grid $\Gamma^* = \{x^x\}$ and an optimal quantizer $\hat{X}_{\Gamma^*} = \Pi_{\Gamma^*} X$.

6.2.2 Optimal tranport approach

This section aims to establish and make precise the close connection between Optimal Quantizer and the *optimal transport* domain. This link has been recently explored in (Liu and Pagès, 2020b) notably showing that there exist an integer n such as that for quantization level $N \geq n$ $L_p(\Omega, \mathcal{A}, \mathbb{P})$ quantization based distance produces the same topology as the Wasserstein distance.

Definition 9 (Wasserstein distance)**.** Let $(E, \|.\|)$ be a vector space equiped with the norm $\|\|$, $\mu, \nu \in \mathcal{M}(E)$ two probability measure on E and $1 \leq p \leq \infty]$. We define the L_p *Wasserstein* distance on $\mathcal{M}(E)$ such as

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint_{E\times E} d^p(x, y) d\pi(x, y) \right)^{1/p}, \tag{6.8}$$

Where $\Pi(\mu, \nu)$ is the set of joint probability measure on $(E^2, \mathcal{B}(E)^{\otimes 2})$ with marginals μ, ν.

The probability measure space $\mathcal{M}(E \times E)$ endowed with this distance is a metric (Villani, 2008) and that $\mathcal{W}_p(\mu, \nu)$ is finite whenever $\mu, \nu \in P_p(E)$. Solving

on the space of probability the measure Π relates to the *optimal transport problem* when one search for an *optimal transport plan* to transfert the all the mass from μ to ν with $c(x, y) = \|x - y\|^p$ the cost associated with transporting an elementary amount of mass from x to y. With this view, for any $A, B \in \mathcal{B}(E)$, the measure $\pi(A \times B)$ represents the amount of mass transported from the region A to B. We refer to Panaretos and Zemel (2019) for more theoretical results on the Wasserstein distance.

The previous interpretation gives a natural way to define the *optimal quantizer* of a random variable $X \in L_p(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ as the closest projection on the set of all measure with support at most n with respect to the Wasserstein distance.

Definition 10. Let $(E, \|.\|)$ be a vector space equiped with the norm $\|.\|$, $\mu \in \mathcal{M}(E)$ a probability measure with p-th finite moment and $n \in \mathbb{N}$ the quantization level. Denoting $\mathcal{M}(n)$ the space of probability measure with support at most n, the optimal quantizer $\hat{\nu}_n$ of μ is defined by

$$\hat{\nu}_n = \underset{\nu \in \mathcal{M}(K)}{\operatorname{argmin}} \mathcal{W}_p(\mu, \nu). \tag{6.9}$$

Taking for instance the empirical measure $\mu = \sum_i^n \delta_i$, we can say that the approximation in term of the Wasserstein can be as best as the level of distortion of the optimal quantizer. The optimal quantizer can be rewritten as a finite weighted sum of *dirac* measure

$$\hat{\nu}_n = \sum_{i=1}^{n} w_i \delta_{x_i}. \tag{6.10}$$

This problem is known as the semi-discrete optimal transport problem (Peyré and Cuturi, 2020) and linked with numerous applications in machine learning. For instance, taking equal weights $w_i = \frac{1}{n}$ it corresponds to the *k-means* problem. The equivalence between the quantizer in definition 10 and the one construction in Equation 6.4 can be established by the following proposition.

Proposition 4. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the propoability space, $X \in L_p(\Omega, \mathcal{A}, \mathbb{P})$ and consider $\hat{X}$ a quantization at level n. The following equality holds*

$$\mathcal{W}_p(\hat{X}, X) = \|X - \hat{X}\|_{L_p(\Omega, \mathcal{A}, \mathbb{P})}. \tag{6.11}$$

Proof of this result can be found in (Graf and Luschgy, 2000)[proposition 4.3].

Remark 1 (Equivalence with the random projection)**.** Consider the following quantization of X

$$\hat{\nu}_n = \sum_{i=1}^{n} w_i \delta_{x_i}. \tag{6.12}$$

It follows from the definition 9 that

$$\mathcal{W}_p(\mu, \nu) = \sum_j w_j \inf_{\gamma(x|y_j)} \int_{\mathcal{X}} c\left(x, y_j\right) \gamma\left(\mathrm{d}x \mid y_j\right). \tag{6.13}$$

It has been shown that for known weights, the optimal transport plan consists of pushing all the mass in μ onto a set of *Laguerre Cells* (Peyré and Cuturi, 2020) which are Voronoi cells with an additive constant factor,

$$L_j(\boldsymbol{g}) = \left\{ x \mid \forall k, \quad c\left(x, y_j\right) - g_j < c\left(x, y_k\right) - g_k \right\}. \tag{6.14}$$

Using the dual formulation it is possible to optimize jointly on the weights and the transporation plan. In this case, the laguerre cells are Voronoi cells ($g = 0$) and the optimal weights are given by

$$\hat{w}_j = \int_{V_j} \nu(dx). \tag{6.15}$$

6.2.3 Properties of the optimal quantizer

On of the main application of an optimal quantizer is to be used for numerical integration. To that end, the optimal quantizer exhibits remarkeable properties that makes it suitable for this task.

Proposition 5 (Stationnarity, Invariance)**.** *Let $X\colon (\Omega, \mathcal{F}) \longrightarrow (E, \mathcal{B}(E))$ a random variable in $L_p\left(\Omega, \mathcal{A}, \mathbb{P}\right)$ and $\widehat{X}^{\Gamma^*}$ the associated optimal quantizer at level n as in definition 8.*

Invariance. *For any scaling function $\mathrm{T}\colon E \longrightarrow E$ with scaling coefficient c as defined in Proposition 2, consider transformation $Y = \mathrm{T}\, X$. The optimal quantization of Y at level n is given by*

$$\hat{Y}_{T(\Gamma^*)} = \Pi_{T(\Gamma^*)}\, Y. \tag{6.16}$$

Stationnarity. *The optimal quantizer has the* stationnary property *in the sense that*

$$\mathbb{E}\big[X \mid \widehat{X}_{\Gamma^*}\big] = \widehat{X}_{\Gamma^*}. \tag{6.17}$$

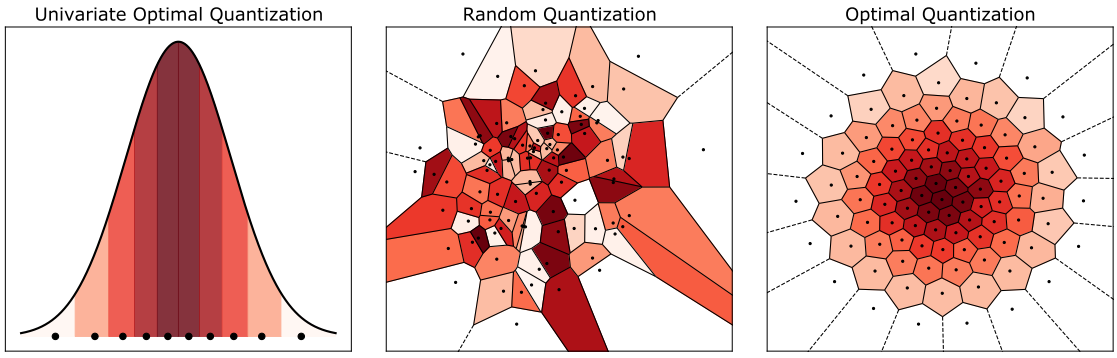


Figure 6.3: Optimal quantization of the standard univariate normal distribution $X \sim \mathcal{N}(0,1)$ at level $n = 10$ (left), random quantization (center) and optimal quantization (right) of the bivariate standard normal distribution $X \sim \mathcal{N}(0, I_2)$. For each point x_i of the quantization grid Γ, the associated weight $\omega_i = \mathbb{P}\big(\widehat{X}_\Gamma = x_i\big)$ gives his relative importance (values are displayed in shades of red).

Cubature Formula. *Given that the elements of the Voronoi diagram $\mathcal{V}(\Gamma^*)$ are real convex cells in $\mathbb{R}^d$ and for any continuous fonction $F\colon \mathbb{R}^d \longrightarrow \mathbb{R}$*

$$\mathbb{E}\big[F\big(\hat{X}^N\big)\big] = \sum_{i=1}^{n} \omega_i F\big(x_i^N\big). \tag{6.18}$$

The invariance property stems from proposition 2 (see (Graf and Luschgy, 2000) for complete proof). Given an optimal quantizer of X, it allows one to compute, at only the cost of evaluating a function, any optimal quantizer of the form $F(X)$ that can be obtained by shifting or scaling operations. This property is key to the *quantized gradient descent* algorithm of section 7.2.2. Note that the stationary property is a necessary but not sufficient condition for a quantizer to be optimal. Other quantizers can have such property has the grid quantization (Pagès, 2018)

The *cubature formula* makes optimal quantization suitable for numerical integration as one can easily substitute $\mathbb{E}\left[F(X)\right]$ with $\mathbb{E}\big[F(\widehat{X}_{\Gamma^*})\big]$.

Proposition 4 shows that the optimal quantizer will weakly converge towards the true measure with respect to the quantization level, and Equation 6.7 gives the pointwise error. In the case of numerical integration, a natural question to ask in order to compare this method with other types of sampling is the rate of convergence at which it occurs. Most of the results are base on Zador's theorem (Zador, 1982).

Theorem 5 ((Pagès, 2018)). *Let $X \in L^{p+\delta}_{\mathbb{R}}(\mathbb{P})$ for some $\delta > 0$.*

Sharp Rate. *Let $\mathbb{P}_z(d\xi) = \varphi(\xi) \cdot \lambda(d\xi) + \nu(d\xi)$, where $\nu \perp \lambda$ i.e. denotes the singular part of $\mathbb{P}_z$ with respect to the Lebesgue measure λ on $\mathbb{R}$. Then,*

$$\lim_{n\to+\infty} n \min_{\Gamma\subset\mathbb{R},|\Gamma|\leqslant n} \left\|X - \widehat{X}_\Gamma\right\|_p = \frac{1}{2^p(p+1)} \left[\int_{\mathbb{R}} \varphi^{\frac{1}{1+p}} d\lambda\right]^{1+\frac{1}{p}}. \tag{6.19}$$

Non Asymptotic Upper-bound. *Let $\delta > 0$. There exists a real constant $C_{1,p,\delta} \in (0, +\infty)$ such that, for every $\mathbb{R}$ -valued random variable X and all $n \geq 1$,*

$$\min_{\Gamma\subset\mathbb{R},|\Gamma|\leqslant n} \left\|X - \widehat{X}_\Gamma\right\|_p \leqslant C_{1,p,\delta}\sigma_{\delta+p}(X)n^{-1}, \tag{6.20}$$

where, for $r \in (0, +\infty)$, $\sigma_r(X) = \min_{a\in\mathbb{R}} \|X - a\|_r < +\infty$.

Example 2 (Univariate standard normal distribution)**.** Consider the real case $E = \mathbb{R}$ equiped with the euclidian norm and $X \sim \mathcal{N}(0, 1)$. The probability measure associated is absolutely continuous with respect to the Lebesgue measure with the log density probability density function

$$\ln \phi(x) = -\ln(\sqrt{2\pi}) - \frac{x^2}{2}.$$

The probability density function ϕ is log-concave, and thus there exist a p optimal quantization of X at any level n (Pagès, 2018)[Theorem 5.3]. The optimal quantization grid of the standard normal distribution is represented in Figure 6.3 along with the induced Voronoi cells and weights.

Example 3 (from (Graf and Luschgy, 2007))**.** Let $X = \mathcal{U}([0, 1]^d)$ and consider a tesselation of $[0, 1]^d$ consisting of $n = k^d$ translates $C_1, \ldots, C_n$ of the cube $\left[0, \frac{1}{k}\right]^d$. Denote by a_i the midpoint of C_i. Then $\Gamma^* = \{a_1, \ldots, a_n\} = \left\{\frac{2i-1}{2k} : i = 1, \ldots, k\right\}^d$.

6.3 Numerical Integration

A proeminent problem in analysis is the computation of the expectation of a random variable

$$I = \mathbb{E}\left[F(X)\right], \tag{6.21}$$

for $F(X)$ in $L_1(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable function $F\colon \mathbb{R}^d \longrightarrow \mathbb{R}^d$. Even for common probability distribution, there is typically no analytical form for this quantity. Most of the techniques resort to numerical computation that provide and good approximation and are asymptotically $\mathbb{P}$-a.s. equal to the true expectation.

As we have seen in proposition 5 it is possible to compute such expectation using the optimal quantizer $\widehat{X}_{\Gamma^*}$ for a random variable with p-th finite moment thanks to the *cubature formula*

$$\begin{aligned} I_n^{OQ} &= \mathbb{E}\big[F\big(\hat{X}^N\big)\big] \\ &= \sum_{i=1}^{N} \omega_i F\big(x_i^n\big). \end{aligned} \tag{6.22}$$

A natural idea is to replace the true expectation with its quantized counterpart. This approximation has several advantages. Let $(X_1, \ldots, X_n)$ be an *i.i.d.* sequence of random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and μ_X distributed. consider the following *Monte-Carlo* estimator

$$I_n^{MC} = \frac{1}{N} \sum_{i=1}^{N} F(X_i). \tag{6.23}$$

By the strong law of large number, I_n^{MC} converges towards I almost surely, and at a rate of $\mathcal{O}(n^{-\frac{1}{2}})$ if $F(X) \in L_2(\Omega, \mathcal{A}, \mathbb{P})$ with a quadratic error

$$\|I_n^{MC} - \mathbb{E}\big[F(X)\big]\|_{L_2(\Omega,\mathcal{A},\mathbb{P})} = \frac{\mathbb{V}F(X)}{\sqrt{n}}, \tag{6.24}$$

and the *MeanSquaredError* in $\mathcal{O}(n^{-1})$. Proposition 4 establishes that the estimator I_n^{OQ} is closest to the true distribution with respect to the Wasserstein distance. An additional and important point is that the OQ estimator does not depend on any event ω in Ω which makes the OQ a variance free but biased estimator. Thanks to proposition 4 it is possible to consider that its the optimal choice among all such estimators.

The error bound is established in (Pagès, 2018)[section 5.2] and considers a measurable Lipschitz function $F\colon (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $F(X) \in L_p(\Omega, \mathcal{A}, \mathbb{P})$. Then

$$\|F(\widehat{X}_{\Gamma^*}) - F(X)\|_{L_1(\Omega,\mathcal{A},\mathbb{P})} \leq F_{\text{Lip}} \|X - \widehat{X}_{\Gamma^*}\|_{L_p(\Omega,\mathcal{A},\mathbb{P})}. \tag{6.25}$$

If F is differentiable, using the stationary property (see proposition 5) gives

$$|I_n^{OQ} - \mathbb{E}\left[F(X)\right]| \leq \frac{1}{2}[\nabla F]_{\text{Lip}} \|X - \widehat{X}_{\Gamma^*}\|_{L_2(\Omega,\mathcal{A},\mathbb{P})}. \tag{6.26}$$

We refer to (Pagès, 2018)[proposition 5.2] for detailed proof. The error is thus given by the distortion function of definition 8. The rate of convergence is obtained by using Zador's theorem (see Theorem 5) but can be improved by considering more regularity on F. let $\alpha \in [0,1], \eta \geq 0$, if F is continuously differentiable on $\mathbb{R}^d$ with α-Hölder gradient and $X \in L^{2+\eta}_{\mathbb{R}^d}(\mathbb{P})$, one has the following bound on the Absolute Error (Pagès, 2015).

$$\left|\mathbb{E}F(X) - \mathbb{E}F\left(\widehat{X}_{\Gamma^*}\right)\right| \leq C_{d,\mu}[\nabla F]_\alpha N^{-\frac{1+\alpha}{d}}. \tag{6.27}$$

The convergence rate obtained is the best finite approximation of X in the sense of the Wasserstein distance (see Proposition 4).

6.4 Construction of the optimal quantizer

The main drawback of this method is the construction of an optimal quantizer. To the best of our knowledge, no method using *semi-discrete optimal transport* has been yet proposed. In the real case with $E = \mathbb{R}^d$ metrizes with euclidian norm for the quadratic optimal quantizer ($p = 2$), the Loyd's algorithm and its variants is used (Lloyd, 1982).

The most used methods use the gradient of the distortion function. Let $E = \mathbb{R}^d$ and consider a random variable $X \in L_2(\Omega, \mathcal{A}, \mathbb{P})$ with known density function μ and an initial random quantizer $\widehat{X}_\Gamma$ with $\Gamma \subset \mathbb{R}^d$. In this case we know from Proposition 3 that the boundary of the cells are $\mathbb{P}$-megligeable and that the distortion function is differentiable (Pagès, 2018)[Theorem 5.1] so that

$$\nabla \mathcal{Q}_{2,\,\mathrm{N}}(x) = 2\left[\int_{C_i(\Gamma_N)} (x_i - \xi)\,\mathbb{P}_X(d\xi)\right]_{i=1,\ldots,N}. \tag{6.28}$$

Stochastic Gradient Descent A simple zero-search algorithm can then be used to find a good approximation of Γ^*. For each step k of the gradient descent procedure, let $\Gamma^{[k]}$ be the n quantization grid and consider the following update scheme

$$\Gamma^{[k+1]} = \Gamma^{[k]} - \gamma_n \nabla_{\Gamma^{[k]}} e_{2,\mu}\left(\Gamma^{[k]}\right), \tag{6.29}$$

with an initial $\Gamma^{[0]}$ thant can be chosen randomly in the support of μ. When the expectation in not analytical or distribution unknown but i.i.d.samples $(X_1, \ldots, X_n)$

Algorithm 3: Quantized Variational Inference.

Input: μ, $\Gamma^{[0]}$.
Result: Approximation of the optimal quantizer Γ^*, optimal weights ω_i, quantization error $e_{2,\mu}(\Gamma^*)$.
while *not converged* **do**
  for $x_i \in \Gamma$ **do**
    $\omega_i = \mu\left(V(x_i, \Gamma)\right)$
    $e_{2,\mu}(x_i) = \mathbb{E}\left[(X - x_i)^2 \mathbb{1}_{X \in V(x_i,\Gamma)}\right]$
  end
  $\Gamma^{[n+1]} = \Lambda\left(\Gamma^{[n]}\right)$
end

Figure 6.4: Randomized Loyd I procedure.

can be obtained, it is direct to obtain a stochastic version of this algorithm known as the Competitive Learning Vector Quantization algorithm (see (Pagès, 2015) for details and convergence proof).

Randomized Loyd I The original Loyd algorithm (Lloyd, 1982) describes a fixed-point search strategy that leverages the stationary property (see proposition 5) on the normalized expectation of each cell. Consider the following quantity

$$\Lambda_i(\Gamma) = \frac{\mathbb{E}\left[X \mathbb{1}_{X \in V(x_i,\Gamma)}\right]}{\mathbb{P}\left(X \in V(x_i, \Gamma)\right)}. \tag{6.30}$$

At each step, the fixed point method writes

$$\Gamma^{[k+1]} = \Lambda\left(\Gamma^{[k]}\right), \tag{6.31}$$

until convergence. Note that the expectation can be directly replace with a Monte Carlo estimator. It is shown in (Pagès, 2018) that this fix point search decreases the quadratic function at each step. The solution to this fix point problem corresponds to the minimum of the distortion function in the sense that if $\Gamma = \Lambda(\Gamma)$ then $\nabla e_{2,\mu}(\Gamma) = 0$. Algorithm 3 describes the complete procedure. Note that the weights ω_i and distortion values are computed as the algorithm progresses with

$$\begin{aligned} \omega_i &= \mu\left(V(x_i, \Gamma)\right) \\ e_{2,\mu}(x_i) &= \mathbb{E}\left[(X - x_i)^2 \mathbb{1}_{X \in V(x_i,\Gamma)}\right]. \end{aligned} \tag{6.32}$$

6.5 Proofs

6.5.1 Proof of Proposition 1

Lemma 3. *Let $(E, \|.\|)$ be a normed space and A a non empty subset. Let $x \in E$ and $a \in A$ such as $d(x, A) = \|x - a\|$. Then, for all $z \in [x, a]$*

$$d(z, A) = \|z - a\|. \tag{6.33}$$

Proof. Let $z \in [x, a]$. Let $t \in [0, 1]$ such as $z = (1 - t)x + ta$,. For all $b \in A$,,

$$\begin{aligned} \|z - a\| &= (1 - t)\|x - a\| \\ &\leq (1 - t)\|x - b\| \\ &= \|(1 - t)x - (1 - t)b\| \\ &= \|(1 - t)x + ta - (1 - t)b - ta\| \\ &= \|z - (1 - t)b - ta\| \\ &= \|(1 - t)z + tz - (1 - t)b - ta\| \\ &\leq (1 - t)\|z - b\| + t\|z - a\|, \end{aligned} \tag{6.34}$$

so that reorganizing the last equation we get $\|z-a\| \leq \|z-b\|$, hence $d(z, A) = \|z - a\|$. □

Proof of proposition 1-(2). It is a direct consequence of applying Lemma 3 to the complementary of each Voronoi cells with $A = E \setminus (V(x_i, \Gamma)$ □

Lemma 4. *Let $\left(E, \langle \cdot, \cdot \rangle\right)$ be a pre-Hilbert space. For all $x, y \in E$*

$$\langle x, y \rangle = \frac{1}{4}\left(\|x + y\|^2 - \|x - y\|^2\right).$$

Proof of proposition 1-(3). Applying Lemma 4 to $y = x-a$ for the separator space we get

$$S(a, b) = \left\{x \in \mathbb{R}^d \colon \left\langle a - b, x - \frac{1}{2}(a + b)\right\rangle = 0\right\}. \tag{6.35}$$

□

Chapter 7

Quantized Variational Inference

This chapter corresponds to the paper (Dib, 2020) published in *Advances in Neural Information Processing Systems 33 Proceedings (NeurIPS 2020).*

Abstract: We present Quantized Variational Inference, a new algorithm for Evidence Lower Bound maximization. We show how Optimal Voronoi Tesselation produces variance free gradients for Evidence Lower Bound (ELBO) optimization at the cost of introducing asymptotically decaying bias. Subsequently, we propose a Richardson extrapolation type method to improve the asymptotic bound. We show that using the Quantized Variational Inference framework leads to fast convergence for both score function and the reparametrized gradient estimator at a comparable computational cost. Finally, we propose several experiments to assess the performance of our method and its limitations.

Key Words: variational inference, Bayesian learning, stochastic optimization, optimal quantization.

7.1 Introduction

Given data y and latent variables z, we consider a model $p(y, z)$ representing our view of the studied phenomenon through the choice of $p(y|z)$ and $p(z)$. The goal of the Bayesian statistician is to find the best latent variable that fits the data, hence the likelihood $p(z|y)$. These quantities are linked by the bayes formula which gives that $p(z|y) = \frac{p(z)p(y|z)}{p(y)}$ where $p(y)$ is the prior predictive distribution (also named marginal distribution or normalizing factor) which is a constant. Given a variational distribution q_λ, the following decomposition can be obtained (Saul, Jaakkola, and Jordan, 1996)

$$\log p(y) = \underbrace{\mathop{\mathbb{E}}_{z\sim q_\lambda}\left[\log \frac{p(z,y)}{q_\lambda(z)}\right]}_{\text{ELBO}(\lambda)} + \underbrace{\text{KL}\left(q_\lambda(z)\|p(z|y)\right)}_{\text{KL-divergence}}. \tag{7.1}$$

It follows that maximizing the ELBO with respect to q_λ leads to find the best approximation of $p(z|y)$ for the Kullback–Leibler (KL) divergence. Intuitively, this procedure minimizes the information loss subsequent to the replacement of the likelihood by q_λ but other distances can be used (Ambrogioni et al., 2018).

The reason for the popularity of such techniques is due to the fact that finding a closed-form for $p(z|y)$ requires to evaluate the prior predictive distribution and thus to integrate over all latent variables which lead to intractable computation (except in the prior conjugate case) even for simple models (Gelman et al., 2013). A common approach is to use methods such as Gibbs Sampling, Monte Carlo Markov Chain or Hamilton Monte Carlo (Betancourt, 2018; Homan and Gelman, 2014; Brooks et al., 2011) which rely solely on the unnormalized posterior distribution (freeing us from the need to compute $p(y)$) and the ability to sample from the posterior. These methods are consistent but associated with heavy computation, high sensitivity to hyperparameters and potential slow to converge to the true target distribution. On the other hand, optimization techniques such as Variational Inference (VI) are generally cheaper to compute, tend to converge faster but are often a crude estimate of the true posterior distribution. Recent work proposes to combine these two strategies to allow for an explicit choice between accuracy and computational time (Salimans, Kingma, and Welling, 2015).

Thanks to approaches such as Black Box Variational Inference (BBVI) (Ranganath, Gerrish, and Blei, 2014; Kingma and Welling, 2014) (which opens the possibility of the generic use of VI), Automatic Variational Inference (AVI) (Kucukelbir et al., 2015) and modern computational libraries, Variational Inference has become one of the most prominent framework for probabilistic inference approximation.

Most of these optimization procedures rely on gradient descent optimization over

the parameters associated with the variational family and subsequently depending heavily on the $\ell_2(\mathbb{R}^K)$ (with K the number of variational parameters) norm of the expected gradient (Bottou, Curtis, and Nocedal, 2018; Domke, 2019). The bias-variance decomposition gives

$$\mathbb{E}|g|_{\ell_2}^2 = \operatorname{tr} \mathbb{V} g + |\mathbb{E} g|_{\ell_2}^2. \tag{7.2}$$

Low variance of the gradient estimators allows for taking larger steps in the parameter space and result in faster convergence if the induced bias can be satisfyingly controlled. Several methods have been used to reduce gradient variance such as filtering (Miller et al., 2017; Roeder, Wu, and Duvenaud, 2017) control variate (Geffner and Domke, 2018) or alternative sampling (Tran, Nott, and Kohn, 2017; Ruiz, Titsias, and Blei, 2016; Buchholz, Wenzel, and Mandt, 2018).

In real-world applications, one would test a large combination of models and hyperparameters associated with multiple preprocessing procedures. A common practice for bayesian modeling on large datasets consists of using VI for model selection before resorting to asymptotically exact sampling methods. More generally, VI is typically the first step towards more complex and demanding sampling. In this work we propose to give more importance to parsimonious computation than accuracy. Our approach is based upon embracing the fundamental bias in resorting to VI approach and finding the best variance free estimator which produces the fastest gradient descent. This work proposes to use Optimal Quantizer (OQ) (also called Optimal Voronoi Tesselation, see (Graf and Luschgy, 2007) for an historical view) in place for the variational distribution. Given a finite subset Γ_N of $\mathbb{R}^d$, the optimal quantizer at level N of a random variable $Z \in L^p_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is defined as the closest finite probability measure on Γ_N for the $L^p_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ distance. Hence, it is by construction the best finite approximation of size N in the $L_p(\Omega, \mathcal{A}, \mathbb{P})$ sense. Recent works have shown that, given a regularity term α, the Absolute Error error induced by such quantization is in $\mathcal{O}(\mathrm{N}^{-\frac{1+\alpha}{d}})$ (Lemaire, Montes, and Pagès, 2019; Pagès, 2018).

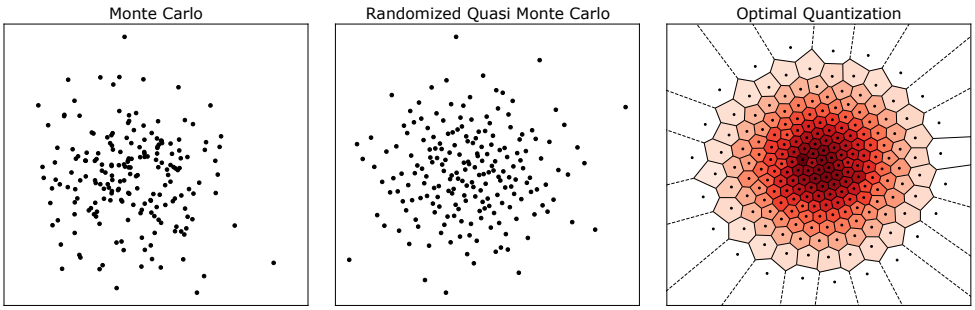


Figure 7.1: Monte Carlo (left), Randomized Monte Carlo (center) and Optimal Quantization with the associated Voronoi Cells (right), for a sampling size $N = 200$ of the bivariate normal distribution $\mathcal{N}(0, I_2)$.

Contribution. We show that: **i)** thanks to invariance under translation and scaling our method can be applied to a large class of variational family at similar computational cost; **ii)** even though biased our estimation is lower than the true lower bound under some assumptions with know theoretical bounds, making it relevant for quick evaluation of model; **iii)** our approach leads to competitive bias-variance trade.

Organisation of the paper. Section 7.2 introduces the idea of using Optimal Quantization for VI and shows how it can be considered as the optimal choice among variance free gradients. Section 7.3 is devoted to the practical evaluation of these methods and show their benefits and limitations. Due to space restrictions, all theoretical proofs and derivations are in the supplementary materials.

Algorithm 4: Monte Carlo Variational Inference.

Input: y, $p(x, z)$, q_{λ_0}.
Result: Optimal VI parameters λ^*.
while *not converged* **do**
  Sample $(X_1^{\lambda_k}, \dots, X_N^{\lambda_k}) \sim q_{\lambda_k}$
  Compute $\widehat{g}_{MC}^N(\lambda_k) = \frac{1}{N}\sum_{i=1}^N \nabla_\lambda H(X^{\lambda_k})$
  $\lambda_{k+1} = \lambda_k - \alpha_k \widehat{g}_{MC}^N(\lambda_k)$
end

7.2 Quantized Variational inference

In this Section, we present Quantized Variational Inference. We review traditional Monte Carlo Variational Inference in 7.2.1. Details of our algorithm along with theoretical results are presented in 7.2.2. Finally, section 7.2.3 proposes an implementation of Richardson extrapolation to reduce the produced bias.

7.2.1 Variational inference

Given a parameter family $\lambda \in \mathbb{R}^K$, exact estimation of equation 7.1 is possible in the conjugate distribution case given some models when closed-forms are avalaible (Blei, Kucukelbir, and McAuliffe, 2017; Winn and Bishop, 2005). Complex or black box models require the use of minimum-search strategy such as Stochastic Gradient Descent (SGD), provided that a suitable form for the gradient can be found. Expressing z as a transformation over a random variable $X \sim q$, which holds all the stochasticity of z, such as $z = h_\lambda(X)$ allows for derivation under the expectation. In this case, the gradient can be expressed as

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{X\sim q}[\nabla_\lambda(\underbrace{\log p\left(y, h_\lambda(X)\right) - \log q\left(h_\lambda(X)\right)|\lambda)}_{H(X,\lambda)}], \tag{7.3}$$

clearing the way for optimization step since one only needs to compute the gradient for a batch of samples and take the empirical expectation. This is known as the reparametrization trick (Kingma, Salimans, and Welling, 2015) . In the following, $H(X^\lambda)$ denotes the stochastic function $H(X, \lambda)$ with $\mathcal{L}(\lambda) = \mathbb{E}[H(X^\lambda)]$ when there is no ambiguity and $g_\lambda(X) = \nabla H(X^\lambda)$ the stochastic gradient for the ELBO maximization problem.
A typical MC procedure at step k samples from an *i.i.d.* sequence $(X_1^{\lambda_k}, \dots, X_N^{\lambda_k}) \sim q_{\lambda_k}$ and computes $\widehat{\mathcal{L}}_{MC}^N(\lambda_k) = \frac{1}{N}\sum_i^N H(X_i^{\lambda_k})$ along with $\widehat{g}_{MC}^N(\lambda_k) = \frac{1}{N}\sum_i^N \nabla H(X_i^{\lambda_k})$. Then, SGD scheme described in algorithm 4 can be used.
The convergence of the procedure typically depends on the expectation of the

quadratic norm of $\mathbb{E}\left[\widehat{g}^N\right]$ (Johnson and Zhang, 2013; Domke, 2019). Equation 7.2 shows that this method results in an MSE error of $\mathcal{O}(N^{-1})$ (by the Law Of Large Number) as the estimator is unbiased. Various methods have already been proposed to improve on this rate (Miller et al., 2017; Roeder, Wu, and Duvenaud, 2017; Geffner and Domke, 2018; Tran, Nott, and Kohn, 2017; Ruiz, Titsias, and Blei, 2016) .
Our work considers the class of variance-free estimator and aims to find the best candidate to improve on this bound, at the cost of introducing a systematic bias in the evaluation which can be reduced using Richardson extrapolation (see section 7.2.2).

7.2.2 Optimal Quantization

In this section we consider the true ELBO $\mathcal{L}(\lambda) = \mathbb{E}\left[H(X^\lambda)\right]$ and construct an optimal quantizer $X^{\Gamma_N,\lambda}$ of X^λ along with an ELBO estimator $\widehat{\mathcal{L}}^N_{OQ}(\lambda) = \mathbb{E}\left[H(X^{\Gamma_N,\lambda})\right]$, such as $\|X^\lambda - X^{\Gamma_N,\lambda}\|_{L_2(\Omega,\mathcal{A},\mathbb{P})}$ is minimized.

Definition 11. Let $\Gamma_N = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ be a subset of size N, $\big(C_i(\Gamma)\big)_{i=1,\dots,N} \subset \mathcal{P}(\mathbb{R}^d)$ and

$$\forall i \in \{1, \dots, N\} \quad C_i(\Gamma) \subset \left\{\xi \in \mathbb{R}^d, |\xi - x_i| \leq \min_{j \neq i} \left|\xi - x_j\right|\right\}, \tag{7.4}$$

then $\big(C_i(\Gamma)\big)_{i=1,\dots,N}$ is a Voronoi partition of $\mathbb{R}^d$ associated with the Voronoi Cells C_i.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the probability space. For $X^\lambda \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$, Optimal Quantization aims to find the best $\Gamma \subset \mathbb{R}^d$ of cordiality at most N in $L^2_{\mathbb{R}^d}$. To that end, the optimal quantizer of X^λ is defined as the projection onto the closest Voronoi cell induced by Γ_N. Formally, if we consider the projection $\Pi : \mathbb{R}^d \to \mathbb{R}^d$ such as $\Pi(x) = \sum_{i=1}^N x_i \mathbb{1}(x)_{C_i(\Gamma)}$, then

$$X^{\Gamma_N,\lambda} = \Pi(X^\lambda). \tag{7.5}$$

The quantizer $\Gamma_N^* = (x_1, \dots, x_N)$ of X^λ at level N is quadratically optimal if it minimizes the quadratic error $\|X^\lambda - X^{\Gamma_N,\lambda}\|_{L_2(\Omega,\mathcal{A},\mathbb{P})} = \mathbb{E}\left[\min_{1\leq i\leq N} \left|X^\lambda - x_i\right|^2\right]$.
The problem can be reformulated as finding the probability measure on the convex subset of probability measure on Γ_N that minimizes the $L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ Wassertein distance (Liu and Pagès, 2020a) .
For illustration, different sampling methods for the bivariate normal distribution $\mathcal{N}(0, I_2)$ are represented in Figure 7.1. It is shown that Randomized Quasi Monte

Carlo produces more concentrated samples in the high density regions where Optimal Quantizer accurately represents the probability distribution. Given a sample from OQ, the associated weights $\mathbb{P}\left(X^{\Gamma_N,\lambda} = x_i\right)$ gives his relative importance (values are displayed in shades of red).

Given N and Γ_N^*, the error rate of such approximation is controlled by Zador's Theorem (Pagès, 2018; Pagès and Printems, 2003; Pagès, 2015)

$$\left\| X^{\lambda} - X^{\Gamma_N^*,\lambda} \right\|_2 \leq \mathcal{O}(N^{-\frac{1}{d}}). \tag{7.6}$$

The key property of the optimal quantizer lays in the simplicity of his cubature formula. For every measurable function f such as $f(X) \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$

$$\mathbb{E}\left[f(X^{\Gamma_N,\lambda})\right] = \sum_{i=1}^{N} \mathbb{P}\left(X^{\Gamma_N,\lambda} = x_i\right) f\left(x_i\right). \tag{7.7}$$

This result opens the possibility for using Optimal Quantizer expectation $\mathbb{E}\left[f(X^{\Gamma_N,\lambda})\right]$ as an approximation for the true expectation. As a deterministic characterization of X^{λ}, equation 7.7 can be compared to its counterpart when one considers Quasi Monte Carlo (QMC) sampling with X^{λ}_{QMC} obtained from evaluating a low discrepancy sequence $\{\mathbf{u}_1, \cdots, \mathbf{u}_N\}$ with the inverse cumulative function of distribution X^{λ}. It results in a similar curbature formula but with equal normalized weights. This method typically produces an absolute error in $\mathcal{O}(\frac{\log(N)}{N})$. By considering relevant weights on each sample, the optimal quantization improves the estimation by a factor $\log(N)$.

Regularity. The precision of the approximation improves with regularity hypothesis. For instance, let $\alpha \in [0, 1], \eta \geq 0$, if F is continuously differentiable on $\mathbb{R}^d$ with α-Hölder gradient and $X \in L^{2+\eta}_{\mathbb{R}^d}(\mathbb{P})$, one has the following bound on the Absolute Error (Pagès, 2015)

$$\left| \mathbb{E}F(X) - \mathbb{E}F\left(\widehat{X}_N^{\Gamma}\right) \right| \leq C_{d,\mu}[\nabla F]_{\alpha} N^{-\frac{1+\alpha}{d}}. \tag{7.8}$$

Getting Optimal Quantization. The main drawback of Optimal Quantization is the computational cost associated with constructing an optimal N-quantizer $X^{\Gamma_N,\lambda}$ compared to sampling from X^{λ}. Even though it is time-consuming in higher dimensions, one must keep in mind that it can be built offline and that efficient methods exist to approximate the optimal quantizer. For instance, K-means are used to obtain such grid at a reasonable cost of $\mathcal{O}(N \log N)$ (Gersho and Gray, 1991). Moreover, in the context of AVI with normal approximation, it is possible to rely solely on D dimensional normal grid to perform optimization since every

Algorithm 5: Quantized Variational Inference.

Input: y, $p(x, z)$, q_{λ_0}.
Result: Optimal Quantized VI parameters λ_q^*.
while *not converged* **do**
  Get $(X_1^{\Gamma_N,\lambda_k}, \dots, X_N^{\Gamma_N,\lambda_k}) \sim q_{\lambda_k}$, $(w_1^k, \dots, w_N^k)$
  Compute $\widehat{g}_{OQ}^N(\lambda_k) = \nabla_\lambda \sum_{i=1}^N w_i^k H(X_i^{\Gamma_N,\lambda_k})$
  $\lambda_{k+1} = \lambda_k - \alpha_k \widehat{g}_{OQ}^N(\lambda_k)$
end

normal distribution can be obtained by shifting and scaling. The same goes for every distribution that can be determined by such transformation of a base random variable. Note that the optimal grid for the normal distribution can be downloaded for dimensions up to 10 (http://www.quantize.maths-fi.com/downloads).

Quantized Variational Inference. The curbature formula 7.7 is used to compute the OQ expectation at a similar cost than regular MC estimation. Replacing the MC term in equation 7.3 by its quantized counterpart is straightforward. The quantized ELBO estimator is defined by

$$\widehat{\mathcal{L}}_{OQ}^N(\lambda) = \sum_{i=1}^N \omega_i H\left(X_i^{\Gamma,\lambda}\right). \tag{7.9}$$

A crucial point is that the quantized ELBO is always lower than the expected one under the assumption of convex ELBO objective. This particular point justifies the usefullness of the method for quick evalutation of model performance.

Proposition 6. *Let $X^\lambda \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ and $X^{\Gamma_N,\lambda}$ the associated optimal quantizer, under the hypothesis that H (Eq. 7.3) is a convex lipschitz function,*

$$\widehat{\mathcal{L}}_{OQ}^N(\lambda) \leq \mathcal{L}(\lambda). \tag{7.10}$$

In fact, for proposition 8 to be true $X^{\Gamma_N,\lambda}$ needs only to fulfill the stationnary property which is defined by $\mathbb{E}\left[X^\lambda | X^{\Gamma_N,\lambda}\right] = X^{\Gamma_N,\lambda}$. Intuitively, the stationnary condition expresses the fact that the quantizer $X^{\Gamma_N,\lambda}$ is the expected value under the subset of events $\mathcal{C} \in \mathcal{A}$ such as $\Pi(X^\lambda) = X^{\Gamma_N,\lambda}$. It can be shown that the optimal quantizer has this property (Graf and Luschgy, 2000; Pagès, 1998) .
Computing the gradient in the same fashion leads to algorithm 5. An immediate consequence of proposition 8 is that for λ_q^* the optimal parameters estimated from algorithm 5 and λ^* the true optimum, we can state the following proposition

Proposition 7. *Let* $\lambda^* = \max_{\lambda \in \mathbb{R}^K} \mathcal{L}(\lambda)$ *and* $\lambda_q^* = \max_{\lambda \in \mathbb{R}^K} \widehat{\mathcal{L}}_{OQ}^N(\lambda)$. *Under the same assumptions than proposition 8,*

$$\mathcal{L}(\lambda^*) - \widehat{\mathcal{L}}_{OQ}^N(\lambda_q^*) \leq C \left[2\|X^{\lambda^*} - X^{\Gamma,\lambda^*}\|_2 + \|X^{\lambda_q^*} - X^{\Gamma,\lambda_q^*}\|_2 \right]. \tag{7.11}$$

The approximation error of the resulting estimation follows from the Zador theorem (Eq. 7.8) and is in $\mathcal{O}(N^{-\frac{2(1+\alpha)}{d}})$ in term of MSE depending on the regularity of H. The crucial implication of proposition 7 is that relative model performance can be evaluated with our method. Poor relative true performance, provided that the difference in terms of ELBO minimum sufficiently large in regard of the approximation error, produces poor relative performance with Quantized Variational Inference.

Performing algorithm 5 implies finding the new optimal quantizer for X^{Γ,λ_k} at each step k. We highlight that the competitiveness of the method in terms of computational time is due to the fact that optimal quantizer derived from the base distribution $\mathbb{P}_X$ can be used to obtain $X^{\Gamma,\lambda}$ when X^{λ} can be obtained through scaling and shifting of X, since optimal quantization is preserved under these operations. For instance, in the case of BBVI with Gaussian distribution, we only need the optimal grid X^{Γ} of $\mathcal{N}(0, I_d)$ and use $X^{\Gamma,\lambda} = \mu + X^{\Gamma}\Sigma^{\frac{1}{2}}$ (given $\Sigma^{\frac{1}{2}}$ the Cholesky decomposition of Σ) to obtain the new optimal quantizer. The same goes for the distributions in the exponential family. Details about the optimal quantization for the gaussian case can be found in (Pagès and Printems, 2003). Thus, this method applies to a large class of commonly used variational distributions.

The previous results imply that quantization is relevant only for $d < 2(1 + \alpha)$ compared to MC sampling. However, numerous empirical studies have shown that this bound may be overly pessimistic, even for a not so sparse class of function in $L^2_{\mathbb{R}^d}$ (Pagès, 2018). Going further, we can implement Richardson extrapolation to improve on this bound.

7.2.3 Richardson Extrapolation

Richardson extrapolation (Richardson and Glazebrook, 1911) was originally used for improving the precision of numerical integration. The extension to optimal quantization was first introduced in (Pagès, 2018; Pagès, 2007) in the finance area to bring an answer to expensive computation of some expectation $\mathbb{E}\left[f(X_T)\right]$ for a diffusion process X_t representing a basket of assets and f an option with maturity T.

Richardson extrapolation leverages the stationary property of an optimal quantizer through error expansion. We illustrate in the one-dimensional case. Let H be

twice differential function with lipschiptz continuous second derivative. By Taylor's expansion

$$\begin{aligned}\mathbb{E}\left[H(X^{\lambda})\right] = \mathbb{E}\left[H(X^{\Gamma_N,\lambda})\right] + \mathbb{E}\left[H'(X^{\Gamma_N,\lambda})(X^{\lambda} - X^{\Gamma_N,\lambda})\right] \\ + \mathbb{E}\left[H''(X^{\Gamma_N,\lambda})(X^{\lambda} - X^{\Gamma_N,\lambda})^2\right] + \mathcal{O}(\mathbb{E}\left[|X^{\lambda} - X^{\Gamma_N,\lambda}|^3\right]).\end{aligned}$$

Then, using the stationnary property, the first order term vanishes since

$$\begin{aligned}\mathbb{E}\left[(X^{\lambda} - X^{\Gamma_N,\lambda})\right] &= \mathbb{E}\left[\mathbb{E}\left[(X^{\lambda} - X^{\Gamma_N,\lambda})|X^{\Gamma_N,\lambda}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X^{\lambda}|X^{\Gamma_N,\lambda}\right] - X^{\Gamma_N,\lambda}\right] \\ &= 0.\end{aligned}$$

Taking two optimal quantizer $X^{\Gamma_N,\lambda}$ and $X^{\Gamma_M,\lambda}$ of X^{λ} at level N, M with $N \geq M$ and using the fact that $\mathbb{E}\left[|X^{\lambda} - X^{\Gamma_N,\lambda}|^3\right] = \mathcal{O}(N^{-3})$ (Graf, Luschgy, and Pagès, 2008/ed), it is possible to eliminate the first order term by combining the two estimators with a factor N^2 and M^2.

$$\mathcal{L}(\lambda) = \frac{N^2\widehat{\mathcal{L}}^N_{OQ}(\lambda) - M^2\widehat{\mathcal{L}}^M_{OQ}(\lambda)}{N^2 - M^2} + \mathcal{O}(N^{-1}\left(N^2 - M^2\right)^{-1}). \tag{7.12}$$

We generally take $\frac{N}{M} = \gamma$ with $\gamma \in [1, 2]$ due to additional computational cost. For instance, taking $N = 2M$ leads to $\mathcal{O}(N^{-3})$ in term of absolute error. Recent results (Pagès, 2018; Lemaire, Montes, and Pagès, 2019) in higher dimension show that the general error is $\mathcal{O}(N^{-\frac{2}{d}}(N^{\frac{2}{d}} - M^{\frac{2}{d}})^{-1})$. Even though $\gamma = 2$ led to satisfying results in our experiments, applying this method to VI can lead to computational instability in higher dimensions and there is no straightforward method for finding the optimal γ.

7.3 Experiments

To demonstrate the validity and effectiveness of our approach, we considered Bayesian Linear Regression (BLR) on various dataset, a Poisson Generalized Linear Model (GLM) on the frisk data and a Bayesian Neural Network (BNN) on the metro dataset. For q_{λ}, we choose the standard Mean-Field variational approximation with Gaussian distributions.

Setup. Experiments are performed using python 3.8 with the computational library Tensorflow (Abadi et al., 2015). Adam (Kingma and Ba, 2015) optimizer

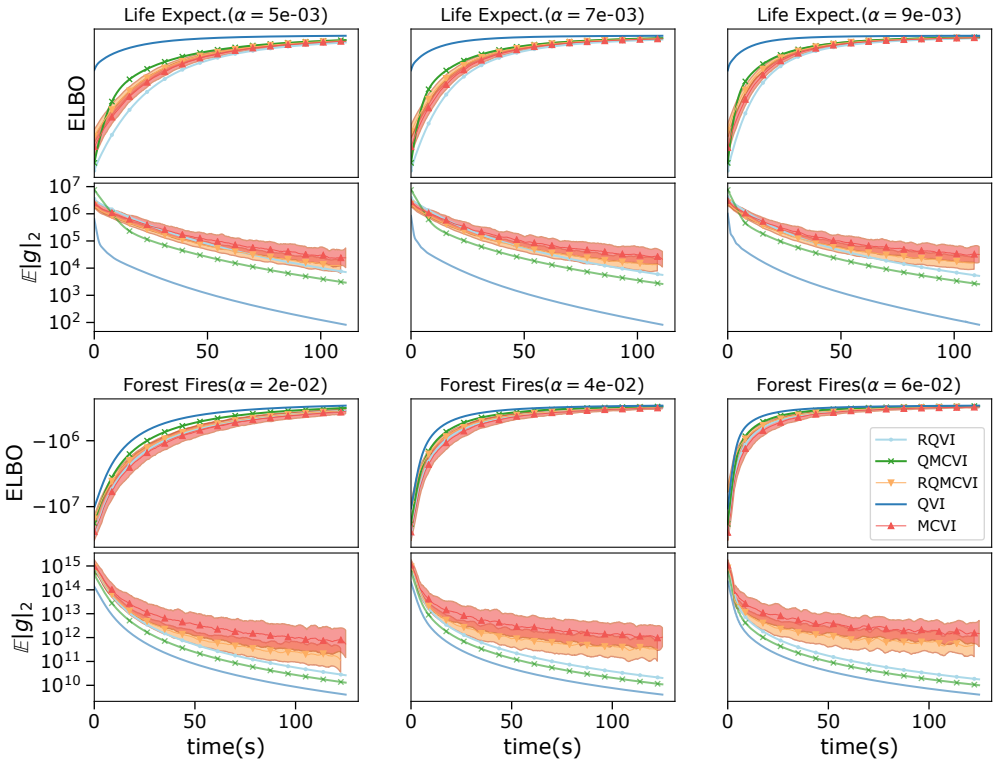


Figure 7.2: **Bayesian Linear Regression**. Evolution of the ELBO (odd rows, log scale) and expect gradient norm (even rows, log scale) during the optimization procedure for datasets reported in Table 7.1 using Adam for MCVI (red), RQMCVI (orange), QMCVI (green), QVI (blue), RQVI (light blue) as function of time. Variance for MC estimator (red area) and RQMC (orange area) are obtained by 20 runs of each experiment.

is used with various learning rates α and default $\beta_1 = 0.9, \beta_2 = 0.999$ values recommended by the author. The benchmark algorithms comprises the traditionnal MCVI described in algorithm 4, RQMC considered in (Buchholz, Wenzel, and Mandt, 2018) and QMC. We underline that (Buchholz, Wenzel, and Mandt, 2018) shows that RQMC outperforms state of the art control variate techniques such as Hessian Vector Product (HPV) (Miller et al., 2017) in a similar setting. We compare it with the implementation of algorithm 5 (QVI) and the Richardson extrapolation RQVI. For all experiments we take a sample size $N = 20$. When $D \leq 10$, precomputed optimal quantizer available online [1] is used. The Optimal Quantization is approximated in higher dimension using the R package muHVT. The number of parameters K along with the number of samples for each dataset

[1] http://www.quantize.maths-fi.com/downloads

is reported in Table 7.1. The complete documented source code to reproduce all experiments is available on GitHub [2].

Table 7.1: Datasets used for the experiments along with the Relative Bias (RB) at the end of execution for QVI and RQVI using the best learning rate.

Dataset	Size	K	QVI RB	RQVI RB
Boston	506	18	**13%**	7%
Fires	517	16	**3%**	1%
Life Expect.	2938	36	**0.3%**	0.04%
Frisk	96	70	**6%**	
Metro	48204	60	**5%**	

Bayesian Linear Regression. Figure 7.2 shows the evolution of the ELBO along with the expected ℓ_2 norm of the gradient $\mathbb{E}|g|_{\ell_2}^2$, both in log-scale. We see that QVI converges faster than vanilla MCVI and the baseline on all datasets. The gradient of both QVI and RQVI is lower than MCVI thanks to the absence of variance. However, only QVI performs better than MCVI on all datasets. For all learning rates α considered, the expected norm of the gradient is significatively lower. In these examples, it appears that the gain obtained from using RQVI is lost in the additional computation required for this method. We observe that using RQMC sampling reduces the gradient variance (odd rows) and improves the convergence rate for all experiments.
In these experiments, the resulting bias after performing a complete Gradient Descent is relatively small compared to the starting value of the ELBO. The resulting biases are reported in Table 7.1 and span from almost 0 for the Life Expectancy dataset to 13% for the Boston dataset. The fact that $\widehat{\mathcal{L}}_{OQ}^N(\lambda) > \mathcal{L}(\lambda)$ is a consequence of proposition 7.

Poisson Generalized Linear Model. Similar results are obtained by QVI for the GLM model on Frisk dataset (see Figure 7.3). QMCVI perform similarly to QVI for all learning rates but produces a larger bias in the ELBO objective function. As mentionned, RQVI can be computationnaly instable as the dimension grows. Indeed, denoting $\gamma = \frac{N}{M}$ and $\epsilon = \frac{2}{D}$, computing ELBO with Richardson extrapolation leads to

$$\widehat{\mathcal{L}}(\lambda) = \frac{\gamma^\epsilon \widehat{\mathcal{L}}_{OQ}^N(\lambda) - \widehat{\mathcal{L}}_{OQ}^M(\lambda)}{\gamma^\epsilon - 1}. \tag{7.13}$$

[2] https://github.com/amirdib/quantized-variational-inference

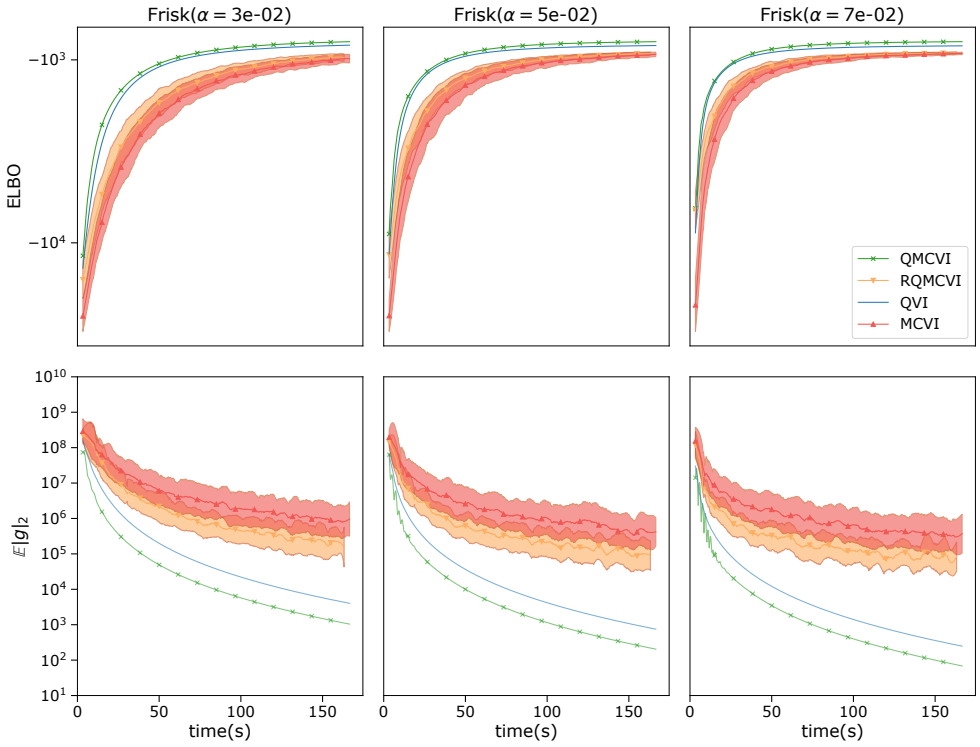


Figure 7.3: **Generalized Poisson Regression**. Evolution of the ELBO in (first row, in log scale) and expect gradient norm (second row, in log scale) during the optimization procedure for the frisk datasets (see Table 7.1) using Adam for MCVI (red), RQMCVI (orange), QMCVI (green), QVI (blue) as function of time. QVI exhibits comparable performance to QMCVI for all selected learning rate α. We use $N = 20$ sample for each experiments. Using QVI produces a relative bias of 6%.

For large D, even a small computational error between $\widehat{\mathcal{L}}^N_{OQ}(\lambda)$ and $\widehat{\mathcal{L}}^M_{OQ}(\lambda)$ can produce a large error in the estimation of $\widehat{\mathcal{L}}(\lambda)$ which led to the failure of the procedure.

Bayesian Neural Network. Finally, Bayesian Neural Network model is tested against the baseline. It consists of a Multi Layer Perceptron composed of 30 ReLu activated neurons with normal prior on weights and Gamma hyperpriors on means and variances. Inference is performed on the metro dataset. Similarly to the other experiments, Figure 7.4 shows that QVI converges faster than the baseline for all hyperparameters considered in only few epochs. Quantitatively, by taking $\alpha = 7\text{e}{-3}$ we can see that a stopping rule on the evolution of the parameters λ_k,

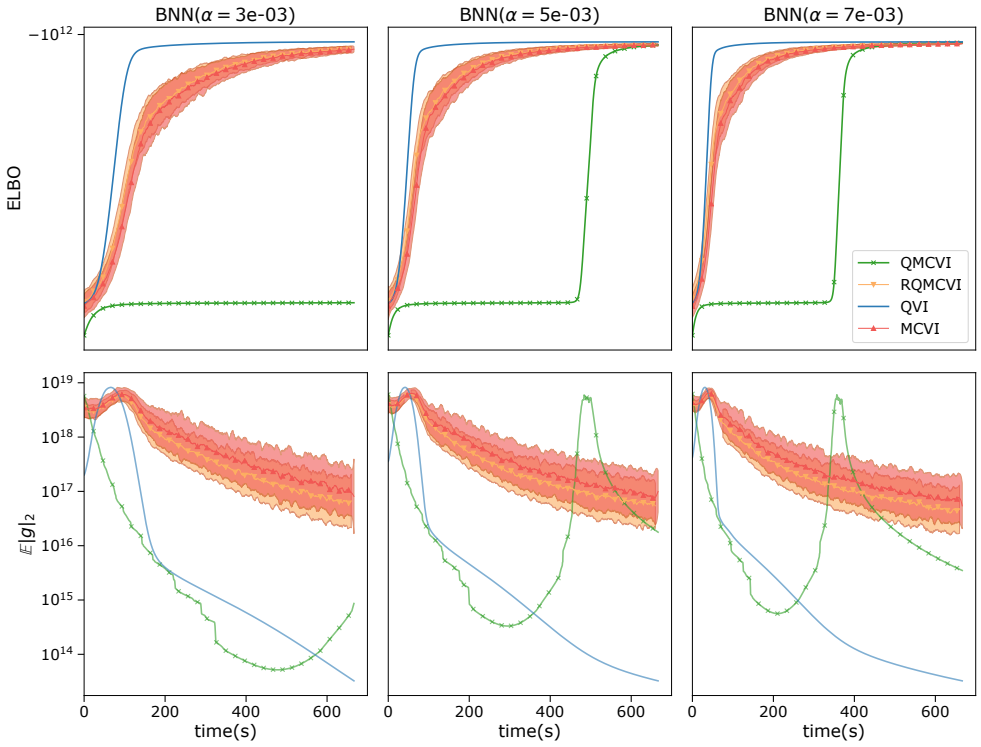


Figure 7.4: **Bayesian Neural Network**. Evolution of the ELBO in (first row, in log scale) and expect gradient norm (second row, in log scale) during the optimization procedure for the metro datasets (see Table 7.1) using Adam for MCVI (red), RQMCVI (orange), QMCVI (green), QVI (blue) as function of time. QVI exhibits superior performance with all selected learning rate α. We use $N = 20$ sample for each experiments. Using QVI produces a relative bias of 10%.

the gradient descent procedure would terminate at $t \approx 100$ seconds for QVI and $t \approx 500$ (seconds) for the MCVI algorithm.

7.4 Conclusion

This work focuses on obtaining a variance-free estimator for the ELBO maximization problem. To that end, we investigate the use of Optimal Quantization and show that it can lead to faster convergence. Moreover, we provide a theoretical guarantee on the bias and regarding its use as an evaluation tool for model selection.
The base QVI algorithm can be implemented with little effort in traditional VI optimization package as one only needs to replace MC estimation with a weighted sum.
Various extensions could be proposed, including a simple quantized control variate using the optimal quantized to reduce variance or Multi-step Richardson extrapolation (Frikha and Huang, 2015). In addition, this method could be applied more broadly to any optimization scheme, where sampling has a central role, such as normalizing flow or Variational Autoencoder. We plan to consider it in future work.

7.5 Broader Impact

Our work provides a method to speed up the convergence of any procedure involving the computation of an expectation on a large distribution class. Such case corresponds to a broad range of applications from probabilistic inference to pricing of financial products (Pagès, 2018). More generally, we hope to introduce the concept of optimal quantizer to the machine learning community and to convince of the value of deterministic sampling in stochastic optimization procedures.

Reducing the computational cost associated with probabilistic inference allows considering a broader range of models and hyperparameters. Improving goodness of fit is the primary goal of any statistician and virtually impacts all aspects of social life where such domain is applied. For instance, we chose to consider the sensitive subject of the New York City Frisk and Search policy in the 1990s. In-depth analysis of the results shows that minority groups are excessively targeted by such measure even after controlling for precinct demographic and ethnic-specific crime participation (Gelman, Fagan, and Kiss, 2007). This study gave a strong statistical argument to be presented to the authorities for them to justify and amend their policies.

Even though environmental benefits could be argued, we do not believe that such benefits can be obtained through increased efficiency of a system due to the rebound effect.

In the paper, we stressed the benefit of using our approach to improve automated machine learning pipelines, which consider large classes of models to find the best fit. This process can remove the practitioner from the modeling process, overlook any ML model's inherent biases, and ignore possible critical errors in the prediction. We strongly encourage practitioners to follow standard practices such as posterior predictive analysis and carefully examine the chosen model's underlying hypothesis.

7.6 Appendix

7.6.1 ELBO derivation

Assumes that we have observations y, latent variables z and a model $p(y, z)$ with p the density fonction for the distribution y. By Bayes' Theorem

$$\begin{aligned} p(z|y) &= \frac{p(y|z)p(z)}{p(y)} \\ &= \frac{p(y|z)p(z)}{\int_z p(z,y)dz}. \end{aligned}$$

Using the definition of KL divergence,

$$\begin{aligned} \mathrm{KL}[q_\lambda(z)\|p(z|y)] &= \int_z q_\lambda(z) \log \frac{q_\lambda(z)}{p(z|y)} dz \\ &= -\int_z q_\lambda(z) \log \frac{p(z|y)}{q_\lambda(z)} dz \\ &= -\int_z q_\lambda(z) \log \frac{p(z,y)}{q_\lambda(z)} dz + \int_z q_\lambda(z) \log p(y) dz \\ &= -\int_z q_\lambda(z) \log \frac{p(z,y)}{q_\lambda(z)} dz + \log p(y) \int_z q_\lambda(z) dz \\ &= -\mathcal{L}(\lambda) + \log p(y). \end{aligned}$$

Rearranging the terms gives equation (1).

7.6.2 Proofs

Let $f(X) \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ and $X^{\Gamma_N,\lambda}$ the the optimal quantizer of X^λ. The general framework of our study can be stated as estimating the quantity

$$I = \mathbb{E}\left[f(X)\right]. \tag{7.14}$$

We define the MC and OQ estimators as

$$I_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \tag{7.15}$$

$$I_{OQ} = \sum_{i=1}^{N} \underbrace{\mathbb{P}\left(X^{\Gamma_N,\lambda} = x_i\right)}_{\omega_i} f\left(x_i\right). \tag{7.16}$$

It is direct to derive $\|I - I_{MC}\|_2 = \mathcal{O}(N^{-\frac{1}{2}})$. In the following we establish the approximation error for the I_{OQ} estimator.

In this part we demonstrates proposition 8 and proposition 7. The former is particularly important since it establishes an asymptomatic bound on the error produced by using QVI. When considering it along with proposition 8 justifies QVI, for ranking models with it will produce true ranking provided that the relative difference in ELBO is lower than the quantization error. In the following we formally demonstrate such result (thorough investigation of optimal quantizer can be found in (Pagès, 2018; Pagès, 2015)). We begin with the definition of a stationnary quantizer.

Definition 12. Let $\Gamma_N = \{x_1, \dots, x_N\}$ be a quantization scheme of X^λ. $X^{\Gamma_N,\lambda}$ is said to be stationary quantizer if the Voronoi partition induced by Γ_N satisfies $\mathbb{P}\left(X \in C_i(x)\right) > 0 \ \forall i \in \{1, \dots, N\}$ and

$$\mathbb{E}\left[X^\lambda | X^{\Gamma_N,\lambda}\right] = X^{\Gamma_N,\lambda}.$$

One of the first question raised by using optimal quantization $\mathbb{E}\left[H(X^{\Gamma_N,\lambda})\right]$ in place for $\mathbb{E}\left[H(X^\lambda)\right]$ is the error produced by such substitution. Let us remind that we denote $\widehat{\mathcal{L}}^N_{OQ}(\lambda) = \mathbb{E}\left[H(X^{\Gamma_N,\lambda})\right]$ the quantized ELBO estimator and $\mathcal{L}(\lambda) = \mathbb{E}\left[H(X^\lambda)\right]$ the true ELBO.

Lemma 5. *Let $X^\lambda \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ and a H a continuous lipschitz function with Lipschitz constant C, we have*

$$\left|\mathcal{L}(\lambda) - \widehat{\mathcal{L}}^N_{OQ}(\lambda)\right| \leq C \left\|X^\lambda - X^{\Gamma_N,\lambda}\right\|_2.$$

Proof.

$$\begin{aligned}
\left|\mathbb{E}\left[H(X^\lambda)\right] - \mathbb{E}\left[H(X^{\Gamma_N,\lambda})\right]\right| &\leq \mathbb{E}\left[\mathbb{E}\left[\left|H(X^\lambda) - H(X^{\Gamma_N,\lambda})\right| | X^{\Gamma_N,\lambda}\right]\right] \quad (7.17)\\
&\leq C \left\|X^\lambda - X^{\Gamma_N,\lambda}\right\|_1 \\
&\leq C \left\|X^\lambda - X^{\Gamma_N,\lambda}\right\|_2. \quad (7.18)
\end{aligned}$$

We use Jensen inequality in equation 7.17 and the monoticity of the $L_p(\Omega, \mathcal{A}, \mathbb{P})$ norm as a function of p in equation 7.18. □

Proposition 8. *Let $X^\lambda \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ and $X^{\Gamma_N,\lambda}$ the associated optimal quantizer, under the hypothesis that H is a convex lipschitz function,*

$$\widehat{\mathcal{L}}^N_{OQ}(\lambda) \leq \mathcal{L}(\lambda).$$

Proof.

$$\begin{aligned}
\widehat{\mathcal{L}}^N_{OQ}(\lambda) &= \mathbb{E}\left[H(X^{\Gamma_N,\lambda})\right] \\
&= \mathbb{E}\left[H\left(\mathbb{E}\left[X^\lambda | X^{\Gamma_N,\lambda}\right]\right)\right] \qquad (7.19)\\
&\leq \mathbb{E}\left[\mathbb{E}\left[H(X^\lambda)|X^{\Gamma_N,\lambda}\right]\right] \\
&= \mathbb{E}\left[H(X^\lambda)\right] \qquad (7.20)\\
&= \mathcal{L}(\lambda)
\end{aligned}$$

When we used Lemma 12 in equation 7.19 and the conditional Jensen inequality to obtain 7.20. □

Proposition 9. *Let $\lambda^* = \min_{\lambda\in\mathbb{R}^K} \mathcal{L}(\lambda)$ and $\lambda^*_q = \min_{\lambda\in\mathbb{R}^K} \widehat{\mathcal{L}}^N_{OQ}(\lambda)$. Under the same assumptions than proposition 8,*

$$\mathcal{L}(\lambda^*) - \widehat{\mathcal{L}}^N_{OQ}(\lambda^*_q) \leq C\left[2\|X^{\lambda^*} - X^{\Gamma,\lambda^*}\|_2 + \|X^{\lambda^*_q} - X^{\Gamma,\lambda^*_q}\|_2\right].$$

Proof. A immediate consequence of proposition 8 is that $\widehat{\mathcal{L}}^N_{OQ}(\lambda^*_q) \leq \mathcal{L}(\lambda^*)$. Then, we can write

$$\begin{aligned}
\mathcal{L}(\lambda^*) - \widehat{\mathcal{L}}^N_{OQ}(\lambda^*_q) &= \mathcal{L}(\lambda^*) - \widehat{\mathcal{L}}^N_{OQ}(\lambda^*) \\
&\quad + \widehat{\mathcal{L}}^N_{OQ}(\lambda^*) - \mathcal{L}(\lambda^*_q) \\
&\quad + \mathcal{L}(\lambda^*_q) - \widehat{\mathcal{L}}^N_{OQ}(\lambda^*_q) \\
&\leq C\|X^{\lambda^*} - X^{\Gamma,\lambda^*}\|_2 \\
&\quad + C\|X^{\lambda^*_q} - X^{\Gamma,\lambda^*_q}\|_2 \\
&\quad + C\|X^{\lambda^*} - X^{\Gamma,\lambda^*}\|_2
\end{aligned}$$

Using Lemma 5 and noting that

$$\widehat{\mathcal{L}}^N_{OQ}(\lambda^*) - \mathcal{L}(\lambda^*_q) \leq \widehat{\mathcal{L}}^N_{OQ}(\lambda^*) - \mathcal{L}(\lambda^*),$$

proposition 7 follows. □

Finally, Zador's theorem is used to derive non-asymptotic bound (see (Luschgy and Pagès, 2008) for a complete proof).

Theorem 6 (Zador's Theorem)**.** *Let $X^{\lambda} \in L^2_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ and $X^{\Gamma_N,\lambda}$ the associated optimal quantizer at level N, there exists a real constant $C_{d,p}$ such that*

$$\forall N \geq 1, \quad \left\| X - \widehat{X}^{\Gamma_x} \right\|_p \leq C_{d,p} N^{-\frac{1}{d}}$$

Where $C_{d,p}$ dependens only d and p. This result can be vastly improved when H exhibits more regularity. For instance, if H is an α hölderian function, we can obtain a bound in $\mathcal{O}(N^{-\frac{1+\alpha}{d}})$ (Pagès, 2015).

7.6.3 Experiments

Bayesian Linear Regression. We used three different real-world dataset, namely Forests Fire, Boston housing datasets from the UCI repository (Dua and Graff, 2017) and Life Expectancy dataset from the Global Health Observatory repository. The generative Bayesian Linear Gaussian Model used is as follow.

$$\begin{aligned} \mathbf{b}_i &\sim \mathcal{N}\left(\mu_\beta, \sigma_\beta\right), && \text{intercepts} \\ y_i &\sim \mathcal{N}\left(\mathbf{x}_i^\top \mathbf{b}_i, \epsilon\right), && \text{output} \end{aligned}$$

Let D be the dimension of the feature space. The dimension of the parameter space for a gaussian variationnal distribution under the mean-field assumption is $K = 2D$.

Poisson Generalized Linear Model. The frisk dataset is a record of stops and searches practice on civilians in New York City for fifteen months in 1998−1999. It contains information about locations, ethnicity and crime statistics for each area. The question is whether these stops targeted particular groups after taking into account population and crime rates in each group for a particular precinct.
We can trace back the use of Poisson Generalized Linear Model for this use case to (Gelman, Fagan, and Kiss, 2007). The model writes as follow

$$\mu \sim \mathcal{N}\left(0, 10^2\right) \qquad \text{mean offset} \tag{7.21}$$

$$\log \sigma_\alpha^2, \log \sigma_\beta^2 \sim \mathcal{N}\left(0, 10^2\right) \qquad \text{group variances} \tag{7.22}$$

$$\alpha_e \sim \mathcal{N}\left(0, \sigma_\alpha^2\right) \qquad \text{ethnicity effect} \tag{7.23}$$

$$\beta_p \sim \mathcal{N}\left(0, \sigma_\beta^2\right) \qquad \text{precinct effect} \tag{7.24}$$

$$\log \lambda_{ep} = \mu + \alpha_e + \beta_p + \log N_{ep} \qquad \text{log rate} \tag{7.25}$$

$$Y_{ep} \sim \text{Poisson}\left(\lambda_{ep}\right) \qquad \text{stops events} \tag{7.26}$$

$$\tag{7.27}$$

Y_{ep} denotes the number of frisk events for the ethnic group e in the precinct p. N_{ep} is the number of arrests for the ethnic group e in the precinct p. Hence, in this model, α_e and α_p represents the ethnicity and precinct effect. The dataset contains three ethnicities and thirty-two precinct, which therefore exhibits $K = 70$ variational parameters.

Bayesian Neural Network. The Bayesian Neural Network (BNN) consists of a Multi Layer Perceptron (MLP) ψ of 30 ReLU activated neurons with normal prior weights and inverse Gamma hyperprior on the mean and variance. Regression is performed on the metro dataset.

$$\alpha \sim \text{Gamma}(1, 0.1) \qquad \text{weights hyper prior} \qquad (7.28)$$

$$\tau \sim \text{Gamma}(1, 0.1) \qquad \text{group variances} \qquad (7.29)$$

$$w \sim \mathcal{N}\left(0, \frac{1}{\alpha}\right), \qquad \text{neural network weights} \qquad (7.30)$$

$$y \sim \mathcal{N}\left(\psi(w, x), \frac{1}{\tau}\right) \qquad \text{output} \qquad (7.31)$$

Thanks to open source libraries

This work and many others would have been impossible without free, open-source computational frameworks and libraries. We particularly acknowledge Python 3 (Van Rossum and Drake, 2009), Tensorflow (Abadi et al., 2015), Numpy (Oliphant, 2006) and Matplotlib (Hunter, 2007).

Conclusion and perspectives

This thesis introduces new methods in the context of Data Mining and Bayesian Learning, starting from the need to design human-readable and relevant methods to study symbolic time series processes in the context of predictive maintenance.

Part II: Anomaly detection for rolling stock maintenance

After a selective review of the anomaly detection domain with a highlight on the railway industry in Chapter 2, we construct an industrial pipeline for the predictive maintenance task with interpretable and human-readable output applied to a large fleet for high-speed trains. This use case is particularly challenging; the industrial system of railway spans across a vast territory with various environments and involves complex heterogeneous and interconnected systems. We designing an industrial prediction pipeline and propose a method to overcome computational complexity that comes with a high number of possible hyperparameters based on a two-sample test to prune the tree of operations to perform. Additionally, the use of pattern extraction method on the temporal signal of error codes by using tools from the Data Mining domain allows retrieving relevant and interpretable patterns, or association of event codes, linked to a specific type of malfunction. This approach provides an approach towards the automatic extraction of rules that experts can directly understand. A possible continuation of this work could be studying ways to transfer these learned rules to similar classes of systems on which we do not have enough data to perform statistical analysis. To that end, various methods in the domain of *transfert learning* and *active learning* offer an exciting venue for future researches.

Part III: Pattern Mining

This part is devoted to the presentation of new methods in the Data Mining field. Data Mining has become one of the most well-studied applied mathematics fields thanks to the broad availability of data. Nevertheless, performing relevant tasks on these high-dimensional databases is typically intractable. We tackle this issue using two different approaches: the first uses models on the underlying process

and the second consists of bounding the empirical estimator of a random variable of interest.

Chapter 4 present a Bayesian Generative Model for the pattern mining and discriminative pattern mining tasks. Notably, we first show that the set of frequent itemsets can be efficiently mined using a stochastic approximation method using variational inference scheme to obtain the space of frequent itemsets with high accuracy. Second, we use a Bayesian Mixture Model to infer with a low computational cost the discriminative itemsets opening the possibility of extracting relevant information on binary labeled datasets. Finally, we present a method for enriching the space of feature variable of any classifier, improving the metric score on various public and industrial datasets.

Chapter 5 tackle the problem of deriving distribution independent bounds on the support of itemsets. Estimating the frequencies of any pattern among the exponential set of possible ones is one of the most fundamental problem in Data Mining. Contrary to previous work that uses global complexity measure on underlying empirical process generating the database, we introduce the first use of local complexities for the task mining low-frequency itemsets. We show theoretically and empirically that our method outperforms asymptotically the most recent approaches. Future work will consider the application of this method to related pattern concepts, wherein new computational routines may be required to bound local Rademacher averages with *infinite pattern families* or *utility pattern families*, and to tasks with other objectives. The computational efficiency of the introduced algorithms could be significantly improved by avoiding the enumeration of all closed patterns by using instead *branching approach* (Pellegrina et al., 2020). In addition, we believe that this method can be relevant to a broader range of tasks where we can partially sample the database on which an algorithm performs and need to evaluate the precision of the produced output.

Part IV: Optimal Quantization for stochastic optimization

Chapter 6 and 7 introduce a new method for variance reduction in stochastic optimization. Stochastic optimization is one of the most prominent problems in statistics and optimization and consists of minimizing the functional of a random variable under expectation. The efficiency of any procedure that performs this task will be highly dependent on the variance of the random variable hence the multiplicity of work aiming at reducing the variance of the considered empirical estimator. In this thesis, we took an entirely new approach toward this problem by considering a variance-free estimator, namely the *optimal quantizer*. Even though biased, it showed superiority over existing methods in terms of convergence speed in the framework of Bayesian learning with a theoretical guarantee over the produced biased.

In Chapter 6 we developed a view of the construction of the optimal quantizer with the tool of optimal transport. To the best of our knowledge, this approach has not yet been explored to compute the optimal quantizer. One of the main drawback when using an optimal quantizer is the computational cost associated with producing it. We believe that several approximation tools from optimal transport can be used to perform this task efficiently.
Chapter 7 introduced Quantized Variational Inference, which is a competitive algorithm for Variational Inference that can be utilized for any inference task in the bayesian settings. Nevertheless, the Quantized Stochastic Approximation could be applied to several domains such as Normalizing Flows, Reinforcement Learning, Variational AutoEncoders and any tasks requiring a stochastic optimization step.

Part V

Appendix

Appendix A

A probabilistic point of view on pattern mining

A priori-based algorithms are commonly used for finding interesting itemsets from transaction databases. We can construct a model which explains from a probabilistic perspective the main results of pattern mining with derived algorithm retrieve itemsets with comparable quality. We can use this reformulation to construct a bayesian probabilistic model for itemsets and to propose a method to significantly reduce exact pattern extraction computation time with control of the error with tight guarantee.

Computational methods for Knowledge Discovery from large Datasets (KDD) aim to extract relevant information structure from a database. This process often involves extracting a collection of interesting patterns $\mathcal{T}(Db, \mathcal{L}, \mathcal{C}) = \{\rho \in \mathcal{L} | \mathcal{C}(\rho, Db)\}$ from a language $\mathcal{L}$ given the data D where $\mathcal{C}$ is constraint function which encodes our criteria of an interesting pattern (Mannila and Toivonen, 1997). Once the problem description set, the computation of such collection can be challenging. If you consider the problem of Frequent Itemset Mining (FIM), the language size is of $\|\mathcal{L}\| = 2^d$ where d is the size of the dictionary (or the number of different items in the database). Even for D relatively small the estimation of each element in the language is unfeasible. The key is to carefully consider constraints function $\mathcal{C}$ as it can be exploited to dramatically reduce the number of computations to perform.

A.1 Background

Agrawal and Srikant (1994) developed the Apiori algorithm which was the the first efficient procedure for FIM which consists of a breadth-first search algorithm over the complete lattice on the associated powerset associated with the inclusion relation leveraging the antimonoticity property of frequent itemsets. This approach

although simple to implement is not optimal since it requires multiple scan of the database and large quantity of memory. Moreover, the structure of the result is himself somewhat redundant; the complete set of frequent itemsets can be generated from a more parsimonious structure on itemsets. One of such structure is the set of closed itemsets which has the so-called *lossless* property. Simple APRIORI like methods are dedicated to recover such sets (Lucchese, Orlando, and Perego, 2006; Zaki and Hsiao, 2005; Cerf, 2010). To reduce the quantity of memory used (Zaki, 2000) proposes the *Eclat* which is a depth-first algorithm and avoid the costly storage of all candidate itemsets of a given length. ECLAT is much faster than APRIORI and perform only one database scan by using a vertical representation of the transaction data. A big leap in term of computation and memory efficiency has been taken with the development of *pattern-growth* methods. The central idea is to avoid the generation of unnecessary by using the projected database given an itemsets to reduce the space to explore as the algorithm searches in larger itemsets. *FP-Growth* (Han et al., 2004) and *LCM* (Uno, Kiyomi, and Arimura, 2004) are two methods based on this principle and are the state of the art in term of computation time for most datasets. An open-source and efficient implementation of these algorithms is available in the SPMF library (Fournier Viger et al., 2016).

A.1.1 Itemset theory

Let $T = (t_1, \dots, t_n)$ and $E = (e_1, \dots, e_d)$ two sets. We denote $\mathcal{E} = \mathcal{P}(E)$ and $\mathcal{T} = \mathcal{P}(T)$ the supersets and $\mathcal{R}$ a binary relation between T and E (*i.e.* a subset of $T \times E$). In this framework, E represents the base dictionary of the patterns and T the transactions in the database and the collection of elements of $T \times E$ the transactions database. To formalize the concept of interesting itemsets, we introduce the following two operators:

Definition 13 (Galois connection)**.** The left and right adjoin f^* and f_* id a pair of function such that for any $E' \in \mathcal{E}$ and $T' \in \mathcal{T}$ we have

$$\begin{array}{rccl} f^*: & \mathcal{E} & \longrightarrow & \mathcal{T} \\ & E' & \longmapsto & \{e \in E' | \forall t \in T', t\mathcal{R}e\}, \\ f_*: & \mathcal{T} & \longrightarrow & \mathcal{E} \\ & T' & \longmapsto & \{t \in T' | \forall e \in E', t\mathcal{R}e\}. \end{array}$$

The pair $\langle f_*, f^* \rangle$ form a Galois connection (Davey and Priestley, 2002) over the posets $(\mathcal{E}, \subseteq)$ and $(\mathcal{T}, \subseteq)$ induced by the binary relation $\mathcal{R}$ on $T \times E$. In this paper, we're particularly interested in the FIM problem which can be defined, given a support threshold $\mu \in [0, 1]$, as the search for an itemsets $E' \in \mathcal{E}$ which appears with a frequency at least μ across the transaction database.

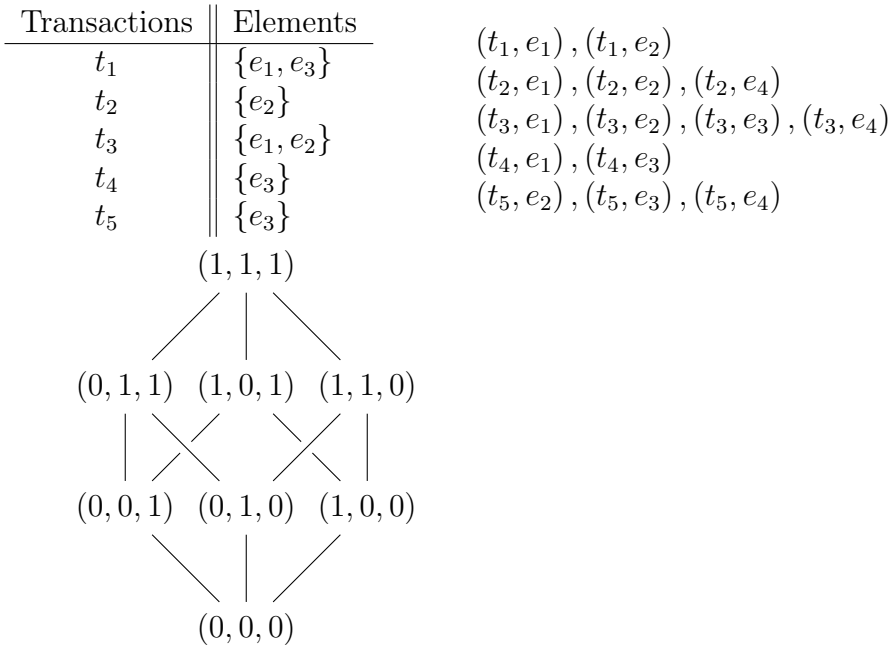


Figure A.1: Different representation of the binary relation $\mathcal{R}$ for $T = (t_1, \ldots, t_4)$ and $E = (e_1, \ldots, e_4)$.

Definition 14 (Frequency constraint)**.** The support of $E' \in \mathcal{E}$ is the function $S : \mathcal{E} \to [0, 1]$ such as

$$s(E') = \frac{|f^*(E')|}{|T|}.$$

Let $\mu \in [0, 1]$ the user-defined treshold, the *frequent itemset constraint* is a boolean function *s.a.*

$$\mathcal{C}_{\mu,\mathrm{s}}(E') = \mu \leq s(E').$$

The goal of FIM is to find all elements of $\mathcal{E}$ that regularly occur in a database viewed as a binary relation $\mathcal{R}$. A formal and general definition reads as follow.

Definition 15 (Frequent Itemset Mining)**.** Let $\mathcal{R}$ a binary relation on $T \times E$, $\mu \in [0, 1]$, $\mathcal{C}_\mu$ a constraint on $\mathcal{E}$. A Frequent Itemset Mining (FIM) algorithm aims to compute the set $\mathcal{TH}(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}) = \{E' \in \mathcal{E}; \mathcal{C}_{\mu,\mathrm{s}}(E') \text{ is true}\} \subseteq \mathcal{E}$.

A brute-force approach to this problem would lead to a prohibitively large number of candidates itemsets to explore. More precisely, such algorithm will have exponential complexity $\mathcal{O}(N2^M)$. We stress that the length database itself can be large and costly to access so that the computation of $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$ can be

infeasible even for reasonable value of d. The key for pruning the set of possible patterns is in the anti-monoticity property of the frequent itemset constraint, which states that every sub-pattern of a $\mathcal{C}_{\mu,\mathrm{s}}$ frequent pattern is frequent:

Proposition 10 (Anti-monoticity)**.** $\forall A, B \in \mathcal{E},$ *if* $A \subseteq B$ *then* $\mathcal{C}_{\mu,s}(B) \Rightarrow \mathcal{C}_{\mu,s}(A)$.

This simple property is at the hear of every algorithm used for pattern mining for pruning the search space. The most notorious, the APRIORI algorithm, starts from the L_1 set of 1-frequent itemsets and generate a set C_2 composed of $\binom{2}{|L_1|}$ candidates derived from L_1 (*i.e.* all the supersets of size 2 containing element of L_1). The support of the candidates itemsets C_2 are then evaluated and we construct L_2 as the ones which achieve the $\mathcal{C}_{\mu,\mathrm{s}}$ criterion. At step k we evaluate the set C_k of $\binom{k}{|L_{k-1}|}$ candidates to construct L_k. The algorithm stops at step K when C_{K+1} is found to be empty. The complexity of such algorithm depends on the support treshold μ, the number of unidentical items M, the number of transaction N and the typical length of the transactions. This last dependence has a substantial effect over the maximum size of μ-frequent itemsets which imply that more candidate itemsets need to be evaluated at each step. However, except for weakly correlated items the set $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$ is very large even when using *loss-less* form of itemsets. An common loss-less representation of the patterns consists of the set of closed itemsets from which any frequent itemset can be retrieved which is defined as follow:

Definition 16 (Closed itemset)**.** Let $\mathcal{R}$ a binary relation on $T \times E$, $\langle f_*, f^* \rangle$ a Galois connection over the posets $(\mathcal{E}, \subseteq)$ and $(\mathcal{T} \subseteq)$. The *closure operator* $f = f_* f^*$ is the galois closure operator over $\mathcal{E}$. The set the set of all closed itemsets is defined as $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{closed}}\right) = \{E \in \mathcal{E}; f(E) = E\}$.

A closed itemset is maximal in the sense that it's support is strictly decreasing with respect to the inclusion. More precisely, for any closed itemset $A \in \mathcal{C}_{\mu,\mathrm{closed}}$ and an itemset $B \in \mathcal{E}$ *s.a.* $A \subset B$ then $s(B) < s(A)$.

Once $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$ has been computed an analysis of interest is the so-called Association Rule Mining (ARM) introduced by (Agrawal, Imielinski, and Swami, 1993) for basket data analysis. Given a binary relation $\mathcal{R}$ the goal is to find couples of itemsets that tend to co-occur.

Definition 17 (Association rule)**.** Given a binary relation $\mathcal{R} \subseteq \mathcal{E} \times \mathcal{O}$ and $X, Y \mathcal{E}$ two disjoint itemsets, an association rule $X \to Y$ is relation on itemsets that is reflexive, antisymmetric. Additionnaly, we define the frequency and confidence associated with such rule by

$$f(X \to Y) = \frac{|f^*(X \cup Y)|}{M},$$
$$c(X \to Y) = \frac{|f^*(X \cup Y)|}{|f^*(X)|}.$$

For a given $\nu, \mu \in [0, 1]$ The goal of ARM is to compute the set of itemsets $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$ amir: Correct defined as the pair of itemset of frequency and confidence respectively greater than μ and ν from the collection of all possible association rules. We can immediately see that the rule $X \to Y$ is $\mu -$ frequent *iff* the itemset $X \cup Y \in \mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$. Moreover, we can use the anti-monocity property of the confidence function to prune the space of possible rules identically to the support for FIM.

Proposition 11. *Given* X, X', *and* $Y \in 2^{\mathcal{P}}, leX \subseteq X' \subseteq Y$, *we have* $c(X \to Y \backslash X) \leq c\left(X' \to Y \backslash X'\right)$.

Therefore, we can divine a simple way to perform ARM by performing FIM then deriving all the association rules from $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$ using the previous remark. There is a great variety of measure of interestingness other that the confidence measure. Most of them are inspired by measure of correlation in statistics ("Interestingness Measures for Data Mining") and each one has it's own implications in term of which type of rule will be extracted.

A.1.2 The probabilistic framework for itemsets

We now turn to the probabilistic framework for pattern mining. The question of a probabilistic framework was raised by the need to tackle the problem of Uncertain Database Mining (UDM). In many application, the data are collected with an uncertainty over the measurement that the user wants to acknowledge in the pattern extraction.

The advantage of generative model are multiple. First, we'll show that the generative model is a good heuristic for several type of pattern mining problem. Second, it's the most parcemonious representation of the set of itemsets support. For large datasets with dense itemsets the memory space needed to store and analyse $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,\mathrm{s}}\right)$ can be prohibitive (it can quickly reach few hundred Gb). In the generative model framework the only memory needed is for storage of the parameters of the model (at most 2^d parameters). Additionnaly, this representation is more versatile as it allows for construction of more complex rules. Moreover, it's easy to introduce element from expert knowledge in the model by adopting, for instance, a bayesian approach. Such account of prior information is difficult to implement in the deterministic pattern mining algorithm. Using adapted techniques,

it's also fast to compute. Last but not least, this approach take into account the uncertainty of a training dataset. Uncertainty from the ignorance of the true underlying model generating the data and uncertainty as a consequence of the finite sampling.

In the following we adapt the common notations in probability theory for pattern extraction. Let $(\Omega, \mathcal{F})$ a measurable space, $\mathcal{X} = \mathcal{P}E$ the set of all itemsets, $\mathcal{G} = \mathcal{X}$ an algebra on $\mathcal{X}$, $X : \Omega \to \mathcal{X}$ a $(\mathcal{F}\text{-}\mathcal{G})$ measurable function, p be a probability distribution on $\mathcal{X}$ (*i.e.* $p : \mathcal{X} \to [0,1]$ and $\sum_{x \in \mathcal{X}} p(x) = 1$). We define the probability distribution on $(\Omega, \mathcal{F})$ as

Definition 18.

$$\begin{array}{rccl} \mathbb{P}: & \mathcal{F} & \longrightarrow & [0,1] \\ & A & \longmapsto & \mathbb{P}(A) = \sum_{\omega \in A} p(X(\omega)). \end{array}$$

Hence $(\Omega, \mathcal{F}, \mathbb{P})$ define a probability space. From the pattern mining perspective, a specific transaction $(t_i, x_i) \in T \times \mathcal{E}$ is viewed as a realization of X under $\mathbb{P}_i$ and the entire transaction database of length n as the sampling model on $(\Omega, \mathcal{G}^n, \mathbb{P}^{\otimes n} = \prod_{i=1}^{n} \otimes \mathbb{P}_i)$. For an event $A \in \mathcal{F}$ and sample $(x_1, \ldots, x_n)$ we can compute empirical probability of A as

$$\hat{\mathbb{P}}_n(A) = \frac{1}{N} \sum_{\omega \in A} \sum_{i=0}^{n} \mathbb{1}_{\{X(\omega) = x_i\}}. \tag{A.1}$$

The convergence is governed by the strong law of large numbers, hence $\hat{\mathbb{P}}A$ converges almost surely towards $\mathbb{P}(A)$. We define the support of a pattern by the function

$$\begin{array}{rccl} s_{\mathbb{P}}: & \mathcal{X} & \longrightarrow & [0,1] \\ & t & \longmapsto & s_{\mathbb{P}}(t) = \mathbb{P}(A_t), \end{array}$$

with $A_t = \{z \in \mathcal{X} | t \subseteq z\}$ the set of all itemsets greater than t (when there's no ambiguity, we'll simply write $s(t)$). Notice the close analogy between this definition and the traditional inverse empirical distribution for classic numerical random variables. However the natural transitive, anti-symmetric and reflexive relation of order on real line $\geq$ is replaced by the order relation on superset $\subseteq$. The previous support random variable has consistent properties expected from the traditional support for a pattern

$$\forall x, y \in \mathcal{X}, \quad x \subseteq y \to s(y) \leq s(x) \quad \text{(anti-monoticity)}.$$

Moreover, following the close relation with the empirical distribution we have the following proposition by using the fact that equality of two measures on $\{A_t; \forall t \in \mathcal{X}\}$ imply equality of the two measures on $\mathcal{X}$

Proposition 12. *Let $\mathbb{P}$ and $\mathbb{Q}$ two probability distribution on $(\Omega, \mathcal{G})$ and $s_{\mathbb{P}}, s_{\mathbb{Q}}$ their respective support functions, if $s_{\mathbb{P}} = s_{\mathbb{Q}}$ then $\mathbb{P} = \mathbb{Q}$.*

In other words, the support function carracterize the probability distribution. We can then reformulate the APRIORI proposition

Proposition 13 (A Priori)**.** *let $\mu \in [0, 1]$ and $\mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,s}\right) = \{x \in \mathcal{X}; \mu \leq s(x)\}$. Then*

$$\forall x \subseteq y, \quad y \in \mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,s}\right) \Rightarrow x \in \mathcal{TH}\left(\mathcal{R}, \mathcal{C}_{\mu,s}\right).$$

This approach allows for interpretation of the KRIMP algorithm base on the Minimum Description Length (MDL) principle

$$L(D) = \min_{H \in \mathcal{H}}(L(H) + L(D|H)).$$

The principle is a formulation of the Occam's razor rule. On an other hand Shannon definition of entropy states that for a probability distribution $p : \mathcal{X} \to [0, 1]$ there is an optimal length for each x that compress any sample of p. Conversly, for a set of values x there is a distribution

$$P(\mathbf{x}) = 2^{-L(\mathbf{x})}, \quad L(\mathbf{x}) = -\log_2 P(\mathbf{x}).$$

From our perspective $L(D, \mathcal{H})$ define a priori over the choice of the model $\mathcal{H}$ and $L(D|\mathcal{H})$ the sampling distribution given the model $\mathcal{H}$.

$$\begin{aligned} L(D, \mathcal{H}) &= -\log P(\mathcal{H}) - \log(P(D|\mathcal{H})) \\ &= -\log P(\mathcal{H}|D) + \text{const.} \end{aligned}$$

In general MDL principle can be interpreted as a probabilistic model as demonstrated by (Smith, Erickson, and Neudorfer, 2013) and seems to have no advantages over the probabilistic approach.

Appendix B

Hidden Markov Model

Markov process models are a class of probability models used to study the evolution of a stochastic system over time. Transition probabilities are used to identify how a system evolves from one time to the next. In our case, a Markov chain tries to characterize the system behavior over time, as described by the transition probabilities matrix, emissions matrix, and the initial state probability. In the following, we give a formal description of Hidden Markov Model.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ the probability space, and let A be the state and U the hidden space, $(X_1, \ldots, X_n)$ a sequence of A-valued random variables, $(U_1, \ldots, U_l)$ a sequence of U-valued random variables. We say that $(X, U) = \big((X_n, U_n)\big)_{n\geq 1}$ is a Hidden Markov Model on $A \times U$ with the following transition law:

$$\begin{aligned} P\Big((x,u),\big(x',u'\big)\Big) &= \mathbb{P}\left(X_{n+1} = x', U_{n+1} = u' | X_n = x, U_n = u\right) \\ &= \rho\left(u, u'\right)\pi_{u'}\left(x, x'\right), \end{aligned} \tag{B.1}$$

where ρ and π_u two transition probability on A and U. The Markov chain is irreducible and recurrent positive. Thus, it converges to a stationary distribution. In practice we have two type of Hidden Markov Model, namely the ergodic and the left-to-right model. The most used one, the left-to-right Hidden Markov Model consists for each hidden state s_i to be linked only to s_i or s_{i+1}. It is a sequential view of the hidden states.

Definition 19. A Hidden Markov Model is defined by the tuple $\Lambda = (A, B, \pi)$ which as the following properties

- The state space $S = \{s_1, s_2, \ldots s_n\}$ with $q_t \in S$ the state of the chain at time t;
- The state of observations $V = (v_1, \ldots, v_M)$ and $O_t \in V$ the observation at time t;

- An ergotic matrix A of transition to represent the probability transitions of the hidden states:

$$a_{ij} = A(i,j) = \mathbf{P}\left(q_{t+1} = s_j | q_t = s_i\right) \forall i,j \in [1 \dots n] \forall t \in [1 \dots T]; \quad \text{(B.2)}$$

- A probability matrix B of observations giving the probability $b_{ij}$of observing v_i in the state s_j

$$b_j(k) = \mathrm{P}\left(O_t = v_k | q_t = s_j\right) \quad 1 \leq j \leq n, 1 \leq k \leq M; \quad \text{(B.3)}$$

- A probability vector π giving the initial probability of the chain state:

$$\pi_i = \mathbf{P}\left(q_1 = s_i\right) \quad 1 \leq i \leq n. \quad \text{(B.4)}$$

Given an observations sample $(O_1, \dots, O_T)$ and a HMM parametrized by $\Lambda = (A, B, \pi)$ there is three fundamental questions that arise:

1. What's the likelihood $\mathcal{P}(O_k|\Lambda)$ for a given sequence O_k ? It can straightfowardly be obtained by summing over the possible sequences:

$$\mathrm{P}(O|\Lambda) = \sum_Q \mathrm{P}(O, Q|\Lambda) = \sum_Q \mathrm{P}(O|Q, \Lambda)\mathrm{P}(Q|\Lambda). \quad \text{(B.5)}$$

   Using the markov property

$$\mathbf{P}(O|\Lambda) = \sum_{Q=q_1,q_2,w_2r_r} \pi_{q_1} b_{q_1}\left(O_1\right) a_{q_1 q_2} b_{q_2}\left(O_2\right) \dots a_{q_{T-1} q_T} b_{q_T}\left(O_T\right). \quad \text{(B.6)}$$

   This quantity is intractable at first sight with complexity in $\mathcal{O}(n^T)$. But it is possible to compute $\mathcal{P}(O_{1:t+1}|\Lambda)$ given $\mathcal{P}(O_{1:t}|\Lambda)$ in $\mathcal{O}(n)$ so that the evaluation of the likelihood can be evaluated in $\mathcal{O}(Tn^2)$.

2. Given a sample of observations O, how do we find the most likely chain of a hidden state producing such a sequence of observable states? Formally, we are in search of a sequence Q that maximize the following quantity:

$$\mathrm{P}(Q, O|\Lambda). \quad \text{(B.7)}$$

   This quantity can be easily obtained using a dynamic programming algorithm (know as the Viterbi in this case) in $\mathcal{O}(Tn^2)$.

3. How do we learn the $\Lambda = (A, B, \pi)$? In this case, the simplest solution consists of using Expectation Maximization procedure (Blume, 2002) taking advantage of the fact that we know the optimal parameters given the latent variable of the location of the chain in the hidden space.

Appendix C

Portée et motivation de la thèse

C.1 Contexte de la thèse

Contexte général Alors que l'adoption de l'apprentissage automatique dans de nombreux contextes appliqués a connu une croissance rapide au cours de la dernière décennie, il reste des défis à relever pour l'utiliser dans certains contextes industriels. La principale raison est le conflit entre les procédures historiques établies et l'incertitude et le manque de transparence du processus de décision d'une chaine d'apprentissage automatique. Une autre raison est que les normes de données nécessaires pour alimenter un modèle d'apprentissage automatique traditionnel ne sont pas adaptées au type ou à la qualité des données disponibles. La plupart des bases de données industrielles n'ont pas été développées pour l'analyse statistique mais pour se conformer aux exigences réglementaires et effectuer des tâches administratives. En particulier, les variables non numériques ou symboliques sont courantes car il s'agit d'un moyen polyvalent d'enregistrer des événements d'intérêt. Des exemples de telles données sont les documents textuels, les séquences d'événements de journaux ou les séquences d'ADN. L'objectif principal de cette thèse est de s'attaquer à ces problèmes en proposant des approches qui peuvent être généralement appliquées à une séquence symbolique de données avec une sortie lisible par l'homme et entraînées à un coût de calcul raisonnable.

Maintenance prédictive pour le parc de trains français. Cette thèse est sponsorisée par la *Compagnie Nationale des Chemins de Fer Français(SNCF)*, l'entreprise ferroviaire publique qui exploite l'ensemble du trafic français en France. Chaque jour en France, 15000 de trains circulent. La seule agglomération parisienne compte $3,2$ millions de voyageurs par jour et 60000 d'arrêts dans les gares. SNCF doivent faire face à un contexte d'augmentation des transports en commun : au cours des dix dernières années, le nombre de voyages à Paris a augmenté

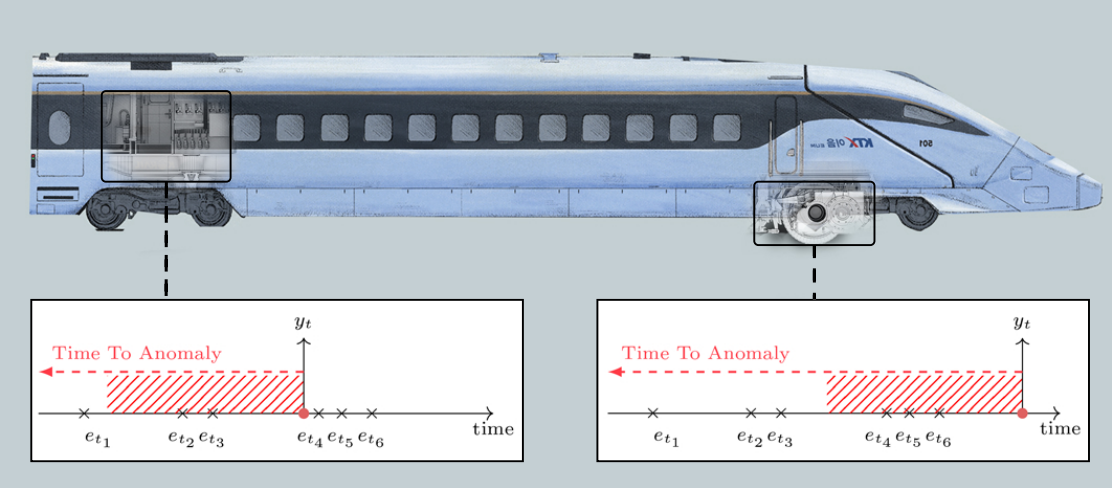


Figure C.1: Un véhicule ferroviaire est un système électromécanique complexe composé de plusieurs sous-systèmes. La figure montre les sous-systèmes du transformateur (à gauche) et du bloc moteur (à droite). Chacun d'eux est composé de nombreux éléments qui émettent des événements de journal horodatés ou des *codes d'erreur* (e_t) à différents moments t. Une panne ou une anomalie Y_t^S au temps t peut être liée à un sous-système spécifique S.

de 30%. Ce contexte exerce une pression croissante sur le réseau ferroviaire et nécessite une approche plus automatisée de la maintenance. Ces dernières années, SNCF a développé un système d'alerte basé sur des règles soigneusement construites par des experts. Bien que réussie, cette approche prend beaucoup de temps et ne permet pas la découverte automatique de nouvelles règles qui ne sont pas déjà connues. De plus, un ensemble de règles conçu par cette méthode est spécifique à une classe de véhicules et ne peut être appliqué à de nouveaux équipements.

C.2 Motivations

La tâche de *maintenance prédictive* vise à anticiper les défaillances critiques d'un système industriel afin de planifier des interventions précoces et efficaces. La méthode pour prévenir la défaillance critique d'un composant en cours d'exploitation était historiquement basée sur la *maintenance préventive*. Connaissant la durée de vie moyenne ou la loi de détérioration du composant, les réparations sont planifiées afin de réduire les risques de défaillances imprévues de l'équipement. C'est un progrès par rapport à la *maintenance réactive*, qui ne remplace et n'entretient les équipements qu'en cas de défaillance observée. La maintenance prédictive est un terme général désignant l'exécution de la maintenance des équipements en fonc-

tion des signes de détérioration observés ou enregistrés. Plus précisément, il s'agit d'une stratégie de maintenance qui surveille l'état de santé des machines en temps réel et prend une décision de maintenance optimale. Même si la maintenance prédictive entraîne une plus grande disponibilité et une réduction des coûts, elle nécessite beaucoup plus de temps, d'efforts et de ressources pour être exécutée. Un haut niveau de compétences est nécessaire pour collecter, modéliser et interpréter les données et réorganiser le processus de maintenance. En réalité, le gestionnaire industriel utilise CBM, RM et la maintenance prédictive.

Maintenance prédictive pour le matériel roulant. L'objectif principal de cette thèse est de construire une solution de maintenance prédictive de bout en bout pour le matériel roulant, de la collecte des données à la prédiction. Les trains sont des systèmes électromécaniques complexes qui utilisent de nombreux composants interconnectés pour offrir aux passagers des trajets courts et sécurisés et qui doivent être économes en énergie. En France, une bonne couverture du territoire implique une exposition à un environnement éventuellement difficile (concernant la topologie des voies, les conditions météorologiques) et est donc exposée à des taux de défaillance élevés. L'intérêt de la maintenance prédictive est particulièrement crucial dans ce contexte puisque l'impact d'une défaillance du matériel roulant a généralement des conséquences globales sur l'ensemble du système ferroviaire. Comme le train fonctionne sur un réseau hautement interconnecté, toute défaillance entraîne l'immobilisation complète du train et propage les retards à une grande partie du réseau de transport. À cet égard, le système ferroviaire constitue un cas particulièrement pertinent pour la valeur ajoutée d'un système de maintenance prédictive.

Dans le contexte de SNCF, l'un des défis était d'identifier un ensemble de caractéristiques pertinentes pouvant informer sur l'état de détérioration du train. La séquence d'un ensemble particulier d'événements, les séquences de *codes d'erreur*, a été identifiée comme étant particulièrement informative. Les codes d'erreur sont des chaînes de texte horodatées émises à intervalles réguliers ou irréguliers par le système spécifique d'un train. L'émission d'un type particulier de code correspond à une règle (parfois obscure) du fabricant. Par exemple, sur le système de la porte du train, une émission de code peut correspondre au franchissement d'un seuil de réponse de la tension du moteur CC de la porte. Notez qu'il y a un léger abus de langage dans l'utilisation du terme *code d'erreur* puisqu'un code d'erreur ne renseigne pas nécessairement sur un dysfonctionnement mais peut indiquer le fonctionnement nominal d'un système. L'un des principaux avantages de ce modèle est que les experts l'utilisent pour *a posteriori* le diagnostic d'une panne. Lorsqu'un train spécifique tombe en panne, il est envoyé à l'usine de maintenance pour être inspecté. Pour déterminer la cause de la panne, les journaux

sont extraits des systèmes et analysés par le responsable de la maintenance. Celui-ci recherche des *patterns* spécifiques et des récurrences connues de codes d'erreur dans ces codes afin de retrouver la cause profonde de la panne. Nous soulignons que cette procédure est largement utilisée en pratique pour la maintenance prédictive dans des contextes industriels au-delà du domaine ferroviaire tels que l'industrie automobile (Sung et al., 2020), les processus de fabrication (Gutschi et al., 2019) ou la détection d'anomalies sur divers systèmes informatiques (Wang et al., 2017a; Wang, Vo, and Ni, 2015; Zhang et al., 2016).

a société nationale des chemins de fer français a développé une plate-forme centralisée pour collecter et traiter les données en temps réel envoyées par le matériel roulant à partir de l'unité informatique embarquée. Ces données comprennent des séries chronologiques électriques, des journaux d'événements et l'état du système conçu par des experts. Pour chaque voiture de chaque train du parc, l'espace des caractéristiques est construit en collectant chaque journal d'événement horodaté associé à un sous-système spécifique.

Dans le cas du trafic ferroviaire, les conséquences d'une panne de locomotive ne se limitent pas à la machine affectée, mais se propagent à travers le réseau ferroviaire et peuvent également affecter les transports publics.

Apprentissage machine pour les données symboliques. La plupart de nos tâches quotidiennes, comme la parole, la lecture ou l'utilisation de la mémoire épisodique, reposent sur des données symboliques plutôt que numériques. Ce qui différencie fondamentalement les données symboliques des données numériques est la propriété *ordering*. Par exemple, il existe un moyen naturel de comparer deux mesures physiques d'un signal électrique mais aucun pour comparer deux symboles. Ce type de données est omniprésent dans un large éventail de domaines tels que la biologie avec la transcription de l'ADN et de l'ARN (Schölkopf, Tsuda, and Vert, 2004; Aubin-Frankowski and Vert, 2020), la chimie pour la prédiction et la classification des structures moléculaires (Elton et al., 2018), l'analyse de graphes (Mansha et al., 2016; Shang et al., 2017; Zheng et al., 2013), et en théorie musicale pour extraire les motifs qui ont la même fonction harmonique (Rompré, Biskri, and Meunier, 2017).

En général, les données symboliques ne conviennent pas à la plupart des algorithmes d'apprentissage automatique, car une hypothèse commune faite dans la théorie de l'apprentissage automatique est que le vecteur de caractéristiques d-dimensionnel est une variable aléatoire évaluée dans $\mathbb{R}^d$. Une première approche consiste à considérer les méthodes à noyau (Kung, 2014) qui étendent l'utilisation des techniques courantes d'apprentissage automatique aux données non numériques. Plus précisément, elle repose sur le choix d'une fonction noyau qui fait correspondre les données des symboles dans un espace structuré. Les

principaux inconvénients des méthodes à noyau sont la difficulté d'interpréter les résultats, ce qui est une exigence pour qu'une solution prédictive puisse être utilisée dans un contexte industriel. Une deuxième approche consiste à transformer le processus en un processus numérique en agrégeant (en comptant ou en considérant certaines statistiques) les événements d'une fenêtre temporelle choisie et a été largement utilisée pour la détection des anomalies : (He et al., 2016; Bogojeski et al., 2020; Aggarwal et al., 2018; Laredo et al., 2019). Bien que populaire (Basora, Olive, and Dubot, 2019), la classification basée uniquement sur cette construction est souvent incapable de capturer les modèles critiques d'événements qui peuvent être très pertinents dans la maintenance prédictive. Plus important encore, elle ne fournit pas directement de résultats explicables en termes d'ensembles d'événements ou de *patterns* de journaux.

En général, une série chronologique brute ne peut pas être considérée comme appropriée pour alimenter un algorithme d'apprentissage automatique pour plusieurs raisons.Dans la théorie de la décision statistique, un vecteur de caractéristiques est habituellement un vecteur de variables aléatoires à valeur réelle telles queX Rm. Dans l'apprentissage supervisé, il existe une variable de sortie, dont le domaine dépend de l'application (par exemple, un ensemble fini de valeursY Cin classification ou Rin régression) telle queXetY sont liées par une distribution inconnuejointePr(X,Y)qui est approximée par une fonctionfstelle quef(X)→Y. La fonction est choisie en fonction de l'hypothèse faite sur la distribution des données etf est ajustée afin d'optimiser une fonction de perte(Y,f(X))pour pénaliser les erreurs de prédiction.Le vecteur de caractéristiques est censé être de faible dimension afin d'éviter le phénomène de malédiction de la dimension qui affecte les performances du fait que les instances sont situées dans un espace de caractéristiques clairsemé [Hastie et al, 2009]. Hegger et al., 1998] discute de l'impact de la haute dimensionnalité pour construire un espace de caractéristiques significatif à partir de séries temporelles afin d'effectuer une analyse de séries temporelles : avec les séries temporelles, la densité des vecteurs est faible et diminue exponentiellement avec la dimension. Pour contrer cet effet, un nombre exponentiellement croissant d'instances dans l'ensemble de données est nécessaire. En outre, la position relative d'une variable aléatoire dans le vecteur de caractéristiques n'est pas prise en compte pour elle.

La société nationale des chemins de fer français a développé une plate-forme centralisée pour collecter et traiter les données en temps réel envoyées par le matériel roulant à partir de l'unité informatique embarquée. Ces données comprennent des séries chronologiques électriques, des journaux d'événements et l'état du système conçu par des experts. Pour chaque voiture de chaque train de la flotte, l'espace des caractéristiques est construit en collectant chaque journal d'événement horodaté associé à un sous-système spécifique.

Comme les réseaux ferroviaires sont de plus en plus fréquentés et développés, les exigences de disponibilité, d'amélioration de la qualité du service et de fiabilité de l'infrastructure sont devenues plus critiques (de Bruin et al., 2017). Avec la détérioration rapide due à l'utilisation intensive, les interventions de maintenance limitées en raison des réductions budgétaires et les demandes de service croissantes, le besoin de maintenance de l'infrastructure augmente continuellement (ERF,2013 ; Agence ferroviaire européenne, 2014). Par conséquent, les gestionnaires d'infrastructure doivent prendre des décisions de maintenance avec pour objectifs d'améliorer l'état des actifs, de dépenser un coût optimal et de maintenir le réseau disponible.

Pour chaque sous-système d'une voiture donnée, l'espace cible est constitué des rapports de maintenance, des pannes non planifiées et des rapports d'opérations de maintenance préventive.

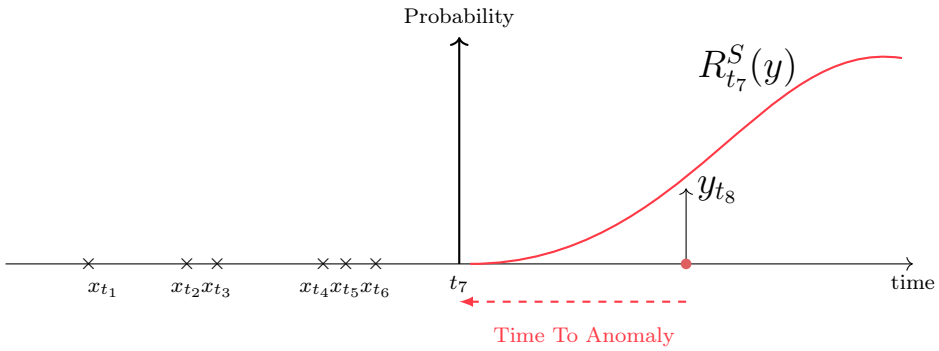


Figure C.2: Regression function $R^S_{t_7}$ at time t_7 (red line) given the events $(x_{t_1}, \ldots, x_{t_6})$ on the subsystem S. At t_7, the past events are used to produce the density probability function of a breakdown appearing in the future. This density is compared with the true occurrence of an anomaly y_{t_8}.

C.3 Contexte

Cette section présente formellement la maintenance prédictive comme une régression statistique basée sur des données symboliques. La tâche d'exploration de motifs est ensuite présentée et reformulée comme un problème d'inférence bayésienne. Enfin, la procédure d'*optimisation stochastique* est décrite en mettant l'accent sur les méthodes de réduction de la variance.

Le cadre bayésien qui sera utilisé tout au long de cette thèse comme modèle génératif pour la tâche d'exploration de motifs et comme cas d'utilisation pour le problème d'optimisation stochastique.

C.3.1 Séries temporelles symboliques pour la maintenance prédictive

Comme mentionné, les données symboliques jouent un rôle crucial dans la maintenance prédictive : leur polyvalence et leur utilisation historique par les mainteneurs. Formellement, les codes d'erreurs sont un dictionnaire ou un ensemble fini $E = (c_1, \ldots, c_d)$ de taille d. À un instant $t \in \mathbb{R}_+$, un événement peut être émis par le sous-système S_i. L'identifiant du sous-système définit le bloc de composant impliqué (tel que le bloc moteur) ainsi que l'identifiant du train et du véhicule. Enfin, l'espace des caractéristiques doit être enrichi par des informations qui sont corrélées avec le processus de dégradation sous-jacent (voir le chapitre 2). Dans notre application et en général, il sera généralement constitué d'un vecteur réel dans $\mathbb{R}^K$. Un exemple d'un tel contexte est, par exemple, le nombre de kilomètres depuis la dernière maintenance, les données météorologiques ou des informations contextuelles supplémentaires au moment de l'émission du code d'erreur. Nous

désignons $X_t^S = \mathcal{E} \times \mathbb{R}_K$ l'espace de description du sous-système S au temps t avec $\mathcal{E} = \mathcal{P}(E)$ étant l'ensemble de tous les sous-ensembles de E. Au temps $t \in \mathbb{R}_+$ on observe l'occurrence d'une panne $Y_t^S \in \{0, 1\}$ sur le sous-système S. L'objectif de tout algorithme de maintenance prédictive est de calculer la *fonction de régression* à chaque instant t définie par

$$R_t^S(y) = \mathbb{P}[Y_t^S = y | (X_{t_0}^S)_{t_0 \leq t}], \tag{C.1}$$

où $y \in \{0, 1\}$ dénote un ensemble de dysfonctionnements. La figure C.2 illustre la construction d'une telle fonction. Au temps $(t_1, \ldots, t_6)$, les codes d'erreur $(e_{t_1}, \ldots, e_{t_6})$ sont émis et enrichis pour produire $(x_{t_1}, \ldots, x_{t_6})$. À t_7, la fonction de régression estime la probabilité d'occurrence d'une panne sur le sous-système S pour chaque instant dans le futur. Un large éventail de techniques basées sur le modèle de processus stochastique (Guan, Tang, and Xu, 2016; Chen et al., 2016; Cha and Pulcini, 2016), les méthodes à noyau (Kung, 2014) ou les approches d'apprentissage profond (Guo et al., 2017; Liu et al., 2018; Karpat et al., 2020) peuvent être utilisées pour modéliser une telle fonction de régression. Comme mentionné, toutes ces méthodes souffrent d'une faible explicabilité et sont incompatibles avec les processus de maintenance établis qui sont basés sur le modèle des codes. L'objectif est donc de construire un modèle basé sur de petits ensembles de codes qui se produisent peu de temps et spécifiquement avant les défaillances, ce qui est une tâche difficile. Trouver ces combinaisons de codes est généralement intraitable en raison du nombre exponentiel de modèles possibles. Il est donc nécessaire de recourir à la classe des techniques de *pattern mining* (Agrawal, Imielinski, and Swami, 1993).

C.3.2 Contexte sur le pattern mining

Le domaine de l'exploration de données est né du besoin d'outils informatiques permettant d'extraire des informations utiles de grandes bases de données collectées par les administrations et les industries. Ces bases de données sont généralement de grands enregistrements de nombreuses variables ou *features* principalement construites pour des tâches administratives telles que la comptabilité et la conformité réglementaire.

Approches déterministes. Les travaux précurseurs de (Agrawal, Imielinski, and Swami, 1993) sur le Frequent Itemset Mining (FIM) pour l'analyse de paniers ont suscité l'intérêt car ils offrent une procédure traçable pour aborder un problème du monde réel avec une vaste application commerciale. Le problème posé était de trouver avec un niveau de précision donné, l'association ou les *patterns* de produits communs qui ont été achetés ensemble à partir d'une base de données

d'achats. Étant donné un nombre d d'articles possibles à acheter, et une base de données de reçus, la complexité associée à l'interrogation de la base de données pour trouver le nombre de fois où chaque motif de produits a été acheté ensemble, ou *support*, est en $\mathcal{O}(2^d)$. Le calcul de tels motifs est donc intraitable même pour un *dictionnaire* d'itemsets de taille modérée. La solution proposée consistait à exploiter la *antimonoticité* de l'ensemble de motifs $\mathcal{E}$: pour deux motifs $x, y \in \mathcal{E}$, si x dérive de y dans le sens que $x \subseteq y$ alors le support de y n'est pas plus grand que le support de x. En fixant un *seuil de support minimal* $\mu \in [0, 1]$, un algorithme peut extraire le support des ensembles d'éléments à la manière d'une recherche en largeur (Zuse, 1972; Moore, 1959) en générant de nouveaux motifs *candidats* à chaque étape et arrêter l'exploration de l'arbre dès qu'il rencontre un motif avec un support inférieur à μ. Cette procédure constitue l'algorithme APRIORI (Agrawal and Srikant, 1994) et a constitué une étape importante pour les tâches liées à l'exploration de données. Même si APRIORI est un algorithme efficace lorsque la taille moyenne des motifs présents dans la base de données n'est pas trop importante (Hegland, 2007), il présente plusieurs inconvénients. Tout d'abord, elle nécessite de multiples balayages de la base de données pour chaque motif évalué, et la nécessité de calculer un nouvel ensemble de motifs à tester pendant la procédure entraîne une complexité mémoire exponentielle de $\mathcal{O}(2^d)$. Des améliorations par rapport à l'algorithme APRIORI telles que ECLAT (Zaki, May-June/2000) propose un algorithme de recherche en profondeur avec un format de données vertical qui allège le besoin de requêtes multiples de la base de données. Une stratégie différente pour FIM a été adoptée par Han et al. (2004) appelée FP-TREE. Les auteurs utilisent une structure arborescente pour coder l'ensemble trié des transactions, ce qui permet de n'effectuer que deux balayages de la base de données. De manière cruciale, la structure arborescente évite de générer des itemsets inutiles, ce qui conduit à une procédure beaucoup plus efficace en termes de mémoire par rapport à APRIORI (Fournier-Viger et al., 2017). Le CP-TREE (Tanbeer et al., 2008) étend l'algorithme FP-TREE en ne nécessitant qu'un seul balayage de la base de données, ce qui réduit d'un facteur N les besoins en calcul. Nous soulignons que FIM est le point de départ de diverses techniques liées aux tâches d'exploration de données. Par exemple, Association Rule Mining (ARM) (Agrawal and Srikant, 1994; Zaki and Hsiao, 2005) considère le problème de la recherche de règles entre itemsets à un niveau de confiance donné. Pour deux motifs $x, y \in \mathcal{E}$, le but est de trouver des règles $x \to y$ telles que le support $s(x \vee y)$ et la mesure de confiance $c(x, y) = \frac{s(x \vee y)}{s(x)}$ ne sont pas supérieurs à deux seuils $\mu, \nu \in [0, 1]$. La mesure de confiance informe sur la co-occurrence de deux motifs en tenant compte de leur fréquence dans la base de données. Episode Rule Mining (Mannila, Toivonen, and Verkamo, 1997; Zimmermann, 2014) considère le problème de trouver les règles de la forme $x \to y$ qui apparaissent régulièrement

dans une fenêtre définie par l'utilisateur. Les applications sont nombreuses dans la détection d'anomalies et de fraudes (Qin and Hwang, 2004; Su, 2010; Wang et al., 2017b), l'analyse de capteurs (Li et al., 2017a), les données de trafic (Fournier-Viger et al., 2017) et dans le domaine médical (Patnaik, Sastry, and Unnikrishnan, 2008). L'objectif du problème de Periodic Pattern Mining est d'extraire les motifs qui se répètent au fil des transactions de la base de données (Venkatesh et al., 2016) et est couramment utilisé pour les applications biomédicales (Zhang et al., 2007) et l'analyse des séquences temporelles (Sirisha, Shashi, and Raju, 2014). Une approche originale a été adoptée par Vreeken, van Leeuwen, and Siebes (2010) en recherchant l'ensemble de motifs qui compresse le mieux la base de données sans perte. L'algorithme KRIMP résultant effectue d'abord un FIM avant d'utiliser le principe Minimum Description Length pour résumer la base de données. Enfin, les tâches de l'exploration progressive de motifs consistent à effectuer un FIM sur un sous-ensemble correctement dimensionné de la base de données afin d'approcher le support de manière uniforme à un niveau de confiance donné : (Riondato and Upfal, 2015). Nous mentionnons d'autres méthodes qui dérivent de FIM telles que l'exploration de sous-graphes (Santhi and Padmaja, 2015), l'exploration de motifs discriminatifs (Hämäläinen and Webb, 2019) et l'exploration de motifs séquentiels (Fournier-Viger et al., 2017).

Approches bayésiennes. Les méthodes mentionnées peuvent extraire avec succès le motif d'une grande base de données avec une utilisation efficace de la mémoire, mais ont toujours une complexité de calcul exponentielle en temps pour un seuil de soutien faible μ, car il a été démontré que le problème est NP-dur (Yang, 2004). De plus, ces modèles ne supposent aucune stochasticité sur le processus sous-jacent générant la base de données. En revanche, dans la grande majorité des cas, les transactions peuvent être considérées comme le résultat d'un processus génératif sous-jacent mais inconnu. Par conséquent, aucun intervalle de confiance probabiliste ne peut être dérivé pour évaluer la signification statistique des résultats.

Des modèles génératifs ont été proposés pour effectuer diverses FIM tâches afin de résoudre ces problèmes fondamentaux. Le modèle de distribution arborescent multivarié (Hegland, 2007) ajuste une distribution de probabilité sur les $\binom{d}{2}$ éléments par paire et une structure arborescente sur les attributs. Fowkes and Sutton (2016) utilisent un modèle de réseau bayésien pour modéliser la base de données des transactions. Comme l'inférence nécessite la résolution de l'intraitable problème de couverture de poids (Korte and Vygen, 2006), les auteurs ont utilisé une approximation gloutonne pour inférer les itemsets intéressants. Pavlov, Mannila, and Smyth (2003) comparent empiriquement plusieurs modèles génératifs tels que le modèle d'indépendance (Hegland, 2007), le modèle de distribution

arborescent multivarié (Chow and Liu, 1968) et le modèle de mélange dans le cadre équivalent de l'interrogation d'ensembles de données binaires clairsemés. Notamment, l'utilisation de ces approches probabilistes va au-delà de FIM et peut servir d'outil pour dériver des limites de convergence pour les algorithmes de type APRIORI (Hegland, 2007). Notez que ces approches sont étroitement liées au principe MDL (Vreeken, van Leeuwen, and Siebes, 2010) puisque l'entropie d'un modèle de probabilité définit la compression sans perte maximale réalisable par tout algorithme de compression (nous renvoyons le lecteur intéressé à (Friedman, Geiger, and Goldszmidt, 1997; Lam and Bacchus, 1994)).

Mode d'inférence. Contrairement à l'approche déterministe, les méthodes probabilistes reposent sur l'hypothèse que la base de données $\mathcal{D}$ est le résultat d'un processus stochastique. Cette hypothèse ouvre la possibilité d'appliquer des outils statistiques courants pour inférer l'ensemble des items fréquents. L'objectif commun de toutes ces méthodes est de trouver pour chaque motif x la distribution de probabilité du support $p(x|z, \mathcal{D})$. Étant donné le modèle génératif, trouver une formule fermée pour calculer $s(x)$ peut s'avérer difficile et implique souvent une énumération intraitable de tous les motifs possibles (Fowkes and Sutton, 2016). En considérant des modèles plus simples tels que *mixture models* (Hegland, 2007) résout ce problème et permet de contrôler la complexité de ce calcul par le choix du nombre de composantes dans la distribution du mélange. Sous cette représentation, la tâche d'extraction du motif le plus fréquent devient une tâche d'optimisation bayesienne. Les prochaines sections décrivent formellement le cadre technique de cette inférence et les stratégies permettant d'accélérer la procédure.

C.3.3 Contexte des statistiques bayésiennes

Dans cette section, nous présentons le cadre des statistiques bayésiennes et les notations de base. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, $(E, \|.\|)$ un espace vectoriel équipé de la distance d et de la norme induite $\|.\|$ et considérons une variable aléatoire $X\colon (\Omega, \mathcal{A}) \longrightarrow (E, \mathcal{B}(E))$. Dans le cadre bayésien, l'espace des paramètres $\mathcal{Z}$ est équipé d'une mesure Π sur $\mathcal{T}$ telle que $(\mathcal{Z}, \mathcal{T}, \Pi)$ est un espace de probabilité et X est distribué selon un modèle paramétrique $\mathcal{P}_z$ de la famille paramétrique de distribution $\mathcal{P} = \{P_z\colon z \in \mathcal{Z}\}$. Dans la plupart des cas, $\mathcal{Z}$ est un sous-ensemble d'un espace euclidien et les applications considèrent souvent le cas réel -dimensionnel $\mathcal{Z} \subset \mathbb{R}^d$. En outre, supposons que pour chaque z dans $\mathcal{Z}$, les mesures P_z et Π admettent une fonction de densité telle que

$$\begin{aligned} dP_z &= p(.|z)d\mu \\ d\Pi &= \pi d\nu, \end{aligned} \tag{C.2}$$

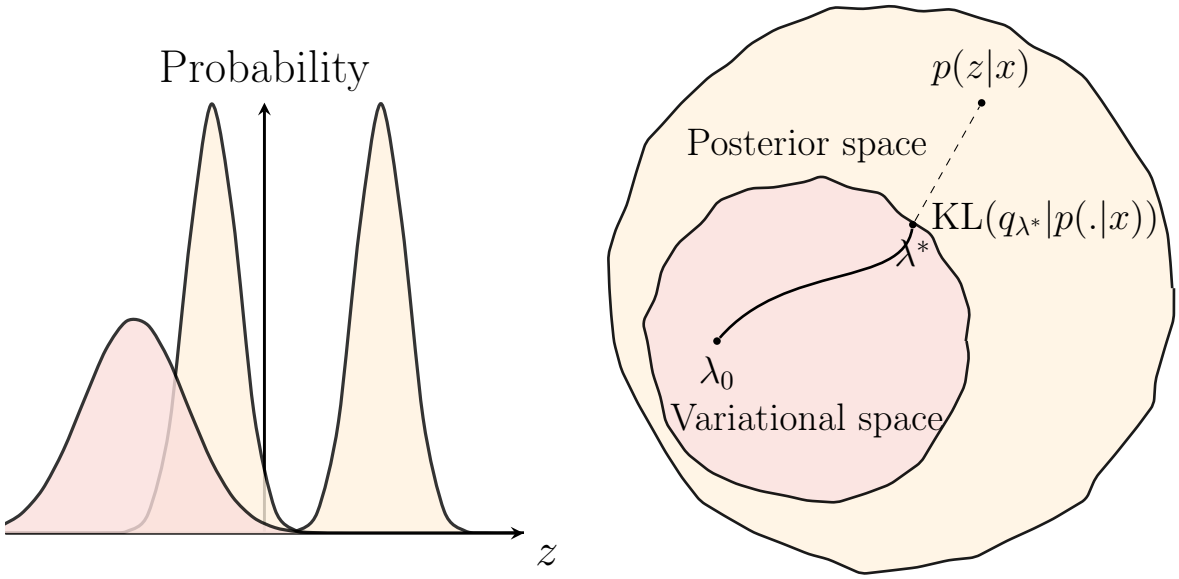


Figure C.3: **Variational inference**. Left: The variational distribution q_λ (orange) parametrized by λ and the true posterior distribution $p(z|x)$ (green). Right: The Variational Inference procedure consists of finding the optimal λ^*, starting from λ_0 to minimize the Kullback–Leibler divergence between the true posterior and the variational distribution (represented as dashed line).

où μ, ν sont des mesures σ-fini sur respectivement $\mathcal{B}(E)$ et $\mathcal{T}$. Alors, la fonction de vraisemblance $z \mapsto p(z|x)$ telle que $p(z|x) = p(x|z)\pi(z)$ est une densité par rapport à la mesure produit $\mu \otimes \nu$. La différence avec le point de vue *fréquentiste* est que le paramètre z est lui-même une variable aléatoire distribuée selon la distribution *prioritaire* π et, conditionnellement aux données x, a la distribution suivante

$$p(z|x) = \frac{p(x|z)\pi(z)}{\int p(x|z)\pi(z)d\nu(z)}. \tag{C.3}$$

Le cadre de l'inférence bayésienne dépend donc de la capacité à simuler z à partir de l'équation C.3. Le calcul de $p(z|x)$ nécessite l'évaluation de la distribution prédictive antérieure et donc d'intégrer sur toutes les variables latentes, ce qui conduit à un calcul intraitable (sauf dans le cas du conjugué antérieur) même pour les modèles simples (Gelman et al., 2013). Une approche courante consiste à utiliser des méthodes telles que Gibbs Sampling, Monte Carlo Markov Chain ou Hamilton Monte Carlo (Betancourt, 2018; Homan and Gelman, 2014; Brooks et al., 2011) qui s'appuient uniquement sur la distribution postérieure non normalisée (nous libérant du besoin de calculer $p(y)$) et sur la capacité à échantillonner à partir de la postérieure. Ces méthodes sont cohérentes mais associées à un calcul lourd,

une sensibilité élevée aux hyperparamètres et une lenteur potentielle à converger vers la vraie distribution cible.

Inférence variationnelle

La distribution postérieure dans l'équation C.3 peut être calculée exactement sous certaines conditions sur la distribution antérieure lorsque la forme fermée est disponible (Gelman et al., 2013). Pour la plupart des applications, cette condition n'est pas remplie et il faut recourir à une procédure asymptotiquement exacte ou s'appuyer sur une approximation. Une approche d'approximation qui est devenue le cadre principal du calcul bayésien approximatif est Variational Inference (VI). Elle repose sur la construction d'une approximation de la distribution postérieure paramétrée par une *famille variationnelle* distribution $Q = \{\boldsymbol{\lambda} \colon \boldsymbol{\lambda} \in \Lambda\}$.

Dans cette méthode, une métrique est choisie de sorte que la distance entre la vraie distribution cible p et la distribution variationnelle q soit minimisée. Un choix courant est la divergence Kullback–Leibler (KL). En désignant x les données, z l'espace des variables latentes et $p(z|x)$ la vraisemblance, et $q_{\boldsymbol{\lambda}}$ la distribution variationnelle paramétrée par $\boldsymbol{\lambda}$, l'inférence variationnelle consiste en un problème de minimisation : (Saul, Jaakkola, and Jordan, 1996)

$$q_{\lambda^*} = \underset{q_{\boldsymbol{\lambda}} \in \mathcal{Q}}{\operatorname{argmin}} \operatorname{KL}\left(q_{\boldsymbol{\lambda}}(z) \| p(z|x)\right), \tag{C.4}$$

avec $\operatorname{KL}\left(q_{\boldsymbol{\lambda}}(z) \| p(z|x)\right) = E_q[\log q_{\boldsymbol{\lambda}}(x) - \log p(z|x)]$ la divergence Kullback–Leibler. Même si KL reste la métrique la plus utilisée, d'autres mesures sur l'espace de distribution ont été étudiées (Ambrogioni et al., 2018). La raison de la popularité de ces techniques est le fait que la divergence KL peut être liée à la Evidence Lower Bound (ELBO) qui ne dépend pas de la distribution postérieure (Saul, Jaakkola, and Jordan, 1996).

$$\log p(y) = \operatorname{ELBO}(\boldsymbol{\lambda}) + \operatorname{KL}\left(q_{\boldsymbol{\lambda}}(z) \| p(z|x)\right), \tag{C.5}$$

où le ELBO est défini comme suit

$$\operatorname{ELBO}(\boldsymbol{\lambda}) = \mathbb{E}_{z \sim q_{\boldsymbol{\lambda}}}\left[\log p(z, x) - \log q_{\boldsymbol{\lambda}}(z)\right]. \tag{C.6}$$

Puisque la vraisemblance marginale $p(y)$ ne dépend pas des paramètres z, il s'ensuit que la maximisation de la ELBO par rapport à $q_{\boldsymbol{\lambda}}$ conduit à trouver la meilleure approximation de $p(z|x)$ pour la divergence Kullback–Leibler (KL). Intuitivement, cette procédure minimise la perte d'information consécutive au remplacement de la vraisemblance par $q_{\boldsymbol{\lambda}}$ mais d'autres distances peuvent être utilisées (Ambrogioni et al., 2018).

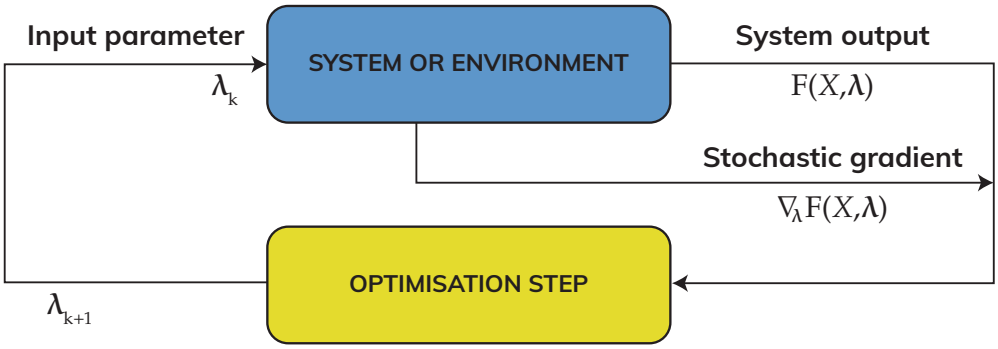


Figure C.4: Un processus d'optimisation stochastique typique composé de deux étapes : la simulation (jaune) et l'optimisation (vert). La première étape produit La phase de simulation produit une simulation du système stochastique ou de l'interaction avec l'environnement, ainsi que des estimateurs sans biais du gradient (adapté de (Mohamed et al., 2020))

En pratique, la classe de distribution $\mathcal{Q}$ est choisie dans une famille de distribution qui peut être facilement échantillonnée. Un choix courant consiste à choisir dans la famille de distribution normale $\mathcal{Q} = \mathcal{N}(\mu, \Sigma) \,|(\mu, \sigma)\mathbb{R}^K \times M_{K\times K}\}$ avec $M_{K\times K}$ l'espace des matrices symétriques à définition positive sur $\mathbb{R}^{K\times K}$. Dans ce cas, l'exécution de VI consiste à trouver l'ensemble optimal de paramètres (μ^*, Σ^*) tel que l'équation C.6 soit minimisée.

Là encore, il n'existe généralement pas de formule fermée pour calculer le ELBO ou son gradient et il faut s'en remettre à une méthode d'optimisation stochastique (Bottou, Curtis, and Nocedal, 2018) pour effectuer cette tâche. Avec cette méthode, la minimisation est effectuée en réalisant une Stochastic Gradient Descent (SGD) procédure sur la ELBO fonction objectif.

C.3.4 Optimisation tochastique

L'un des problèmes d'optimisation les plus importants de la statistique moderne consiste à trouver la racine d'une *fonction objectif* qui est une espérance d'une variable aléatoire (Bottou, Curtis, and Nocedal, 2018). Ce problème a des applications vastes et connues en apprentissage automatique (Bottou, Curtis, and Nocedal, 2018; Sutton and Barto, 2018; Gelman et al., 2013; Simsekli et al., 2019) mais aussi en finance pour l'analyse de sensibilité (Pagès, 2018; Glasserman, 2013), la gestion des réseaux de transport CITE et la chaîne logistique. Étant donné une μ-distributed variable aléatoire $X : \Omega \longrightarrow E$ sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$,

le problème général d'optimisation stochastique se lit comme suit

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^K}{\operatorname{argmin}} f(\boldsymbol{\lambda}) &= \mathbb{E}\left[F(X,\boldsymbol{\lambda})\right] \\ &= \int_E F(x,\boldsymbol{\lambda})\mu(dx), \end{aligned} \tag{C.7}$$

où $F\colon E\times\mathbb{R}^K \longrightarrow \mathbb{R}$ est une fonction réelle dans le $L_1\left(\Omega,\mathcal{A},\mathbb{P}\right)$. Sous la condition de régularité que f est différentiable continue (ou a leat qu'un *sous-gradient* peut être calculé), ce problème peut être résolu en trouvant les points où le gradient $\boldsymbol{g} = \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda})$ disparaît puisque $\boldsymbol{\lambda}^* \in \operatorname{argmin}\{\boldsymbol{g} = 0\}$. ℕ- Nouvelle ligne Ce problème peut être interprété comme l'optimisation d'une *cost* ou *loss function* F par rapport à $\boldsymbol{\lambda}$ avec une interférence bruyante distribuée selon μ. Dans les applications d'apprentissage automatique (comme l'entraînement d'un réseau neuronal), F représente la *loss* attendue d'un modèle paramétré par $\boldsymbol{\lambda}$ pour un ensemble d'entraînement distribué selon μ. Dans ce cas, il a été montré que trouver l'ensemble optimal de paramètres $\boldsymbol{\lambda}^*$ est NP-hard même pour un modèle de classification binaire simple (Feldman et al., 2012). Plus généralement, la principale difficulté pour trouver une solution à C.7 est qu'elle implique le calcul d'une espérance potentiellement de haute dimension, ce qui est prohibitif. Même lorsque la distribution est connue, il n'existe généralement pas de forme fermée pour calculer le gradient. De nos jours, les méthodes de quadrature (Leader, 2004) pour calculer l'intégrale à une précision donnée ne sont réalisables que pour une dimension allant jusqu'à dix ou vingt, ce qui les rend inutilisables pour la plupart des applications modernes. De plus, dans la plupart des cadres tels que l'apprentissage statistique, la distribution μ est inconnue et seuls des échantillons de la distribution μ sont disponibles.

Même lorsque la distribution est connue, équation C.7, il n'existe généralement pas de forme fermée pour calculer le gradient car cela nécessite une dérivation sous l'attente d'un espace de paramètres potentiellement de haute dimension.

Échantillonnage alternatif pour l'estimateur de la moyenne.

L'échantillonnage alternatif a été introduit pour accélérer les procédures d'optimisation stochastique. La recherche d'une approximation pour le problème d'optimisation dans C.7 dépend essentiellement de la capacité à calculer efficacement une approximation de l'espérance dépendant de l'échantillon.

Monte Carlo. La procédure numérique la plus couramment utilisée pour approximer l'espérance dans C.7 est basée sur la Law of Large Number. Elle repose sur le remplacement de l'espérance par un estimateur de moyenne empirique. Soit $(X_1,\ldots,X_n)$ une séquence *i.i.d.* de variables aléatoires distribuées μ_X, F toute

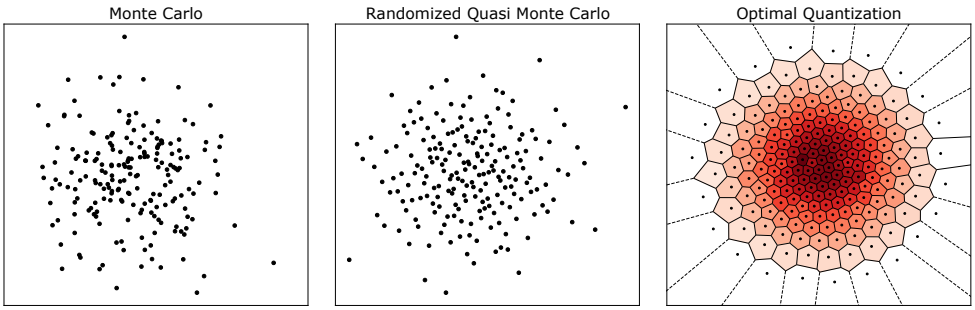


Figure C.5: Monte Carlo (gauche), Monte Carlo aléatoire (centre) et Quantification optimale avec les cellules de Voronoï associées (droite), pour une taille d'échantillonnage $N = 200$ de la distribution normale bivariée $\mathcal{N}(0, I_2)$. (Dib, 2020)

fonction mesurable à valeur réelle, et considérons l'estimateur *Monte-Carlo* suivant

$$I_n^{MC} = \frac{1}{n} \sum_{i=1}^{n} F(X_i). \tag{C.8}$$

Par la loi forte des grands nombres, I_n^{MC} converge vers $\mathbb{E}\big[F(X)\big]$ μ-almost surely et, à condition que $F(X) \in L_2(\Omega, \mathcal{A}, \mathbb{P})$, à un taux de $\mathcal{O}(n^{-\frac{1}{2}})$ avec une erreur quadratique

$$\| I_n^{MC} - \mathbb{E}\big[F(X)\big] \|_{L_2(\Omega,\mathcal{A},\mathbb{P})} = \frac{\mathbb{V}F(X)}{\sqrt{n}}. \tag{C.9}$$

La méthode de Monte-Carlo repose uniquement sur la possibilité de tirer de la distribution μ à un coût raisonnable. En outre, le Central Limit Theorem peut être utilisé pour produire un intervalle de confiance asymptotique.

Quasi Monte-Carlo. Des méthodes ont été conçues pour améliorer le taux de convergence, principalement en considérant des méthodes d'échantillonnage alternatives pour générer le $(X_1, \ldots, X_n)$. Les plus utilisées sont les méthodes Quasi Monte Carlo (Dick, Kuo, and Sloan, 2013). Ces méthodes sont basées sur la génération de séquences de *nombre pseudo-aléatoire* qui imitent les propriétés statistiques d'une séquence d'échantillons *i.i.d.* cible. Plus précisément, laissons X être une variable aléatoire qui admet une densité ψ par rapport à la mesure de Lebesgue de dimension d et considérons une variable aléatoire uniformément distribuée $U \sim \mathcal{U}([0,1]^d)$. Alors, la variable aléatoire $\psi^{-1}(U)$ est distribuée selon X et pour toute fonction mesurable H on a que $\mathbb{E}\left[H(X)\right] = \mathbb{E}\left[H \circ \psi^{-1}(U)\right]$. Une séquence *low-discrepancy* $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, avec $(\boldsymbol{u}_i)_{i=1}^n$ évaluée dans l'hypercube

d-dimensionnel $[0,1]^d$, est produit et évalué par la fonction de distribution de probabilité à densité inverse ψ^{-1} (Pagès, 2018). Puisque $\boldsymbol{u}$ converge faiblement vers la mesure de Lebesgue sur $[0,1]^d$, ce qui suit s'applique à l'estimateur QMC.

$$\begin{aligned} I_n^{QMC} &= \frac{1}{n}\sum_{i=1}^{n} F \circ \psi^{-1}(\boldsymbol{u}_i) \\ &\xrightarrow[n\to\infty]{} \mathbb{E}\big[F(X)\big]. \end{aligned} \tag{C.10}$$

Intuitivement, si $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ est similaire à la réalisation d'une séquence *i.i.d.* d'une variable aléatoire uniformément distribuée, la séquence $(\psi^{-1}(\boldsymbol{u}_1), \ldots, \psi^{-1}(\boldsymbol{u}_n))$ sera similaire à l'ensemble *i.i.d.* d'échantillons cible $(X_1, \ldots, X_n)$. La qualité d'une telle approximation est contrôlée par la mesure de la *disparité étoile* qui est définie comme la distance ℓ_∞ entre la distribution cumulative de la mesure empirique et la mesure de Lebesgue

$$D_n^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \sup_{\boldsymbol{b}\in[0,1]^d} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{u}\in[0,\boldsymbol{b}]\}} - \lambda_d([0,\boldsymbol{b}]) \right|. \tag{C.11}$$

Pour la séquence $\boldsymbol{u}$ dont, l'inégalité de Hlawka-Koksma (Koksma, 1942; Hlawka, 1961) stipule que l'erreur d'approximation de C.10 est limitée supérieurement par sa mesure de divergence pour h avec une variation finie. Puisqu'il existe plusieurs séquences $\boldsymbol{u}$ qui présentent une mesure de divergence telle que

$$D_n^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \leq c_d \frac{(\log n)^{d-1}}{n}, \tag{C.12}$$

l'estimateur QMC I_n^{QMC} peut donc atteindre un bien meilleur taux de convergence que l'estimateur MC I_n^{MC} de l'équation C.9. Il existe plusieurs méthodes pour calculer de telles séquences à faible écart telles que les séquences de Halton, Faure ou Sobol. Nous mentionnons également qu'il existe une version stochastique de la méthode QMC, la Randomized Quasi Monte Carlo (RQMC), qui est obtenue en introduisant soigneusement le caractère aléatoire dans la séquence $\boldsymbol{u}$ (Owen, 2008; Gerber, 2015). L'estimateur RQMC est obtenu comme précédemment par moyenne de la séquence produite. Contrairement à l'estimateur I_n^{QMC}, l'estimateur produit est sans biais et il a été récemment démontré qu'il atteint un taux d'intégration $\mathcal{O}(n^{-1})$ sous l'hypothèse d'intégrabilité carrée (Gerber, 2015).

Descente de gradient stochastique

La méthode SGD introduite par (Robbins and Monro, 1951) a été spécifiquement conçue comme une procédure stochastique de recherche zéro de premier ordre pour

une fonction objectif bruyante. Cette classe d'algorithmes et ses variantes (Polyak and Juditsky, 1992; Kingma and Ba, 2015; Duchi, Hazan, and Singer, 2011a; McMahan and Streeter, 2010) ont rapidement attiré l'attention en raison de leur simplicité et de leur large éventail d'applications. Dans les problèmes modernes, elle se rapporte à de nombreuses applications en statistiques et en apprentissage automatique ("Stochastic Approximation Approach to Stochastic Programming"; Bottou and Le Cun, 2005). La méthode originale de descente de gradient (Cauchy, 1847; Hadamard, 1908; Rumelhart, Hinton, and Williams, 1985) utilise une estimation du gradient pour mettre à jour de façon récursive λ au temps t comme suit

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \alpha_t \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_t). \tag{C.13}$$

Dans le cadre décrit dans C.7, nous n'avons pas accès à l'espérance totale $f(\boldsymbol{\lambda})$ mais seulement à un estimateur bruité. L'essentiel de la méthode *stochastic gradient descent* consiste à remplacer le gradient réel par son estimateur, ce qui aboutit à

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \alpha_t \boldsymbol{g}(\boldsymbol{\lambda}_t). \tag{C.14}$$

Le choix du *taux d'apprentissage* α_t est crucial car il contrôle la taille des mises à jour. Un ensemble de conditions suffisantes connues sous le nom de *conditions de Robbins-Monro* assure que la procédure C.14 converge si le programme de mise à jour décroissante est tel que $\sum_{t=1}^{\infty} \alpha_t = \infty$ et $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$. Le choix du taux d'apprentissage est en soi un défi et influence grandement le taux de convergence : (Bottou, Curtis, and Nocedal, 2018). Un choix simple consiste à prendre $\alpha_t = ct^a$ pour une puissance réelle a et c une constante réelle. Les méthodes modernes utilisent *taux d'apprentissage adaptatif* pour régler le taux d'apprentissage, comme AdaDelta (Zeiler, 2012), AdaGrad (Duchi, Hazan, and Singer, 2011b) ou Adam (Kingma and Ba, 2015). Une garantie théorique sur le taux de convergence peut être obtenue en donnant une certaine hypothèse de régularité sur f. Par exemple, en supposant la régularité et la forte convexité, Bottou, Curtis, and Nocedal (2018) montrent que l'erreur $f(\boldsymbol{\lambda}_t) - f(\boldsymbol{\lambda}^*) = \epsilon$ est dans $\mathcal{O}(t^{-1})$.

Variance du gradient. Si deux estimateurs de gradient sont disponibles au même coût de calcul, celui dont la variance est la plus faible doit généralement être préféré car la convergence des méthodes d'optimisation stochastique dépend crucialement de la variance. La plupart de ces procédures d'optimisation reposent sur une optimisation par descente de gradient sur les paramètres associés à la famille variationnelle et dépendent ensuite fortement de la norme $\ell_2(\mathbb{R}^K)$ (avec K le nombre de paramètres variationnels) du gradient attendu (Bottou, Curtis, and

Nocedal, 2018; Domke, 2019). Une faible variance des estimateurs de gradient permet de prendre de plus grands pas dans l'espace des paramètres et d'obtenir une convergence plus rapide si le biais induit peut être contrôlé de manière satisfaisante. Plusieurs méthodes ont été utilisées pour réduire la variance du gradient, comme le filtrage (Miller et al., 2017; Roeder, Wu, and Duvenaud, 2017), la variante de contrôle (Geffner and Domke, 2018) ou l'échantillonnage alternatif (Tran, Nott, and Kohn, 2017; Ruiz, Titsias, and Blei, 2016; Buchholz, Wenzel, and Mandt, 2018). Ces méthodes souffrent généralement de plusieurs inconvénients. Premièrement, elles nécessitent généralement des hypothèses restrictives sur la distribution variationnelle. Par exemple, QMCVI n'est valable que pour une distribution avec une fonction de densité inversible. Ensuite, la plupart du temps, la garantie théorique sur la bonté de la solution n'est pas correctement établie. Enfin, elle implique souvent un cadre de calcul complexe et peut être difficile à mettre en œuvre.

C.4 Contributions

L'apprentissage à base de motifs appliqué à la maintenance prédictive. Nous proposons une vue d'ensemble du domaine de la maintenance prédictive en mettant l'accent sur les avancées récentes de la maintenance prédictive dans le contexte de l'industrie ferroviaire. Ce cas d'utilisation est particulièrement difficile ; le système industriel du chemin de fer s'étend sur un vaste territoire avec des environnements variés et implique des systèmes complexes hétérogènes et interconnectés. La deuxième contribution consiste à concevoir un pipeline de prédiction industriel pour aborder le problème de la maintenance prédictive dans un contexte industriel. Pour surmonter la complexité informatique qui découle d'un nombre élevé d'hyperparamètres possibles, nous concevons une méthode basée sur un test à deux échantillons pour élaguer l'arbre des opérations à effectuer. Divers algorithmes et ensembles d'hyperparamètres sont testés et comparés sur les deux classes de la flotte de trains français sur une période de deux ans.

Modèle génératif bayésien pour l'exploration de motifs . Nous développons des méthodes utilisant un modèle génératif bayésien pour l'exploration de motifs et montrons leur supériorité sur les méthodes déterministes traditionnelles pour diverses tâches. Tout d'abord, nous montrons que l'ensemble des items fréquents peut être extrait efficacement à l'aide de méthodes d'approximation stochastique. Nous proposons une approche bayésienne avec un schéma d'inférence variationnelle pour obtenir l'espace des items fréquents avec une grande précision.

Ensuite, nous utilisons un Bayesian Mixture Model pour déduire avec un faible coût de calcul les itemsets discriminants (Hämäläinen and Webb, 2019) avec une preuve empirique de l'utilisation générale de ces motifs discriminants en les considérant comme des caractéristiques pour la tâche de classification. Il en résulte une méthode capable d'extraire un ensemble d'attributs interprétables et d'améliorer considérablement tout classificateur. En outre, le modèle génératif bayésien permet de calculer la distribution postérieure et d'estimer les intervalles de confiance. Enfin, des connaissances expertes supplémentaires peuvent être naturellement introduites dans le modèle *via* le choix des antériorités (Gelman et al., 2013). Cette méthode est appliquée à la tâche de maintenance prédictive et améliore significativement le score de classification de manière interprétable.

Une partie de ce travail correspond à l'article (Dib et al., 2021) publié dans *29th IEEE European Signal Processing Conference (EUSIPCO) proceedings*.

Complexité locale de Rademacher pour l'exploration de motifs peu fréquents. La tâche d'échantillonnage progressif consiste à calculer la taille du sous-ensemble de la base de données n nécessaire pour obtenir une estimation de toute fréquence à la précision $\varepsilon \in [0, 1]$ avec une probabilité d'au moins $1 - \delta$. Il s'agit donc de borner un processus empirique généré par une distribution inconnue indexée sur un espace fonctionnel fini (Boucheron, Lugosi, and Massart, 2013).

Les méthodes existantes utilisent les moyennes de Rademacher (globales) pour extraire les *fréquents* ou les *top-k* itemsets, ce qui est approprié, car nous n'avons pas besoin de limites précises sur les itemsets à basse fréquence. Notamment, Riondato and Upfal, 2015 utilise un argument de comptage analytique pour obtenir une limite libre sur la moyenne empirique globale de Rademacher. De la même manière, Pellegrina et al., 2020 a suivi cette voie en utilisant une stratégie d'approximation de Monte-Carlo pour obtenir une limite plus nette au prix d'un calcul supplémentaire.

Ce travail marque la première utilisation de la complexité de Rademacher localisée au problème de l'exploration de motifs à basse fréquence. Nous montrons que les moyennes de Rademacher localisées sont suffisantes pour obtenir des estimations d'intervalles de confiance relatifs sur les fréquences des motifs, ainsi que d'autres mesures d'intérêt, telles que le *lift*, le *confidence*, ou le *odds ratio*, alors que les techniques précédentes n'y parviennent pas pour les motifs à basse fréquence.

Nos méthodes s'appuient sur des outils standard dans le domaine de l'exploration de motifs, tels que les familles de motifs fermées, l'antimonotonicité et les moyennes de Rademacher de Monte-Carlo, ainsi que sur de nouvelles techniques que nous introduisons pour relever les défis informatiques spécifiques au problème découlant de l'évaluation de la moyenne de Rademacher localisée. Les performances de notre approche sont démontrées empiriquement sur des ensembles de données du monde réel, dans lesquels nous présentons des taux de convergence rapides pour la sous-classe de motifs considérée, ce qui contraste fortement avec les travaux existants.

Ce travail correspond au préprint (Cousins* and Dib*, 2021)[1] soumis à la *IEEE International Conference on Data Mining (ICDM 2021)*.

Échantillonnage alternatif pour l'optimisation stochastique. Nous développons une nouvelle approche pour la technique d'optimisation stochastique basée sur Optimal Quantizer (OQ) (Graf and Luschgy, 2000; Pagès, 2018). Nous montrons que l'utilisation de OQ produit une estimation optimale du gradient sans gradient au prix de l'introduction d'un biais asymptotiquement décroissant avec une garantie théorique. Nous appliquons la méthode au cadre de l'apprentissage

[1] contributions égales.

bayésien pour la maximisation de Evidence Lower Bound (ELBO) et montrons que l'utilisation du cadre d'inférence variationnelle quantifiée conduit à une convergence rapide à la fois pour la fonction de score et l'estimateur de gradient reparamétré à un coût de calcul comparable à celui de la méthode traditionnelle Monte Carlo Variational Inference. Par la suite, nous proposons une méthode de type extrapolation de Richardson (Richardson and Glazebrook, 1911; Pagès, 2007) pour améliorer la borne asymptotique et réduire le biais produit. Deux nouveaux algorithmes, QVI et RQVI, sont évalués sur plusieurs expériences à grande échelle et présentent des performances supérieures par rapport aux méthodes de pointe (Miller et al., 2017; Buchholz, Wenzel, and Mandt, 2018).

Une partie de ce travail correspond à l'article (Dib, 2020) publié dans *Advances in Neural Information Processing Systems 33 Proceedings (NeurIPS 2020).*

C.5 Outline de la thèse

- Part II: Anomaly detection for rolling stock maintenance.
  - Chapitre 2 : Examen systématique de la maintenance prédictive.
  - Chapitre 3 : Extraction de motifs pour la détection d'anomalies dans le cadre de la maintenance du matériel roulant. Ce chapitre décrit l'approche adoptée pour aborder la question complexe de la maintenance prédictive sur le parc français de trains à grande vitesse.
- Part III: Pattern Mining.
  - Chapitre 4 : Vue probabiliste pour le problème d'extraction de motifs. Une approche bayésienne du problème d'extraction des itemset célèbres est décrite avec diverses expériences.
  - Chapitre 5 : Complexité localisée pour l'échantillonnage progressif. Ce chapitre décrit l'utilisation des moyennes de Rademacher localisées pour aborder le problème de l'extraction progressive. Nous montrons comment cette méthode peut conduire à une extraction de motifs plus rapide avec une garantie théorique.
- Part IV: Optimal Quantization for stochastic optimization.
  - Chapitre 6 : Contexte de la quantification optimale. Nous donnons un contexte théorique sur la Tesselation de Voronoï et proposons d'utiliser cet échantillonnage alternatif pour l'optimisation stochastique. Des résultats théoriques sur la qualité de l'approximation sont développés.

– Chapitre 7 : Inférence variationnelle quantifiée. Nous introduisons un nouvel algorithme pour la maximisation d'ELBO. Nous montrons que grâce au gradient sans variance, cette méthode surpasse l'état de l'art sur diverses expériences du monde réel, notamment le cas du problème d'extraction de motifs bayésiens.

Bibliography

[1] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. 2015.

[2] Karan Aggarwal et al. "Two Birds with One Network: Unifying Failure Event Prediction and Time-to-Failure Modeling". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1308–1317.

[3] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. "Mining Association Rules between Sets of Items in Large Databases". In: *In: Proceedings of the 1993 Acm Sigmod International Conference on Management of Data, Washington Dc (Usa*. 1993, pp. 207–216.

[4] Rakesh Agrawal and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules". In: *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Vol. 1215. Citeseer, 1994, pp. 487–499.

[5] Abdullah M. Al-Ghamd and David Mba. "A Comparative Experimental Study on the Use of Acoustic Emission and Vibration Analysis for Bearing Defect Identification and Estimation of Defect Size". In: *Mechanical Systems and Signal Processing* 20.7 (Oct. 2006), pp. 1537–1571.

[6] Zaharah Allah Bukhsh et al. "Predictive Maintenance Using Tree-Based Classification Techniques: A Case of Railway Switches". In: *Transportation Research Part C: Emerging Technologies* 101 (Apr. 2019), pp. 35–54.

[7] Luca Ambrogioni et al. "Wasserstein Variational Inference". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 2473–2482.

[8] Ido Amihai et al. "An Industrial Case Study Using Vibration Data and Machine Learning to Predict Asset Health". In: *2018 IEEE 20th Conference on Business Informatics (CBI)*. Vienna: IEEE, July 2018, pp. 178–185.

[9] Nagdev Amruthnath and Tarun Gupta. *A Research Study on Unsupervised Machine Learning Algorithms for Fault Detection in Predictive Maintenance*. Apr. 2018.

[10] Pierre-Cyril Aubin-Frankowski and Jean-Philippe Vert. "Gene Regulation Inference from Single-Cell RNA-Seq Data with Linear Differential Equations and Velocity Inference". In: *Bioinformatics* 36.18 (2020), pp. 4774–4780.

[11] C. Badulescu et al. "Applying the Grid Method and Infrared Thermography to Investigate Plastic Deformation in Aluminium Multicrystal". In: *Mechanics of Materials* 43.1 (Jan. 2011), pp. 36–53.

[12] S. Bagavathiappan et al. "Infrared Thermography for Condition Monitoring – A Review". In: *Infrared Physics & Technology* 60 (Sept. 2013), pp. 35–55.

[13] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. "Local Rademacher Complexities". In: *The Annals of Statistics* 33.4 (Aug. 2005), pp. 1497–1537.

[14] Luis Basora, Xavier Olive, and Thomas Dubot. "Recent Advances in Anomaly Detection Methods Applied to Aviation". In: *Aerospace* 6.11 (2019), p. 117.

[15] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep Learning*. Vol. 1. MIT press Massachusetts, USA: 2017.

[16] James Bergstra and Yoshua Bengio. "Random Search for Hyper-Parameter Optimization." In: *Journal of machine learning research* 13.2 (2012).

[17] James Bergstra et al. "Algorithms for Hyper-Parameter Optimization". In: *25th Annual Conference on Neural Information Processing Systems (NIPS 2011)*. Vol. 24. Neural Information Processing Systems Foundation, 2011.

[18] Michael Betancourt. "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]* (July 2018).

[19] M Bevilacqua and M Braglia. "The Analytic Hierarchy Process Applied to Maintenance Strategy Selection". In: *Reliability Engineering & System Safety* 70.1 (Oct. 2000), pp. 71–83.

[20] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877.

[21] Moritz Blume. "Expectation Maximization: A Gentle Introduction". In: *Technical University of Munich Institute for Computer Science* (2002).

[22] Mihail Bogojeski et al. "Forecasting Industrial Aging Processes with Machine Learning Methods". In: *Computers & Chemical Engineering* 144 (2020), p. 107123.

[23] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Review* 60.2 (Jan. 2018), pp. 223–311.

[24] Léon Bottou and Yann Le Cun. "On-Line Learning for Very Large Data Sets". In: *Applied Stochastic Models in Business and Industry* 21.2 (Mar. 2005), pp. 137–151.

[25] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. 1. ed. Oxford: Oxford Univ. Press, 2013.

[26] Andrew P Bradley. "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[27] F. Braghin et al. "A Mathematical Model to Predict Railway Wheel Profile Evolution Due to Wear". In: *Wear* 261.11 (Dec. 2006), pp. 1253–1264.

[28] Steve Brooks et al. *Handbook of Markov Chain Monte Carlo*. CRC Press, May 2011.

[29] Alexander Buchholz, Florian Wenzel, and Stephan Mandt. "Quasi-Monte Carlo Variational Inference". In: *International Conference on Machine Learning*. July 2018. Chap. Machine Learning, pp. 668–677.

[30] Eduardo Calixto. "Chapter 4 - Reliability, Availability, and Maintainability (RAM Analysis)". In: *Gas and Oil Reliability Engineering (Second Edition)*. Ed. by Eduardo Calixto. Boston: Gulf Professional Publishing, Jan. 2016, pp. 269–470.

[31] M. Canizo et al. "Real-Time Predictive Maintenance for Wind Turbines Using Big Data Frameworks". In: *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*. June 2017, pp. 70–77.

[32] F Cartella et al. "Online Adaptive Learning of Left-Right Continuous HMM for Bearings Condition Assessment". In: *Journal of Physics: Conference Series*. Vol. 364. IOP Publishing, 2012, p. 012031.

[33] Thyago P. Carvalho et al. "A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance". In: *Computers & Industrial Engineering* 137 (Nov. 2019), p. 106024.

[34] Augustin Cauchy. "Méthode Générale Pour La Résolution Des Systemes d'équations Simultanées". In: *Comp. Rend. Sci. Paris* 25.1847 (1847), pp. 536–538.

[35] Loıc Cerf. "Constraint-Based Mining of Closed Patterns in Noisy n-Ary Relations". In: (2010).

[36] Ji Hwan Cha and Gianpaolo Pulcini. "Optimal Burn-in Procedure for Mixed Populations Based on the Device Degradation Process History". In: *European Journal of Operational Research* 251.3 (2016), pp. 988–998.

[37] Nitesh V Chawla et al. "SMOTE: Synthetic Minority over-Sampling Technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[38] Chia-Mei Chen et al. "Anomaly Network Intrusion Detection Using Hidden Markov Model". In: *Int. J. Innov. Comput. Inform. Control* 12 (2016), pp. 569–580.

[39] Junwen Chen et al. "Automatic Defect Detection of Fasteners on the Catenary Support Device Using Deep Convolutional Neural Network". In: *IEEE Transactions on Instrumentation and Measurement* 67.2 (2017), pp. 257–269.

[40] Xavier Chiementin, Fabrice Bolaers, and J Dron. "Early Detection of Fatigue Damage on Rolling Element Bearings Using Adapted Wavelet". In: *Journal of vibration and acoustics* 129 (Aug. 2007), pp. 495–506.

[41] CKCN Chow and Cong Liu. "Approximating Discrete Probability Distributions with Dependence Trees". In: *IEEE transactions on Information Theory* 14.3 (1968), pp. 462–467.

[42] M. R Clark, D. M McCann, and M. C Forde. "Application of Infrared Thermography to the Non-Destructive Testing of Concrete and Masonry Bridges". In: *NDT & E International*. Structural Faults and Repair 36.4 (June 2003), pp. 265–275.

[43] Alice Consilvio, Angela Di Febbraro, and Nicola Sacco. "A Rolling-Horizon Approach for Predictive Maintenance Planning to Reduce the Risk of Rail Service Disruptions". In: *IEEE Transactions on Reliability* (2020), pp. 1–13.

[44] Francesco Corman and Pavle Kecman. "Stochastic Prediction of Train Delays in Real-Time Using Bayesian Networks". In: *Transportation Research Part C: Emerging Technologies* 95 (2018), pp. 599–615.

[45] Cyrus Cousins* and Amir Dib*. "Fast Convergence Rates for Low-Frequency Pattern Mining with Localization". In: *IEEE International Conference on Data Mining (ICDM 2021)*. 2021.

[46] Martin Crowder and Jerald Lawless. "On a Scheme for Predictive Maintenance". In: *European Journal of Operational Research* 176.3 (2007), pp. 1713–1722.

[47] Guglielmo D'Amico, Jacques Janssen, and Raimondo Manca. "Semi-Markov Reliability Models with Recurrence Times and Credit Rating Applications". In: *Journal of Applied Mathematics and Decision Sciences* 2009 (2009).

[48] Ilesanmi Daniyan et al. "Artificial Intelligence for Predictive Maintenance in the Railcar Learning Factories". In: *Procedia Manufacturing* 45 (2020), pp. 13–18.

[49] Brian Davey and Hilary Priestley. *Introduction to Lattices and Order (2nd Ed.)* Jan. 2002.

[50] Tim de Bruin, Kim Verbert, and Robert Babuska. "Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.3 (Mar. 2017), pp. 523–533.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38.

[52] Amir Dib. "Quantized Variational Inference". In: *Advances in Neural Information Processing Systems* 33 (2020).

[53] Amir Dib et al. "Bayesian Feature Discovery for Predictive Maintenance". In: *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, Mar. 2021.

[54] Josef Dick, Frances Y Kuo, and Ian H Sloan. "High-Dimensional Integration: The Quasi-Monte Carlo Way". In: *Acta Numerica* 22 (2013), p. 133.

[55] Justin Domke. "Provable Gradient Variance Guarantees for Black-Box Variational Inference". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 329–338.

[56] Ming Dong and Zhi-bo Yang. "Dynamic Bayesian Network Based Prognosis in Machining Processes". In: *Journal of Shanghai Jiaotong University (Science)* 13.3 (2008), pp. 318–322.

[57] Ying Du, Tonghai Wu, and Viliam Makis. "Parameter Estimation and Remaining Useful Life Prediction of Lubricating Oil with HMM". In: *Wear* 376–377 (Apr. 2017), pp. 1227–1233.

[58] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. 2017.

[59] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." In: *Journal of machine learning research* 12.7 (2011).

[60] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." In: *Journal of machine learning research* 12.7 (2011).

[61] D. Dyer and R. M. Stewart. "Detection of Rolling Element Bearing Damage by Statistical Vibration Analysis". In: *Journal of Mechanical Design* 100.2 (Apr. 1978), pp. 229–235.

[62] Daniel C. Elton et al. "Applying Machine Learning Techniques to Predict the Properties of Energetic Materials". In: *Scientific Reports* 8.1 (Dec. 2018), p. 9059.

[63] Vitaly Feldman et al. "Agnostic Learning of Monomials by Halfspaces Is Hard". In: *SIAM Journal on Computing* 41.6 (2012), pp. 1558–1590.

[64] Philippe Fournier-Viger et al. "A Survey of Itemset Mining". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7.4 (2017), e1207.

[65] Philippe Fournier-Viger et al. "A Survey of Sequential Pattern Mining". In: *Data Science and Pattern Recognition* 1.1 (2017), pp. 54–77.

[66] Philippe Fournier Viger et al. *The SPMF Open-Source Data Mining Library Version 2*. Sept. 2016.

[67] Jaroslav Fowkes and Charles Sutton. "A Bayesian Network Model for Interesting Itemsets". In: *arXiv:1510.04130 [cs, stat]* 9852 (2016), pp. 410–425.

[68] Nir Friedman, Dan Geiger, and Moises Goldszmidt. "Bayesian Network Classifiers". In: *Machine learning* 29.2 (1997), pp. 131–163.

[69] N. Frikha and L. Huang. "A Multi-Step Richardson–Romberg Extrapolation Method for Stochastic Approximation". In: *Stochastic Processes and their Applications* 125.11 (Nov. 2015), pp. 4066–4101.

[70] Yong Fu et al. "Reliability Analysis of Bogie System Based on Complex Network and Failure Propagation". In: *2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE)*. Zhangjiajie, China: IEEE, Aug. 2019, pp. 564–569.

[71] Emanuele Fumeo, Luca Oneto, and Davide Anguita. "Condition Based Maintenance in Railway Transportation Systems Based on Big Data Streaming Analysis". In: *Procedia Computer Science* 53 (2015), pp. 437–446.

[72] Fuchang Gao. "Non-Zero Boundaries of Leibniz Half-Spaces". In: *Proceedings of the American Mathematical Society* 133.6 (2005), pp. 1757–1762.

[73] A. Garg et al. "Model Development Based on Evolutionary Framework for Condition Monitoring of a Lathe Machine". In: *Measurement* 73 (Sept. 2015), pp. 95–110.

[74] Amulya K Garga et al. "Hybrid Reasoning for Prognostic Learning in CBM Systems". In: *2001 IEEE Aerospace Conference Proceedings (Cat. No. 01TH8542)*. Vol. 6. IEEE, 2001, pp. 2957–2969.

[75] Nagi Z Gebraeel and Mark A Lawley. "A Neural Network Degradation Model for Computing and Updating Residual Life Distributions". In: *IEEE Transactions on Automation Science and Engineering* 5.1 (2008), pp. 154–163.

[76] Nagi Z Gebraeel et al. "Residual-Life Distributions from Component Degradation Signals: A Bayesian Approach". In: *IiE Transactions* 37.6 (2005), pp. 543–557.

[77] Tomas Geffner and Justin Domke. "Using Large Ensembles of Control Variates for Variational Inference". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 9960–9970.

[78] Andrew Gelman, Jeffrey Fagan, and Alex Kiss. "An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias". In: *Journal of the American Statistical Association* 102.479 (Sept. 2007), pp. 813–823.

[79] Andrew Gelman et al. *Bayesian Data Analysis, Third Edition*. CRC Press, Nov. 2013.

[80] Liqiang Geng, Howard, and J. Hamilton. "Interestingness Measures for Data Mining: A Survey". In: *ACM Computing Surveys* (), p. 2006.

[81] Mathieu Gerber. "On Integration Methods Based on Scrambled Nets of Arbitrary Size". In: *Journal of Complexity* 31.6 (2015), pp. 798–816.

[82] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. USA: Kluwer Academic Publishers, 1991.

[83] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Boston, MA: Springer US, 1992.

[84] Ilya Gertsbakh. *Reliability Theory: With Applications to Preventive Maintenance*. Springer, 2013.

[85] Faeze Ghofrani et al. "Recent Applications of Big Data Analytics in Railway Transportation Systems: A Survey". In: *Transportation Research Part C: Emerging Technologies* 90 (May 2018), pp. 226–246.

[86] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Vol. 53. Springer Science & Business Media, 2013.

[87] David S. González-González, Rolando J. Praga-Alejo, and Mario Cantú-Sifuentes. "A Non-Linear Fuzzy Degradation Model for Estimating Reliability of a Polymeric Coating". In: *Applied Mathematical Modelling* 40.2 (Jan. 2016), pp. 1387–1401.

[88] Steven Goodman. "A Dirty Dozen: Twelve P-Value Misconceptions". In: *Seminars in Hematology*. Interpretation of Quantitative Research 45.3 (July 2008), pp. 135–140.

[89] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Berlin Heidelberg: Springer-Verlag, 2000.

[90] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*. Springer, May 2007.

[91] Siegfried Graf, Harald Luschgy, and Gilles Pagès. "Distortion Mismatch in the Quantization of Probability Measures". In: *ESAIM: Probability and Statistics* 12 (2008/ed), pp. 127–153.

[92] Siegfried Graf, Harald Luschgy, and Gilles Pagès. "Optimal Quantizers for Radon Random Vectors in a Banach Space". In: *Journal of Approximation Theory* 144.1 (2007), pp. 27–53.

[93] Ravi Teja Grandhi and N. Krishna Prakash. "Machine-Learning Based Fault Diagnosis of Electrical Motors Using Acoustic Signals". In: *Data Intelligence and Cognitive Informatics*. Ed. by I. Jeena Jacob et al. Algorithms for Intelligent Systems. Singapore: Springer, 2021, pp. 663–671.

[94] Arthur Gretton et al. "A Fast, Consistent Kernel Two-Sample Test." In: *NIPS*. Vol. 23. 2009, pp. 673–681.

[95] Arthur Gretton et al. "A Kernel Method for the Two-Sample Problem". In: *arXiv:0805.2368 [cs]* (May 2008).

[96] E Grinzato et al. "Monitoring of the Scrovegni Chapel by IR Thermography: Giotto at Infrared". In: *Infrared Physics & Technology* 43.3 (June 2002), pp. 165–169.

[97] Qiang Guan, Yincai Tang, and Ancha Xu. "Objective Bayesian Analysis Accelerated Degradation Test Based on Wiener Process Models". In: *Applied Mathematical Modelling* 40.4 (2016), pp. 2743–2755.

[98] Liang Guo et al. "A Recurrent Neural Network Based Health Indicator for Remaining Useful Life Prediction of Bearings". In: *Neurocomputing* 240 (2017), pp. 98–109.

[99] Clemens Gutschi et al. "Log-Based Predictive Maintenance in Discrete Parts Manufacturing". In: *Procedia CIRP*. 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 18-20 July 2018, Gulf of Naples, Italy 79 (Jan. 2019), pp. 528–533.

[100] Jacques Hadamard. *Mémoire Sur Le Problème d'analyse Relatif à l'équilibre Des Plaques Élastiques Encastrées*. Vol. 33. Imprimerie nationale, 1908.

[101] Wilhelmiina Hämäläinen and Geoffrey I Webb. "A Tutorial on Statistically Sound Pattern Discovery". In: *Data Mining and Knowledge Discovery* 33.2 (2019), pp. 325–377.

[102] Jiawei Han et al. "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach". In: *Data Mining and Knowledge Discovery* 8.1 (Jan. 2004), pp. 53–87.

[103] Shilin He et al. "Experience Report: System Log Analysis for Anomaly Detection". In: *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2016, pp. 207–218.

[104] Markus Hegland. "The Apriori Algorithm–a Tutorial". In: *Mathematics and computation in imaging science and information processing* (2007), pp. 209–262.

[105] M. Heidarysafa et al. "Analysis of Railway Accidents' Narratives Using Deep Learning". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Dec. 2018, pp. 1446–1453.

[106] Edmund Hlawka. "Funktionen von Beschränkter Variatiou in Der Theorie Der Gleichverteilung". In: *Annali di Matematica Pura ed Applicata* 54.1 (1961), pp. 325–333.

[107] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 13–30.

[108] Matthew D. Homan and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *The Journal of Machine Learning Research* 15.1 (Jan. 2014), pp. 1593–1623.

[109] Can Hu and Xiang Liu. "Modeling Track Geometry Degradation Using Support Vector Machine Technique". In: *2016 Joint Rail Conference*. Columbia, South Carolina, USA: American Society of Mechanical Engineers, Apr. 2016, V001T01A011.

[110] John D Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in science & engineering* 9.3 (2007), pp. 90–95.

[111] Khac Tuan Huynh et al. "Modeling Age-Based Maintenance Strategies with Minimal Repairs for Systems Subject to Competing Failure Modes Due to Degradation and Shocks". In: *European journal of operational research* 218.1 (2012), pp. 140–151.

[112] Mohd Shawal Jadin and Soib Taib. "Recent Progress in Diagnosing the Reliability of Electrical Equipment by Using Infrared Thermography". In: *Infrared Physics & Technology* 55.4 (July 2012), pp. 236–245.

[113] Seyedahmad Jalili Hassankiadeh. "Failure Analysis of Railway Switches and Crossings for the Purpose of Preventive Maintenance." In: (2011).

[114] Olivier Janssens et al. "Thermal Image Based Fault Diagnosis for Rotating Machinery". In: *Infrared Physics & Technology* 73 (Nov. 2015), pp. 78–87.

[115] Yi Jiang et al. "Fast Classification for Rail Defect Depths Using a Hybrid Intelligent Method". In: *Optik* 180 (2019), pp. 455–468.

[116] Wittawat Jitkrittum et al. "Interpretable Distribution Features with Maximum Testing Power". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 181–189.

[117] Rie Johnson and Tong Zhang. "Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 315–323.

[118] Sushma Kamlu and V Laxmi. "Condition-Based Maintenance Strategy for Vehicles Using Hidden Markov Models". In: *Advances in Mechanical Engineering* 11.1 (Jan. 2019), p. 1687814018806380.

[119] Mehmet Karakose and Orhan Yaman. "Complex Fuzzy System Based Predictive Maintenance Approach in Railways". In: *IEEE Transactions on Industrial Informatics* 16.9 (Sept. 2020), pp. 6023–6032.

[120] F. Karpat et al. "A Novel AI-Based Method for Spur Gear Early Fault Diagnosis in Railway Gearboxes". In: *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. Oct. 2020, pp. 1–6.

[121] Sebastian Kauschke, Johannes Fürnkranz, and Frederik Janssen. "Predicting Cargo Train Failures: A Machine Learning Approach for a Lightweight Prototype". In: *International Conference on Discovery Science*. Springer, 2016, pp. 151–166.

[122] A Khadersab and S Shivakumar. "Vibration Analysis Techniques for Rotating Machinery and Its Effect on Bearing Faults". In: *Procedia Manufacturing* 20 (2018), pp. 247–252.

[123] J Kieffer. "Exponential Rate of Convergence for Lloyd's Method I". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 205–210.

[124] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[125] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014.

[126] Durk P Kingma, Tim Salimans, and Max Welling. "Variational Dropout and the Local Reparameterization Trick". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 2575–2583.

[127] Aaron Klein et al. "Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets". In: *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 528–536.

[128] Simon Kocbek and Bogdan Gabrys. "Automated Machine Learning Techniques in Prognostics of Railway Track Defects". In: *2019 International Conference on Data Mining Workshops (ICDMW)*. Beijing, China: IEEE, Nov. 2019, pp. 777–784.

[129] Jurjen Ferdinand Koksma. "A General Theorem from the Theory of Uniform Distribution modulo 1". In: *Mathematica B (Zutphen)* 11 (1942), pp. 7–11.

[130] N. Kolokas et al. "Forecasting Faults of Industrial Equipment Using Machine Learning Classifiers". In: *2018 Innovations in Intelligent Systems and Applications (INISTA)*. July 2018, pp. 1–6.

[131] Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Algorithms*. Third. Algorithms and Combinatorics. Berlin Heidelberg: Springer-Verlag, 2006.

[132] Sofia Koukoura et al. "Wind Turbine Gearbox Vibration Signal Signature and Fault Development through Time". In: *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece: IEEE, Aug. 2017, pp. 1380–1384.

[133] Alp Kucukelbir et al. "Automatic Variational Inference in Stan". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, Dec. 2015, pp. 568–576.

[134] Max Kuhn and Kjell Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.

[135] Sun Yuan Kung. *Kernel Methods and Machine Learning*. Cambridge University Press, 2014.

[136] Wai Lam and Fahiem Bacchus. "Learning Bayesian Belief Networks: An Approach Based on the MDL Principle". In: *Computational intelligence* 10.3 (1994), pp. 269–293.

[137] David Laredo et al. "A Neural Network-Evolutionary Computational Framework for Remaining Useful Life Estimation of Mechanical Systems". In: *Neural networks* 116 (2019), pp. 178–187.

[138] Jeffery J Leader. *Numerical Analysis and Scientific Computation*. Sirsi) i9780201734997. 2004.

[139] Seungchul Lee, Lin Li, and Jun Ni. "Online Degradation Assessment and Adaptive Fault Detection Using Modified Hidden Markov Model". In: *Journal of Manufacturing Science and Engineering* 132.2 (Apr. 2010).

[140] Yaguo Lei et al. "Machinery Health Prognostics: A Systematic Review from Data Acquisition to RUL Prediction". In: *Mechanical Systems and Signal Processing* 104 (May 2018), pp. 799–834.

[141] Vincent Lemaire, Thibaut Montes, and Gilles Pagès. *New Weak Error Bounds and Expansions for Optimal Quantization*. Mar. 2019.

[142] Li Li et al. "Sequential Behavior Pattern Discovery with Frequent Episode Mining and Wireless Sensor Network". In: *IEEE Communications Magazine* 55.6 (2017), pp. 205–211.

[143] Lisha Li et al. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6765–6816.

[144] Zhibin Li et al. "Sample Adaptive Multiple Kernel Learning for Failure Prediction of Railway Points". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: ACM, July 2019, pp. 2848–2856.

[145] Zhiguo Li and Qing He. "Prediction of Railcar Remaining Useful Life by Multiple Data Source Fusion". In: *IEEE Transactions on Intelligent Transportation Systems* 16.4 (Aug. 2015), pp. 2226–2235.

[146] Haitao Liao and Elsayed A Elsayed. "Reliability Inference for Field Conditions from Accelerated Degradation Testing". In: *Naval Research Logistics (NRL)* 53.6 (2006), pp. 576–587.

[147] Padraig Liggan and David Lyons. "Applying Predictive Maintenance Techniques to Utility Systems". In: *Pharm Eng* 31.6 (2011), pp. 1–7.

[148] Stan Lipovetsky and Michael Conklin. "Analysis of Regression in Game Theory Approach". In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pp. 319–330.

[149] Zachary C. Lipton, John Berkowitz, and Charles Elkan. "A Critical Review of Recurrent Neural Networks for Sequence Learning". In: *arXiv:1506.00019 [cs]* (May 2015).

[150] Han Liu et al. "Unsupervised Fault Diagnosis of Rolling Bearings Using a Deep Neural Network Based on Generative Adversarial Networks". In: *Neurocomputing* 315 (Nov. 2018), pp. 412–424.

[151] Yating Liu and Gilles Pagès. "Characterization of Probability Distribution Convergence in Wasserstein Distance by Lp-Quantization Error Function". In: *Bernoulli* 26.2 (May 2020), pp. 1171–1204.

[152] Yating Liu and Gilles Pagès. "Convergence Rate of Optimal Quantization and Application to the Clustering Performance of the Empirical Measure". In: *Journal of Machine Learning Research* 21.86 (2020), pp. 1–36.

[153] Stuart Lloyd. "Least Squares Quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[154] Ariane Lorton, Mitra Fouladirad, and Antoine Grall. "A Methodology for Probabilistic Model-Based Prognosis". In: *European Journal of Operational Research* 225.3 (2013), pp. 443–454.

[155] Shengfang Lu, Zhen Liu, and Yuan Shen. "Automatic Fault Detection of Multiple Targets in Railway Maintenance Based on Time-Scale Normalization". In: *IEEE Transactions on Instrumentation and Measurement* 67.4 (Apr. 2018), pp. 849–865.

[156] C. Lucchese, S. Orlando, and R. Perego. "Fast and Memory Efficient Mining of Frequent Closed Itemsets". In: *IEEE Transactions on Knowledge and Data Engineering* 18.1 (Jan. 2006), pp. 21–36.

[157] Gabor Lugosi and Marten Wegkamp. "Complexity Regularization via Localized Random Penalties". In: *The Annals of Statistics* 32.4 (Aug. 2004), pp. 1679–1697.

[158] B. Luo et al. "Early Fault Detection of Machine Tools Based on Deep Learning and Dynamic Identification". In: *IEEE Transactions on Industrial Electronics* 66.1 (Jan. 2019), pp. 509–518.

[159] Minh Phong Luong. "Fatigue Limit Evaluation of Metals Using an Infrared Thermographic Technique". In: *Mechanics of Materials* 28.1 (July 1998), pp. 155–163.

[160] Harald Luschgy and Gilles Pagès. "Functional Quantization Rate and Mean Regularity of Processes with an Application to Lévy Processes". In: *Annals of Applied Probability* 18.2 (Apr. 2008), pp. 427–469.

[161] Shuai Ma et al. "Deep Learning for Track Quality Evaluation of High-Speed Railway Based on Vehicle-Body Vibration Prediction". In: *IEEE Access* 7 (2019), pp. 185099–185107.

[162] Chandrabhanu Malla and Isham Panigrahi. "Review of Condition Monitoring of Rolling Element Bearing Using Vibration Analysis and Other Techniques". In: *Journal of Vibration Engineering & Technologies* 7.4 (Aug. 2019), pp. 407–414.

[163] Heikki Mannila and Hannu Toivonen. "Levelwise Search and Borders of Theories in Knowledge Discovery". In: *Data Mining and Knowledge Discovery* 1.3 (Sept. 1997), pp. 241–258.

[164] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. "Discovery of Frequent Episodes in Event Sequences". In: *Data mining and knowledge discovery* 1.3 (1997), pp. 259–289.

[165] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkarno. "Efficient Algorithms for Discovering Association Rules". In: (), p. 12.

[166] Sameen Mansha et al. "A Self-Organizing Map for Identifying Influential-communities in Speech-Based Networks". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2016, pp. 1965–1968.

[167] H. Brendan McMahan and Matthew Streeter. "Adaptive Bound Optimization for Online Convex Optimization". In: *COLT* (July 2010).

[168] Kamal Medjaher, Diego Alejandro Tobon-Mejia, and Noureddine Zerhouni. "Remaining Useful Life Estimation of Critical Components With Application to Bearings". In: *IEEE Transactions on Reliability* 61.2 (June 2012), pp. 292–302.

[169] Hector Mendoza et al. "Towards Automatically-Tuned Neural Networks". In: *Workshop on Automatic Machine Learning*. PMLR, 2016, pp. 58–65.

[170] Carosena Meola. "Infrared Thermography of Masonry Structures". In: *Infrared Physics & Technology* 49.3 (Jan. 2007), pp. 228–233.

[171] Sophie Mercier, Carolina Meier-Hirmer, and Michel Roussignol. "Bivariate Gamma Wear Processes for Track Geometry Modelling, with Application to Intervention Scheduling". In: *Structure and Infrastructure Engineering* 8.4 (Apr. 2012), pp. 357–366.

[172] Andrew Miller et al. "Reducing Reparameterization Gradient Variance". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3708–3718.

[173] Shakir Mohamed et al. "Monte Carlo Gradient Estimation in Machine Learning." In: *J. Mach. Learn. Res.* 21.132 (2020), pp. 1–62.

[174] Edward F Moore. "The Shortest Path through a Maze". In: *Proc. Int. Symp. Switching Theory, 1959*. 1959, pp. 285–292.

[175] Ahmed Mosallam, Kamal Medjaher, and Noureddine Zerhouni. "Component Based Data-Driven Prognostics for Complex Systems: Methodology and Applications". In: *2015 First International Conference on Reliability Systems Engineering (ICRSE)*. IEEE, 2015, pp. 1–7.

[176] Krikamol Muandet et al. "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141.

[177] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[178] Roberto Nappi et al. "A Predictive-Based Maintenance Approach for Rolling Stocks Vehicles". In: *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. Vol. 1. IEEE, 2020, pp. 793–798.

[179] A Nemirovski et al. "Stochastic Approximation Approach to Stochastic Programming". In: *SIAM J. Optim.* Citeseer.

[180] Robin P Nicolai, Rommert Dekker, and Jan M Van Noortwijk. "A Comparison of Models for Measurable Deterioration: An Application to Coatings on Steel Structures". In: *Reliability Engineering & System Safety* 92.12 (2007), pp. 1635–1650.

[181] Alfredo Nunez, Ali Jamshidi, and Hongrui Wang. "Pareto-Based Maintenance Decisions for Regional Railways With Uncertain Weld Conditions Using the Hilbert Spectrum of Axle Box Acceleration". In: *IEEE Transactions on Industrial Informatics* 15.3 (Mar. 2019), pp. 1496–1507.

[182] Travis E Oliphant. *A Guide to NumPy*. Vol. 1. Trelgol Publishing USA, 2006.

[183] Luca Oneto. *Model Selection and Error Estimation in a Nutshell*. Springer, 2020.

[184] Luca Oneto et al. "Local Rademacher Complexity: Sharper Risk Bounds with and without Unlabeled Samples". In: *Neural Networks* 65 (2015), pp. 115–125.

[185] Art B Owen. "Local Antithetic Sampling with Scrambled Nets". In: *The Annals of Statistics* 36.5 (2008), pp. 2319–2343.

[186] Gilles Pagès. "A Space Quantization Method for Numerical Integration". In: *Journal of Computational and Applied Mathematics* 89.1 (Mar. 1998), pp. 1–38.

[187] Gilles Pagès. "Introduction to Vector Quantization and Its Applications for Numerics". In: *ESAIM: Proceedings and Surveys* 48 (Jan. 2015), pp. 29–79.

[188] Gilles Pagès. "Multi-Step Richardson-Romberg Extrapolation: Remarks on Variance Control and Complexity". In: *Monte Carlo Methods and Applications* 13 (Jan. 2007).

[189] Gilles Pagès. *Numerical Probability: An Introduction with Applications to Finance*. Universitext. Springer International Publishing, 2018.

[190] Gilles Pagès and Jacques Printems. "Optimal Quadratic Quantization for Numerics: The Gaussian Case". In: *Monte Carlo Methods and Applications* 9.2 (Apr. 2003), pp. 135–165.

[191] Victor M Panaretos and Yoav Zemel. "Statistical Aspects of Wasserstein Distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431.

[192] M. L. Pastor et al. "Applying Infrared Thermography to Study the Heating of 2024-T3 Aluminium Specimens under Fatigue Loading". In: *Infrared Physics & Technology* 51.6 (Oct. 2008), pp. 505–515.

[193] Kuruppulage Asela Buddhika Pathirathna et al. "Use of Thermal Imaging Technology for Locomotive Maintenance in Sri Lanka Railways". In: *2018 International Conference on Intelligent Rail Transportation (ICIRT)*. Singapore: IEEE, Dec. 2018, pp. 1–4.

[194] Debprakash Patnaik, PS Sastry, and KP Unnikrishnan. "Inferring Neuronal Network Connectivity from Spike Data: A Temporal Data Mining Approach". In: *Scientific Programming* 16.1 (2008), pp. 49–77.

[195] Dmitry Pavlov, Heikki Mannila, and Padhraic Smyth. "Beyond Independence: Probabilistic Models for Query Approximation on Binary Transaction Data". In: *IEEE Transactions on Knowledge and Data Engineering* 15.6 (2003), pp. 1409–1421.

[196] Karl Pearson. "Contributions to the Mathematical Theory of Evolution". In: *Philosophical Transactions of the Royal Society of London. A* 185 (1894), pp. 71–110.

[197] Dana Pe'er. "Bayesian Network Analysis of Signaling Networks: A Primer". In: *Science's STKE* 2005.281 (2005), pl4–pl4.

[198] Leonardo Pellegrina, Cinzia Pizzi, and Fabio Vandin. "Fast Approximation of Frequent K-Mers and Applications to Metagenomics". In: *Journal of Computational Biology* 27.4 (2020), pp. 534–549.

[199] Leonardo Pellegrina, Matteo Riondato, and Fabio Vandin. "SPuManTE: Significant Pattern Mining with Unconditional Testing". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 1528–1538.

[200] Leonardo Pellegrina et al. "MCRapper: Monte-Carlo Rademacher Averages for Poset Families and Approximate Pattern Mining". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 2165–2174.

[201] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *arXiv:1803.00567 [stat]* (Mar. 2020).

[202] B. T. Polyak and A. B. Juditsky. "Acceleration of Stochastic Approximation by Averaging". In: *SIAM Journal on Control and Optimization* 30.4 (July 1992), pp. 838–855.

[203] T Praveenkumar et al. "Fault Diagnosis of Automobile Gearbox Based on Machine Learning Techniques". In: *Procedia Engineering* 97 (2014), pp. 2092–2098.

[204] Min Qin and Kai Hwang. "Frequent Episode Rules for Intrusive Anomaly Detection with Internet Datamining". In: *USENIX Security Symposium*. Citeseer, 2004, pp. 1–15.

[205] Yongyi Ran et al. "A Survey of Predictive Maintenance: Systems, Purposes and Approaches". In: *arXiv:1912.07383 [cs, eess]* (Dec. 2019).

[206] Rajesh Ranganath, Sean Gerrish, and David Blei. "Black Box Variational Inference". In: *Artificial Intelligence and Statistics*. Apr. 2014. Chap. Machine Learning, pp. 814–822.

[207] John T. Renwick and Paul E. Babson. "Vibration Analysis—A Proven Technique as a Predictive Maintenance Tool". In: *IEEE Transactions on Industry Applications* IA-21.2 (Mar. 1985), pp. 324–332.

[208] Lewis Fry Richardson and Richard Tetley Glazebrook. "IX. The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations, with an Application to the Stresses in a Masonry Dam". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 210.459-470 (Jan. 1911), pp. 307–357.

[209] Matteo Riondato. "Sampling-Based Randomized Algorithms for Big Data Analytics". PhD thesis. Citeseer, 2014.

[210] Matteo Riondato and Eli Upfal. "Mining Frequent Itemsets through Progressive Sampling with Rademacher Averages". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. Sydney, NSW, Australia: ACM Press, 2015, pp. 1005–1014.

[211] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407.

[212] Bernard Robles et al. "HMM Framework, for Industrial Maintenance Activities". In: 2013.

[213] Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. "Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 6925–6934.

[214] Tomasz Rolski et al. *Stochastic Processes for Insurance and Finance*. Vol. 505. John Wiley & Sons, 2009.

[215] Louis Rompré, Ismaïl Biskri, and Jean-Guy Meunier. "Using Association Rules Mining for Retrieving Genre-Specific Music Files." In: *FLAIRS Conference*. 2017, pp. 706–711.

[216] Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. "Overdispersed Black-Box Variational Inference". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'16. Jersey City, New Jersey, USA: AUAI Press, June 2016, pp. 647–656.

[217] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning Internal Representations by Error Propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[218] Nazmus Sakib and Thorsten Wuest. "Challenges and Opportunities of Condition-Based Predictive Maintenance: A Review". In: *Procedia CIRP*. 6th CIRP Global Web Conference – Envisaging the Future Manufacturing, Design, Technologies and Systems in Innovation Era (CIRPe 2018) 78 (Jan. 2018), pp. 267–272.

[219] Tim Salimans, Diederik P. Kingma, and Max Welling. "Markov Chain Monte Carlo and Variational Inference: Bridging the Gap". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France, July 2015, pp. 1218–1226.

[220] S Santhi and P Padmaja. "A Survey of Frequent Subgraph Mining Algorithms for Uncertain Graph Data". In: *International Research Journal of Engineering and Technology (IRJET)* 2.2 (2015), pp. 688–696.

[221] B Santoso et al. "Synthetic over Sampling Methods for Handling Class Imbalanced Problems: A Review". In: *IOP Conference Series: Earth and Environmental Science*. Vol. 58. IOP Publishing, 2017, p. 012031.

[222] L. K. Saul, T. Jaakkola, and M. I. Jordan. "Mean Field Theory for Sigmoid Belief Networks". In: *Journal of Artificial Intelligence Research* 4 (Mar. 1996), pp. 61–76.

[223] Abhinav Saxena et al. "Metrics for Offline Evaluation of Prognostic Performance". In: *International Journal of Prognostics and Health Management* 1 (Jan. 2010), pp. 2153–2648.

[224] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel Methods in Computational Biology*. 2004.

[225] Richard F. Serfozo. "Semi-Stationary Processes". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 23.2 (June 1972), pp. 125–132.

[226] Bobak Shahriari et al. "Taking the Human out of the Loop: A Review of Bayesian Optimization". In: *Proceedings of the IEEE* 104.1 (2015), pp. 148–175.

[227] Shuo Shang et al. "Searching Trajectories by Regions of Interest". In: *IEEE Transactions on Knowledge and Data Engineering* 29.7 (2017), pp. 1549–1562.

[228] Claude E Shannon. "Prediction and Entropy of Printed English". In: *Bell system technical journal* 30.1 (1951), pp. 50–64.

[229] John W Sheppard and Mark A Kaufman. "Bayesian Diagnosis and Prognosis Using Instrument Uncertainty". In: *IEEE Autotestcon, 2005.* IEEE, 2005, pp. 417–423.

[230] Xiao-Sheng Si et al. "Remaining Useful Life Estimation – A Review on the Statistical Data Driven Approaches". In: *European Journal of Operational Research* 213.1 (Aug. 2011), pp. 1–14.

[231] António Simões et al. "The State of the Art of Hidden Markov Models for Predictive Maintenance of Diesel Engines: HMM for Predictive Maintenance of Diesel Engines". In: *Quality and Reliability Engineering International* (Jan. 2017).

[232] Umut Simsekli et al. "On the Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks". In: *arXiv preprint arXiv:1912.00018* (2019).

[233] GNVG Sirisha, M Shashi, and GV Padma Raju. "Periodic Pattern Mining–Algorithms and Applications". In: *Global Journal of Computer Science and Technology* (2014).

[234] C. R. Smith, G. Erickson, and Paul O. Neudorfer. *Maximum Entropy and Bayesian Methods: Seattle, 1991*. Springer Science & Business Media, June 2013.

[235] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. "An Instance Level Analysis of Data Complexity". In: *Machine learning* 95.2 (2014), pp. 225–256.

[236] Alex Smola et al. "A Hilbert Space Embedding for Distributions". In: *International Conference on Algorithmic Learning Theory*. Springer, 2007, pp. 13–31.

[237] Le Song et al. "Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 961–968.

[238] Ravi Srinivasan et al. "Preventive Maintenance of Centralized HVAC Systems: Use of Acoustic Sensors, Feature Extraction, and Unsupervised Learning". In: *Proceedings of the 15th IBPSA Conference*. 2017, pp. 2518–2524.

[239] Bharath K. Sriperumbudur et al. "Hilbert Space Embeddings and Metrics on Probability Measures". In: *arXiv:0907.5309 [math, stat]* (Jan. 2010).

[240] Erik Štrumbelj and Igor Kononenko. "Explaining Prediction Models and Individual Predictions with Feature Contributions". In: *Knowledge and information systems* 41.3 (2014), pp. 647–665.

[241] Ming-Yang Su. "Discovery and Prevention of Attack Episodes by Frequent Episodes Mining and Finite State Machines". In: *Journal of Network and Computer Applications* 33.2 (2010), pp. 156–167.

[242] Kyungbok Sung et al. "A Formal and Quantifiable Log Analysis Framework for Test Driving of Autonomous Vehicles". In: *Sensors* 20.5 (2020), p. 1356.

[243] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[244] Hwa-yaw Tam et al. "Intelligent Optical Fibre Sensing Networks Facilitate Shift to Predictive Maintenance in Railway Systems". In: *2018 International Conference on Intelligent Rail Transportation (ICIRT)*. Singapore: IEEE, Dec. 2018, pp. 1–4.

[245] Syed Khairuzzaman Tanbeer et al. "CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2008, pp. 1022–1027.

[246] Danesh Tarapore, Anders Lyhne Christensen, and Jon Timmis. "Generic, Scalable and Decentralized Fault Detection for Robot Swarms". In: *PLOS ONE* 12.8 (Aug. 2017), e0182058.

[247] Albert Thomas et al. "Learning Hyperparameters for Unsupervised Anomaly Detection." In: *ICML, Anomaly Detection Workshop*. 2016.

[248] Zhigang Tian. "An Artificial Neural Network Method for Remaining Useful Life Prediction of Equipment Subject to Condition Monitoring". In: *Journal of Intelligent Manufacturing* 23.2 (2012), pp. 227–237.

[249] Zhigang Tian, Lorna Wong, and Nima Safaei. "A Neural Network Approach for Remaining Useful Life Prediction Utilizing Both Failure and Suspension Histories". In: *Mechanical Systems and Signal Processing* 24.5 (2010), pp. 1542–1555.

[250] Zhigang Tian and Ming J Zuo. "Health Condition Prediction of Gears Using a Recurrent Neural Network Approach". In: *IEEE transactions on reliability* 59.4 (2010), pp. 700–705.

[251] DA Tobon-Mejia et al. "Estimation of the Remaining Useful Life by Using Wavelet Packet Decomposition and HMMs". In: *2011 Aerospace Conference*. IEEE, 2011, pp. 1–10.

[252] Hannu Toivonen. "Sampling Large Databases for Association Rules". In: *Proceedings of the 22th International Conference on Very Large Data Bases*. VLDB '96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Sept. 1996, pp. 134–145.

[253] Minh-Ngoc Tran, David J. Nott, and Robert Kohn. "Variational Bayes With Intractable Likelihood". In: *Journal of Computational and Graphical Statistics* 26.4 (Oct. 2017), pp. 873–882.

[254] C. I. Ugechi et al. "Condition-Based Diagnostic Approach for Predicting the Maintenance Requirements of Machinery". In: *Engineering* 01 (Nov. 2009), p. 177.

[255] Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. "LCM Ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets". In: *Fimi*. Vol. 126. 2004.

[256] George J Vachtsevanos and George J Vachtsevanos. *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. Vol. 456. Wiley Hoboken, 2006.

[257] Jan M van Noortwijk. "A Survey of the Application of Gamma Processes in Maintenance". In: *Reliability Engineering & System Safety* 94.1 (2009), pp. 2–21.

[258] Jan N Van Rijn and Frank Hutter. "An Empirical Study of Hyperparameter Importance Across Datasets." In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*. 2017, pp. 91–98.

[259] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[260] JN Venkatesh et al. "Discovering Periodic-Frequent Patterns in Transactional Databases Using All-Confidence and Periodic-All-Confidence". In: *International Conference on Database and Expert Systems Applications*. Springer, 2016, pp. 55–70.

[261] Cédric Villani. *Optimal Transport: Old and New*. Vol. 338. Springer Science & Business Media, 2008.

[262] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. "Krimp: Mining Itemsets That Compress". In: *Data Mining and Knowledge Discovery* 23.1 (2010), pp. 169–214.

[263] Chen Wang, Hoang Tam Vo, and Peng Ni. "An IoT Application for Fault Diagnosis and Prediction". In: *2015 IEEE International Conference on Data Science and Data Intensive Systems*. IEEE, 2015, pp. 726–731.

[264] J Wang et al. "Predictive Maintenance Based on Event-Log Analysis: A Case Study". In: *IBM Journal of Research and Development* 61.1 (2017), pp. 11–121.

[265] Lei Wang et al. "Web Anomaly Detection Based on Frequent Closed Episode Rules". In: *2017 IEEE Trustcom BigDataSE ICESS*. IEEE, 2017, pp. 967–972.

[266] Ning Wang et al. "A Hidden Semi-Markov Model with Duration-Dependent State Transition Probabilities for Prognostics". In: *Mathematical Problems in Engineering* 2014 (2014).

[267] Qi Wang, Siqi Bu, and Zhengyou He. "Achieving Predictive and Proactive Maintenance for High-Speed Railway Power Equipment With LSTM-RNN". In: *IEEE Transactions on Industrial Informatics* 16.10 (Oct. 2020), pp. 6509–6517.

[268] John Winn and Christopher M. Bishop. "Variational Message Passing". In: *Journal of Machine Learning Research* 6.Apr (2005), pp. 661–694.

[269] Sze-jung Wu et al. "A Neural Network Integrated Decision Support System for Condition-Based Optimal Predictive Maintenance Policy". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 37.2 (Mar. 2007), pp. 226–236.

[270] P. Xu et al. "Condition Monitoring of Wheel Wear for High-Speed Trains: A Data-Driven Approach". In: *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*. June 2018, pp. 1–8.

[271] Jihong Yan, Muammer Koç, and Jay Lee. "A Prognostic Algorithm for Machine Performance Assessment and Its Application". In: *Production Planning & Control* 15.8 (Dec. 2004), pp. 796–801.

[272] Guizhen Yang. "The Complexity of Mining Maximal Frequent Itemsets and Maximal Frequent Patterns". In: *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*. Seattle, WA, USA: ACM Press, 2004, p. 344.

[273] Wenxian Yang et al. "Research on a Simple, Cheap but Globally Effective Condition Monitoring Technique for Wind Turbines". In: *2008 18th International Conference on Electrical Machines*. Sept. 2008, pp. 1–5.

[274] Shun-Zheng Yu. "Hidden Semi-Markov Models". In: *Artificial Intelligence* 174.2 (Feb. 2010), pp. 215–243.

[275] Shun-Zheng Yu and Hisashi Kobayashi. "An Efficient Forward-Backward Algorithm for an Explicit-Duration Hidden Markov Model". In: *IEEE signal processing letters* 10.1 (2003), pp. 11–14.

[276] Shun-Zheng Yu and Hisashi Kobayashi. "Practical Implementation of an Efficient Forward-Backward Algorithm for an Explicit-Duration Hidden Markov Model". In: *IEEE Transactions on Signal Processing* 54.5 (2006), pp. 1947–1951.

[277] Paul Zador. "Asymptotic Quantization Error of Continuous Signals and the Quantization Dimension". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 139–149.

[278] M.J. Zaki. "Scalable Algorithms for Association Mining". In: *IEEE Transactions on Knowledge and Data Engineering* 12.3 (May-June/2000), pp. 372–390.

[279] Mohammed Zaki and C.-J Hsiao. "Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure". In: *Knowledge and Data Engineering, IEEE Transactions on* 17 (May 2005), pp. 462–478.

[280] Mohammed Zaki et al. "Evaluation of Sampling for Data Mining of Association Rules". In: (Sept. 1998).

[281] Mohammed Javeed Zaki. "Scalable Algorithms for Association Mining". In: *IEEE transactions on knowledge and data engineering* 12.3 (2000), pp. 372–390.

[282] Matthew D Zeiler. “Adadelta: An Adaptive Learning Rate Method”. In: *arXiv preprint arXiv:1212.5701* (2012).

[283] Ke Zhang et al. “Automated IT System Failure Prediction: A Deep Learning Approach”. In: *2016 IEEE International Conference on Big Data (Big Data)*. Washington DC,USA: IEEE, Dec. 2016, pp. 1291–1300.

[284] Minghua Zhang et al. “Mining Periodic Patterns with Gap Requirement from Sequences”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.2 (2007), 7–es.

[285] Weiting Zhang, Dong Yang, and Hongchao Wang. “Data-Driven Methods for Predictive Maintenance of Industrial Equipment: A Survey”. In: *IEEE Systems Journal* 13.3 (Sept. 2019), pp. 2213–2227.

[286] Xinyu Zhao et al. “Semi-Supervised Constrained Hidden Markov Model Using Multiple Sensors for Remaining Useful Life Prediction and Optimal Predictive Maintenance:” in: *Proceedings of the Annual Conference of the PHM Society* 11.1 (Sept. 2019).

[287] Yang Zhao, Tian-hua Xu, and Wang Hai-feng. “Text Mining Based Fault Diagnosis of Vehicle On-Board Equipment for High Speed Railway”. In: *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. Qingdao: IEEE, Oct. 2014, pp. 900–905.

[288] Kai Zheng et al. “Towards Efficient Search for Activity Trajectories”. In: *2013 IEEE 29Th International Conference on Data Engineering (ICDE)*. IEEE, 2013, pp. 230–241.

[289] Alexander Zien et al. “The Feature Importance Ranking Measure”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Wray Buntine et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 694–709.

[290] Albrecht Zimmermann. “Understanding Episode Mining Techniques: Benchmarking on Diverse, Realistic, Artificial Data”. In: *Intelligent Data Analysis* 18.5 (2014), pp. 761–791.

[291] Konrad Zuse. “Der Plankalkül”. In: *It-Information Technology* (1972).

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Apprentissage de motifs en grande dimension appliqué aux séries temporelles

Mots clés: Apprentissage automatique, serie temporelle, statistiques bayesiennes, detection d'anomalie, quantization optimale, maintenance predictive.

Résumé: Bien que l'application des méthodes apprentissage automatique dans divers contextes ait connu une croissance rapide au cours de la dernière décennie, son utilisation dans les environnements industriels reste problématique. La raison principale tient au conflit entre les procédures historiques établies et le manque de transparence du processus de décision d'une chaîne d'apprentissage automatique. Par ailleurs, la nature et la qualité des données disponibles ne permet pas l'utilisation directe des modèles d'apprentissage statistiques traditionnels. La plupart des bases de données industrielles n'ont pas été construites dans l'objectif de satisfaire aux standards du traitement automatique mais pour se conformer a des exigences réglementaires et assister aux tâches administratives. En particulier, les données non numériques ou symboliques sont couramment utilisées pour leur versatilité. Des exemples de telles données sont les documents textuels, les séries d'événements d'un ordinateur de bord ou encore les séquences ADN.

La motivation première de cette thèse est la conception d'approches humainement interprétable pour la maintenance prédictive du parc ferroviaire français. Nous proposons d'aller au-delà des approches standards par l'utilisation de méthodes associant techniques d'extraction de motifs et approches statistiques pour la détection d'anomalie. Le contenu de cette thèse trouve une application plus large dans n'importe quel domaine d'application nécessitant le traitement de séries temporelles symboliques.

La première contribution consiste en une solution complète d'apprentissage automatique pour la maintenance prédictive d'une large flotte de trains. Comme seconde contribution, nous proposons une nouvelle méthode pour les ensembles de données symboliques basée sur un modèle génératif bayésien qui permet l'amélioration des métriques de références de façon interprétable pour un ensemble de données symboliques. Dans une troisième contribution, nous introduisons une nouvelle méthode d'extraction progressive basée sur les complexités locales afin d'obtenir des intervals de confiance sur la fréquence des motifs. Finalement, une nouvelle méthode générale d'optimisation stochastique basée sur un échantillonage alternatif est proposé. Cette méthode s'applique au cas spécifique de l'apprentissage bayésien dans le cadre de l'inférence variationnelle. Dans ce cadre, nous fournissons une preuve théorique et empirique de la supériorité de cette approche par rapport aux méthodes les plus avancées.

Title: High dimensional pattern learning applied to symbolic time-series

Keywords: Machine learning, temporal series, bayesian learning, anomaly detection, optimal quantization, predictive maintenance.

Abstract: While the adoption of machine learning in many applied contexts has been growing rapidly in the last decade, there remain challenges to use it in certain industrial settings. The main reason is the clash between established historical procedures with the uncertainty and lack of transparency of a machine learning pipeline's decision process. Another reason is that the input needed to feed a traditional machine learning model does not fit the available type or quality of available data. Most industrial databases have not been developed for statistical analysis but to comply with the regulatory requirements and to perform administrative tasks. In particular, non-numerical or symbolic features are common as it is a versatile way of recording events of interest. Examples of such data are textual documents, sequence of log-events or DNA sequences. The exponential number of possible patterns typically dominates the complexity associated with learning relevant information from symbols.

This thesis's applicative framework and primary motivation is to design efficient, human-readable and computationally tractable methods for predictive maintenance on the french train fleet. To that end, we propose to go beyond standard approaches by using a combination of traditional machine learning algorithms with pattern mining techniques to allow human experts to understand and interact with the algorithmic layer of the predictive maintenance pipeline. This thesis's main objective is to tackle these issues by proposing approaches that can be generally applied to a symbolic sequence of data with a human-readable output and trained at a reasonable computational cost. To that end, we begin by constructing a complete machine learning pipeline solution for predictive maintenance on a large fleet of rail vehicles that can be computed at a reasonable cost and provides valuable insight on the underlying symbol dynamic of the degradation process. As a second contribution, we propose a new method for symbolic data set based on a Bayesian generative model for patterns that can increases score accuracy in an interpretable fashion for any symbolic data set. As a third contribution, we introduce a new progressive mining method based on local complexities to obtain sharper statistical bounds on the pattern frequency. Finally, a new and general stochastic optimization method based on alternative sampling is proposed. This method can be applied to the specific use case of Bayesian learning through the Variational Inference setting. In this instance, we provide theoretical and empirical proof of the superiority of this approach compared to the most advanced methods.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France