Statistical modeling and analysis of Internet latency traffic data

Alexis Fremond

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Université Paris-Dauphine

Statistical modeling and analysis of Internet latency traffic data

Soutenue par
Alexis Frémond
Le 02/10/2020

École doctorale n°543
École doctorale SDOSE

Spécialité
Mathématiques

Composition du jury :

Gérard BIAU Professeur, Sorbonne Université	*Président*
Arnak DALALYAN Professeur, ENSAE Paris - IP Paris	*Rapporteur*
Mathieu ROSENBAUM Professeur, École Polytechnique	*Rapporteur*
Gérard BIAU Professeur, Sorbonne Université	*Examinateur*
Vincent RIVOIRARD Professeur, Université Paris-Dauphine	*Examinateur*
Marc HOFFMANN Professeur, Université Paris-Dauphine	*Directeur de thèse*

Dauphine | PSL
UNIVERSITÉ PARIS

Remerciements

En premier lieu, je tiens naturellement à remercier Marc, qui a accepté d'encadrer cette thèse et qui, par sa disponibilité et son soutien, m'a permis de la mener à bien. Tout au long de nos échanges et de nos nombreuses sessions de travail, ce sont sa patience, son exigence et sa rigueur qui m'ont permis de me dépasser. Merci de m'avoir guidé scientifiquement, accompagné chaleureusement et poussé quand cela était nécessaire.

Merci à Arnak Dalalyan et Mathieu Rosenbaum d'avoir bien voulu rapporter ma thèse, j'en suis très honoré. Je remercie également Gérard Biau et Vincent Rivoirard d'avoir accepté d'être membres de mon jury, je suis heureux de leur présence.

Je tiens à remercier l'ensemble des équipes du CEREMADE qui m'a accueilli durant ces trois années. J'ai eu la chance de bénéficier de l'aide précieuse et de la patience des membres de l'équipe administrative. Je remercie également Sophie Donnet et Julien Stoehr, qui m'ont confié leurs étudiants en TD durant plusieurs semestres. Je n'oublie pas les autres doctorants pour nos échanges réguliers, et ceux du DRM, avec qui j'ai partagé de nombreuses pauses déjeuner et discussions enflammées.

Bien entendu, je remercie Cedexis, puis Citrix, qui m'ont donné l'opportunité de réaliser cette thèse. Merci à Paris Zafiris et Rony Lerner qui, de la Grèce aux Etats-Unis, ont, par leurs retours et commentaires, apporté leur contribution scientifique à ce travail.

Enfin, je remercie mes proches : mes amis et ma famille qui, durant ces trois années, m'ont soutenu et encouragé. Merci d'être parvenus, par moment, à me faire penser à autre chose qu'aux statistiques !

Ysé, merci pour tout.

Contents

Chapter 1

Introduction

1.1 Présentation

Cette thèse a été réalisée dans le cadre d'une Convention Industrielle de Formation par la REcherche (CIFRE). Le travail présenté dans ce manuscrit est issu d'une collaboration entre le CEREMADE de l'Université Paris Dauphine (UMR 7534) et la société Citrix, spécialisée dans l'optimisation du traffic Internet. Dans ces travaux, nous nous intéressons à la problématique de l'aiguillage automatique des utilisateurs dans le Réseau Internet dans un cadre de modélisation de la latence. Cette durée, qui mesure le temps nécessaire à la transmission d'informations dans le Réseau Internet, est au centre de ce manuscrit.

1.2 Problématique

Avant de présenter les résultats de cette thèse, une présentation de l'activité de la société Citrix s'impose. Le réseau Internet se compose d'ordinateurs interconnectés capables d'échanger de l'information à distance. En première approximation, le réseau Internet a la structure d'un graphe $G = (V, E)$. Chaque noeud $x \in V$ représente un ordinateur, et chaque arrête $e \in E$, $e = (u, v) \in V^2, u \neq v$, représente une connection possible entre les ordinateurs u et v, c'est à dire qu'un échange d'information est possible entre ces deux machines, appelés plus généralement serveurs.

Le contenu d'un site Internet est stocké physiquement sur un serveur, dit *Origine*, lequel est mis à disposition des autres utilisateurs. À cause de contraintes physiques que nous ne détaillerons pas dans ce manuscrit, la quantité maximale d'information qu'un serveur donné peut débiter par unité de temps, appelée *bande passante*, est finie. Ainsi, pour un serveur donné à un instant donné, la *bande passante* doit être partagée entre les différents utilisateurs connectés à ce serveur. Cela signifie que plus le traffic augmente sur le serveur, plus la fraction de *bande passante* par utilisateur diminue: le temps que chaque utilisateur doit patienter pour obtenir les ressources du site, comme le corps du texte ou les images par exemple, augmente. Au delà d'un certain nombre d'utilisateurs connectés simultanément, le serveur devient indisponible et cesse de fonctionner, on

parle de défaillance ou de déni de service [7].

Une solution pour éviter le déni de service consiste à recopier le contenu de l'*Origine* sur d'autres serveurs afin de réduire le nombre de connexions par serveur pour augmenter la *bande passante* par utilisateur. Cette opération peut s'effectuer de deux manières : soit en construisant une infrastructure capable de supporter le traffic, soit en louant cette infrastructure à un tiers. La première option pose des difficultés pratiques de par son coût et son manque de flexibilité. En effet, si un site Web voit son traffic diminuer, une partie de son infrastructure devient inutile et donc une partie de l'investissement est perdue. Au contraire, si l'affluence du site augmente, il faut agrandir en conséquence l'infrastructure existante, c'est à dire le parc de serveurs, opération généralement complexe et coûteuse. Si une partie de l'audience est délocalisée dans un autre pays par exemple, y construire une infrastructure pose également des difficultés légales, d'organisation et de gestion: on parle de manque de scalabilité. Par conséquent, une grande partie des sites Web sont hébergés sur des serveurs loués à des tiers, appelés *CDN* pour Content Delivery Network, littéralement réseau de distribution de contenu. Un *CDN* est une entreprise qui propose à la location des serveurs pour héberger du contenu. Un site Web qui ne souhaite pas bâtir son infrastructure peut avoir recours à un *CDN* afin de dupliquer le contenu de son *Origine*. Dans ce cas, le site Web loue un sous-ensemble de serveurs au *CDN*. Chaque serveur loué est appelé serveur *Edge*, et l'ensemble des serveurs *Edge* est appelé *Map*.

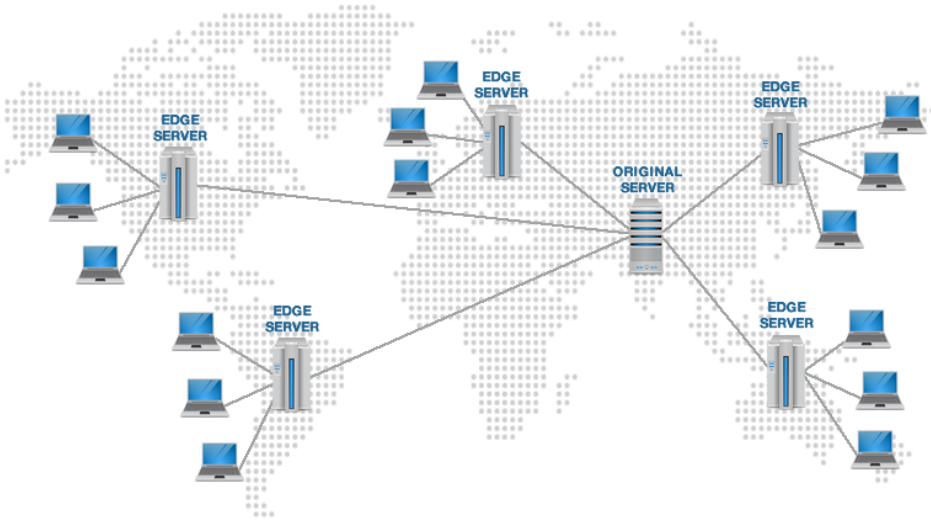

Figure 1.1: *Illustration du fonctionnement d'un CDN. Le serveur Origine duplique son contenu sur des serveurs Edge constitutifs d'une map afin de mieux répartir la charge de traffic. Source: HighCharts* `https://www.highcharts.com/blog/news/50-codehighchartscom-moves-to-cdn/`.

Le site Web n'est pas totalement libre dans le choix des serveurs *Edge* qu'il loue, notamment en ce qui concerne leurs nombres et leurs localisations géographiques. En effet, pour des raisons

d'efficacité et de simplicité, le *CDN* propose à ses clients des *maps* pré-définies. Les différentes *maps* proposées par un *CDN* peuvent correspondre à différentes couvertures géographiques ou différents niveaux de prestation. Il existe aujourd'hui dans le monde des dizaines de *CDN* différents, proposant chacun des dizaines de *maps* différentes. Le nombre de *maps* proposées par un *CDN* étant fini, deux clients peuvent souscrire à la même *map* M sur un *CDN* C donné. Il est naturel de penser que des mesures de latence effectuées chez ces deux clients doivent avoir la même distribution. La latence est l'observable qui détermine la performance d'un réseau Internet, voir définition 1.

Definition 1. *Soient $x, y \in V$ tel que $(x, y) \in E$ sont deux serveurs distincts connectés au réseau Internet. La latence de x vers y, notée $L(x, y)$, est une grandeur exprimée en milli-secondes (ms) et définie comme le temps écoulé entre l'envoi du premier bit d'information d'une requête émise par x à destination de y, et la reception du premier bit d'information de la réponse envoyée par y à destination de x.*

Remark 1. *La latence L n'est pas symétrique: pour deux serveurs x, y, en général $L(x, y) \neq L(y, x)$.*

Remark 2. *La latence est parfois définie comme le temps écoulé entre l'envoi du premier bit d'information d'une requête et la reception de ce bit d'information par la destination.*

Dans les faits, les distributions des mesures de latence pour deux clients souscrivant à la même *map* M du *CDN* C peuvent être différentes. En effet, les *CDN* n'allouent pas exactement les mêmes serveurs aux clients ayant souscrits à la *map* M. La raison s'explique par la nature dynamique de la duplication des ressources au sein d'un *CDN*. Chaque ressource constitutive du site est stockée sur l'*Origine*. Le *CDN* effectue des copies de l'*Origine* sur différents serveurs de la map, mais cette cette copie est limitée dans le temps. Chaque ressource dupliquée possède un Time To Live, ou TTL. Le TTL est une durée généralement exprimée en secondes. Si une ressource sur un serveur du *CDN* n'est pas demandée pendant un laps de temps supérieur au TTL, la ressource est écrasée pour optimiser l'espace de stockage. Dès qu'un utilisateur a besoin de cette resource, le serveur doit la re-télécharger depuis l'*Origine* avant de la restituer à l'utilisateur. La ressource restera alors disponible pendant la durée de son TTL, puis sera de nouveau écrasée, et ainsi de suite: on parle de mise en cache temporaire. Ainsi, l'ensemble des serveurs constitutifs d'une *map* qui hébergent le contenu d'une page Web fluctue au cours du temps. On comprend dès lors que le terme *map* ne désigne pas fondamentalement un ensemble figé de serveurs, mais correspond à la garantie d'un certain niveau de prestation. Par exemple, une *map premium* pourra garantir un TTL plus long, et une mise en cache sur un plus grand nombre de serveurs, par exemple. Le *CDN* est libre d'organiser la copie et la distribution des ressources comme il le souhaite tant qu'il garantit ce niveau de prestation.

Le *CDN* met à disposition de son client une *map* afin de copier le contenu de son *Origine* pour réduire la charge de trafic, mais chaque utilisateur n'a besoin que d'un seul serveur pour télécharger le contenu: il est donc nécessaire de désigner un serveur spécifique pour fournir le contenu à l'utilisateur. Ce processus est appelé *load-balacing*, littéralement le *partage de charge*. On parle de routage ou d'aiguillage en français. De façon générale, le *load-balacing* désigne l'action de répartir les utilisateurs d'un réseau sur ses serveurs constitutifs. Cette répartition peut se faire de manière déterministe (par exemple en envoyant l'utilisateur sur le serveur géographiquement le plus proche) ou stochastique (par exemple en envoyant l'utilisateur sur un serveur sélectionné aléatoirement dans la *map*). Lorsqu'un utilisateur d'un site Web qui héberge son contenu sur un *CDN* requête le contenu d'une page Internet, le *load-balacing* est effectué par le *CDN* lui-même.

Un *CDN* se différencie de ses concurrents notamment grâce à l'efficacité de ses algorithmes de *load-balacing*. Ces derniers sont donc tenus secrets. À notre niveau, il s'agit donc d'une boîte noire à laquelle nous n'avons pas accès.

La seule chose que le client décide est la portion de son traffic qu'il fait transiter par le *CDN*. Beaucoup de sites Web devant gérer un grand traffic ont recours à plusieurs *CDN*: on parle d'architecture *multi-CDN*. Il existe trois raisons principales de recourir à une architecture *multi-CDN*: limiter le risque d'indisponibilité du site en reportant le traffic d'un *CDN* en panne sur l'autre, exploiter les spécificités de chaque *CDN*, et augmenter la couverture géographique. Dans ce cas, se pose la question de la répartition du traffic sur les différents *CDN*. Autrement dit, un client ayant opté pour une infrastructure *multi-CDN* a besoin de mettre en place un algorithme de *load-balacing* pour répartir optimalement son audience entre ses différents *CDN*. Citrix est une société spécialisée précisément dans cette activité de *load-balacing* sur les architectures *multi-CDN*, on parle alors de *load-balancer*.

Afin d'effectuer ce *load-balacing*, Citrix doit être en mesure de comparer $K \in \mathbb{N}$ *CDN*. Sans perte de généralités, nous nous limiterons dans cette discussion à $K = 2$, dans la mesure où, comme nous le verrons, le critère de comparaison des *CDN* est transitif: si le *CDN* A est meilleur que B, et B est meilleur que C, alors A est meilleur que C. Le critère de performance sur lequel nous nous attarderons jusqu'à la fin de ce manuscrit est la latence.

Nous décrivons maintenant comment Citrix procède pour évaluer la latence des différents *CDN*. Citrix loue aux différents *CDN* présents sur le marché des centaines de *maps* correspondant à leurs différentes offres commerciales: haute performance, entrée de gamme, *map* spécialisée pour le contenu vidéo, etc. Citrix déploie sur chacune de ces *maps* un unique objet, appelé *objet test*. Ces *maps* sont dites *Publiques*. Lorsqu'un utilisateur est connecté au site d'un client de Citrix, plusieurs *maps Publiques* sont sélectionnées aléatoirement et l'*objet test* qu'elles hébergent est téléchargé afin d'effectuer des mesures de latence. Ces mesures s'effectuent uniquement lorsque toutes les ressources de la page consultée par l'utilisateur sont téléchargées afin de ne pas interférer avec le chargement de la page elle-même. Citrix collecte alors ces mesures ainsi que des informations relatives à la géolocalisation de l'utilisateur et son fournisseur d'accès à Internet. De part le grand nombre de clients répartis dans le monde, Citrix peut alors mesurer en temps réel la latence des différentes *maps* des *CDN* en chaque point du globe et pour chaque fournisseur d'accès à Internet. Les *maps Publiques* sont ainsi nommées car tout utilisateur d'un site de tout client de Citrix peut faire une mesure sur une *map Publique*.

Les *maps Publiques* se différencient des *maps Privées* qui correspondent aux *maps* souscrites par les clients de Citrix aux *CDN*. Avec l'accord du client, Citrix déploie l'*objet test* sur les *maps Privées* de ses clients pour effectuer des tests de latence selon un protocole identique aux mesures effectuées sur les *maps Publiques*. Comme les *maps Privées* sont la propriété du client, seuls les visiteurs du site du client en question peuvent faire ces mesures. Elles sont dites *Privées* pour cette raison: un utilisateur du site A ne peut pas faire une mesure de latence du site B sans l'accord de ce dernier. On parle de *mesures Publiques* (resp. *Privées*) lorsque l'*objet test* est hébergé sur une *map Publique* (resp. *Privée*). Voir Table 1.1.

Pour les raisons expliquées au début de cette introduction, des différences peuvent exister dans la distribution des *mesures Publiques* et *Privées*. Lorsqu'un utilisateur cherche à accéder au site

	map Publique	*map Privée*
Qui loue la *map* au *CDN* ?	Citrix	Le client
Quels objets y sont déployés ?	L'*objet test*	L'*objet test* + tous les objets qui composent le site du client (textes, images, feuilles de style etc.)
Qui peut télécharger le contenu de cette *map*?	Tous les utilisateur des sites des clients de Citrix	Uniquement les utilisateurs du site du client
Quels objets sont utilisés pour les tests de latence ?	L'*objet test*	L'*objet test*

Table 1.1: *Différences entre maps Publiques et maps Privées.*

d'un client possédant une infrastructure *multi-CDN*, Citrix identifie les *maps Privées* du client souscrites chez les différents *CDN*, dits en compétition, et utilise préférentiellement les mesures *Privées* effectuées sur ces *maps* par les autres utilisateurs du site. À l'aide de ces mesures, une prediction du *CDN* le plus rapide parmi ceux en compétition est effectuée. Le *CDN* sélectionné est celui avec la plus petite latence prédite. Dans certain cas, le site en question ne fournit pas assez de mesures, et Citrix utilise alors les *mesures Publiques* pour effectuer la prédiction. Ces *mesures Publiques* agissent comme un *proxy* pour les *mesures Privées* du client. Lorsque l'utilisateur a chargé sa page, il peut alors commencer les tests de latence mentionnés ci-avant. Deux grandes problématiques se dégagent, et forment le coeur des deux premiers chapitres de cette thèse:

1) Les *maps* d'un *CDN* donné ne sont jamais explicitement révélées. Ainsi, les *maps Publiques* de Citrix peuvent ne pas coïncider exactement avec les *maps Privées* de ses clients. Pouvons nous caractériser la qualité de ce *proxy*? Nous décrivons dans le premier chapitre de cette thèse une méthode d'ajustement distributionnel afin de modéliser et estimer le lien entre les distributions de mesures *Publiques* et *Privées*.

2) Dans un second temps, nous nous intéressons à la modélisation des données de latence elle-mêmes. Dans le second chapitre, nous décrivons l'objet d'intérêt dans l'industrie, à savoir le processus de latence médian à l'échelle temporelle $\Delta > 0$, c'est-à-dire la série temporelle régulière obtenue après le calcul de la médiane des mesures de latence obtenues sur chaque élément d'une partition régulière de taille Δ de l'intervalle de temps $[0, T]$. Nous proposons une nouvelle méthode de modélisation de ce processus, et analysons sa performance prédictive sur données réelles. L'importance de Δ est discutée. Enfin, nous introduisons un nouvel outil de mesure de predictabilité basé sur un certain critère entropique.

Dans le troisième chapitre, nous nous intéressons à la détection de pannes dans une certaine

sous-classe de réseaux. Pour les sites hébergés chez des *CDN*, les pannes sont très rares. Cela est du au nombre important de serveurs contenus dans un *CDN*. Lorsqu'un serveur cesse de fonctionner, le reste de l'infrastructure absorbe sans conséquences le traffic normalement assuré par le serveur défaillant. Dans le cas d'une infrastructure *personnelle*, c'est à dire une infrastructure physiquement possédée par le client par opposition à une infrastructure louée à un *CDN*, le nombre de serveur est souvent plusieurs ordres de grandeur inférieur à celui d'un *CDN*: l'impact d'une panne est donc plus important. Malgré sa plus grande fragilité, une infrastructure *personnelle* de haut de gamme fourni généralement une performance plus stable qu'un *CDN*. En allouant une *bande passante* suffisamment élevée à son infrastructure *personnelle*, un site Internet peut garantir une latence quasiment constante dans le temps. Ce surplus de *bande passante* est un luxe que le *CDN* ne peut généralement pas s'offrir si bien que la latence d'un *CDN* évolue au gré de l'affluence sur le site: un nombre élevé de connexions s'accompagne généralement d'une hausse de la latence, et réciproquement. La problématique qui nous intéresse dans ce troisième chapitre concerne ces infrastructures *personnelles*:

3) Dans les réseaux *personnels* haut de gamme, caractérisés par des mesures de latence plus stables, exposés au risque de panne, comment mettre en place un système de détection online de changement dans la distribution des mesures de latence? Dans un second temps, nous nous interrogeons sur comment exploiter cette stabilité des mesures de latence pour diminuer la taille des échantillons d'entraînement des algorithmes prédictifs sans détériorer la précision. L'intérêt est principalement économique: traiter moins de données représente un coût opérationnel plus faible.

1.3 Estimation d'un transport entre distributions de probabilités

Comme brièvement introduit dans la section 1.2, les mesures *Privées* et *Publiques* de latence d'une *map* donnée chez un *CDN* donné peuvent être différentes. Les ingénieurs au sein de l'entreprise pensent que les *maps Privées* et *Publiques* partagent des propriétés communes fortes. La caractérisation de la dépendance statistique entre mesures *Privées* et *Publiques* représente un intérêt stratégique pour Citrix. Lorsqu'une prédiction est effectuée en utilisant les mesures *Publiques*, la prédiction est biaisée. Pouvoir corriger ce biais représente un bon levier pour améliorer les prédictions du *CDN* le plus performant. Nous considérons une période de temps $[0, T]$ au cours de laquelle nous observons des données de latence *Privées* et *Publiques* issues d'une même map, qu'on appellera plus généralement *source* et *proxy* respectivement. La distribution des mesures de latence évolue très lentement au cours de la journée, si bien qu'il est raisonnable de considérer l'existence d'une partition $(t_i)_{0\leq i\leq K}$ de $[0, T]$ où $t_0 = 0$, $t_K = T$, $t_{i+1} - t_i = h > 0$ $\forall 0 \leq i \leq K-1$ telle que les mesures de latence du *proxy* et de la *source* peuvent être idéalisées comme des échantillons indépendents et identiquement distribués (i.i.d.) sur chaque élément de la partition $]t_i, t_{i+1}]$. Dans ce contexte, nous nous intéressons au problème de l'inférence de la dépendance statistique entre le *proxy* et la *source* lorsqu'on observe des paires d'échantillons i.i.d.

Formellement, nous supposons l'existence d'une transformation déterministe $f : \mathbb{R} \to \mathbb{R}$ qui transporte la distribution du *proxy* vers la la distribution de la *source*. Soit $M \in \mathbb{N}$, et $(P_i)_{1\leq i\leq M}$ des lois de probabilités où les P_i sont tirées aléatoirement de manière indépendante selon une mesure $\mathbb{M}$ à valeurs dans un certain ensemble $\mathcal{Z}$. Le problème d'ajustement auquel nous nous

intéressons dans ce chapitre est le suivant:

$$\text{trouver } f \text{ tel que: } Q_i = f_{\#}P_i \quad \forall 1 \leq i \leq M$$

où $f_{\#}P_i$ désigne la mesure push-forward de P_i sous f, c'est à dire que pour tout ensemble mesurable A on a

$$f_{\#}P_i(A) := P_i(f^{-1}(A)).$$

Dans le cadre de ce modèle, nous supposons que les lois P_i et Q_i ne sont pas observées, seuls des échantillons $(\mathcal{X}_1, \mathcal{Y}_1), \ldots, (\mathcal{X}_M, \mathcal{Y}_M)$, où $\mathcal{X}_i$ est un échantillon indépendant de loi P_i et , indépendamment, $\mathcal{Y}_i$ est un échantillon indépendant de loi Q_i, sont observés. Chaque couple $(\mathcal{X}_i, \mathcal{Y}_i)$ est appelé un *batch*. Le choix de ce modèle repose sur le fait que les données que nous observons ne sont pas appareillées, comme ce serait le cas dans un problème de regression classique, de sorte que l'hypothèse d'un lien statistique entre les mesures ne peut être formulé qu'au niveau distributionnel. L'hypothèse d'un transport entre les distributions de la *source* et du *proxy* est alors naturelle.

Il est aisé de se convaincre que f ne peut pas être estimée raisonnablement sur tout son domaine de définition. Supposons f croissante et intéressons nous à l'estimation de $f(x)$ avec $x \in \mathbb{R}$. Si l'ensemble $\mathcal{Z}$ contient une mesure P telle que $P(\mathcal{R}) = 0$ où $\mathcal{R}$ est un voisinage de x, alors il suit immédiatement de l'égalité $Q = f_{\#}P$ et par croissance de f que $Q(f(\mathcal{R})) = 0$. Autrement dit, l'observation d'échantillons distribuées selon P et Q ne pourra jamais nous renseigner sur la valeur de f en x. Il est donc nécessaire d'imposer des contraintes locales et d'uniformité sur les mesures contenues dans $\mathcal{Z}$ autour du point d'estimation. Formellement, pour estimer $f(x)$ avec $x \in \mathbb{R}$, on demande que $\mathcal{Z}$ vérifie

$$\mathcal{Z} = \Big\{ P(du) = p(u)du | p \in \mathcal{F}_{\mathcal{R}}(A, B, C) \Big\}$$

où

$$\mathcal{F}_{\mathcal{R}}(A, B, C) = \Big\{ p \in \mathcal{C}^1(\mathcal{R}), p \geq 0, \int p(u)du = 1, A \leq p(u) \leq B, |p'(u)| \leq C, \text{ pour tout } u \in \mathcal{R} \Big\}$$

avec A, B, C trois réels positifs et $\mathcal{R}$ un intervalle contenant x. Autrement dit, $\mathcal{F}_{\mathcal{R}}(A, B, C)$ est un ensemble de densités de probabilités différentiables sur $\mathcal{R}$, minorée et majorée sur $\mathcal{R}$ et dont la dérivée est continue et bornée sur $\mathcal{R}$.

Le processus de *load-balacing* reposant intensivement sur l'estimation des quantiles de la distribution des mesures de latence, nous cherchons par cette représentation un moyen d'ajuster les quantiles de la *source* aux quantiles du *proxy*. On peut observer immédiatement que si f n'est pas croissante, cet objectif est irréalisable. Pour s'en convaincre, soit $P \in \mathcal{Z}$ et $Q = f_{\#}P$. Notons F, G les fonctions de répartition de P, Q respectivement, supposées inversibles, et soient X, Y des variables aléatoires de lois respectives P, Q. Notons que l'égalité

$$Q = f_{\#}P$$

implique que $f(X)$ et Y sont identiquement distribuées. Ajuster les quantiles de F aux quantiles de G revient à trouver une fonction h vérifiant

$$h \circ F^{-1}(\alpha) = G^{-1}(\alpha)$$

pour tout $P \in \mathcal{Z}$. Autrement dit

$$h = G^{-1} \circ F.$$

La fonction h recherchée est nécessairement croissante sur $\mathcal{R}$ et $G^{-1} \circ F$ est indépendante du choix de $P \in \mathcal{Z}$. Si f est strictement croissante, on pourra se convaincre aisément que $h = f$ et que $G^{-1} \circ F$ est bien indépendante du choix de $P \in \mathcal{Z}$. Au contraire, si f n'est pas croissante, une telle fonction h ne peut exister en général. Supposons par exemple que $f\colon x \mapsto |x|$, et choisissons $P \in \mathcal{Z}$ symétrique. Alors pour un certain $x \in \mathcal{R}$, nous avons par un calcul élémentaire

$$G(x) = 2F(x) - 1.$$

Or les applications F et $x \mapsto 2x + 1$ sont inversibles et on obtient alors l'égalité:

$$G^{-1}(\alpha) = F^{-1}\left(\frac{\alpha + 1}{2}\right).$$

pour tout $\alpha \in (0, 1)$. L'ajustement des quantiles de F aux quantiles de G est alors donné par la fonction h vérifiant

$$\begin{aligned} h(x) &= G^{-1} \circ F(x) \\ &= F^{-1}\left(\frac{F(x) + 1}{2}\right). \end{aligned}$$

Clairement, cette expression dépend du choix de P, et il n'existe alors aucune fonction h vérifiant $h \circ F^{-1} = G^{-1}$ pour tout $P \in \mathcal{Z}$. Nous posons les questions suivantes:

> **Question 1** Sous l'hypothèse d'un transport croissant, comment estimer ponctuellement f? En quel sens et à quelle vitesse?
>
> **Question 2** Comment vérifier en pratique que f est croissante?
>
> **Question 3** Transporter les données du *proxy* par notre estimateur améliore-t-il la prédiction du meilleur *CDN*?

La question 1 pose les bases du problème de recherche de lien entre le *proxy* et la *source*. Nous supposons f croissante. Notre estimateur pour f prend la forme:

$$\hat{f}_{n,M}(x) = \frac{1}{M} \sum_{i=1}^{M} (\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) \tag{1.1}$$

où:

- $\widehat{F_n^i}(x) = \frac{1}{n} \sum_{l=1}^{n} \mathbf{1}_{\{X_l^i \leq x\}}$
- $\widehat{G_n^i}(x) = \frac{1}{n} \sum_{l=1}^{n} \mathbf{1}_{\{Y_l^i \leq x\}}$
- $(\widehat{G_n^i})^{-1}(x_0) = \inf\{x \in \mathbb{R}, (\widehat{G_n^i})(x) \geq x_0\}$

sont respectivement les fonctions de répartitions empiriques et la fonction quantile empirique des échantillons $(X_1^i, \dots, X_n^i)$ i.i.d. P_i et $(Y_1^i, \dots, Y_n^i)$ i.i.d. $Q_i = f_\# P_i$. La forme de cet estimateur est motivée par le théorème d'inversion: si X est une variable aléatoire réelle de distribution F, alors $F^{-1}(X) \sim \mathcal{U}(0,1)$, voir par exemple [29]. Ce résultat permet de construire un exemple explicite de transport entre deux mesures de probabilités. En effet, soient X, Y deux variables aléatoires réelles de distribution F, G respectivement. Alors $Y \overset{d}{=} G^{-1} \circ F(X)$, et il suit $\mathcal{L}(Y) = f_\# \mathcal{L}(X)$ où $f = G^{-1} \circ F$. Ce couplage est souvent appelé *réarrangement croissant*, voir par exemple [90]. Nous avons le résultat principal suivant:

Theorem 1. *Soient $M, n \in \mathbb{N}$ des entiers positifs et $\hat{f}_{n,M}(x)$ définie en (1.1). Soient $a < b$ et $\delta > 0$ tel que $\delta < (b-a)/2$. Posons $\mathcal{R} = [a,b]$, $\mathcal{R}_\delta = [a+\delta, b-\delta]$. Supposons f croissante, deux fois différentiables et $\|1/f'\|_{L^\infty(\mathcal{R})} < \infty$. Si de plus $\mathbb{E}(Y^2) < \infty$ où $Y \sim Q = f_\# P$ et $P \sim \mathbb{M}$, alors*

$$\left| \hat{f}_{n,M}(x) - f(x) \right| \lesssim \frac{1}{\sqrt{Mn}} + \frac{\log(n)^{3/2}}{n^{3/4}} + M e^{-n\delta^2 A^2/2} \tag{1.2}$$

en probabilité, uniformément en $x \in \mathcal{R}_\delta$ et où $\lesssim$ est l'inégalité à une constante positive près ne dépendant que des constantes $A, B, C, \|f'\|_{L^\infty(\mathcal{R})}, \|f''\|_{L^\infty(\mathcal{R})}, \|1/f'\|_{L^\infty(\mathcal{R})}$ et $\mathbb{E}(Y^2)$.

La croissance de la fonction f posée en question 2 est essentielle pour espérer l'estimer. Il est aisé de se convaincre que si f est croissante stricte et continue, alors pour tout $P \in \mathcal{Z}$, $G^{-1} \circ F(x)$ ne dépend pas du choix de P. Par contraposition, s'il existe 2 distributions $P_1, P_2 \in \mathcal{Z}$ telles que $G_1^{-1} \circ F_1 \neq G_2^{-1} \circ F_2$ où F_1, F_2, G_1, G_2 sont les fonctions de répartition de $P_1, P_2, f_\# P_1, f_\# P_2$ respectivement, alors f n'est pas croissante. La réciproque n'est en général pas vraie cependant. Par exemple, dans le cas où f: $x \mapsto -x$ et $\mathcal{Z}$ contient uniquement des distributions symétriques alors $G^{-1} \circ F$ est indépendante du choix de P et pourtant f n'est pas croissante. En effet dans ce cas, si F, G dénotent les fonctions de répartition de $P \in \mathcal{Z}$ et $Q = f_\# Q$ et $X \sim P$, $Y \sim Q$ alors pour tout x on a:

$$G(x) = \mathbb{P}(f(X) \leq x) = F(x)$$

si bien que $h = G^{-1} \circ F$ est égale à l'identité et donc indépendante de P, mais $f \neq h$ n'est pas croissante. Néanmoins nous avons la proposition suivante:

Proposition 1. *Soit $f \in \mathcal{C}^2(\mathbb{R})$, $P \in \mathcal{Z}$, et on note F, G les distributions de P et $f_\# P$ respectivement. Si la fonction*

$$h = G^{-1} \circ F$$

est indépendante du choix de P, alors

$$f_\# P = h_\# P$$

pour tout $P \in \mathcal{Z}$ et notre estimateur converge vers h comme dans le théorème (2.2).

Le fait de converger vers h au lieu de f dans ce cas est sans conséquence car dans la mesure où $f_\# P = h_\# P$ pour tout $P \in \mathcal{Z}$, les deux fonctions jouent des rôles indifférenciables. La Proposition 1 suggère d'introduire la relation d'équivalence $\sim$ sur l'ensemble des fonctions $\mathcal{C}^2(\mathbb{R})$ vérifiant:

$$f \sim h \text{ si et seulement si } f_\# P = h_\# P \text{ pour tout } P \in \mathcal{Z}.$$

De cette manière, s'il existe h croissante telle que $h \in [f] = \{h \in \mathcal{C}^2(\mathbb{R}) | h \sim f\}$, alors nous garantissons que notre estimateur converge vers h, et l'existence de ce h est donnée par l'indépendance de $G^{-1} \circ F$ en $P \in \mathcal{Z}$ où F, G sont les fonctions de répartition de P et $f_{\#}P$ respectivement. Pour une collection $P_1, \dots, P_M \in \mathcal{Z}$ et un certain $x \in \mathbb{R}$, nous proposons un test d'égalité des $G_i^{-1} \circ F_i(x)$ qui est une adaptation du test de Wald [45] pour l'égalité de paramètres multiples en utilisant une caractérisation appropriée de la convergence en loi du processus $\sqrt{n}(G_n \circ F_n(x) - G^{-1} \circ F(x))$ en terme d'équicontinuité stochastique. L'objet de la proposition suivante est de tester

$$\begin{aligned} H_0^x &: G_1^{-1} \circ F_1(x) = \cdots = G_M^{-1} \circ F_M(x) \\ \text{contre } H_1^x &: \exists\, i \neq j, G_i^{-1} \circ F_i(x) \neq G_j^{-1} \circ F_j(x). \end{aligned}$$

Proposition 2. *Posons:*

$$\boldsymbol{\theta}_x = \begin{bmatrix} G_1^{-1} \circ F_1(x) \\ G_2^{-1} \circ F_2(x) \\ \vdots \\ G_M^{-1} \circ F_M(x) \end{bmatrix}$$

$$\boldsymbol{\theta}_{n,x} = \begin{bmatrix} (G_n^1)^{-1} \circ F_n^1(x) \\ (G_n^2)^{-1} \circ F_n^2(x) \\ \vdots \\ (G_n^M)^{-1} \circ F_n^M(x) \end{bmatrix}$$

$$\boldsymbol{\Sigma_n} = \begin{bmatrix} \sigma_n(F_1)^2 & & \\ & \ddots & \\ & & \sigma_n(F_M)^2 \end{bmatrix}$$

et

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 & & \\ \vdots & & \ddots & \\ 1 & & & -1 \end{bmatrix}$$

où $\boldsymbol{R}$ est une matrice $(M-1) \times M$ dont la première colonne est composée uniquement de 1 et où les autres colonnes forment une matrice carrée de taille $(M-1) \times (M-1)$ où chaque élément diagonal vaut -1 et 0 partout ailleurs. Enfin, soit

$$\sigma_n(F)^2 = 2 \frac{F_n(x)(1 - F_n(x))}{\left(g_n \circ G_n^{-1} \circ F_n(x)\right)^2}$$

avec g_n un estimateur à noyau de la densité de G, alors:

$$W_n = n \left\| \left(\boldsymbol{R \Sigma_n R}^t\right)^{-1/2} \boldsymbol{R \theta}_{n,x} \right\|^2 \xrightarrow{d} \chi^2(M-1)$$

sous H_0^x lorsque $n \to \infty$, où $\chi^2(M-1)$ désigne la distribution du χ^2 à $M-1$ degrés de libertés. Ainsi, pour $\alpha \in (0,1)$

$$\Phi(\alpha) = \mathbf{1}_{\mathcal{R}(\alpha)}$$

est un test asymptotique de niveau α *de* H_0^x *avec zone de rejet* $\mathcal{R}(\alpha) = \{W_n > q_\alpha^{\chi^2(M-1)}\}$ *où* $q_\alpha^{\chi^2(M-1)}$ *est le quantile d'ordre* α *de* $\chi^2(M-1)$.

L'implémentation de l'estimateur est faite sur des données réelles correspondant à 3 jours de mesures de latence d'une *source* et d'un *proxy* avec $M = 46$ couples d'échantillons, chacun contenant $n = 2082 \approx M^2$ mesures de latence. Le test de H_0^x est effectué pour $x \in \{20, \dots, 150\}$. Cette gamme de valeurs est sélectionnée pour des raisons de stabilité d'estimation. Les résultats sur données réelles sont présentées en Figure 1.2 et l'estimateur est représenté en Figure 1.3. Pour

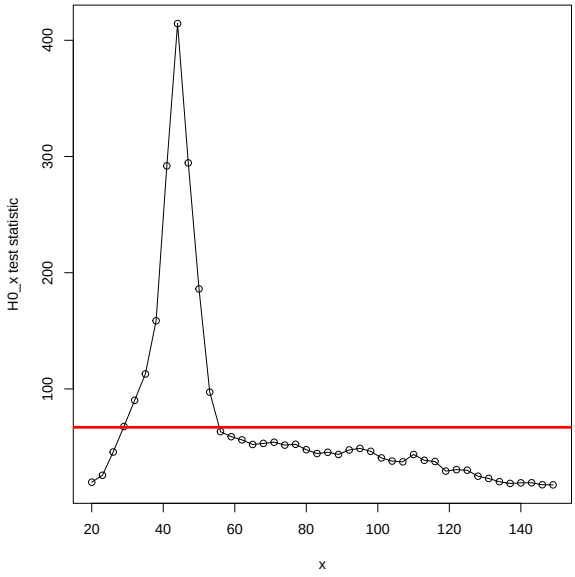

Figure 1.2: *Statistique du test de* H_0^x *contre* x *avec la valeur critique* $q_{1-\alpha}^{\chi^2}(M)$ *en rouge. Le test rejette* H_0^x *sur l'intervalle* $[35, 55]$, *suggérant que les* $G_i^{-1} \circ F_i(x)$ *prennent des valeurs différentes sur cet intervalle. L'hypothèse d'un transport croissant entre la distribution du proxy et de la source sur cet intervalle doit être rejetée.*

répondre à la question 3, l'évaluation de la qualité du transport des données du *proxy* vers la *source* à l'aide de notre estimateur est effectuée par comparaison des distances de Wasserstein entre les distributions des mesures de la *source* et du *proxy* avec et sans le transport. Le choix de la distance de Wasserstein est motivé pour son lien avec les quantiles des distributions et sa connexion avec le transport optimal [90], [36]. L'utilisation de notre estimateur permet une réduction très importante de la distance entre la distribution du *proxy* et la distribution de la *source*, voir Figure 1.4 et Table 1.2.

1.4 Modélisation des données de latence

Dans ce second chapitre, nous nous intéressons à la description des données de latence, et à la modélisation de l'objet d'intérêt dans l'industrie pour la prédiction: le processus médian. Il existe

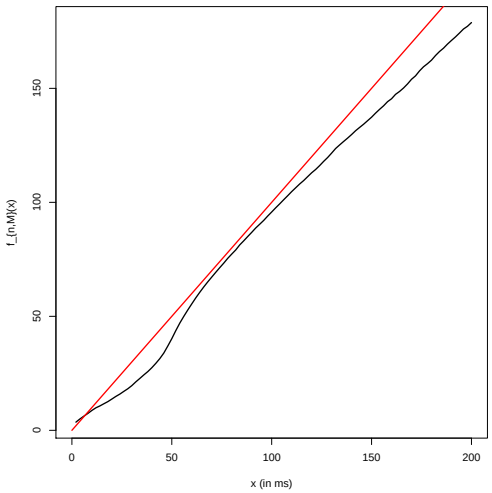

Figure 1.3: $\hat{f}_{n,M}$ *en noir plotté contre l'identité en rouge.*

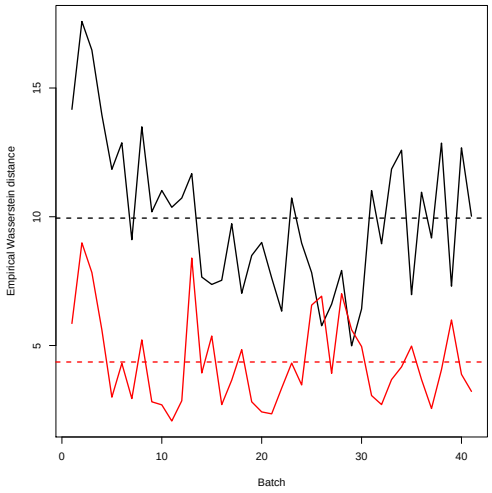

Figure 1.4: *Distance de Wasserstein calculée sur 42 batchs de test entre la source et le proxy en noir, et entre la source et le proxy transporté par* $\hat{f}_{n,M}$ *en rouge. Les lignes pointillées indiquent les moyennes respectives.*

a priori deux solutions pour aiguiller les individus dans le réseau: l'approche collective et l'approche individuelle. Afin de décrire brièvement ces deux approches, nous nous plaçons dans le cas où un individu I se connecte à un site Web dont le contenu peut être fourni par les *CDN* $C_1, \ldots, C_K$. Ces *CDN* sont dits en concurrence.

$\widehat{W}_1(\mathcal{X}, \mathcal{Y})$	$\widehat{W}_1(\hat{f}_{n,M}(\mathcal{X}), \mathcal{Y})$	Changement relatif	$\widehat{W}_1(\hat{f}_{n,M}(\mathcal{X}), \mathcal{Y}) < \widehat{W}_1(\mathcal{X}, \mathcal{Y})$
9.95	4.36	-56%	95%

Table 1.2: *Distance de Wasserstein empirique entre la source et le proxy (colonne 1), Distance de Wasserstein empirique entre la source et le proxy transporté par notre estimateur $\hat{f}_n$ (colonne 2), réduction relative due au transport du proxy (colonne 3) et pourcentage de batchs pour lesquels une réduction absolue a été reportée (colonne 4).*

Dans l'approche collective le *load-balancer* partitionne géographiquement la surface du globe, et génère des prédictions de la latence à venir des *CDN* $C_1, \dots C_K$ en fonction de la localisation et du fournisseur d'accès à Internet, en utilisant les données de latence des *CDN* $C_1, \dots C_K$ générées par les autres utilisateurs du réseau. Lorsque le nouvel individu I se présente, le *load-balancer* identifie son fournisseur d'accès à Internet ainsi qu'un voisinage géographique pour lequel il a en base de données des prédictions pour les *CDN* $C_1, \dots C_K$. Le *load-balancer* classe alors par ordre croissant les prédictions de latence des *CDN* en concurrence et sélectionne le *CDN* ayant la valeur prédite la plus faible. Les prédictions étant effectuées en amont, le temps nécessaire pour aiguiller un utilisateur dans l'approche collective se limite à cette requête des prédictions pour chaque *CDN*, qui est de l'ordre de quelques millisecondes.

L'approche individuelle repose sur l'idée de faire tester la latence des *CDN* $C_1 \dots, C_K$ directement à l'utilisateur I et de sélectionner le *CDN* ayant fourni la mesure la plus faible. Il n'est plus question ici d'utiliser des mesures d'utilisateurs proches en temps et en espace, ni d'utiliser des méthodes prédictives. Le temps nécessaire pour effectuer ces tests de latence est de l'ordre de quelques secondes, même dans le cas $K = 2$. Ce temps non négligeable est précisément la raison pour laquelle les tests de latence des utilisateurs sont effectués après le chargement de la page, comme décrit dans la section 1.2.

L'approche individuelle peut paraître plus attractive que l'approche collective car elle garantit de toujours sélectionner le *CDN* le plus rapide. Néanmoins, les trois ordres de grandeurs de différence dans le temps nécessaire pour choisir un *CDN* entre les deux méthodes la rendent totalement inopérante en pratique. En effet, même en cas d'erreur de prédiction dans l'approche collective, il est très improbable que l'utilisateur ait chargé la page Web plus vite par la méthode individuelle. L'approche collective est donc largement privilégiée.

La nature intrinsèquement distributionnelle de cette approche ne permet pas de définir de manière absolue le *CDN* avec la plus faible latence, tout comme il est impossible de parler dans l'absolu d'un rythme cardiaque plus faible qu'un autre. Parce qu'il n'existe pas de relation d'ordre totale sur l'ensemble des mesures de probabilité, le critère doit reposer sur la comparaison d'une certaine fonctionnelle. Le choix le plus répandu dans l'industrie est la comparaison des quantiles, et plus précisément, la médiane, des mesures de latence.

Soit $T > 0$ et $[0, T]$ un intervalle de temps. Les horodatages, ou *timestamps* en anglais, des mesures collectées par Citrix sont arrondies à la seconde, donc une structure naturelle pour le processus générateur des données est celle d'un processus stochastique à temps discret avec l'indice

de temps exprimé en secondes. Formellement, nous observons le processus $Z = \{Z_t | t \in \{0, \dots, T\}\}$, où Z_t est le processus empirique défini par

$$Z_t = \frac{1}{N_t} \sum_{k=1}^{N_t} \delta_{Y_k^t}$$

où δ_x est la mesure de Dirac en $x \in \mathbb{R}$, N_t est le nombre (aléatoire) de mesures de latence collectées à l'instant t et $(Y_k^t)_{K \in \{1,\dots,N_t\}}$ sont les mesures de latence reçues à l'instant t. Un exemple de données récoltées sur 8 secondes est présenté en Figure 1.5. L'objet d'intérêt dans l'industrie n'est pas Z,

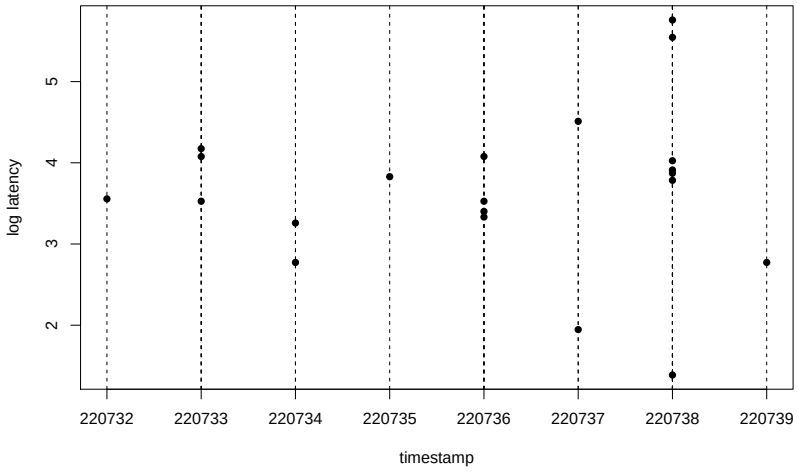

Figure 1.5: *Réalisation du processus Z sur une fenêtre de 8 secondes.*

mais une fonctionnelle de Z. Pour un certain $\Delta > 0$, on se donne $n \in \mathbb{N}$, $t_n = n\Delta$ et on définit la série temporelle régulière suivante:

$$X_{t_n}^{\Delta} = \text{Médiane} \left(Y_k^t\right)_{\substack{t \in]t_{n-1}, t_n] \\ k \in \{1,\dots,N_t\}}}.$$

$X_{t_n}^{\Delta}$ est la médiane de toutes les mesures reçues dans l'intervalle $]t_{n-1}, t_n]$. $(X_{t_n}^{\Delta})_{0 \leq n\Delta \leq T}$ est appelé le processus médian à la fréquence Δ. Dans ce contexte, nous posons le problème de la modélisation et de la prédiction du processus médian.

Le paramètre Δ joue un rôle primordial. À mesure que Δ diminue, la variance du processus médian explose, voir Figure 1.6. Pour des valeurs faibles, de l'ordre de la minute ou moins, les effets saisonniers et d'autocorrelation notamment diminuent voire disparaissent. Au contraire, lorsque Δ augmente, la variance du processus médian diminue, et un signal périodique clair révélant les cycles d'activité jour/nuit sur Internet devient aisément identifiable. Cela suggère que pour de grandes valeures du paramètre Δ, le processus médian peut être prédit avec précision, alors que pour de petites valeurs du paramètre Δ, le rapport signal sur bruit est trop faible et le processus médian n'est pas prédictible. Ceci est réminiscent des phénomènes de bruit de microstructure en

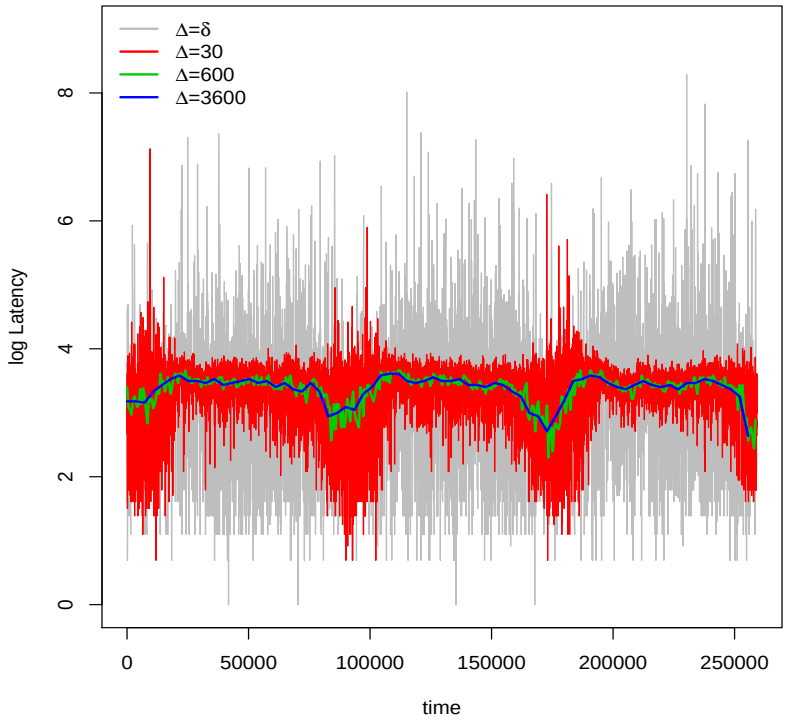

Figure 1.6: *3 jours de données de latence consécutives agrégées à 4 niveaux différents: 1s, 30s, 1min et 1h.*

finance [44] [5] [75] [76].

Le paramètre Δ représente l'horizon de prédiction : si t_n est l'instant présent, $X^{\Delta}_{t_n}$ représente la médiane des dernières mesures de latence reçues dans l'intervalle $]t_{n-1}, t_n]$, c'est à dire la médiane des mesures reçues au cours des dernières Δ secondes par définition de t_n. Ainsi, la prédiction du processus pour la période suivante, notée $\widehat{X^{\Delta}_{t_{n+1}}}$, représente donc une estimation de la latence des mesures qui seront reçues dans l'intervalle $]t_n, t_{n+1}]$, c'est à dire dans les Δ prochaines secondes. Dans une optique de *load-balacing*, il est important de faire des prédictions avec l'horizon de temps le plus court possible. En effet, à l'instant t_n, lorsqu'un utilisateur doit être aiguillé vers A ou B, savoir que A sera plus performant que B dans $\Delta = 2$ heures n'a généralement aucune valeur. Nous posons les questions suivantes :

Question 1 Quel modèle pour le processus médian à l'échelle Δ ?

Question 2 Comment se comporte ce modèle vis à vis des modèles de bases utilisés dans l'industrie ?

Question 3 Comment quantifier la quantité d'information résiduelle non captée par le

modèle ?

Nous mettons en évidence dans ce chapitre une forte dynamique de la moyenne et la variance conditionnelle du processus médian. L'analyse spectrale du processus à l'échelle Δ révèle des périodes significatives à 8,12 et 24 heures et suggère donc une décomposition de Fourier à 3 regresseurs, voir Figure 1.7.

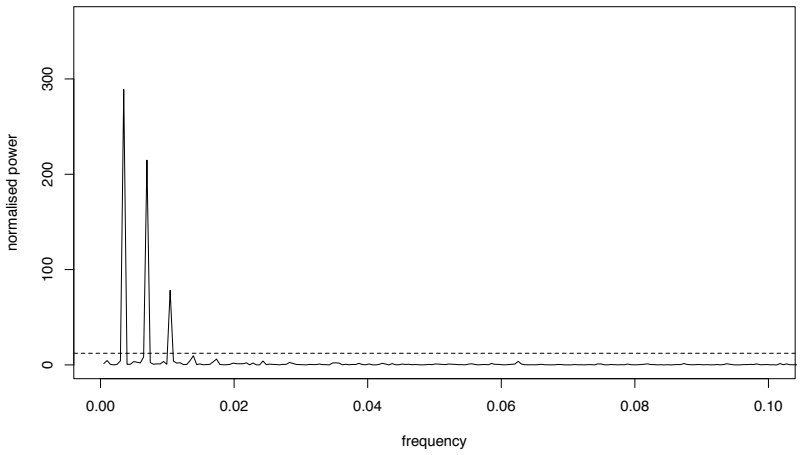

Figure 1.7: *Estimation de la densité spectrale du processus médian,* $\Delta = 300$.

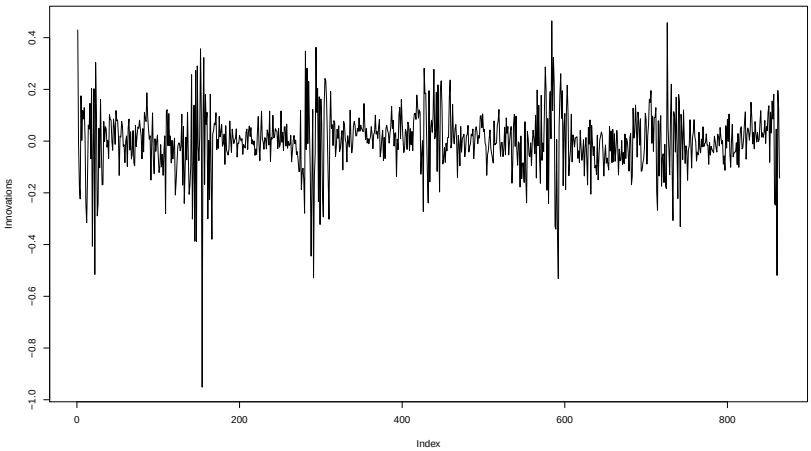

Figure 1.8: *Innovations de la décomposition de Fourier sur un échantillon test comportant 6 jours de données.* $\Delta = 600$ *secondes.*

Les résidus de cette décomposition ne présentent plus de saisonnalité mais une forte auto-

correlation à l'ordre deux et des clusters de volatilité, voir Figure 1.8. Ces clusters présentent des caractéristiques saisonnières qui peuvent être précisément décrites par un ARMA Seasonal-GARCH, c'est à dire un processus ARMA-GARCH avec l'ajout de composantes saisonnières dans la partie GARCH [81] [64] [49]. L'étude empirique du processus médian à travers les échelles nous conduit à adopter le modèle suivant pour X_n^Δ :

$$
\begin{aligned}
&X_n^\Delta = \mu + \sum_{k=1}^{K} \alpha_k \sin\left(\frac{2k\pi t_n}{\phi}\right) + \sum_{k=1}^{K} \beta_k \cos\left(\frac{2k\pi t_n}{\phi}\right) + \varepsilon_n \\
&\varepsilon_n = \nu + \kappa\varepsilon_{n-1} + u_n + \rho u_{n-1} \\
&u_n = \sigma_n z_n \\
&\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta\sigma_{n-1}^2 + \sum_{k=1}^{R} \lambda_{1,k} \left|\cos\left(\gamma_{1,k} + \frac{k\pi t_n}{\phi}\right)\right|^a + \lambda_{2,k} \left|\sin\left(\gamma_{2,k} + \frac{k\pi t_n}{\phi}\right)\right|^a \\
&(z_n)_{n\geq 1} \text{ i.i.d. de moyenne 0 et variance 1}
\end{aligned}
$$

où $K = R = 3$ est fixé et correspondent respectivement au nombre de périodes saisonnières du processus à l'ordre un et des innovations à l'ordre deux. Le paramètre ϕ est la période fondamentale du signal, correspondant à la longueur du plus grand cycle saisonnier, et n'est pas estimée. On fixe $\phi = 24$h, valeur suggérée par l'analyse spectrale. D'un point de vue pratique, ϕ correspond à la durée d'un cycle jour/nuit de l'activité sur Internet. ϕ, K, R sont constants à travers les échelles. Pour ne pas alourdir l'écriture, tous les paramètres du modèle à estimer, à savoir

$$\mu, (\alpha_k)_{1\leq k\leq K}, (\beta_k)_{1\leq k\leq K}$$

pour la moyenne conditionnelle et

$$\nu, \kappa, \rho, \omega, \alpha, \beta, (\lambda_{1,k})_{1\leq k\leq R}, (\lambda_{2,k})_{1\leq k\leq R}, (\gamma_{1,k})_{1\leq k\leq R}, (\gamma_{2,k})_{1\leq k\leq R}, a$$

pour la partie variance conditionnelle dépendent implicitement de Δ. La loi Skewed-Student [35] centrée réduite est choisie pour les z_n afin de tenir compte des queues lourdes et asymétriques des résidus. La loi Skewed-Student englobe plusieurs distributions connues comme la loi de Student, Normal ou Laplace, entre autres. Elle se caractérise par des paramètres de forme et d'asymétrie qui lui offre une grand flexibilité dans l'ajustement à de nombreuses données réelles, telles que financières [87]. Cette loi est absolument continue par rapport à la mesure de Lebesgue et admet pour densité de probabilité la fonction

$$f(x; m, \tau, \nu, \xi) = \frac{2s\sigma}{\tau(\xi + \xi^{-1})} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\pi\nu}} \left(1 + \frac{s^2}{\nu} \frac{(\frac{x-m}{\tau}\sigma + \mu)^2}{\xi^{2\text{sign}(\frac{x-m}{\tau}\sigma+\mu)}}\right)^{-\frac{\nu+1}{2}},$$

avec

$$
\begin{aligned}
m_1 &= \frac{2\sqrt{\nu-2}}{(\nu-1)B(\frac{1}{2}, \frac{\nu}{2})}, \\
\mu &= m_1(\xi - \xi^{-1}), \\
\sigma &= \sqrt{(1 - m_1^2)(\xi^2 + \xi^{-2}) + 2m_1^2 - 1}, \\
s &= \sqrt{\frac{\nu}{\nu-2}},
\end{aligned}
$$

où

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx, \quad a,b>0$$

est la fonction Beta et

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \quad x>0$$

est la fonction Gamma. Les paramètres $m \in \mathbb{R}$ and $\tau > 0$ représentent la moyenne et l'écart-type de la distribution, tandis que les paramètres $\nu > 2$ et $\xi > 0$ contrôlent la forme et l'asymétrie de la distribution. Le paramètre de forme ν contrôle l'épaisseur des queues de distributions, et le paramètre ξ contrôle l'orientation de l'asymétrie: gauche pour $\xi < 1$, droite pour $\xi > 1$, et symétrique pour $\xi = 1$.

Les coefficients de la décompositions de Fourier, qui caractérisent la moyenne conditionnelle, sont estimés par regression linéaire. Pour les coefficients qui caractérisent les innovations, on choisit pour l'estimation une approche par Quasi Maximum de Vraisemblance à une seule étape: les parties ARMA et Seasonal-GARCH sont estimées dans le même programme d'optimisation.

Les paramètres d'asymétrie et de forme utilisés dans les tests d'ajustement, notés désormais Sk(Δ) et Sh(Δ) respectivement, sont sélectionnés par apprentissage sur un jeu d'entraînement. En effet, une des propriétés remarquable du modèle est l'existence de régularités dans l'évolution de ces paramètres à travers les échelles, voir Figure 1.9. Des modèles logistique et linéaire sont utilisés

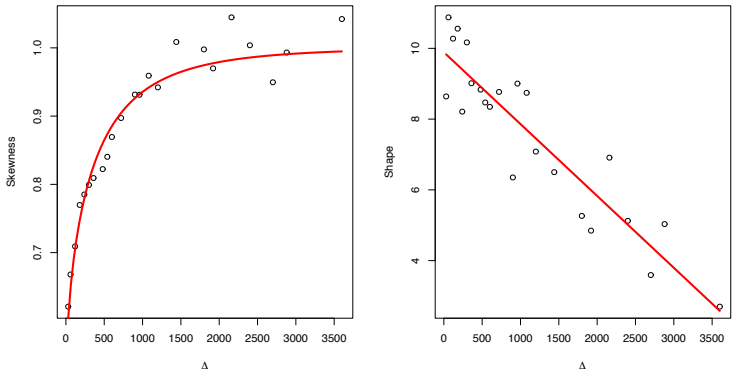

Figure 1.9: *Evolution des paramètres d'asymétrie et de forme des résidus du modèle sur un jeu d'entraînement en fonction de* Δ.

pour les paramètres d'asymétrie et de forme respectivement, suggérant:

$$\text{Sk}(\Delta) = \frac{1}{1+e^{-0.07\sqrt{\Delta}}}$$

et

$$\text{Sh}(\Delta) = 9.9 - 0.002\Delta$$

avec $R^2 = 0.89$ et 0.84 respectivement. L'ajustement du modèle pour pour X_n^Δ avec les paramètres $\mathrm{Sk}(\Delta)$ et $\mathrm{Sh}(\Delta)$ pour l'asymétrie et la forme des résidus aux données réelles fournit d'excellents résultats mesurés par des tests d'ajustement de Anderson Darling, Cramer-Von Mises et Kolmogorov Smirnov [4] [92] [22] [65] échouant simultanément à rejeter l'hypothèse d'adéquation à la loi Skewed-Student. On note de plus un échec de rejet d'auto-corrélation dans les résidus et résidus au carré, ainsi qu'un échec de rejet de la présence de cycles saisonniers dans les résidus.

Ce modèle nous apporte deux renseignements importants sur la nature du processus médian:

1. L'indépendance en Δ des paramètres de la moyenne conditionnelle pour des petites valeurs de Δ.
2. L'existence de 2 régimes distincts dans la variabilité des mesures.

Le premier point nous indique l'existence d'une dynamique dans la moyenne conditionnelle qui ne dépend pas de l'échelle Δ pour les valeurs testées, de l'ordre de l'heure au maximum. Ce fait remarquable nous montre l'extrême lenteur de l'évolution de la latence au cours d'une journée. Naturellement ce fait disparait si l'on considère un Δ de l'ordre d'une journée, soit la période fondamentale. Dans ce cas le processus obtenu a toutes les caractéristiques d'un bruit blanc. Le deuxième point concerne les clusters de volatilités dans les innovations de notre modèle, voir Figure 1.8. Ces hausses saisonnières nous renseignent sur l'existence d'une alternance entre un régime de basse volatilité et un régime de haute volatilité, clairement observés dans l'estimation de l'écart-type conditionnel des innovations, voir Figure 1.10. Comme nous le verrons dans l'analyse des

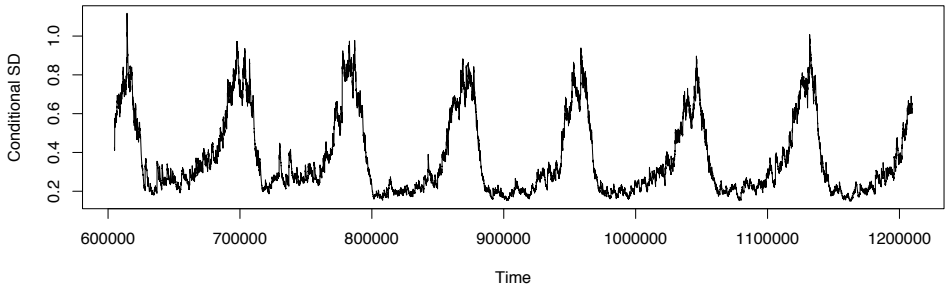

Figure 1.10: *Estimation de l'écart-type conditionnel du processus médian pour* $\Delta = 60$.

résultats de prédiction du modèle, ces phases alternées de haute et basse volatilités représentent un levier d'action possible pour améliorer la performance du *load-balancer*.
La performance du modèle est mesurée par la racine carrée de la moyenne quadratique (RMSE). Pour une série temporelle $(x_t)_{t=1,\dots,T}$ et des prédicteurs $(\hat{x}_t)_{t=1,\dots,T}$, la RMSE est définie par :

$$\mathrm{RMSE}(\hat{x}_t) = \sqrt{\frac{\sum_{t=1}^{T}(\hat{x}_t - x_t)^2}{T}}.$$

La performance du modèle est comparée à celle de deux modèles basiques communément utilisés dans l'industrie : la prédiction NAIVE et AVG. On parlera de *baselines* à partir de maintenant.

Pour une série temporelle $(y_t)_{t\in\mathbb{N}}$, la prédiction NAIVE correspond à la dernière valeur observée :

$$\hat{y}_{t+1}^{NAIVE} = y_t.$$

La prédiction AVG correspond à la moyenne des k dernières valeurs observées:

$$\hat{y}_{t+1}^{AVG} = \frac{1}{k}\sum_{i=1}^{k} y_{t+1-i}.$$

Ces deux baselines sont utilisées pour juger de l'efficacité des algorithmes prédictifs spécifiquement pour de petites valeurs de Δ, lorsque le ratio signal sur bruit est maximal. Battre ces baselines pour les plus petites valeurs de Δ est un objectif important car cela permet de raccourcir l'horizon de la prédiction. La performance de notre modèle en terme de précision est uniformément meilleure que

	$\hat{f}^{\Delta}$	prédiction NAIVE	prédiction AVG
$\Delta = 30$	0.42	0.57	0.43
$\Delta = 60$	0.32	0.43	0.34
$\Delta = 120$	0.25	0.32	0.27
$\Delta = 180$	0.21	0.27	0.23
$\Delta = 240$	0.19	0.24	0.21
$\Delta = 300$	0.18	0.22	0.22
$\Delta = 360$	0.16	0.2	0.23
$\Delta = 480$	0.14	0.17	0.21
$\Delta = 540$	0.14	0.16	0.21
$\Delta = 600$	0.13	0.15	0.2
$\Delta = 720$	0.13	0.15	0.2
$\Delta = 900$	0.12	0.14	0.2
$\Delta = 960$	0.12	0.13	0.19
$\Delta = 1080$	0.12	0.13	0.19
$\Delta = 1200$	0.11	0.12	0.19

Table 1.3: *RMSE de notre modèle, ici abrégé f^{Δ}, et des deux baselines. En colonne, de gauche à droite: RMSE associée à notre modèle, à la prediction NAIVE et la prédiction AVG, en fonction de Δ.*

les baselines à travers toutes les valeurs de Δ testées, bien que faible pour les plus petites valeurs de Δ, voir Table 1.3. Le gain marginal de précision en haute fréquence couplé aux hausses de volatilités mis en évidence par la modélisation de la variance conditionnelle exhibée en Figure 1.10 nous permet d'envisager une stratégie prédictive plus efficace combinant notre modèle au modèle AVG. L'identification des plages horaires de haute volatilité peut se réaliser à l'aide d'un seuil sur l'écart-type conditionnel. Pour une valeur du seuil donné, on constate que le gain relatif de notre modèle sur la prédiction AVG en haute fréquence est réalisé uniquement dans les périodes de haute volatilité, voir Table 1.4. Cette observation, couplée au très bon ajustement de notre modèle aux données, nous fait penser qu'il existe un horizon infranchissable pour la prédiction en haute fréquence, particulièrement lors des épisodes journaliers de basse volatilité. En haute fréquence, nous recommandons donc de n'utiliser le modèle que sur les période de haute fréquence, et d'utiliser un modèle AVG pour les périodes de basse volatilité, afin de réduire le coût computationnel global

	Ratio global	Ratio en haute volatilité	Ratio en basse volatilité
$\Delta = 30$s	1.020	1.032	0.99
$\Delta = 60$s	1.039	1.057	1.007
$\Delta = 120$s	1.067	1.092	1.015
$\Delta = 180$s	1.089	1.115	1.024
$\Delta = 240$s	1.109	1.159	1.056

Table 1.4: *Ratio d'erreur de prédiction entre notre modèle et la prédiction en AVG globale (gauche), sur les période de haute variabilité (milieu) et basse volatilité (droite).*

de la méthode sans impacter la qualité de la prédiction.

Concernant la question 3, commençons par noter qu'à mesure que Δ diminue, notre modèle devient de moins en moins attractif relativement à des modèles plus simples. Cela peut être expliqué de deux manières: notre modèle n'exploite pas toute la structure des données, ou alors il n'y a aucune structure supplémentaire à exploiter. La question est alors de savoir s'il est possible de quantifier l'information résiduelle dans les données après avoir appliqué le modèle. Pour répondre, nous proposons une nouvelle approche basée sur un critère entropique pour évaluer la quantité d'information non captée par le modèle. Nous développons un nouveau test à cet effet basé sur la Sample Entropy (SE) [74].

Definition 2. *Soit $m \in \mathbb{N}$, $r > 0$ and $X = (x_1, \dots, x_N)$ une série temporelle. On pose $X_m(i) = (x_i, \dots, x_{i+m-1})$ et soit $d = \|\cdot\|_\infty$ la sup norme. Alors la Sample Entropy de X est définie par:*

$$SE_n^X = -\log \frac{A}{B}$$

où: $A = \#\{i \neq j, d(X_{m+1}(i), X_{m+1}(j)) < r\}$, $B = \#\{i \neq j, d(X_m(i), X_m(j)) < r\}$.

Comme $A \leq B$, SE est toujours un nombre positif. SE est une approximation de la probabilité conditionnelle que deux séries consécutives d'observations de longueur $m+1$ restent à une distance r sachant que les deux sous séries contenant les m premiers points l'étaient. Une valeur élevée de SE indique que cette probabilité est faible, suggérant ainsi une certaine non-prédictabilité de la série. Cette notion a été introduite par Richman et Moormanis [74] et a trouvé depuis de nombreuses applications notamment en médecine, voir par exemple [17] [84] [56].

Dans le cas où $X = (X_1, \dots, X_N)$ sont des variables aléatoires i.i.d., alors $-\log(A/B)$ est la version empirique de:

$$\begin{aligned} -\log \frac{\mathbb{P}\left(d((Z_1, \dots, Z_{m+1}), (Y_1, \dots, Y_{m+1})) < r\right)}{\mathbb{P}\left(d((Z_1, \dots, Z_m), (Y_1, \dots, Y_m)) < r\right)} &= -\log \frac{\mathbb{P}\left(\max\limits_{1 \leq i \leq m+1} |Y_i - Z_i| < r\right)}{\mathbb{P}\left(\max\limits_{1 \leq i \leq m} |Y_i - Z_i| < r\right)} \\ &= -\log\left(\mathbb{P}(|Z_1 - Y_1| < r)\right) \end{aligned}$$

où les variables aléatoires $(Y_1, \dots, Y_{m+1}, Z_1, \dots, Z_{m+1})$ sont des copies i.i.d. de X_1. Autrement dit, dans le cas d'observations i.i.d. de loi F, SE estime la concentration de F et mesure le volume moyen d'une boule aléatoire de rayon r. On montre alors la propriété suivante:

Proposition 3. *Soit* $\theta = -\log(\mathbb{P}(|X_1 - X_2| < r))$. *Alors il existe une matrice* Σ *de taille* 2×2 *telle que:*

$$\begin{cases} SE_n^X \xrightarrow{P} \theta \\ \sqrt{n}(SE_n^X - \theta) \xrightarrow{d} \mathcal{N}(0, \nabla g(\theta) \Sigma \nabla g(\theta)^t) \end{cases}$$

où g*:* $(x, y) \mapsto \log(x) - \log(y)$

Cette proposition permet alors de tester l'hypothèse que la série X est i.i.d. de loi cible F. En particulier, nous appliquons ce résultat aux résidus de notre modèle obtenons la distribution des p-valeurs en Figure 1.11, et obtenons des résultats corroborant les tests de diagnostics de notre modèle obtenus précédemment. En particulier, nous obtenons une indication empirique que notre modèle exploite bien toute la structure des données.

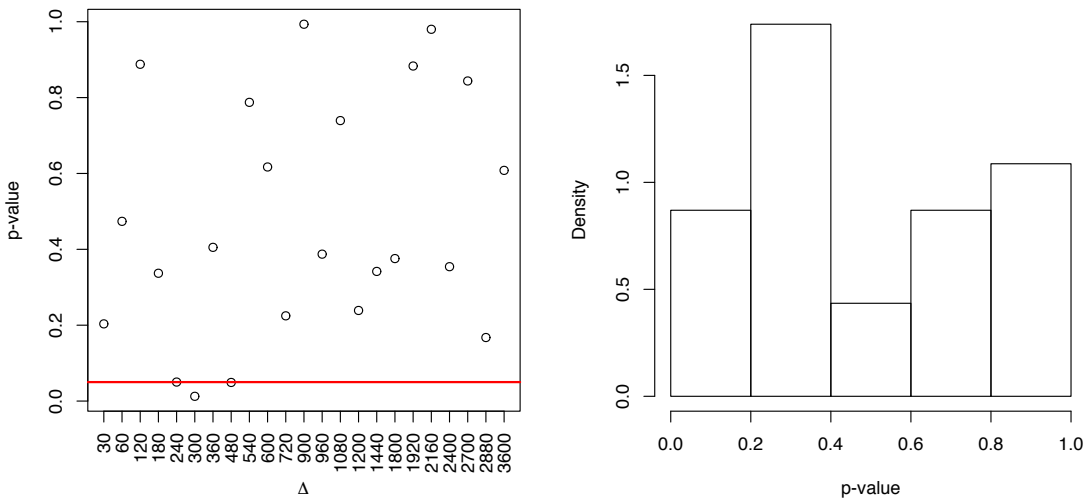

Figure 1.11: *P-valeurs des tests dérivés de la Proposition 3 pour* $\Delta \leq 3600$ *avec seuil critique à* 5% *en rouge (gauche). Histogramme des p-valeurs (droite).*

1.5 Prediction et détection de panne dans les réseaux stables

À la différence du chapitre 2 où nous nous sommes intéressés à des processus médian riches en structure, dans le troisième chapitre nous nous sommes intéressés à une classe particulière de réseaux, qu'on appelle les réseaux $\varepsilon-$Stable, abrégé $\varepsilon - SN$ pour ε-Stable Networks. Un tel réseau se caractérise par une absence d'effets saisonniers ou d'autocorrelation notables dans le processus $X_{t_n}^{\Delta}$. Ces réseaux sont plus rares et correspondent à des infrastructures particulièrement haut de gamme. Il ne s'agit presque jamais de *CDN*, mais souvent d'infrastructures privées. Des entreprises comme Google par exemple n'ont pas recours aux *CDN*, et construisent leurs propres

infrastructures. De façon heuristique, un $\varepsilon - SN$ se caractérise par un processus median $(X_n^{\Delta})_{n\in\mathbb{N}}$ stationnaire pour lequel des modèles sophistiqués ne permettent pas un gain significatif de précision prédictive. Plus spécifiquement, un $\varepsilon - SN$ est défini comme un réseau produisant un processus médian purement non déterministe, voir définition 4, pour lequel le ratio des risques quadratiques des prédictions à une période des modèles AVG et du meilleur $ARMA(p,q)$, noté $\widehat{X_{t+1}}^{ARMA}$, est borné par $1+\varepsilon$:

$$\frac{\text{RMSE}\left(\widehat{X_{t+1}}^{AVG}\right)}{\text{RMSE}\left(\widehat{X_{t+1}}^{ARMA}\right)} \leq (1+\varepsilon)$$

Afin de définir proprement les réseaux $\varepsilon - SN$, nous rappelons quelques faits concernant les séries temporelles, voir par exemple [43].

Definition 3 (Processus déterministes)**.** *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus du second ordre. Pour $t \in \mathbb{Z}$ on pose:*

$$H_{t-1} = \overline{Vect\{X_{t-1}, X_{t-2}, \ldots\}}$$

la fermeture dans L^2 de l'espace vectoriel $Vect\{X_{t-1}, X_{t-2}, \ldots\}$, c'est-à-dire toutes les combinaisons linéaires de la forme $\sum_{k=0}^{\infty} \lambda_k X_{t-k}$ qui convergent dans L^2. On dit alors que $(X_t)_{t\in\mathbb{Z}}$ est déterministe si et seulement si:

$$X_t \in H_{t-1}$$

autrement dit si et seulement si

$$X_t = \text{proj}(X_t, H_{t-1})$$

où

$$\text{proj}(X_t, H_{t-1}) = \arg \min_{Y\in H_{t-1}} \|X_t - Y\|_2$$

est la projection orthogonale dans L^2 de X_t sur le sous espace vectoriel H_{t-1}.

Rappelons maintenant le théorème de Wold [3] :

Theorem 2 (Décomposition de Wold)**.**
Soit $(X_t)_{t\in\mathbb{N}}$ un processus de moyenne nulle faiblement stationnaire. Alors il existe des processus aléatoire $(\varepsilon_t)_{t\in\mathbb{N}}$ et $(d_t)_{t\in\mathbb{N}}$ et des nombres réels $(\psi_t)_{t\in\mathbb{N}}$ tel que:

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} + d_t \quad \forall t \in \mathbb{Z}$$

où:

i) $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$,

ii) $(\varepsilon_t)_{t\in\mathbb{N}}$ *est un bruit blanc, i.e.* $\mathbb{E}(\varepsilon_t) = 0$ *et* $\mathbb{E}(\varepsilon_t \varepsilon_s) = \sigma^2 \mathbf{1}_{\{s=t\}}$.

iii) $(d_t)_{t\in\mathbb{N}}$ *est un processus déterministe.*

iv) $\forall s,t,\ \mathbb{E}(d_s\varepsilon_t)=0$.

De plus, cette décomposition est unique.

Definition 4 (Processus purement non déterministe)**.**
Soit $(X_t)_{t\in\mathbb{N}}$ *un processus de moyenne nulle faiblement stationnaire, alors* $(X_t)_{t\in\mathbb{N}}$ *est dit purement non déterministe si et seulement si* $d_t=0$ *pour tout* t*, où* $(d_t)_{t\in\mathbb{N}}$ *est le processus déterministe qui apparait dans la décomposition de Wold de* $(X_t)_{t\in\mathbb{N}}$*.*

Afin de vérifier empiriquement qu'un processus est purement non déterministe, nous nous reposons sur le théorème suivant, dû à Kolmogorov [53] :

Theorem 3 (Kolmogorov)**.** *Soit* $(X_t)_{t\in\mathbb{N}}$ *un processus du second ordre de fonction d'auto-covariance* γ*. Alors* $(X_t)_{t\in\mathbb{N}}$ *est purement non déterministe si et seulement si les conditions suivantes sont vérifiées:*

1) F_X *est absolument continue par rapport à la mesure de Lebesgue*

2) f_X *est positive presque partout*

3) $\log f_X$ *est intégrable*

où F_X *et* f_X *sont les distributions et densités spectrales respectivement de* $X=(X_t)_{t\in\mathbb{N}}$ *c'est-à-dire que* F_X *est la fonction de répartition de la mesure spectrale de* $(X_t)_{t\in\mathbb{N}}$*, autrement dit de la mesure de probabilité dont les coefficients de Fourier sont* $\gamma(h)$*,* $h\in\mathbb{Z}$ *:*

$$\gamma(h)=\int_{-1/2}^{1/2} e^{2\pi i\nu h}F_X(d\nu)$$

et

$$f_X(\nu)=\sum_{h\in\mathbb{Z}}\gamma(h)e^{-2\pi i\nu h}.$$

Il est aisé de voir que la condition $f_X>0$ partout implique que $\log(f_X)$ est intégrable. En effet dans ce cas par continuité de f_X sur l'intervalle fermé $[-1/2,1/2]$, on a directement que f est minorée et majorée par des constantes strictement positives, donc $\log(f_X)$ est intégrable. À notre connaissance, aucun test dans la littérature n'est proposé pour le vérifier en pratique. Nous nous basons donc sur une inspection visuelle de l'estimation de la densité spectrale du processus médian, voir Figures 1.12 et 1.13. Il suit immédiatement de la décomposition de Wold que tout processus purement non déterministe peut être approché arbitrairement par un processus ARMA(p,q). Or si $(X_t)_{t\in\mathbb{Z}}$ est un ARMA(p,q), c'est à dire que X_t est solution de l'équation:

$$X_t=u_t+\sum_{i=1}^{p}\phi_i X_{t-i}+\sum_{j=1}^{q}\theta_j u_{t-j}$$

où $(u_t)_t$ est un bruit blanc et $|\sum\phi_i|<1$, alors la prédiction du processus à une période donnée par

$$\widehat{X_{t+1}}^{ARMA}=\sum_{i=1}^{p}\hat{\phi}_i X_{t-i}+\sum_{j=1}^{q}\hat{\theta}_j\hat{u}_{t-j}$$

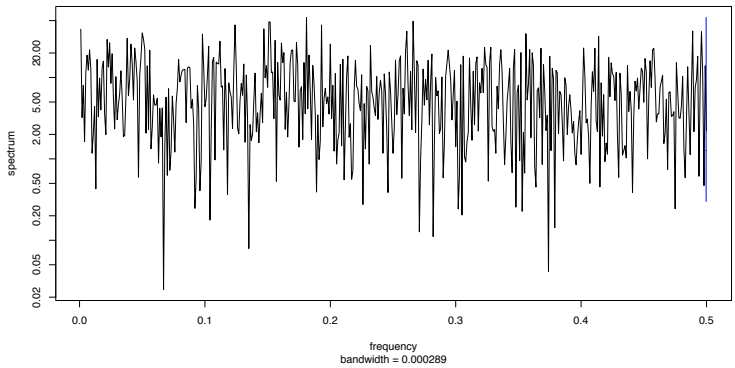

Figure 1.12: *Densité spectrale estimée d'un processus médian* $(X_n^{\Delta})_n$. *Le domaine des fréquence varie de 0 et 1/2, et la puissance est en échelle logarithmique*

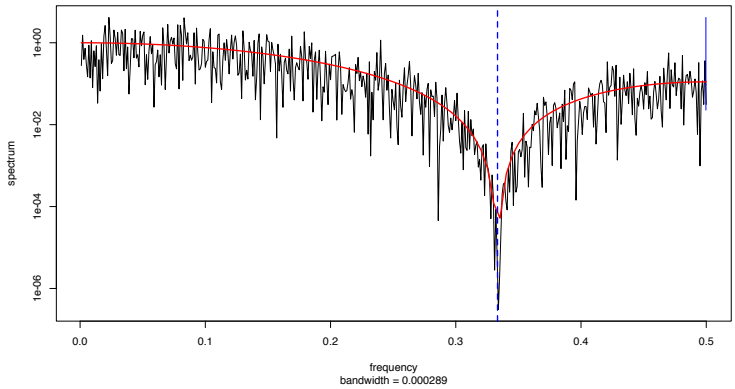

Figure 1.13: *Densité spectrale estimée du processus MA(3)* $X_t = (\varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t)/3$, *avec* ε_t *un bruit blanc gaussien. La densité spectrale admet un unique zéro en* $\nu = 1/3$. *En rouge figure la vraie densité, et la droite verticale est la droite d'équation* $x = 1/3$. *Le domaine des fréquence varie de 0 et 1/2, et la puissance est en échelle logarithmique.*

où les $\widehat{u}_t$ sont les résidus estimés et les paramètres ϕ_i, θ_j sont estimés par minimisation de la somme des carrés des résidus, est un estimateur de l'espérance conditionnelle

$$\mathbb{E}(X_{t+1}|X_s, u_s, s \leq t) = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j u_{t-j}$$

qui est optimale dans le sens où elle minimise l'erreur quadratique. La prédiction du modèle AVG, qui est simplement égale à la moyenne empirique des k dernières observations, est un estimateur de l'espérance. On remarque alors qu'un réseau $\varepsilon - SN$ est tel que la prédiction optimale utilisant tout l'historique du processus n'améliore que marginalement la prédiction optimale utilisant

l'historique immédiat.

Par définition, des modèles prédictifs sophistiqués ne sont donc pas pertinents pour la prédiction des $\varepsilon - SN$: plutôt que d'entraîner un algorithme prédictif sur un jeu de données volumineux, des modèles simples ne demandant que peu de données d'entraînement ont des performances prédictives similaires. L'intérêt est principalement économique: entraîner un modèle avec un gros volume de données nécessite plus de calculs, et s'accompagne donc de coûts supplémentaires.

Pour des raisons de confidentialité, l'algorithme prédictif dans les $\varepsilon - SN$ que nous avons développé chez Citrix sera traité comme une boîte noire. Au temps t, notons $\widehat{X_t^{\Delta}}^{CITRIX}$ la prédiction associée. Nous ne traiterons que la détermination de l'ensemble d'entraînement optimal. Ce prédicteur de la latence médiane utilise au plus les N dernières mesures les plus récentes à condition qu'elles aient été reçues dans les dernières M minutes, même si cela implique de considérer moins de N mesures. On notera $\widehat{X_t^{\Delta}}^{CITRIX} = \widehat{X_{t,N,M}^{\Delta}}^{CITRIX}$. Nous nous intéressons au problème de la calibration de (N, M) afin de minimiser la taille du jeu d'entraînement sans dégrader la qualité de la prédiction au dela d'une certaine tolérance τ choisie par le praticien.

La valeur de Δ sera fixée dans ce chapitre. Nous choisissons $\Delta = \lambda$, où λ est le Time To Live ou TTL. TTL est un terme générique pour quantifier la durée de vie de toute donnée stockée pendant une durée limitée dans un réseau avant d'être mises à jour ou écrasée. Un TTL est utilisé pour les prédictions de Citrix, car la mise à jour des prédictions en temps réel est trop exigeante d'un point de vue calculatoire. Cela signifie que si une prédiction est mise à jour à l'instant t, chaque nouvel utilisateur arrivant entre les instants t et $t + \lambda$ utilisera la même prédiction. Ce n'est qu'au moment $t + \lambda$ que la prédiction sera mise à jour. Par conséquent, une prédiction au temps t avec la durée de vie λ doit prédire la valeur médiane des mesures de latence sur l'intervalle $[t, t + \lambda]$. $\lambda = 60$ s est la valeur par défaut définie par les ingénieurs.

Parallèlement, nous nous penchons sur un problème qui touche principalement les réseaux *personnels*, dont les $\varepsilon - SN$ sont une sous classe: ces derniers sont généralement plus propices aux pannes que les *CDN*. En effet, un *CDN* rassemble typiquement des dizaines voire centaines de milliers de serveurs. En cas de panne d'une partie de l'infrastructure, les serveurs restants peuvent prendre le relais. Dans un réseau *personnel*, qui compte généralement considérablement moins de serveurs, le risque de panne est plus grand. Nous nous intéressons donc dans ce chapitre à une méthode de détection de panne dans les $\varepsilon - SN$.

Question 1 Comment sélectionner le jeu d'entraînement le plus petit possible sans impacter la qualité de la prédiction?

Question 2 Comment construire un algorithme de détection de panne?

Afin de répondre à ces questions, la première étape consiste à identifier les réseaux $\varepsilon - SN$. Des routines classiques de stationnarité sont d'abord effectuées, à l'aide des tests de Dickey-Fuller (ADF), Kwiatkowski-Phillips-Schmidt-Shin (KPSS) et Phillips-Perron (PP) [30], [38], [55], [60]. Ensuite, une inspection visuelle permet d'écarter la présence d'un zéro dans la densité spectrale.

Enfin, un modèle ARMA(p,q) est ajusté sur un jeu d'entraînement et le ratio

$$\text{RMSE}(\widehat{X_{t+1}}^{AVG})/\text{RMSE}(\widehat{X_{t+1}}^{ARMA})$$

est calculé sur un échantillon de test.

Sur les données présentées dans ce chapitre, nous étudions trois $\varepsilon - SN$ avec une valeur du paramètre $\varepsilon = 2\%$. Pour répondre à la question 1, les paramètres (N, M) sont estimés par grid searching [18] [8] sur un jeu d'entraînement. Pour chaque couple sélectionné, on effectue des prédictions à des instants pré-définis et l'erreur, notée $\mathcal{E}(N, M)$ est calculée. On définit ensuite

$$(N^*, M^*) = \arg\min_{N,M} \mathcal{E}(N, M).$$

Le couple (N^*, M^*) qui minimize l'erreur de prédiction correspond aux plus grandes valeurs testées des paramètres N et M: la prédiction $\widehat{X_t^{\Delta}}^{CITRIX}$ est d'autant meilleure que les paramètres N, M sont grands. Or, on constate dans les $\varepsilon - SN$ que la précision de cette prédiction a un profil particulier: l'erreur converge très rapidement vers $\mathcal{E}(N^*, M^*)$ quand N, M augmentent. Nous proposons de limiter la taille de l'échantillon relativement à une tolérance $\tau > 0$ de la manière suivante: soit $\mathcal{C}(\tau)$ l'ensemble des couples (N, M) vérifiant

$$\mathcal{C}(\tau) = \Big\{(N, M) | \mathcal{E}(N, M) < (1 + \tau)\mathcal{E}(N^*, M^*)\Big\},$$

alors le couple $(N, M) = (N_\tau, M_\tau)$ retenu est donné par:

$$M_\tau = \min\Big\{M | (N, M) \in \mathcal{C}(\tau)\Big\}$$
$$N_\tau = \min\Big\{N | (N, M_\tau) \in \mathcal{C}(\tau)\Big\}$$

autrement dit, parmi l'ensemble des couples (N, M) vérifiant $\mathcal{E}(N, M) < (1 + \tau)\mathcal{E}(N^*, M^*)$, on choisit le couple (N_τ, M_τ) en minimisant d'abord en M, puis en N. Cette méthodologie permet de réduire considérablement la taille des données d'entraînement. Les valeurs de (N^*, M^*) sont présentées en Tables 1.5, et la sélection des paramètres (N_τ, M_τ) en Table 1.6.

	N^*	M^*	$\mathcal{E}(N^*, M^*)$
$P1$	940	9min	1.27
$P2$	950	57min	2.34
$P3$	850	45min	4.62

Table 1.5: Résultats du calibrage des paramètres optimaux (N^*, M^*) pour trois réseaux *personnels* $P1$, $P2$ et $P3$.

Concernant la question 2 et la détection de panne, nous proposons un algorithme basé sur une comparaison online de deux fenêtres glissantes. Une panne est caractérisée par une explosion soudaine de la variance dans les mesures et d'une hausse du niveau de ces dernières comme illustré en Figure 1.14. La détection de panne se réduit alors à un problème de comparaison de deux distributions empiriques. Les deux fenêtres glissantes seront dénommées Ref et Shift. La fenêtre

	τ	N_τ	M_τ	$\mathcal{E}(N_\tau, M_\tau)/\mathcal{E}(N^*, M^*)$
$P1$	10%	300	3min	1.093
	5%	400	4min	1.048
	1%	820	6min	1.009
$P2$	10%	150	3min	1.087
	5%	300	4min	1.049
	1%	600	17min	1.009
$P3$	10%	45	2min	1.099
	5%	100	6min	1.049
	1%	500	20min	1.009

Table 1.6: (N_τ, M_τ) pour différentes valeurs du paramètre de tolérance τ pour trois réseaux *personnels* $P1$, $P2$ et $P3$.

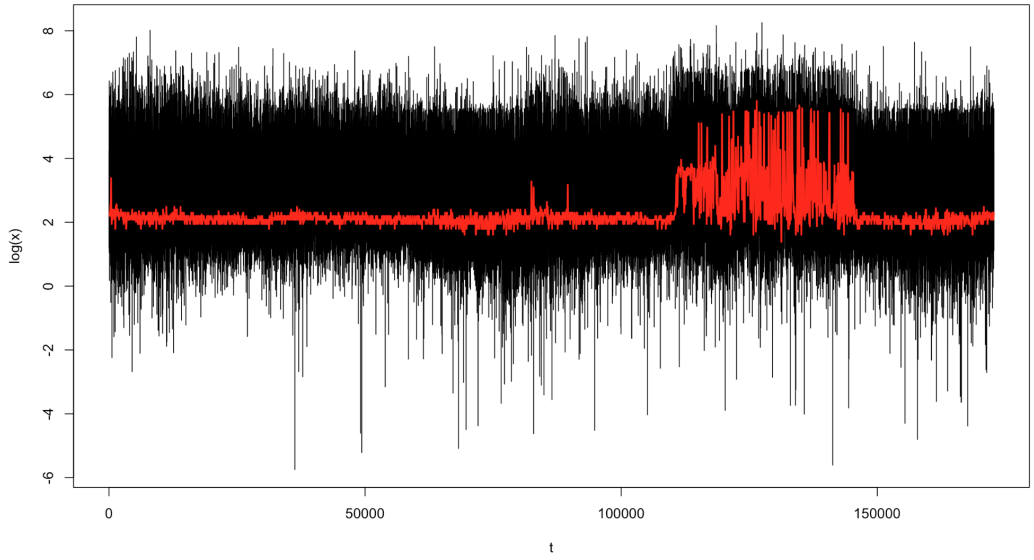

Figure 1.14: *Example de panne. Données de latences brutes en noir, et processus médian en rouge. Echelle logarithmique en ordonnées. La panne se traduit par une hausse soudaine de la moyenne et de la variance dans la distribution des mesures de latence.*

Ref contient des observations du processus X_n^Δ qui ont été reçues dans un passé proche durant lequel aucune panne n'a été détectée. Le fenêtre Shift ingère les nouvelles observations de ce processus en temps réel. Les deux fenêtres gardent une taille constante égale à $C \in \mathbb{N}$. À chaque nouvelle observation reçue, la distance de Wasserstein [90] est calculée entre Ref et Shift. La distance de Wasserstein d'ordre $p \geq 1$ entre deux mesures de probabilités P et Q sur l'ensemble des réels, notée $W_p(P, Q)$ est définie par:

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int |x-y|^p \pi(dx, dy) \right)^{1/p}$$

où $\Pi(P,Q)$ est l'ensemble des mesures de probabilité ayant pour marginales P et Q. La distance W_p peut se réécrire:

$$W_p(P,Q) = \left(\int_{\mathbb{R}} |F(x) - G(x)|^p dx\right)^{1/p}$$
$$= \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du\right)^{1/p}$$

où F, G sont les c.d.f. de P et Q respectivement, et F^{-1}, G^{-1} leurs inverses généralisés. Pour des échantillons $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de loi respective F et G, la distance de Wasserstein empirique est définie comme la distance de Wasserstein entre les distributions empiriques:

$$W_{p,n}(F,G) = W_p(F_n, G_n) = \left(\int_0^1 |F_n^{-1}(u) - G_n^{-1}(u)|^p du\right)^{1/p}$$

où $F_n(u) = n^{-1}\sum_{i=1}^n 1_{\{X_i \leq u\}}$ et $G_m(u) = n^{-1}\sum_{j=1}^m 1_{\{Y_j \leq u\}}$ sont les fonctions de répartition empiriques et F_n^{-1} et G_n^{-1} sont leurs inverses généralisés respectivement. La distance de Wasserstein empirique se réécrit:

$$W_{p,n}(F,G) = \left(\frac{1}{n}\sum_{i=1}^n |X_{(i)} - Y_{(i)}|^p\right)^{1/p}$$

où $(X_{(i)})_{1\leq i\leq n}$ et $(Y_{(i)})_{1\leq i\leq n}$ sont les statistiques d'ordre des deux échantillons.

Dès que la distance de Wasserstein entre les échantillons Ref et Shift excède un certain seuil Q, une panne est déclarée. L'ensemble des instants où aucune panne n'est déclarée constitue la zone Verte. Au contraire, l'ensemble des instants où une panne est déclarée constitue la zone Rouge: durant ces périodes, la confiance dans le modèle est perdue, le client doit être notifié instantanément, et le traffic transitant vers ce réseau doit, si possible, être aiguillé sur un autre réseau pouvant livrer le contenu. L'algorithme est décrit en 1 et illustré en Figure 1.15.

Initialization;
Ref = First C points of stream;
Shift = Next C points of stream ;
while *not at end of stream* **do**
  Slide Shift by 1 point;
  if $W_{p,n}(Ref, Shift) > Q$ **then**
    $t_0 \leftarrow$ current time ;
    Report change at t_0;
    Ref = First C points starting at t_0;
    Shift = First C points starting at $t_0 + C$;
  else
    $t_0 \leftarrow$ current time ;
    Report confidence in model at time t_0;
  end
end

Algorithm 1: *Algorithme de détection de changement.*

Un problème important se pose: les données de latence sont très hétérogènes et présentent une distribution en loi de puissance. Cette propriété rend fréquente l'apparition d'événements

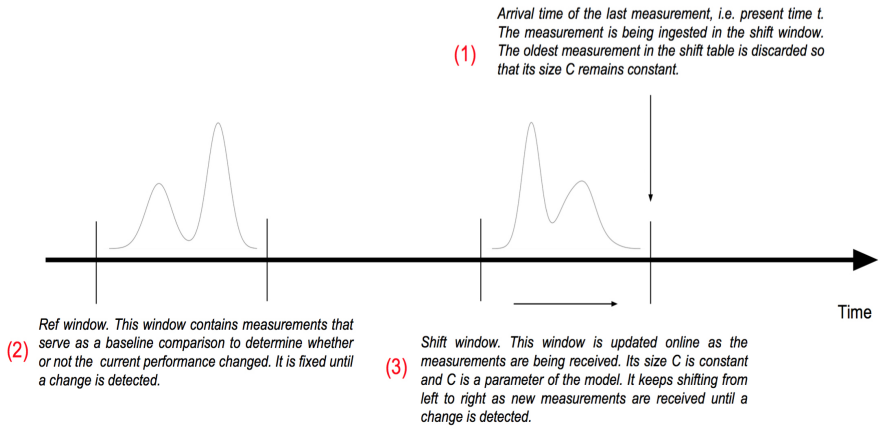

Figure 1.15: *Illustration des fenêtres Ref et Shift*

catastrophiques caractérisés par des mesures très élevées. Nous décrivons précisément dans ce chapitre ces distributions typique en loi de puissance, et comment cela affecte la détection de pannes. En particulier, ces événements catastrophiques ont tendance à déclencher des faux positifs: l'algorithme déclare une panne à tort. Pour diminuer le taux de faux positifs, nous introduisons une version pondérée de la distance de Wasserstein afin de limiter l'impact des valeurs extrêmes. Pour des échantillons $\boldsymbol{X}, \boldsymbol{Y}$, nous introduisons la distance de Wasserstein pondérée:

$$\overline{W}_p(\boldsymbol{X}, \boldsymbol{Y}) = \left(\int_0^1 |F^{-1}_{w^{\boldsymbol{X}}}(u) - G^{-1}_{w^{\boldsymbol{Y}}}(u)|^p du \right)^{1/p}$$

où:

$$F_{w^{\boldsymbol{X}}}(u) = \sum_{i=1}^{n} w_i^{\boldsymbol{X}} 1_{\{X_i \leq u\}} \text{ et } G_{w^{\boldsymbol{Y}}}(u) = \sum_{j=1}^{m} w_j^{\boldsymbol{Y}} 1_{\{Y_j \leq u\}}$$

sont les fonctions de répartition empiriques pondérées de $\boldsymbol{X}$ et $\boldsymbol{Y}$. Les poids $w^{\boldsymbol{X}}$ et $w^{\boldsymbol{Y}}$ vérifient:

$$w_i^{\boldsymbol{X}} \geq 0,\ w_j^{\boldsymbol{Y}} \geq 0,\ \sum_{i=1}^{n} w_i^{\boldsymbol{X}} = 1,\ \sum_{i=1}^{m} w_i^{\boldsymbol{Y}} = 1.$$

Nous choisissons des poids qui décroissent exponentiellement avec la distance à un quantile donné, typiquement la médiane dans le cas du *load-balancer*. On pose:

$$\begin{cases} u^{\boldsymbol{X}}_{\pi(i)} = e^{-\lambda \left| X_i - q^{\boldsymbol{X}}_\beta \right|} \\ u^{\boldsymbol{Y}}_{\sigma(j)} = e^{-\lambda \left| Y_j - q^{\boldsymbol{Y}}_\beta \right|} \end{cases}$$

où π et σ sont des permutations de $\{1, \ldots, n\}$ telles que $X_{\pi^{-1}(1)} \leq \ldots \leq X_{\pi^{-1}(n)}$ et $Y_{\pi^{-1}(1)} \leq \ldots \leq Y_{\pi^{-1}(n)}$, $\lambda > 0$ et $q^{\boldsymbol{X}}_\beta, q^{\boldsymbol{Y}}_\beta$ sont les quantiles empiriques d'ordre $\beta \in (0,1)$ de $\boldsymbol{X}$ et $\boldsymbol{Y}$ respectivement. Les poids sont alors définis par:

$$\begin{cases} w_i^X = \frac{u_i^{\boldsymbol{X}}}{\sum_j u_j^{\boldsymbol{X}}} \\ w_i^Y = \frac{u_i^{\boldsymbol{Y}}}{\sum_j u_j^{\boldsymbol{Y}}} \end{cases}$$

Ces poids pénalisent les observations dans l'échantillon qui sont trop éloignés d'un quantile donné. Notre algorithme est testé sur données réelles, en le back-testant sur des réseaux dont on a pu identifier des pannes. Notre algorithme se montre particulièrement efficace dans la détection de panne, voir Figure 1.16. La version pondérée permet de limiter grandement le taux de faux positifs tout en gardant un taux de vrais positifs proche de 100%.

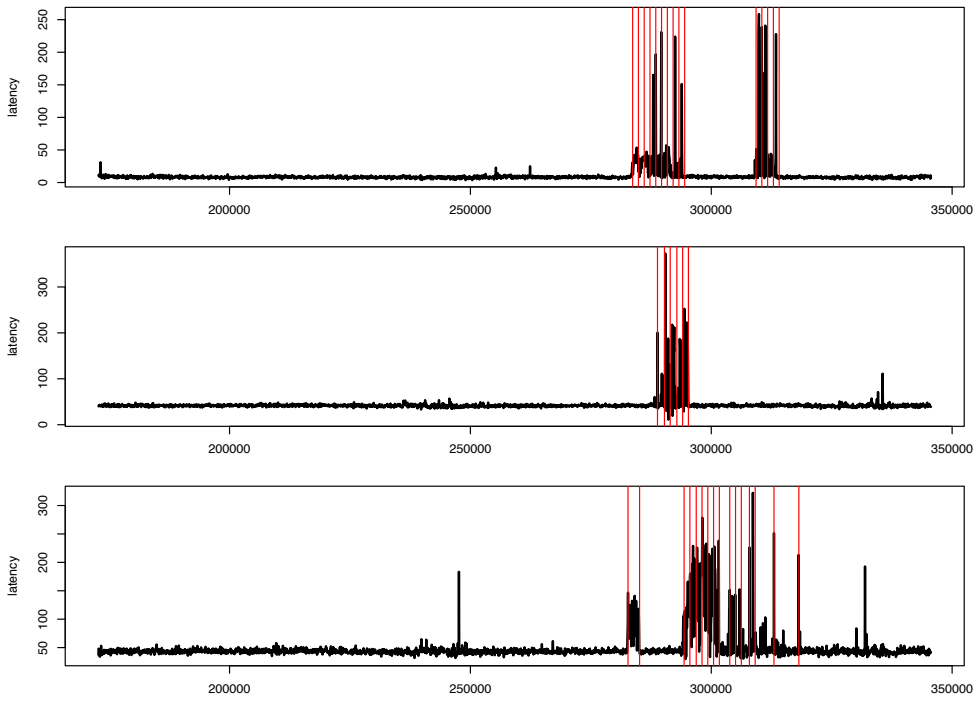

Figure 1.16: *De haut en bas: trois réseaux P_1, P_2 et P_3 qui ont expérimenté une panne. Le processus médian est en noir, et les droites verticales rouges représentent les instants de detection de changement dans la distribution des observations. Tous ces instants sont correctement associés à la panne réelle. La distribution en loi de puissance des mesures de latences se manifeste au travers d'observations particulièrement élevées, notamment autour des instants $t = 250000$ et $t = 330000$ pour P_3 et $t = 340000$ pour P_2, sans que ces mesures ne soient pour autant le résultat d'une panne. L'algorithme utilise la distance de Wasserstein pondérée et permet de ne déclarer aucun faux positifs, tout en identifiant correctement chaque panne.*

Chapter 2

Pointwise estimation of a transport between two probability distributions

Abstract

Citrix is a technology company that optimizes network packets travel to accelerate the loadings of web pages. Citrix collects in real time latency measurements from a large number of interconnected servers, called *CDN*'s, capable of receiving and transmitting information. It is suspected by the operational engineers that the complex configuration of those networks makes it likely that certain specific subsets of servers, called *maps*, have overlapping infrastructures either because they are maintained by the same vendor or installed in the same data center. It is conjectured that the overlap between certain *maps* should be reflected in the distribution of latency measurements. Being able to characterize the statistical dependency between certain *maps* is of great strategic importance for Citrix because it can help overcome major issues in the *load-balancing* process, i.e. the action of distributing Internet users across the network to spread the audience in order to accelerate page load time, such as data scarcity. *Maps* that generate distributions of latency measurements that share statistical properties will be called *related maps*. In this chapter we are interested in mapping the distribution of latency measurements of *related maps* into one another. We formulate this problem as a distribution matching problem where the outputs are the transported probability distributions of the inputs under an unknown function f. Under some regularity assumptions on the densities of the input probability distributions and f in a neighborhood of some point $x \in \mathbb{R}$, we propose an estimator of $f(x)$ and derive uniform convergence properties in that neighborhood when the only observables are independent samples drawn from the input and output distributions. We provide empirical evidence of the existence of a deterministic transport between the distributions of latency measurements of *related maps* on real data and show that the numerical implementations are consistent with the theoretical rates of convergence.

2.1 Introduction

2.1.1 Motivation

Load balancing Internet users across the Internet is similar to road navigation. What Waze-like applications do for road networks, Citrix does it for the Internet: both estimate the optimal route in a network based on data generated by other users close in space and time. The key performance metric for *load-balancing* users across the network is latency: a measurement expressed in milliseconds (ms) of the time it takes to establish a connexion between two distant servers. For reasons of privacy protection, we will anonymize all data presented in this chapter.

Difficulty may arise when not enough data is collected from a specific *map* in the network, called the *source*. In such situation Citrix uses a *proxy* in order to estimate the *source* status. A *proxy* in this context refers to an *map related* to the *source*, meaning that it shares multiple common properties with it: typically both *source* and *proxy* are maintained by the same vendor and correspond to an equivalent requirement level in terms of performance. We will consider a period of time $[0, T]$ over which we will observe latency measurements collected from both channels, *source* and *proxy*. Since the distribution of latency measurements slowly evolves through time, we will assume that there exists a uniform partition $(t_i)_{0\leq i\leq K}$ of $[0, T]$ where $t_0 = 0$, $t_K = T$, $t_{i+1} - t_i = h > 0$ $\forall 0 \leq i \leq K-1$ such that latency measurements from both channels with timestamps falling in any subinterval $[t_i, t_{i+1}]$ form i.i.d. samples. To determine h, we can test the hypothesis that two samples of measurements with timestamps falling in consecutive intervals of length $h/2$ come from the same distribution. The choice of h will be discussed in details in the numerical implementation section. In this context, we address the inference of the statistical dependency between the *source* and *proxy* when they are observed through pairs of data sets collected from both channels on uniform subintervals of the period $[0, T]$. The two samples of measurements from the *source* and *proxy* falling in a sub-interval will be called a batch.

Let $M \in \mathbb{N}$, we consider the following model of distribution adjustment:

$$\text{find } f : \mathbb{R} \to \mathbb{R} \text{ such that } Q_i = f_{\#}P_i \quad \forall 1 \leq i \leq M \tag{2.1}$$

where the input variables P_i are drawn independently from a measure $\mathbb{M}$, where $\mathbb{M}$ belongs to a given set of probability distributions $\mathcal{Z}$. In this model, the output variables $Q_i = f_{\#}P_i$ are the push-forward measures of the input variables P_i under a function f meaning that for all Borel set A we have

$$f_{\#}P_i(A) := P_i(f^{-1}(A)).$$

The inference of f when one only observes M *batches* of measurements formed by i.i.d. n-samples $\mathcal{X}_i = (X_1^i, \ldots, X_n^i) \sim P_i$ and, independently, $\mathcal{Y}_i = (Y_1^i, \ldots, Y_n^i) \sim Q_i = f_{\#}P_i$ is the topic of this chapter. The main difference with a classical regression setting lies in the fact that the data samples are not paired since the pairing occurs at the probability distributions level, not at the measurement level.

Quantiles of latency measurements play an important role for *load-balancing* purposes because users can not be routed individually so the distributional approach using quantiles is a powerful

tool. The *load-balancing* industry heavily relies on the estimation of certain quantiles, hence being able to estimate $f(x)$ for certain key values of x could lead significant improvements of the *load-balancing* prediction algorithm by correcting the bias in the *proxy*. This approach encompasses many other real world applications. For instance since the publication of the Framingham Heart Study [63], heart rate is known to be a major predictor for a wide variety of health complications. A lot of classical regression settings or machine learning tasks will infer mappings between finite dimension spaces, reducing a distribution to a scalar or real valued vector, despite the fact that the observed phenomenon, like heart rate, is distributional in essence.

In terms of Citrix data, P_i (resp Q_i) will be the distribution of measurements of the *proxy* (resp *source*) over the interval $[t_i, t_{i+1}]$. Here f reflects the assumption of a dependency between the two channels caused by the commun physical structure of interconnected servers at the cables level in the network. In other words the distribution of measurements of one channel should be a deterministic transport of the distribution of the other.

2.1.2 Main results and organization of the chapter

Our estimator for f takes the form:

$$\widehat{f}_{n,M}(x) = \frac{1}{M}\sum_{i=1}^{M}(\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) \tag{2.2}$$

where:

- $\widehat{F_n^i}(x) = \frac{1}{n}\sum_{l=1}^{n} \mathbf{1}_{\{X_l^i \leq x\}}$
- $\widehat{G_n^i}(x) = \frac{1}{n}\sum_{l=1}^{n} \mathbf{1}_{\{Y_l^i \leq x\}}$
- $(\widehat{G_n^i})^{-1}(x_0) = \inf\{x \in \mathbb{R}, (\widehat{G_n^i})(x) \geq x_0\}$

are the cumulative distribution functions of the samples $(X_1^i, \dots, X_n^i)$ and $(Y_1^i, \dots, Y_n^i)$ and the generalized inverse of $(Y_1^i, \dots, Y_n^i)$ respectively. The form of the estimator is motivated by the well known inverse transform theorem: if X is a random variable on $\mathbb{R}$ with cumulative function F, then $F^{-1}(X) \sim \mathcal{U}(0,1)$, see for instance [29]. If X, Y are two independent real random variables with increasing and continuous cumulative distribution functions F and G respectively, there exists an increasing function h such that $Y \overset{\mathcal{L}}{=} h(X)$, where $h = G^{-1} \circ F$. This coupling is often called *increasing rearrangement*, and has connexions to transport theory, see [90]. By construction, $\widehat{f}$ is a non decreasing function. This suggests that the true function f must be an increasing function in order to be estimated. It might be the case that $\widehat{f}_{n,M}(x)$ converges to $f(x)$ when f is non increasing for some x when specific conditions are met, but we will mainly focus on increasing f in this chapter. The general case will be examined in the section 2.6. We will derive a sufficient condition on f and $\mathcal{Z}$ that guarantees that $\widehat{f}_{n,M}(x)$ converges to $f(x)$, but also that in general the problem of estimating f in the non increasing case is ill-posed.

The analysis of $\widehat{f}_{n,M}$ in this chapter is strongly motivated by the fact that the computation of $\widehat{f}_{n,M}(x)$ for all reasonable values of x – mainly those inside the range of our latency measurement values – reveal a clear smooth function on real data, see 2.1, which is stable across time and *batches*, with good convergence properties. This is highly indicative that the *source* distribution of latency

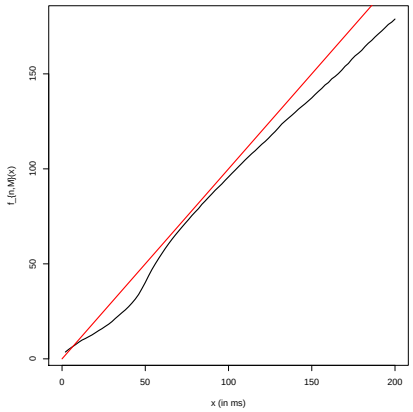

Figure 2.1: *Graph of* $\widehat{f}_{n,M}$ *(black line) plotted against the identity (red line) for* x *ranging from 20 to 200.*

measurements in the network is the push forward distribution of the *proxy* measurements.

This function is of particular interest because it does not change over time: the estimator in (2.2) always converges to the same values. Independently of the time of the day or the state of the *source* or *proxy* when looked at separately, the distribution of the *source* can always be estimated by transporting the distribution of the *proxy* with that function. This function quantifies the missing information that is lost when one only looks at the *proxy*.

The Bahadur decomposition of empirical quantiles in [6] will be a key ingredient in establishing the rates of convergence of the estimator uniformly in the vicinity of x in probability for increasing f in Theorem 4 of Section 2.2. In Section 2.3 we numerically study the estimator on simulated data that mimic real data where f is chosen to be a function that has a similar profile as the one revealed on real data in Figure 2.1. In Section 2.4 we conduct a detailed experiment on real data and provide empirical evidence that the estimator does indeed converge on real data and can be used to correct the bias and predict accurately the *source* distribution.

Since we do not have the theoretical guarantee of the existence of an increasing f on real data a priori, we will observe that a sufficient condition for our estimator to converge withe rates as in Theorem 4 is that $G^{-1} \circ F(x)$ must be constant across all $P \in \mathcal{Z}$ where F is the cumulative function of P and G is the cumulative distribution function of $f_{\#}P$. We propose a test for equality of $G_i^{-1} \circ F_i(x)$ to gain deeper insight that Figure 2.1 does actually reveals the true transport. The test is an adaptation to the Wald test for equality of multiple parameters [45], using an appropriate characterization of weak convergence of the stochastic process $\sqrt{n}(G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x))$ in terms of asymptotic equicontinuity [71].

We then demonstrate that using the estimator to correct the *proxy* distribution leads to significant improvements in the estimation of the *source* quantiles. We also propose a model for the convergence in order to estimate the rates. Finally, we show adequacy between experimental

results and theoretical results.

In section 2.6 we open a discussion concerning non increasing f. We investigate the existence of sufficient conditions on f and $\mathcal{Z}$ that guarantee the uniqueness of f and the fact the our estimator does converge to it. One major issue with non increasing f is that for such f, if P,Q are two probability distributions such that $Q = f_{\#}P$, there always exists a function h such that $h \neq f$ and $Q = f_{\#}P = h_{\#}P$. If one observes samples from P and Q, there is no way to determine which of f or h generated the data, hence there is no guarantee that f can be estimated. In fact, we will show that in general, if f is not increasing, it cannot be estimated.

2.1.3 Data

Before going into the experiments and theoretical analysis, we shall provide the reader with detailed elements of contexts concerning Citrix and its data. The collection of computer files that compose a website are stored in a server that is connected to the network, called the *Origin*. Large audience websites typically need more than just the *Origin* server to handle the trafic in order to avoid high latency, poor throughput or even unavailability of the website. One solution to this problem is to use a *CDN* (Content Delivery network). A *CDN* is a company that rents to Websites systems of distributed servers that deliver the webpages to their Internet users depending, for instance, on their geographic locations. A *CDN* makes a copy of the *Origin* server on a particular subset of servers owned by the *CDN*, called a *map*. The idea is to spread the audience to avoid overloading one specific server that can not handle all the trafic at once, while avoiding building a whole infrastructure from scratch. These kind of technologies handle a large portion of the Internet traffic today.

Citrix collects data to measure the performance delivered by the *CDN*'s. Each time an end user has loaded a webpage of one of Citrix' customer, a javascript tag in the source code of the webpage initiates a ping test on numerous *maps* monitored by Citrix. Those tests consist in downloading a test object of 43 octets on some specific *map* and measuring several performance metrics aggregated in a *report*. *Reports* contain the latency measurement of the test object, along with the user's location, Internet provider, timestamp, identification number of the *map* etc. Latency measurements are expressed in ms. Those reports allow Citrix to measure in real time the performance of the numerous networks from all over the world.

Two kind of measurements are possible: *public* and *private* . A *private* measurement, if enabled by the customer, directly measures the *map* of the said customer: the object is physically on the customer's *map* along with other assests of the website. Such *maps* will be called *private maps*. Only the audience of that customer can trigger *private* measurements, because they require direct connexion to the infrastructure of the customer. A *public* measurement differs only in the location of the test object. It consists in doing the ping test on a *map* that belongs to Citrix, and not one belonging to a customer. Such *maps* will be called *public maps*. Those measurements can be triggered by anyone surfing any Website of any Citrix's customer. Citrix owns hundreds of *maps* in order to have a global overview of the *CDN* market offer. Of particular interest are *related public* and *private maps*. In this case, the *public map* is the *proxy*, and the *private map* is the *source*. In this chapter, *proxy* and *private map* will by synonymous, as well as *source* and *public map*. *Public* data are often used in place of *private* data for *load-balancing* purposes. *Public* measurements are much more numerous than *private* measurements because the entire Citrix community, that

is the sum of the audiences of all its customers can trigger them, but they are a *proxy*, hence potentially less accurate since they do not measure directly the infrastructure of the customer. In later sections, we will preferably use the term *proxy* (resp *source*) than *private map* (resp *public map*) for sake of generality.

2.2 Pointwise estimation of the transport

2.2.1 Presentation of the estimator

Let $a, b \in \mathbb{R}$ such that $a < b$ and $\delta > 0$ such that $\delta < (b-a)/2$. Denote $\mathcal{R} = [a, b]$ and $\mathcal{R}_\delta = [a + \delta, b - \delta]$. For $A, B, C > 0$ three positive reals, define:

$$\mathcal{F}_{\mathcal{R}}(A, B, C) = \Big\{ p \in \mathcal{C}^1(\mathcal{R}), p \geq 0, \int p(u)du = 1, A \leq p(u) \leq B, |p'(u)| \leq C, \text{ for all } u \in \mathcal{R} \Big\}.$$

In other words $\mathcal{F}_{\mathcal{R}}(A, B, C)$ is a set of probability densities on $\mathbb{R}$ that are differentiable, bounded above and below and have first derivative continuous and bounded on $\mathcal{R}$. Define

$$\mathcal{Z} = \Big\{ P(du) = p(u)du | p \in \mathcal{F}_{\mathcal{R}}(A, B, C) \Big\}$$

the set of probability distributions whose densities are in $\mathcal{F}_{\mathcal{R}}(A, B, C)$. Let $M, n \in \mathbb{N}$ be positive integers. We place ourselves on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where

$$\Omega = \mathcal{Z}^M \times \mathbb{R}^n \times \mathbb{R}^n$$

$$\mathcal{A} = \mathcal{B}(\Omega) \quad \text{the Borel } \sigma\text{-algebra on } \Omega$$

and

$$\mathbb{P}(d\omega_1, \ldots, d\omega_M, dx_1, \ldots, dx_n, dy_1, \ldots, dy_n) = \bigotimes_{i=1}^{M} \left[\bigotimes_{j=1}^{n} P(\omega_i, dx_j) Q(\omega_i, dy_j) \right] \mathbb{M}(d\omega_i)$$

where $\mathbb{M}$ is a random measure with value in $\mathcal{Z}$, and $P(\omega, \cdot), Q(\omega, \cdot)$ are probability measures on $\mathbb{R}$ that belong to $\mathcal{Z}$. For $1 \leq i \leq M$, abusing notation slightly, the observables of this random experiment are constructed as follows: first draw

$$P(\omega_i, dx) \sim \mathbb{M}(d\omega_i)$$

then let

$$Q(\omega_i, dx) = f_{\#} P(\omega_i, dx)$$

where $f : \mathbb{R} \to \mathbb{R}$. $Q(\omega_i, dx)$ is the pushforward measure of $P(\omega_i, dx)$ under f, i.e. for all Borel sets A

$$Q(\omega_i, A) = P(\omega_i, f^{-1}(A)).$$

Finally draw independently two samples $(X_1^i, \dots, X_n^i)$ and $(Y_1^i, \dots, Y_n^i)$ with common distributions $P(\omega_i, dx)$ and $Q(\omega_i, dy)$ respectively. The random samples $(X_1^i, \dots, X_n^i)$ and $(Y_1^i, \dots, Y_n^i)$ will be called a *batch*. When no confusion is possible, we shall write

$$P_i := P(\omega_i, dx), \quad Q_i := Q(\omega_i, dy).$$

Let $F(\omega_i, \cdot) := F_i$ and $G(\omega_i, \cdot) := G_i$ the cumulative distribution functions of X_1^i and Y_1^i respectively. The last notation we need is the conditional distribution

$$\Lambda_n(\omega_1, dx_1, \dots, dx_n, dy_1, \dots, dy_n) = \bigotimes_{j=1}^{n} P(\omega_i, dx_j) Q(\omega_i, dy_j). \tag{2.3}$$

In other words, $\Lambda_n(\omega_1, dx_1, \dots, dx_n, dy_1, \dots, dy_n)$ is the joint probability distribution of the samples $(X_1^1, \dots, X_n^1)$ and $(Y_1^1, \dots, Y_n^1)$ conditional on $P(\omega_1, dx)$.

Remark 3. *It is easy to notice that f in 2.1 can not be estimated over its entire domain. Suppose for instance that f is increasing and that we want to estimate $f(x)$ for some $x \in \mathbb{R}$ from the observation of M batches. If the set $\mathcal{Z}$ contains a measure P such that $P(\mathcal{R}) = 0$, then it follows immediately from $Q = f_{\#}P$ and because f is increasing that $Q(f(\mathcal{R})) = 0$. In other words, the observation of random samples distributed from P and Q respectively will give no information on the value $f(x)$. It is necessary to impose local constraints on the measures in $\mathcal{Z}$ in a vicinity of x.*

Suppose f is not known and that one only observes the M *batches*, i.e. the independent random samples $(X_1^i, \dots, X_n^i)$ and $(Y_1^i, \dots, Y_n^i)$ for all $1 \leq i \leq M$. In this setting, we are interested in the estimation of f over the set $\mathcal{R}_\delta$.

Assumption 1. *f is strictly increasing, twice differentiable, and*

$$\|1/f'\|_{L^\infty(\mathcal{R})} < \infty.$$

In other words, f' must be bounded away from 0 on $\mathcal{R}$.

Remark 4. *Since f is twice differentiable, f' and f'' are continuous. Because $\mathcal{R}$ is a compact set, it follows that*

$$\|f'\|_{L^\infty(\mathcal{R})} < \infty, \quad \|f''\|_{L^\infty(\mathcal{R})} < \infty.$$

Assumption 2. *Let $P \sim \mathbb{M}$ and $Y \sim Q = f_{\#}P$. Then $\mathbb{E}(Y^2) < \infty$.*

Theorem 4. *Let $M, n \in \mathbb{N}$ be positive integers and define $\widehat{f}_{n,M}(x)$ as in (2.2) for $x \in \mathcal{R}_\delta$. Work under Assumptions 1 and 2, we have:*

$$\left|\widehat{f}_{n,M}(x) - f(x)\right| \lesssim \frac{1}{\sqrt{Mn}} + \frac{\log(n)^{3/2}}{n^{3/4}} + Me^{-n\delta^2 A^2/2} \tag{2.4}$$

in probability, uniformly over $\mathcal{R}_\delta$, and where $\lesssim$ means inequality up to a positive constant that depends only on $A, B, C, \|f'\|_{L^\infty(\mathcal{R})}, \|f''\|_{L^\infty(\mathcal{R})}, \|1/f'\|_{L^\infty(\mathcal{R})}$ and $\mathbb{E}(Y^2)$.

2.2.2 Assessing the monotonicity of the transport

As stated in section 2.1.2, when used on real data, our estimator reveals a smooth function f, see Figure 2.1. Theorem 4 is valid for increasing functions only. Can we guarantee that our estimator converges on real data to the true transport? If a true transport f exists and is not increasing, it is possible that our estimator defined on (2.2) converges to a function $h \neq f$.

To illustrate, let us give an examples where our estimator would fail to estimate the true transport, but still be converging to a smooth increasing function. Denote by $\mathcal{U}(a, b)$ the continuous Uniform distribution over the interval $[a, b]$ where $a < b$. Let $f(x) = x^2$, $A_i \sim \mathcal{U}(1, 2)$, $P_i = \mathcal{U}(-A_i, A_i)$, $Q_i = f_{\#}P_i$. In this case, $\widehat{f}_{n,M}$ will not converge towards f but, as the result of the law of large numbers, to $x \mapsto \frac{1}{4}\mathbb{E}(x + A)^2$, see Figure 2.2.

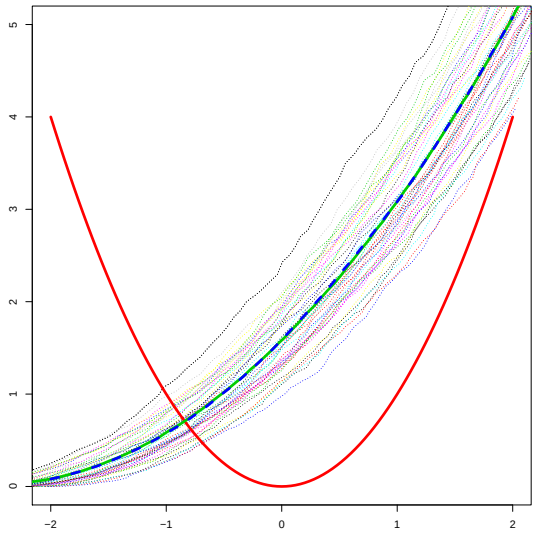

Figure 2.2: $f(x) = x^2$, $A_i \sim \mathcal{U}(1, 2)$, $P_i = \mathcal{U}(-A_i, A_i)$ *and* $Q_i = f_{\#}P_i$. *Red curve is* f, *dashed blue is* $x \mapsto \frac{1}{4}\mathbb{E}(x + A)^2$, *green is* $\widehat{f}_N$, *and thin dashed lines are* $G_{n,i}^{-1} \circ F_{n,i}$ *for all* i.

In this example, $\widehat{f}_{n,M}$ converges to an increasing function: the expected value of $G^{-1} \circ F(x)$, but typically at a rate smaller than the rate in Theorem 4. The reason for this is because in this example, each choice of distribution P_i leads to a different value for $G_i^{-1} \circ F_i(x)$. The convergence and rate of convergence obtained in Theorem 4 comes from aggregating terms that individually converge to 0. The increasing assumption precisely serves that purpose: it guarantees that $G^{-1} \circ F$ is independent of the choice of $P \in \mathcal{Z}$, see Proposition 8. Conversely, the independence of $G^{-1} \circ F$ of the choice of $P \in \mathcal{Z}$ implies the existence of an increasing function h such that $h_{\#}P = f_{\#}P$ for all $P \in \mathcal{Z}$, see in Proposition 6. In this case, one can easily see that our estimator will converge towards h instead of f, and at the rate in Theorem 4. From a practical perspective it is of no consequence since $h_{\#}P = f_{\#}P$ for all $P \in \mathcal{Z}$ implies that in our model, f and h play identical role and are perfectly indistinguishable: no experience can be conducted to determine that indeed the

estimator did not converge towards f but h. Think about symmetric distributions with respect to 0 and $f = -\mathrm{id}_{\mathbb{R}}$ for instance. In this case the estimator will converge to $h = \mathrm{id}_{\mathbb{R}}$, and there is no way to determine that the true transport that generated the data was $-\mathrm{id}_{\mathbb{R}}$. We can safely ignore this problem by introducing an equivalence relation on the set of functions $\mathcal{C}^2(\mathbb{R})$ such that

$$f \sim g \text{ if and only if } f_{\#}P = h_{\#}P \text{ for all } P \in \mathcal{Z}.$$

If $f \in \mathcal{C}^2(\mathbb{R})$ is the true transport and if $G^{-1} \circ F$ is independent of $P \in \mathcal{Z}$ where F, G are the cumulative distribution functions of $P, Q = f_{\#}P$ respectively, then there exists an increasing h in $[f] = \{g \in \mathcal{C}^2(\mathbb{R}) | g \sim f\}$, and we guarantee that the estimator converges to h at the rates in Theorem 4, this is the purpose of Proposition 6.

For $1 \leq i \leq M$, when we observe the samples $(X_1^i, \dots, X_n^i)$ and $(Y_1^i, \dots, Y_n^i)$ drawn from P_i and $Q_i = f_{\#}P_i$ respectively where $(P_1, \dots, P_M)$ are independent with common distribution $\mathbb{M}$, testing for equality of all $G_i^{-1} \circ F_i(x)$ for different values of x is a reasonable way to determine whether or not $\widehat{f}_{n,M}$ is an estimate of the true transport. For $x \in \mathbb{R}$, we want to test:

$$H_0^x : G_1^{-1} \circ F_1(x) = \cdots = G_M^{-1} \circ F_M(x)$$
$$\text{against } H_1^x : \exists\, i \neq j, G_i^{-1} \circ F_i(x) \neq G_j^{-1} \circ F_j(x).$$

In order to test those hypothesis, we will use the fact that $G_{n,i}^{-1} \circ F_{n,i}(x)$ is asymptotically Gaussian.

Theorem 5. *Work under Assumptions 1 and 2. Let*

$$P \sim \mathbb{M} \text{ and } Q = f_{\#}P$$

and $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ be independent samples with common distributions P and Q respectively. Denote F, G the cumulative distributions of P, Q respectively, F_n, G_n their empirical counterparts and g the density of Q. We have

$$\frac{\sqrt{n}}{\sigma(F)}(G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)) \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}(0,1) \tag{2.5}$$

where

$$\sigma(F)^2 = 2\frac{F(x)(1-F(x))}{(g \circ G^{-1} \circ F(x))^2}.$$

If $\sigma(F)^2$ is unknown, we show that it can be estimated and that the limiting distribution in Theorem 5 is preserved, see Proposition 4.

Proposition 4. *Work under Assumptions 1 and 2. Let*

$$P \sim \mathbb{M} \text{ and } Q = f_{\#}P$$

and $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ be independent samples with common distributions P and Q respectively. Denote F, G the cumulative distributions of P, Q respectively, F_n, G_n their empirical counterparts, and g be the density distribution of Q. Let

$$\sigma_n(F)^2 = 2\frac{F_n(x)(1-F_n(x))}{\left(g_n \circ (\widehat{G_n^1})^{-1} \circ F_n(x)\right)^2}$$

where:

$$g_n(t) = \frac{1}{nh_n}\sum_{j=1}^{n} K\left(\frac{Y_j - t}{h_n}\right), \quad h_n > 0$$

is a kernel density estimator of g [82], [93]. Suppose moreover that the following conditions on K and h_n hold:

i) K is right continuous.

ii) K is compactly supported.

iii) $\sum_{n\leq 1} e^{-\gamma n h_n^2} < \infty$ *for all* $\gamma > 0$.

Then $\sigma_n(F)^2$ *converges in probability to* $\sigma(F)^2$ *where*

$$\sigma(F)^2 = 2\frac{F(x)(1-F(x))}{\left(g \circ G^{-1} \circ F(x)\right)^2}.$$

Finally, we present in Proposition 5 a test for H_0^x versus H_1^x at level $\alpha \in (0,1)$.

Proposition 5. *Let:*

$$\boldsymbol{\theta}_x = \begin{bmatrix} G_1^{-1} \circ F_1(x) \\ G_2^{-1} \circ F_2(x) \\ \vdots \\ G_M^{-1} \circ F_M(x) \end{bmatrix} \tag{2.6}$$

$$\boldsymbol{\theta}_{n,x} = \begin{bmatrix} (G_n^1)^{-1} \circ F_n^1(x) \\ (G_n^2)^{-1} \circ F_n^2(x) \\ \vdots \\ (G_n^M)^{-1} \circ F_n^M(x) \end{bmatrix} \tag{2.7}$$

$$\boldsymbol{\Sigma_n} = \begin{bmatrix} \sigma_n(F_1)^2 & & \\ & \ddots & \\ & & \sigma_n(F_M)^2 \end{bmatrix}$$

and

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 & & \\ \vdots & & \ddots & \\ 1 & & & -1 \end{bmatrix}$$

where $\boldsymbol{R}$ *is a* $(M-1)\times M$ *matrix where the first column is a vector of* 1*'s and the remaining columns is a* $(M-1)\times(M-1)$ *matrix with diagonal elements equal to* -1 *and 0 everywhere else. Then:*

$$W_n := n\left\|\left(\boldsymbol{R\Sigma_n R}^t\right)^{-1/2}\boldsymbol{R\theta}_{n,x}\right\|^2 \xrightarrow{d} \chi^2(M-1)$$

under H_0^x *as* $n \to \infty$*, where* $\chi^2(M-1)$ *denotes the* χ^2 *distribution with* $M-1$ *degrees of freedom. Hence for* $\alpha \in (0,1)$*, the decision rule*

$$\Phi(\alpha) = \mathbf{1}_{\mathcal{R}(\alpha)}$$

is an asymptotic test of level α *of the null hypothesis* H_0^x *with rejection zone*

$$\mathcal{R}(\alpha) = \{W_n > q_\alpha^{\chi^2(M-1)}\}$$

where $q_\alpha^{\chi^2(M-1)}$ *is the quantile of order* α *of the* $\chi^2(M-1)$ *distribution.*

2.3 Illustration on a toy example

We propose in this section to implement the estimator on a toy example where the samples are log normally distributed, because real latency measurements from the *source* and *proxy* can be roughly seen as log normal realizations as seen in Figure 2.3. We recall that X has a log normal distribution with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$ if $X > 0$ a.s. and $\log(X) \sim \mathcal{N}(\mu, \sigma^2)$. In this case we write $X \sim \mathcal{LN}(\mu, \sigma^2)$.

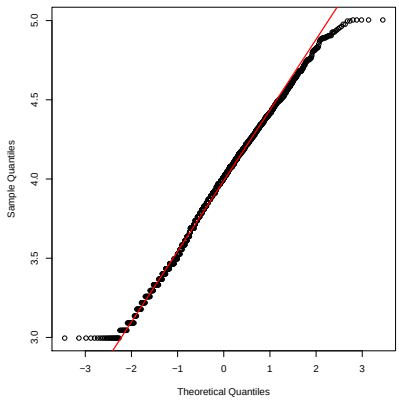

Figure 2.3: *Example of normal Q-Q plot where the quantiles of a log proxy sample is compared to the theoretical quantiles of the normal distribution. Reveals a Gaussian behavior with slightly heavier tails.*

The function f that we will consider is chosen to have similarities to the function in Figure 2.1. We use $f : x \mapsto x/2 - e^{-x^3/3} + 1$, see Figure 2.4, and propose the pointwise estimation of f at $x = 2.5$. It is easily verified that f meets Assumption 1 for $\mathcal{R} = (2,3)$ for instance.
Data generation: We fix a sample size n and the number of *batch* m is set to $m = \sqrt{n}$. For the i-th batch, $1 \leq i \leq m$, we draw $u \sim \mathcal{U}([0,1])$ and independently $s \sim \mathcal{U}([1/2,1])$. Then we generate an i.i.d. n-sample $\mathcal{X}_i$ from $\mathcal{LN}(u,s)$ and independently we generate an i.i.d. n-sample $\mathcal{Y}_i$

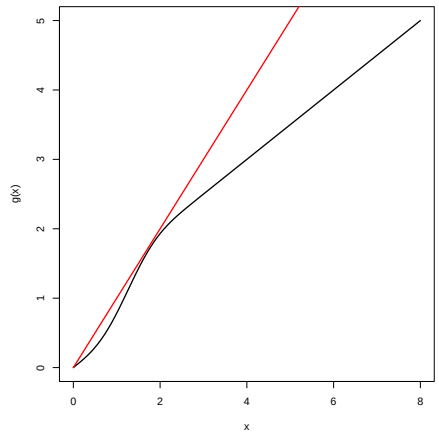

Figure 2.4: *Graph of f for x ranging from 0 to 8.*

from $f(\log \mathcal{N}(u, s))$ and compute the estimator $\widehat{f}_{n,m}(x)$. To estimate the rates of convergence, we repeat this process for n ranging from 20 to 1.000. The rate of convergence is then estimated from a log-log plot of the absolute difference between $\widehat{f}_{n,m}(x)$ and $f(x)$. Results are shown in Figures 2.5, 2.6 and 2.7.

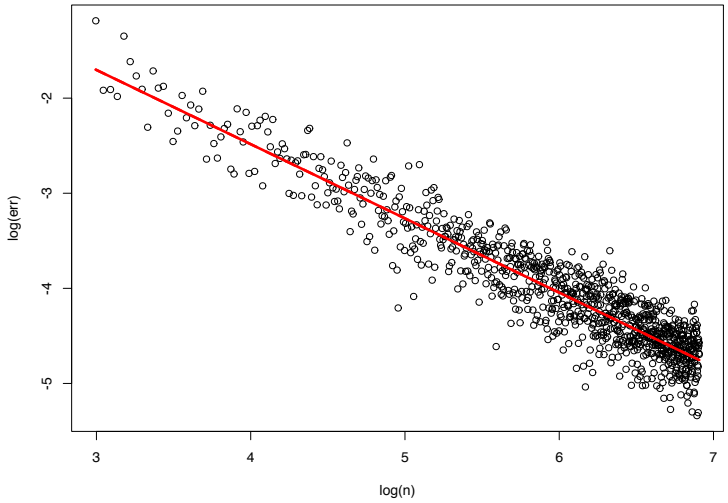

Figure 2.5: *log-log plot of $|\widehat{f}_{n,m}(x) - f(x)|$, with $m = \sqrt{n}$ and n ranging from 20 to 1.000 for $x = 2.5$ with fitted linear model (solid red). $R^2 = 0.87$, Slope $= -0.78$.*

The log-log plots of the error against n shows strong evidence of a linear relation. For $x = 2.5$, fitting a linear model across the 200 simulations led to a median R^2 of 0.87 and median slope of -0.78.

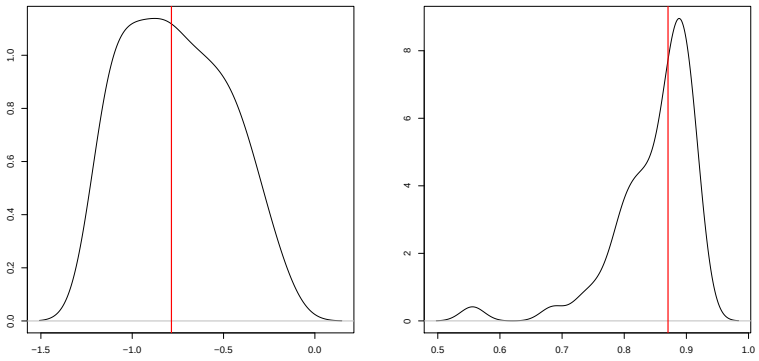

Figure 2.6: *Left: distribution of the slopes in the linear models fitted across the 200 simulations with median in red. Right: distribution of the R^2 coefficients in the linear models fitted across the 200 simulations with median in red. $x = 2.5$.*

Outside this range, $\widehat{f}_{n,m}$ deviates from f because of data scarcity.

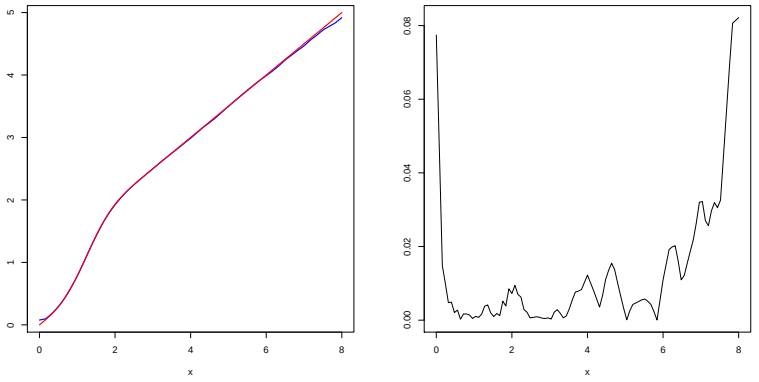

Figure 2.7: *Graph of f and $\widehat{f}_{n,m}$ (left) and absolute differences between f and $\widehat{f}_{n,m}$ (right) for x ranging from 0 to 8. The estimation is less accurate on the edges, because of data scarcity.*

2.4 Empirical results

2.4.1 The experiment

Given a *proxy* – *private map* – and *source* – *public map* – we want to determine whether or not the distribution of the *source* is a monotonic transport of the distribution of the *proxy*.

2.4.2 Presentation of the data

The latency measurements collected by Citrix from a *map* are modeled as realizations of random variables whose underlying distributions may vary depending on the time of the day. In this experiment we collected 7 consecutive days of latency data from a *source* and *proxy* that were generated by Internet users located in Paris on the Orange ISP between 05/14/18 and 05/20/18. Denote this period of time $[0, T]$ where $T = 604800$ is the number of seconds in a week, and $0 = t_0 < t_1 < \cdots < t_K = T$ a partition of $[0, T]$ such that $t_{i+1} - t_i = h > 0$ for all i. The *CDN* performance varies across time, so the choice of h is important in order to obtain i.i.d. samples.

To choose h, we proceed as follows. Fix $h > 0$ and let S_i^h be the collection of all measurements with timestamps falling in the interval $]ih/2, (i+1)h/2]$ for $i = 0, \ldots, 2T/h$. We then perform a hypothesis testing that the samples S_i^h and S_{i+1}^h come from the same distribution using the two-samples Kolmogorov-Smirnov test [65] and let p_i^h be the corresponding p-value. Finally we analyze the distribution of those p-values for varying h and choose the larger h so that the distribution of the p_i^h is uniform, see Figure 2.8. The value $h = 60$ min is chosen and aligns with operational engineers recommandations.

Once the value for h is chosen, we fix the partition $(t_i)_{0 \le i \le K}$ of $[0, T]$. Measurements with timestamps falling between times t_{i-1} and t_i will form the samples. We then perform uniform subsampling without replacement to force the same number of observations N per sample to meet the model assumptions. The number of measurements N is chosen so that the number of *batches* containing at least N measurements is $K' = 2\sqrt{N}$ leading to $N = 1685$ and $K' = 82$. The other samples are discarded because of too few measurements, corresponding to the middle night periods. Denote the selected samples from both *maps*by $(\mathcal{X}_i)_{1 \le i \le 82}$ and $(\mathcal{Y}_i)_{1 \le i \le 82}$ respectively. Now let $M = K'/2 = 41 = \sqrt{N}$, and use $(\mathcal{X}_i)_{1 \le i \le M}$ and $(\mathcal{Y}_i)_{1 \le i \le M}$ as training data to build $\widehat{f}_{N,M}(x)$ for $x = 1, \ldots, 200$ according to (2.2), see Figure 2.9. Since $M = \sqrt{N}$, we shall just write $\widehat{f}_N$. The remaining *batches* are used for testing our estimator.

2.4.3 Monotonicity of the true transport

As explained in Section 2.2.2, $\widehat{f}_N$ will converge to the true transport only if f is increasing. To assess the monotonicity of the true transport f on real data we use the test derived in Proposition 5. The test statistic

$$W_n = n \left\| \left(\boldsymbol{R} \boldsymbol{\Sigma}_{\boldsymbol{n}} \boldsymbol{R}^t \right)^{-1/2} \boldsymbol{R} \boldsymbol{\theta}_{n,x} \right\|^2$$

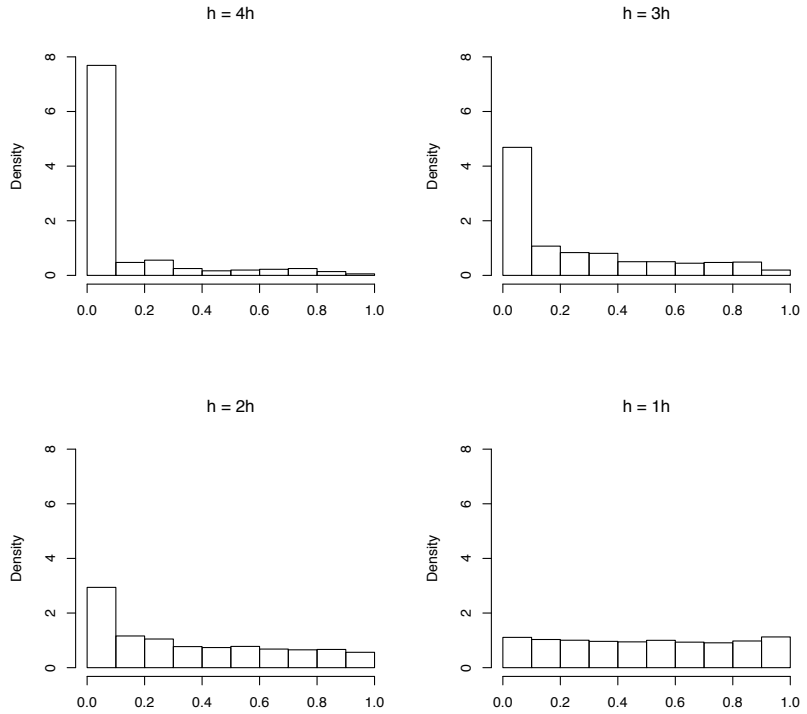

Figure 2.8: *Histogram of the p_i^h, for varying h. For large values of h, the p-values concentrate near 0, suggesting that the samples S_i^h and S_{i+1}^h containing latency measurements with timestamps falling in consecutive time intervals of length $h/2$ do not come from the same distribution. This suggests that over periods of time of length $h/2+h/2 = h$, the distribution of latency measurements changes. As h decreases, less p-values p_i^h concentrate near 0 and a uniform distribution is revealed.*

depends of the kernel estimation of the density of the measurements of the *source*. We choose the triangle kernel and the Silverman rule of thumb [82] for bandwidth selection, i.e. $h = h_n$ is chosen to be $0.9n^{-1/5}$ times the minimum between the sample standard deviation and the interquartile range divided by 1.34. In order to guarantee sufficient density of points, we tested H_0^x for x ranging from 20 to 150 since there were too few data points outside that range. 17% of the tests rejected H_0^x on our data, with $\alpha = 0.05$, see Figure 2.10.

The values x for which the test was rejected is concentrated in a region of the real line corresponding on average to quantiles of order 0.2 to 0.4 for the *proxy* distribution, suggesting that the estimation of f in that region is not as reliable as it is outside that range. Denote $I_{\text{Reject}} = [35, 55]$ this range, and $I_{\text{Accept}} = I_{\text{Reject}}^c$. The average p-values is 0.006 for the tests where H_0^x was rejected, and 0.61 for the tests where H_0^x failed to be rejected, suggesting that

$$\forall\ x \in I_{Accept},\ \forall\ 1 \leq i \leq M\ ,\ G_i^{-1} \circ F_i(x) \text{ is independent of } i.$$

Moreover, even for the values of x where H_0^x was rejected, the estimator still converges. Heuristi-

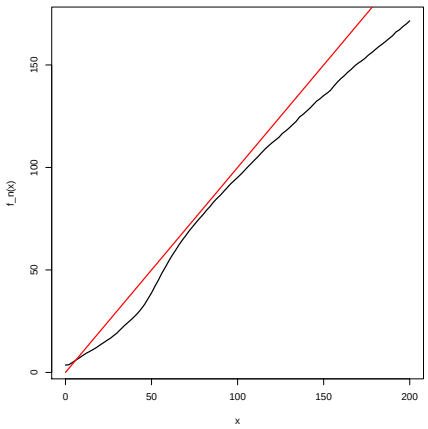

Figure 2.9: *Graph of $\widehat{f}_N$ (black line) plotted against the identity (red line).*

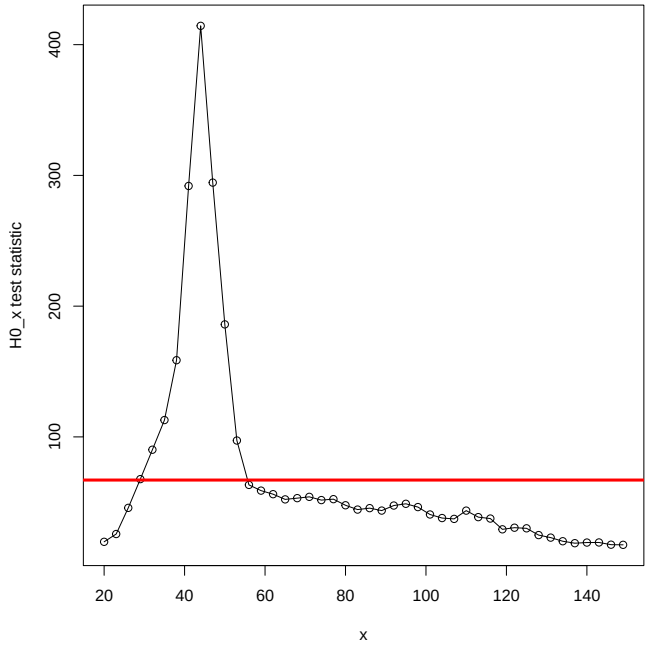

Figure 2.10: *Test statistic for H_0^x against x with critical value $q_{1-\alpha}^{\chi^2}(M)$ in red.*

cally, under the hypothesis that $G_1^{-1} \circ F_1(x)$ is integrable and if one allows n to be arbitrarily large in front of M, then $\widehat{f}_n(x)$ is close to the expected value of $G_1^{-1} \circ F_1(x)$. This suggests that the

distribution of *proxy* and *source* measurements are indeed transported versions of one another by a deterministic transport, and this transport may be increasing only on I_{Accept}. $\widehat{f}_n(x)$ is an estimate of that transport on I_{Accept}, and an estimate of the average behavior of $G_i^{-1} \circ F_i$ on I_{Reject}. As seen above, if $G_i^{-1} \circ F_i(x)$ is not independent of i, the rates of convergence in Theorem 4 should not be met. We will present empirical evidence of this fact in Section 2.4.5 about the analysis of the rates of convergence.

2.4.4 Performance of the estimator

The goal of the estimation of $\widehat{f}_N$ is to reproduce the behavior of the *source* when one only has access to the *proxy*. The estimator $\widehat{f}_N$ is built on the train sets $(\mathcal{X}_i)_{1\leq i\leq M}$ and $(\mathcal{Y}_i)_{1\leq i\leq M}$. We can assess the performance of $\widehat{f}_N$ on the test sets $(\mathcal{X}_i)_{M+1\leq i\leq 2M}$ and $(\mathcal{Y}_i)_{M+1\leq i\leq 2M}$ by comparing $d(\mathcal{X}_i, \mathcal{Y}_i)$ and $d(\widehat{f}_N(\mathcal{X}_i), \mathcal{Y}_i)$ for $M+1 \leq i \leq 2M$ where d is a distance between distributions. As seen in section 2.1.1 *load-balancers* heavily rely on the quantiles of the latency distribution to route the users. Hence a natural choice for d is the Wasserstein distance because it acts as the average between the absolute difference of the quantiles of the underlying two distributions. We briefly recall the definition and some facts about this distance, see for instance [90].

Definition 5. *Let P, Q be two probability distributions on $\mathbb{R}$ and $p \geq 1$. The Wasserstein distance of order p between P and Q is defined as:*

$$W_p(P,Q) := \left(\inf_{\pi \in \Pi(P,Q)} \int |x-y|^p \pi(dx,dy) \right)^{1/p}$$

where $\Pi(P,Q)$ is the set of probability measures with marginal P and Q respectively.

When P, Q are probability distributions over the real line, the W_p has the closed form:

$$\begin{aligned} W_p(P,Q) &= \left(\int_{\mathbb{R}} |F(x) - G(x)|^p dx \right)^{1/p} \\ &= \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p} \end{aligned}$$

where F, G are the c.d.f. of P and Q respectively, and F^{-1}, G^{-1} their generalized inverse. If one observes the i.i.d. samples $(X_1, \dots, X_n)$ with distribution F and $(Y_1, \dots, Y_n)$ with distribution G, the empirical Wasserstein distance is defined defined as the Wasserstein distance between the two empirical distributions:

$$W_{p,n}(F,G) := W_p(F_n, G_n) = \left(\int_0^1 |F_n^{-1}(u) - G_n^{-1}(u)|^p du \right)^{1/p}$$

where $F_n(u) = n^{-1} \sum_{i=1}^n 1_{\{X_i \leq u\}}$ and $G_n(u) = n^{-1} \sum_{j=1}^n 1_{\{Y_j \leq u\}}$ are the empirical cumulative distributions and F_n^{-1} and $(\widehat{G_n^1})^{-1}$ denote their general inverse, or empirical quantile function. The empirical Wasserstein distance has the closed form:

$$W_{p,n}(F,G) = \left(\frac{1}{n} \sum_{i=1}^n |X_{(i)} - Y_{(i)}|^p \right)^{1/p}$$

where $(X_{(i)})_{1\leq i\leq n}$ and $(Y_{(i)})_{1\leq i\leq n}$ are the order statistics of the two samples, ie $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$.

We choose the Wasserstein distance of order $p = 1$ as the criteria to assess the performance of our estimator by comparing $W_{1,N}(\mathcal{X}_i, \mathcal{Y}_i)$ with $W_{1,N}(\widehat{f}_N(\mathcal{X}_i), \mathcal{Y}_i)$ for $M+1 \leq i \leq 2M$. Results are presented in Figure 2.11 and Table 2.1.

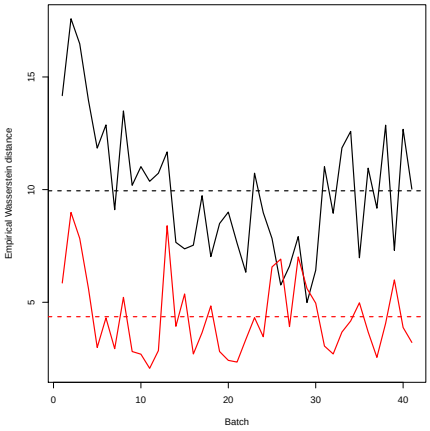

Figure 2.11: *Wasserstein distances computed for the 42 test batches between $\mathcal{X}_i$ and $\mathcal{Y}_i$ in black, and $\widehat{f}_N(\mathcal{X}_i)$ and $\mathcal{Y}_i$ in red. Dashed lines are the respective means.*

$W_{1,n}(\mathcal{X}, \mathcal{Y})$	$W_{1,n}(\widehat{f}_N(\mathcal{X}), \mathcal{Y})$	Relative change	$W_{1,n}(\widehat{f}_N(\mathcal{X}), \mathcal{Y}) < W_{1,n}(\mathcal{X}, \mathcal{Y})$
9.95	4.36	-56%	95%

Table 2.1: *Empirical Wasserstein distance between the source and proxy (column 1), empirical Wasserstein distance between the source and transported proxy using our estimator $\widehat{f}_n$ (column 2), relative change in the empirical Wasserstein distance resulting from transporting the proxy (colonne 3), percentage of batchs for which the empirical Wasserstein distance distance between the source and proxy was reduced by transporting the proxy.*

The average distance between the *source* and *proxy* over the 42 test *batches* is

$$\frac{1}{42}\sum_{i=43}^{82} W_{1,N}(\mathcal{X}_i, \mathcal{Y}_i) = 9.95$$

whereas the average distance between the *source* and the transported *proxy* using our estimator $\widehat{f}_{n,M}$ over the 42 test *batches* is

$$\frac{1}{42}\sum_{i=43}^{82} W_{1,N}(\widehat{f}_N(\mathcal{X}_i), \mathcal{Y}_i) = 4.36$$

hence reducing this distance on average by 56%. Moreover, for 95% of the test *batches*, we observed a reduction in the distance:

$$\frac{1}{42}\sum_{i=43}^{82} \mathbf{1}_{W_{1,N}(\widehat{f}_N(\mathcal{X}_i),\mathcal{Y}_i)<W_{1,N}(\mathcal{X}_i,\mathcal{Y}_i)} = 95\%$$

For *load-balancing* purposes this means that the prediction error made when estimating the quantiles of the *source* by using those of the *proxy* be cut in half by transforming the measurements from the *proxy* with $\widehat{f}_N$ beforehand, which represents a significant improvement. The correction of the *proxy* distribution using $\widehat{f}_N$ actually improves on all quantiles, not just on average, as seen in Figure 2.12. An example of output on one test *batch* is presented in Figure 2.13.

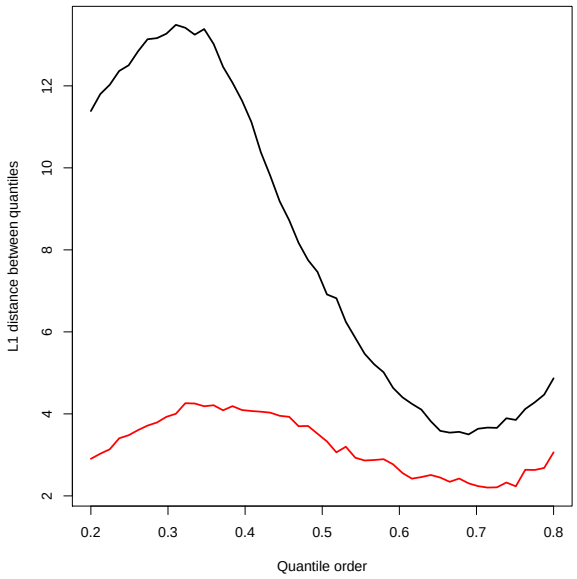

Figure 2.12: *Black line: average L_1 distance between proxy and source quantiles. Red line: average L_1 distance between the transformation of the proxy under $\widehat{f}_{n,M}$ and source quantiles.*

2.4.5 Convergence rate estimation

We now focus in modeling $\widehat{f}_N$ and estimating rates of convergence. The problem we are facing here is that we ignore the limit f. If f were known we could perform a linear analysis in log-log scale of the error $|\widehat{f}_n - f|$ against n. Since f is unknown, we propose to first compute $\widehat{f}_N(x)$ with N as large as possible, compute successive approximation $\widehat{f}_n(x)$, then perform a log-log analysis of the absolute differences. We keep the grid $(t_i)_i$ such that $t_{i+1} - t_i = h = 3600$s, but we

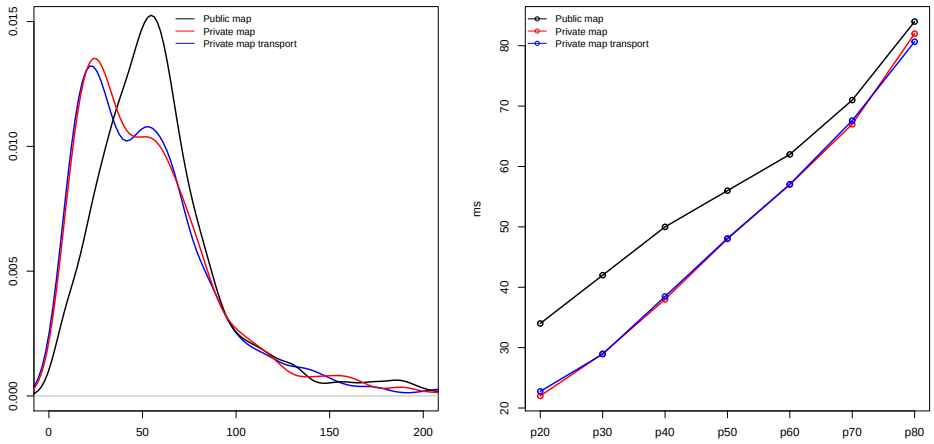

Figure 2.13: *Examples of output on a test sample. Empirical distributions (left) and deciles (right) for the private and public maps, along with the modification of the public map.*

choose N such that the number of samples left is $M = \sqrt{N}$. Here $N = 2082$ and $M = 46$. No train or test sets here because we only focus in the convergence properties. $\widehat{f}_n(x)$ is computed for $x \in \{20, \dots, 150\}$, $n \in \{30, \dots, N = 2082\}$. For each n and x, we compute the absolute difference $|\widehat{f}_n(x) - \widehat{f}_N(x)|$. For each n we randomly selected $m = \sqrt{n}$ samples among the M available, and subsample n observations in those samples among the N available. Results are presented in Figure 2.14.

All log-log plots exhibit the same pattern. A linear relation between $\log|\widehat{f}_n(x) - \widehat{f}_N(x)|$ and $\log(n)$ is present up until $n \approx N/2$, then the decay accelerates. This acceleration is due to the fact that we are looking at $\log|\widehat{f}_n(x) - \widehat{f}_N(x)|$ instead of $\log|\widehat{f}_n(x) - f(x)|$, hence when n approaches N, $|\widehat{f}_n(x) - \widehat{f}_N(x)|$ approaches (and even reaches) 0 much faster than $|\widehat{f}_n(x) - f(x)|$. On the other hand, when n is small in front of N, $|\widehat{f}_n(x) - \widehat{f}_N(x)|$ and $|\widehat{f}_n(x) - f(x)|$ are expected to behave similarly. Because we use only an approximation of the limit instead of the true limit, fitting a linear model and extracting the slope here will overestimate the rate. It seems natural to fit a classic linear model only for small n to estimate the rate of convergence.

To do so, we propose the following approach to have a quantitative insight on the cut off point. Using the results of Theorem 4, a linear model between $\log|\widehat{f}_n(x) - f(x)|$ and $\log(n)$ for $M = \sqrt{n}$ is reasonable. We suppose that there exists $a > 0$, $b \in \mathbb{R}$ and i.i.d. centered Gaussian random variables E_n with variance σ^2 such that

$$\log|\widehat{f}_n(x) - f(x)| = -a\log(n) + b + E_n$$

which can be rewritten

$$\widehat{f}_n(x) = CR_n\mathcal{E}_n n^{-a} + f(x)$$

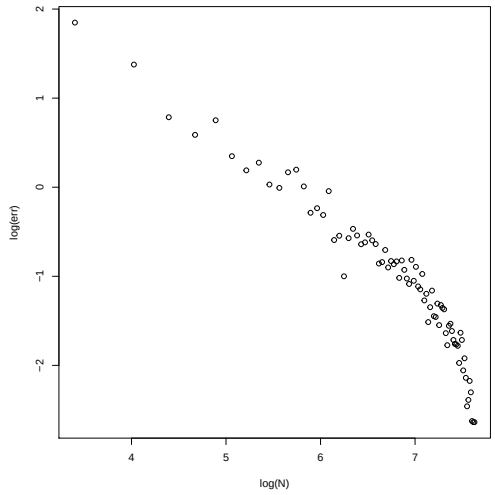

Figure 2.14: *Example of* log-log *plot for* $x = 71$*, when* n *ranges from 30 to* N*.*

where $C = e^b > 0$, $R_n = \text{sign}(\widehat{f}_n(x) - f(x))$ and $\mathcal{E}_n = e^{E_n}$ are i.i.d. log-Normal with parameters $(0, \sigma^2)$. For ease of analysis, we will assume that the variables R_n are i.i.d. Rademacher with parameter $p \in (0,1)$, denoted $\text{Rad}(p)$, i.e. $\mathbb{P}(R_n = 1) = 1 - \mathbb{P}(R_n = -1) = p$. We will use this model for $\widehat{f}_n(x)$. It follows that

$$\log|\widehat{f}_n(x) - \widehat{f}_N(x)| = -a\log(n) + \log(C) + \log(\mathcal{E}_n) + \log\left|1 - \left(\frac{n}{N}\right)^a \tilde{R}_n\tilde{\mathcal{E}}_n\right|$$

where

$$\tilde{R}_n = R_N/R_n \sim \text{Rad}(2p(p-1)+1), \tilde{\mathcal{E}}_n = \mathcal{E}_N/\mathcal{E}_n \sim \mathcal{LN}(0, 2\sigma^2), \quad \text{and } \log(\mathcal{E}_n) \sim \mathcal{N}(0, \sigma^2).$$

In other words, the term $\log|1 - n^a/N^a\tilde{R}_n\tilde{\mathcal{E}}_n|$ is the price to pay for replacing $f(x)$ with $\widehat{f}_N(x)$. If σ^2 is reasonably small, this cost is close to 0 for small n, i.e. $\log|\widehat{f}_n(x) - \widehat{f}_N(x)|$ behaves linearly in $\log(n)$ for values of n below a certain threshold in this model, as observed from the data. We can compute the expected value of $\log|1 - n^a/N^a\tilde{R}_n\tilde{\mathcal{E}}_n|$ as a function of $x = n/N \in (0,1)$. Define $h : [0,1] \mapsto \mathbb{R}$ such that

$$h(x) = \mathbb{E}\log|1 - x^a R\mathcal{E}|$$

where $R \sim \text{Rad}(q)$ and $\mathcal{E} \sim \mathcal{LN}(0, s^2)$ is independent of R. h is not analytical, but can be computed numerically with arbitrary precision.

Figure 2.15 provides insight that indeed for $n < N/2$, the cost $\log|1 - n^a/N^a\tilde{R}_n\tilde{\mathcal{E}}_n|$ plays minimal role in average. Small values of s exhibit an accelerated decay as x goes to 1, when large values of s reveal an inflection point with abscissa getting smaller when s gets larger. This model suggests that the value of σ^2 is indeed small as all graphs of $\log|\widehat{f}_n(x) - \widehat{f}_N(x)|$ against $\log(n)$ exhibit this precise accelerated decay pattern. Assuming $\sigma^2 < 0.05$, we will neglect the random effects of the log-Normal distribution in $h(x) = \mathbb{E}\log|1 - x^a R\mathcal{E}|$. This simplification allows us to compute

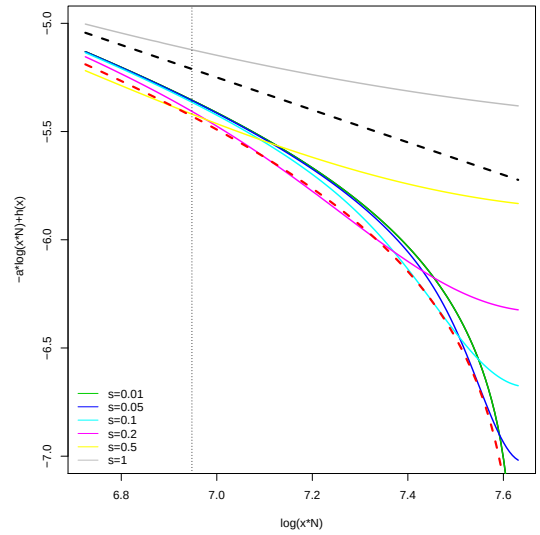

Figure 2.15: *Graphs of* $-a\log(xN)+h(x)$ *against* $\log(xN)$ *for different values of* s *and fixed* $q=1/2$, $a=0.75$. *Dashed red is* $s=0$, *dashed black is the linear part* $-a\log(xN)$. *Vertical light dashed black is* $\log(N/2)$.

analytically h:

$$h(x)=p\log|1-x^a|+(1-p)\log|1+x^a|.$$

Our model for $\log|\widehat{f}_n(x)-\widehat{f}_N(x)|$ now takes the form:

$$\log|\widehat{f}_n(x)-\widehat{f}_N(x)|=C-a\log(n)+p\log\left|1-\left(\frac{n}{N}\right)^a\right|+(1-p)\log\left|1+\left(\frac{n}{N}\right)^a\right|+\mathcal{N}(0,\sigma^2).$$

Using the fact that for $n<N/2$, $\log|\widehat{f}_n(x)-\widehat{f}_N(x)|$ is approximately linear in $\log(n)$, we estimate C by $\hat{C}=\hat{C}(x)$, the offset of the linear fit between $\log|\widehat{f}_n(x)-\widehat{f}_N(x)|$ and $\log(n)$ with slope a for $n<N/2$, i.e.

$$\hat{C}(x)=\sum_{n=1}^{N/2}(\log|\widehat{f}_n(x)-\widehat{f}_N(x)|+a\log(n))$$

Define the objective function:

$$L(a,p)=\sum_x\sum_{n=1}^{N}\left(\log|\widehat{f}_n(x)-\widehat{f}_N(x)|-\hat{C}(x)+a\log(n)-p\log\left|1-\left(\frac{n}{N}\right)^a\right|-(1-p)\log\left|1+\left(\frac{n}{N}\right)^a\right|\right)^2.$$

Our estimator of (a,p) will be:

$$(\hat{a},\hat{p})=\underset{p\in(0,1),a\in(0,2)}{\arg\min}\ L(a,p)$$

We estimate the solution by grid searching the optimal couple of parameters. For instance, numerical implementations of the above problem gives $(\hat{a}, \hat{p}) = (0.78, 0.38)$ for $x = 71$ see Figure 2.16.

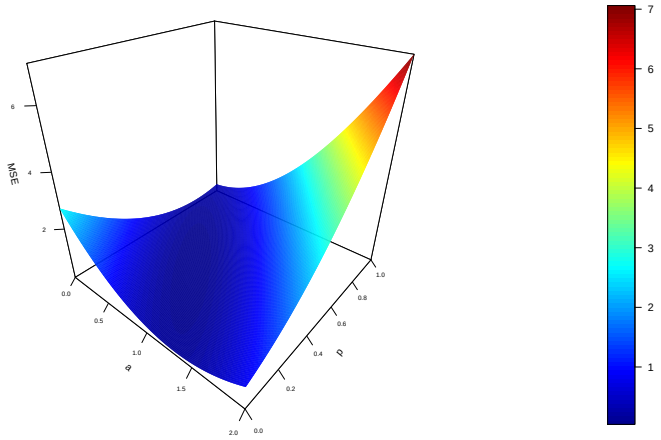

Figure 2.16: *Surface plot of L. The minimum is attained for* $(a, p) = (0.78, 0.38)$. $x = 71$.

Setting $(\hat{a}, \hat{p}) = (0.78, 0.38)$ for $x = 71$, our model for $\log|\widehat{f}_n(x) - \widehat{f}_N(x)|$ finally takes the form:

$$\log|\widehat{f}_n(x) - \widehat{f}_N(x)| = \hat{C}(x) - 0.78\log(n) + 0.38\log\left|1 - \left(\frac{n}{N}\right)^{0.78}\right| + 0.62\log\left|1 + \left(\frac{n}{N}\right)^{0.78}\right|$$

where $\hat{C}(x) = 2/N\sum_{n=1}^{N/2}\log|\widehat{f}_n(x) - \widehat{f}_N(x)| + 0.78\log(n)$ is the offset in the linear regression between $\log|\widehat{f}_n(x) - \widehat{f}_N(x)|$ and $\log(n)$. The fit of the model is shown in Figure 2.17 and shows very good adequation with real data. The average R^2 for this model across all values of x was 0.96, leading in addition uncorrelated residuals with Gaussian distribution. The values of the exponent a for $x \in I_{Accept}$ are consistent with theoretical analysis and the convergence rate obtained in Theorem 4. The estimated rate of convergence across x is presented in Figure 2.18.

2.5 Proofs

2.5.1 Proof of Theorem 4

What we want to do is upper bound $\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K)$ for $x \in \mathcal{R}_\delta$ and $K > 0$ and show that it decreases at a rate as in Theorem 4. One key ingredient of the proof consists in decomposing this probability using the event

$$\xi_{n,M} = \bigcap_{i=1}^{M}\left\{\widehat{F_n^i}(x) \in F_i(\mathcal{R}_{\delta/2})\right\}$$

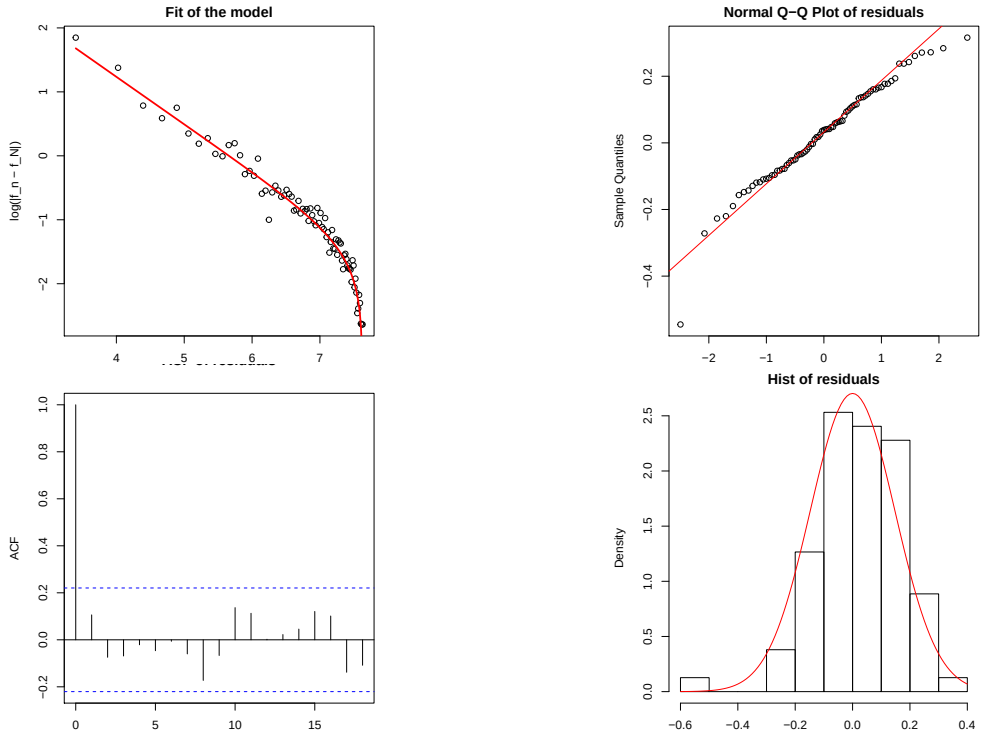

Figure 2.17: *Left to right, top to bottom: fitted model, normal Q-Q plot of the residuals, ACF of the residuals and histogram of the residuals.*

where $\mathcal{R}_{\delta/2} = [a + \delta/2, b - \delta/2]$, and its complementary as follows:

$$\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K) \leq \mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K, \xi_{n,M}) + \mathbb{P}(\xi^c_{n,M}). \qquad (\star)$$

Observe that:

$$\begin{aligned}
\mathbb{P}(\xi^c_{n,M}) &= \mathbb{P}\left(\left[\bigcap_{i=1}^{M} \{\widehat{F^i_n}(x) \in F_i(\mathcal{R}_{\delta/2})\}\right]^c\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{M} \left\{\widehat{F^i_n}(x) \notin F_i(\mathcal{R}_{\delta/2})\right\}\right) \\
&\leq M\mathbb{P}\left(\widehat{F^1_n}(x) \notin F_1(\mathcal{R}_{\delta/2})\right).
\end{aligned}$$

Since $\mathcal{R}_{\delta/2}$ is an interval and F_1 is strictly increasing:

$$\begin{aligned}
\mathbb{P}\left(\widehat{F^1_n}(x) \notin F_1(\mathcal{R}_{\delta/2})\right) &= \mathbb{P}\left(\widehat{F^1_n}(x) \notin [F_1(a + \delta/2), F_1(b - \delta/2)]\right) \\
&= \mathbb{P}\left(\widehat{F^1_n}(x) < F_1(a + \delta/2)\right) + \mathbb{P}\left(\widehat{F^1_n}(x) > F_1(b - \delta/2)\right).
\end{aligned}$$

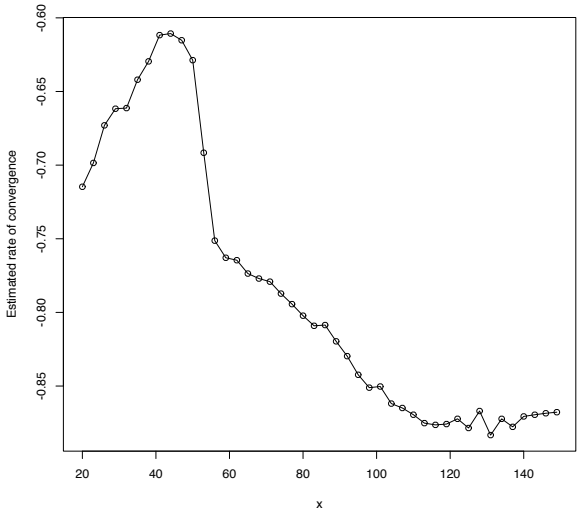

Figure 2.18: *Estimated value for the parameter a across x. Notice that the estimated slope is greater for the values $x \in I_{Reject}$, and consistent with the theoretical rates for $x \in I_{Accept}$. The average estimated rate for $x \in I_{Accept}$ is -0.60, whereas the average estimated rate for $x \in I^c_{Accept}$ is -0.81.*

Let us first focus on $\mathbb{P}\left(\widehat{F_n^1}(x) > F_1(b-\delta/2)\right)$. Since F_1 is the cumulative distribution function of $P_1 \in \mathcal{Z}$, it is increasing and differentiable. In particular, for any $x \in \mathcal{R}_\delta = [a+\delta, b-\delta]$ there exists $\tau_b \in (b-\delta, b-\delta/2) \subset \mathcal{R}$ such that:

$$\begin{aligned} F_1(b-\delta/2) - F_1(x) &\geq F_1(b-\delta/2) - F_1(b-\delta) \\ &= \frac{\delta}{2} F_1'(\tau_b). \end{aligned}$$

Since $F_1' \in \mathcal{F}_{\mathcal{R}}(A, B, C)$, $F_1'(\tau_b) \geq A$. It follows:

$$\begin{aligned} \mathbb{P}\left(\widehat{F_n^1}(x) > F_1(b-\delta/2)\right) &= \mathbb{P}\left(\widehat{F_n^1}(x) - F_1(x) > F_1(b-\delta/2) - F_1(x)\right) \\ &\leq \mathbb{P}\left(\widehat{F_n^1}(x) - F_1(x) > F_1(b-\delta/2) - F_1(b-\delta)\right) \\ &\leq \mathbb{P}\left(\widehat{F_n^1}(x) - F_1(x) > \frac{\delta A}{2}\right). \end{aligned}$$

We recall Hoeffding's Lemma [48].

Lemma 1 (Hoeffding's Lemma)**.** *If X is a random variable on $\mathbb{R}$ such that there exists two real numbers a, b such that $\mathbb{P}(a \leq X \leq b) = 1$ then for all $\lambda > 0$*

$$\log \mathbb{E}\left(e^{\lambda X}\right) \leq \lambda \mathbb{E}(X) + \frac{\lambda^2}{8}(b-a)^2.$$

Hoeffding's Lemma can be stated in terms of conditional expectation.

Lemma 2 (Conditional Hoeffding's Lemma)**.** *If X is a random variable on $\mathbb{R}$ such that there exists two real numbers a, b such that $\mathbb{P}(a \leq X \leq b) = 1$ then for all $\lambda > 0$ and $\sigma-$algebra $\mathcal{G}$*

$$\log \mathbb{E}\left(e^{\lambda X}|\mathcal{G}\right) \leq \lambda \mathbb{E}(X|\mathcal{G}) + \frac{\lambda^2}{8}(b-a)^2.$$

The proof of the conditional Hoeffding's Lemma is identical to the proof of the unconditional Hoeffding's Lemma. Let

$$S_n = \sum_{i=1}^{n} 1_{X_i^1 \leq x}$$

so that $\widehat{F_n^1}(x) = S_n/n$. Moreover notice $F_1(x) = \mathbb{E}(S_n|P_1)/n$. Let

$$Z_i = 1_{X_i^1 \leq x} - \mathbb{E}(1_{X_i^1 \leq x}|P_1)$$

and observe that for any $\lambda > 0$:

$$\begin{aligned}
\mathbb{P}\left(\widehat{F_n^1}(x) - F_1(x) > \frac{\delta A}{2}|P_1\right) &= \mathbb{P}\left(n\widehat{F_n^1}(x) - nF_1(x) > n\frac{\delta A}{2}|P_1\right) \\
&= \mathbb{P}\left(S_n - \mathbb{E}(S_n|P_1) > n\frac{\delta A}{2}|P_1\right) \\
&= \mathbb{P}\left(e^{\lambda(S_n - \mathbb{E}(S_n|P_1))} > e^{n\lambda\frac{\delta A}{2}}|P_1\right) \\
&\leq e^{-n\lambda\frac{\delta A}{2}}\mathbb{E}\left(e^{\lambda(S_n - \mathbb{E}(S_n|P_1))}|P_1\right) \quad \text{using Markov's inequality} \\
&\leq e^{-n\lambda\frac{\delta A}{2}}\prod_{i=1}^{n}\mathbb{E}\left(e^{\lambda Z_i}|P_1\right) \quad \text{by conditional independence of the } Z_i\text{'s} \\
&\leq e^{-n\lambda\frac{\delta A}{2}}\prod_{i=1}^{n}e^{\lambda\mathbb{E}(Z_i|P_1)+\lambda^2/8} \quad \text{by the conditional Hoeffding's Lemma} \\
&= e^{-n\lambda\frac{\delta A}{2}}e^{n\lambda^2/8} \quad \text{since } \mathbb{E}(Z_i|P_1) = 0 \\
&\leq e^{-n\delta^2A^2/2}.
\end{aligned}$$

Hence

$$\mathbb{P}\left(\widehat{F_n^1}(x) > F_1(b - \delta/2)\right) \leq e^{-n\delta^2A^2/2}.$$

The proof that

$$\mathbb{P}\left(\widehat{F_n^1}(x) < F_1(a + \delta/2)\right) \leq e^{-n\delta^2A^2/2}$$

is identical. Hence

$$\mathbb{P}(\xi_{n,M}^c) \leq 2Me^{-n\delta^2A^2/2}. \tag{2.8}$$

This completes the first part of the proof. Recall that we want to upper bound the term on the right side in ($\star$). Now, let us focus on $\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)|> K, \xi_{n,M})$. Let $x \in \mathcal{R}_\delta$, $K > 0$. We have:

$$\begin{aligned}
\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)|> K, \xi_{n,M}) &= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}(\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) - f(x)\Big| > K, \xi_{n,M}\Big) \\
&= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}(\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) - G_i^{-1} \circ F_i(x)\Big| > K, \xi_{n,M}\Big) \\
&= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}(\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) \pm G_i^{-1} \circ \widehat{F_n^i}(x) - G_i^{-1} \circ F_i(x)\Big| > K, \xi_{n,M}\Big) \\
&\leq I + II \qquad (2.9)
\end{aligned}$$

where:

$$\begin{aligned}
I &= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}(\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) - G_i^{-1} \circ \widehat{F_n^i}(x)\Big| > K/2, \xi_{n,M}\Big) \\
II &= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M} G_i^{-1} \circ \widehat{F_n^i}(x) - G_i^{-1} \circ F_i(x)\Big| > K/2, \xi_{n,M}\Big).
\end{aligned}$$

- The term I.

For any $\nu \in (0,1)$ define:

$$A_{n,i}(\nu) := (\widehat{G_n^i})^{-1}(\nu) - G_i^{-1}(\nu) - \frac{\nu - \widehat{G_n^i}(G_i^{-1}(\nu))}{g_i(G_i^{-1}(\nu))},$$

so that we have:

$$\begin{aligned}
I &= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}(\widehat{G_n^i})^{-1} \circ \widehat{F_n^i}(x) - G_i^{-1} \circ \widehat{F_n^i}(x)\Big| > K/2, \xi_{n,M}\Big) \\
&= \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}\frac{\widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x)))}{g_i \circ G_i^{-1} \circ \widehat{F_n^i}(x)} + A_{n,i}(\widehat{F_n^i}(x))\Big| > K/2, \xi_{n,M}\Big) \\
&\leq \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}\frac{\widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x)))}{g_i \circ G_i^{-1} \circ \widehat{F_n^i}(x)}\Big| > K/4, \xi_{n,M}\Big) + \mathbb{P}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M} A_{n,i}(\widehat{F_n^i}(x))\Big| > K/4, \xi_{n,M}\Big) \\
&=: I_1 + I_2 \quad \text{say.} \qquad (2.10)
\end{aligned}$$

First, let us upper bound I_1. For any $u \in \mathcal{R}$, if $X \sim F \in \mathcal{Z}$, $\mathbb{P}(X \leq u) = \mathbb{P}(f(X) \leq f(u))$ because f is increasing. Moreover $Y = f(X)$ is a random variable with distribution $Q = f_\# P$. Denoting by G its c.d.f. we have:

$$F(u) = G \circ f(u) \qquad (2.11)$$

so for any $v \in f(\mathcal{R})$, we have:

$$G(v) = F \circ f^{-1}(v) \qquad (2.12)$$

or equivalently for $w \in \mathcal{R}$:

$$f(w) = G^{-1} \circ F(w) \tag{2.13}$$

By assumption, F and f are differentiable, hence:

$$G'(v) = \frac{F' \circ f^{-1}(v)}{f' \circ f^{-1}(v)}$$

so for $v \in f(\mathcal{R})$, using that $F \in \mathcal{F}_{\mathcal{R}}(A, B, C)$ and Assumption 2 it follows that:

$$0 < A \left\| f' \right\|_{L^\infty(\mathcal{R})}^{-1} \leq G'(v) \leq B \left\| 1/f' \right\|_{L^\infty(\mathcal{R})}, \tag{2.14}$$

meaning that if $\widehat{F_n^i}(x) \in F_i(\mathcal{R}_{\delta/2})$, it holds that $g_i \circ G_i^{-1} \circ \widehat{F_n^i}(x)^{-1} \leq \left\| f' \right\|_{L^\infty(\mathcal{R})} A^{-1}$ by (2.14). Hence:

$$\begin{aligned} I_1 &= \mathbb{P}\Big(\Big| \frac{1}{M} \sum_{i=1}^{M} \frac{\widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x)))}{g_i \circ G_i^{-1} \circ \widehat{F_n^i}(x)} \Big| > K/4, \xi_{n,M} \Big) \\ &\leq \mathbb{P}\Big(\Big| \frac{1}{M} \sum_{i=1}^{M} \widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x))) \Big| > \frac{AK}{4 \left\| f' \right\|_{L^\infty(\mathcal{R})}}, \xi_{n,M} \Big) \end{aligned} \tag{2.15}$$

Now let:

$$Z_{n,i} = \widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x))).$$

Let $\mathcal{A}_i$ be the σ-algebra generated by $\{P_i, X_1^i, \ldots, X_n^i\}$. We plan to condition on $\mathcal{A}_i$ to obtain first and second moments of $Z_{n,i}$ and apply Markov's inequality. Note that conditioning on $\mathcal{A}_i$ freezes all random variables except the sample $(Y_1^i, \ldots, Y_n^i)$.

$$\begin{aligned} \mathbb{E}(Z_{n,i}) &= \mathbb{E}(\mathbb{E}(Z_{n,i} | \mathcal{A}_i)) \\ &= \mathbb{E}\left(\mathbb{E}\left(\widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x))) | \mathcal{A}_i \right) \right). \end{aligned}$$

Since $\widehat{G_n^i}$ is the empirical c.d.f. of $(Y_1^i, \ldots, Y_n^i)$ we have

$$\begin{aligned} \mathbb{E}\left(\widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x))) | \mathcal{A}_i \right) &= \mathbb{E}\left(\mathbf{1}_{\{Y_1^i \leq G_i^{-1}(\widehat{F_n^i}(x))\}} | \mathcal{A}_i \right) \\ &= G_i \circ G_i^{-1} \circ \widehat{F_n^i}(x) \\ &= \widehat{F_n^i}(x) \end{aligned}$$

hence we obtain:

$$\mathbb{E}\left(\widehat{F_n^i}(x) - \widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x))) | \mathcal{A}_i \right) = 0,$$

i.e. :

$$\mathbb{E}(Z_{n,i}) = 0.$$

Moreover, using the well-known form for the variance of the empirical c.d.f. we obtain:

$$\begin{aligned}\mathbb{E}\left(\left[\widehat{F_n^i}(x)-\widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x)))\right]^2|\mathcal{A}_i\right) &= \frac{\widehat{F_n^i}(x)(1-\widehat{F_n^i}(x))}{n}\\ &\leq \frac{1}{4n}\end{aligned}$$

meaning that:

$$\begin{aligned}\mathbb{E}(Z_{n,i}^2) &= \mathbb{E}(\mathbb{E}(Z_{n,i}^2|\mathcal{A}_i))\\ &= \mathbb{E}\left(\mathbb{E}\left(\left[\widehat{F_n^i}(x)-\widehat{G_n^i}(G_i^{-1}(\widehat{F_n^i}(x)))\right]^2|\mathcal{A}_i\right)\right)\\ &\leq \frac{1}{4n}.\end{aligned}$$

It follows by Markov's inequality and 2.15:

$$\begin{aligned}I_1 &\leq \frac{16\left\|f'\right\|_{L^\infty(\mathcal{R})}^2}{A^2K^2}\mathbb{E}\left(\left|\frac{1}{M}\sum_{i=1}^M Z_{n,i}\right|^2\right)\\ &\leq \frac{16\left\|f'\right\|_{L^\infty(\mathcal{R})}^2}{A^2MK^2}\mathbb{E}\left(Z_{n,1}^2\right)\\ &\leq \frac{4\left\|f'\right\|_{L^\infty(\mathcal{R})}^2}{A^2}\frac{1}{MnK^2}.\end{aligned} \tag{2.16}$$

Now, concerning I_2, we want to prove that for any $\epsilon>0$, there exists $\tilde{K}>0$ and $N\in\mathbb{N}$ such that, for all $n\geq N$:

$$\mathbb{P}\left(\frac{1}{M}\sum_{i=1}^M|A_{n,i}(F_{n,i}(x))|>\frac{\log(n)^{3/2}}{n^{3/4}}\tilde{K},\xi_{n,M}\right)\leq\epsilon.$$

To do this, first observe that using Markov's inequality and the triangle inequality we obtain:

$$\begin{aligned}I_2 &= \mathbb{P}\left(\left|\frac{1}{M}\sum_{i=1}^M A_{n,i}(\widehat{F_n^i}(x))\right|>\frac{K}{4},\xi_{n,M}\right)\\ &\leq \frac{4}{K}\frac{1}{M}\sum_{i=1}^M\mathbb{E}\left(|A_{n,i}(\widehat{F_n^i}(x))|\mathbf{1}_{\xi_{n,M}}\right).\end{aligned}$$

Now, recall that $\xi_{n,M}=\bigcap_{i=1}^M\{\widehat{F_n^i}(x)\in F_i(\mathcal{R}_{\delta/2})\}$, hence $\mathbf{1}_{\xi_{n,M}}\leq\mathbf{1}_{\{\widehat{F_n^j}(x)\in F_j(\mathcal{R}_{\delta/2})\}}$ $\mathbb{P}-$a.s. for all $1\leq j\leq M$. Furthermore, $\left(|A_{n,i}(\widehat{F_n^i}(x))|\mathbf{1}_{\{\widehat{F_n^i}(x)\in F_i(\mathcal{R}_{\delta/2})\}}\right)_{1\leq i\leq M}$ are independent and identically distributed, meaning that:

$$I_2\leq\frac{4}{K}\mathbb{E}\left(|A_{n,1}(\widehat{F_n^1}(x))|\mathbf{1}_{\{\widehat{F_n^1}(x)\in F_1(\mathcal{R}_{\delta/2})\}}\right). \tag{2.17}$$

Proving that this bound is controlled by $n^{-3/4}\log(n)^{3/2}$ uniformly in $x \in \mathcal{R}_\delta$ will complete the proof. Instead of directly bounding I_2, we will instead focus on bounding the expectation conditional on the σ-algebra generated by $(\widehat{F_n^1}(x), P)$. Observe that:

$$\begin{aligned}\mathbb{E}\left(|A_{n,1}(\widehat{F_n^1}(x))|\mathbf{1}_{\{\widehat{F_n^1}(x)\in F_1(\mathcal{R}_{\delta/2})\}}\right) &= \mathbb{E}\left[\mathbb{E}\left(|A_{n,1}(\widehat{F_n^1}(x))|\mathbf{1}_{\{\widehat{F_n^1}(x)\in F_1(\mathcal{R}_{\delta/2})\}}|\widehat{F_n^1}(x), P_1\right)\right] \\ &= \mathbb{E}\left(h_n(\widehat{F_n^1}(x), P_1)\right) \quad \text{say,}\end{aligned}$$

where $h_n : [0,1] \times \mathcal{Z} \longrightarrow (0,\infty)$ is defined by

$$h_n(\nu, \tilde{P}) = \mathbb{E}\left(|A_{n,1}(\nu)|\mathbf{1}_{\{\nu\in F_1(\mathcal{R}_{\delta/2})\}}|\widehat{F_n^1}(x) = \nu, P_1 = \tilde{P}\right).$$

In the definition of h_n we condition on the event $P_1 = \tilde{P}$, hence the cumulative functions F_1 and G_1 are fixed. In other words, in this part of the proof, F_1 and G_1 are to be understood as the cumulative functions of the measures $\tilde{P}$ and $f_\#\tilde{P}$ respectively, where $\tilde{P}$ is a fixed element of $\mathcal{Z}$.

By assumption, $(X_1^1, \dots, X_n^1)$ and $(Y_1^1, \dots, Y_n^1)$ are independent conditional on P_1 and the variables $(X_1^1, \dots, X_n^1)$ are embedded only through the measurable functional $T(X_1, \dots, X_n) = \widehat{F_n}(x)$. Hence:

$$h_n(\nu, \tilde{P}) = \mathbb{E}\left(|A_{n,1}(\nu)|\mathbf{1}_{\{\nu\in F_1(\mathcal{R}_{\delta/2})\}}|P_1 = \tilde{P}\right).$$

If one proves that $n^{3/4}\log(n)^{-3/2}h_n(\nu, \tilde{P})$ is uniformly bounded in $n \in \mathbb{N}$, $\nu \in [0,1]$ and $\tilde{P} \in \mathcal{Z}$, the result will follow. For any choice of n, ν and $\tilde{P}$, observe that $h_n(\nu, \tilde{P}) = 0$ if $\nu \notin F_1(\mathcal{R}_{\delta/2})$, and $h_n(\nu, \tilde{P}) = \mathbb{E}\left(|A_{n,1}(\nu)|\,|P_1 = \tilde{P}\right)$ if $\nu \in F_1(\mathcal{R}_{\delta/2})$. Let $n \in \mathbb{N}$, $\nu \in [0,1]$ and $\tilde{P} \in \mathcal{Z}$ such that $\nu \in F_1(\mathcal{R}_{\delta/2})$. Let:

$$J_n = \left(G_1^{-1}(\nu) - \frac{\log(n)}{\sqrt{n}}, G_1^{-1}(\nu) + \frac{\log(n)}{\sqrt{n}}\right), \tag{2.18}$$

and recall that:

$$A_{n,1}(\nu) = (\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu) - \frac{\nu - \widehat{G_n^1}(G_1^{-1}(\nu))}{g_1(G_1^{-1}(\nu))}. \tag{2.19}$$

We will decompose the expectation on the events $\{(\widehat{G_n^1})^{-1}(\nu) \in J_n\}$ and $\{(\widehat{G_n^1})^{-1}(\nu) \notin J_n\}$. We will upper bound the former using the results provided by Bahadur in [6], and the latter will be easily controlled by $\mathbb{P}((\widehat{G_n^1})^{-1}(\nu) \notin J_n)$. We shall recall the main result of Bahadur in [6].

Theorem 6 (Bahadur)**.** *Let $(Z_n)_{n\geq 1}$ be a sequence of independent real valued random variables with common distribution F. Let $x \in \mathbb{R}$ and let*

$$F(x) = p$$

and suppose that F has two derivatives in some neighborhood of x, that F'' is bounded in that neighborhood and that $F'(x) =: f(x) > 0$. These assumptions imply that F is invertible at x and

that $x = F^{-1}(p)$ *with* $0 < p < 1$ *is the unique p-quantile of* F*. Let*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i \leq x}$$

the empirical cumulative distribution function and for $\alpha \in (0,1)$ *let*

$$(F_n)^{-1}(\alpha) = \inf\{u \in \mathbb{R} | F_n(u) \geq \alpha\}$$

be the generalized inverse of F_n*. Then*

$$(F_n)^{-1}(p) = F^{-1}(p) + \frac{p - F_n(F^{-1}(p))}{f \circ F^{-1}(p)} + R_n$$

where

$$R_n = \mathcal{O}_{a.s.}(n^{3/4}\log(n)).$$

We know that $\nu \in F_1(\mathcal{R}_{\delta/2})$ implies $G_1^{-1}(\nu) \in G_1^{-1} \circ F_1(\mathcal{R}_{\delta/2}) = f(\mathcal{R}_{\delta/2}) \subset f(\mathcal{R})$. Moreover there exists $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$, we have $n^{-1/2}\log(n) < \delta/2$. For such a n, observe that $J_n \subset f(\mathcal{R})$. Choose such a n. There are 2 cases:

Case 1: $(\widehat{G_n^1})^{-1}(\nu) \in J_n$
We follow up closely on Bahadur's ideas developed in the proof of his theorem on the almost sure representation of empirical quantiles [6] to derive the upper bound. By a Taylor expansion we obtain:

$$\begin{aligned} G_1((\widehat{G_n^1})^{-1}(\nu)) = G_1(G_1^{-1}(\nu)) &+ \left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right) G_1'(G_1^{-1}(\nu)) \\ &+ \frac{G_1''(\theta_n)}{2}\left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right)^2 \end{aligned}$$

where $\theta_n \in \left(G_1^{-1}(\nu), (\widehat{G_n^1})^{-1}(\nu)\right)$. Since $(\widehat{G_n^1})^{-1}(\nu) \in J_n$, we have $|G_1^{-1}(\nu) - \theta_n| \leq |G_1^{-1}(\nu) - (\widehat{G_n^1})^{-1}(\nu)| \leq n^{-1/2}\log(n)$ using (2.18), meaning that $\theta_n \in f(\mathcal{R})$ because $n \geq N_1$. Moreover, using (2.12) and the fact that F_1 and f are twice differentiable we have:

$$G_1''(u) = \frac{F_1'' \circ f^{-1}(u)}{f' \circ f^{-1}(u)^2} - \frac{f'' \circ f^{-1}(u)}{f' \circ f^{-1}(u)^3} \cdot F_1' \circ f^{-1}(u),$$

meaning that for $u \in f(\mathcal{R})$ it holds:

$$|G_1''(u)| \leq C \left\|1/f'\right\|^2_{L^\infty(\mathcal{R})} + B \left\|f''\right\|_{L^\infty(\mathcal{R})} \left\|1/f'\right\|^3_{L^\infty(\mathcal{R})} =: \Theta_1. \tag{2.20}$$

Using (2.20), we know that $|G_1''(u)| \leq \Theta_1$ for all $u \in f(\mathcal{R})$, so in particular there exists $\phi_{n,1}$ such that $|\phi_{n,1}| \leq 1$, $G_1''(\theta_n) = \phi_{n,1}\Theta_1$. Moreover $(\widehat{G_n^1})^{-1}(\nu) \in J_n$ implies that there exists $\phi_{n,2}$ such that $|\phi_{n,2}| \leq 1$, $((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu))^2 = \phi_{n,2} n^{-1}\log(n)^2$. So we have:

$$G_1((\widehat{G_n^1})^{-1}(\nu)) = G_1(G_1^{-1}(\nu)) + \left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right) G_1((G_1^{-1}(\nu)) + \frac{\Theta_1}{2}\phi_{n,1}\phi_{n,2}\frac{\log(n)^2}{n}.$$

Let $\phi_n = \phi_{n,1}\phi_{n,2}$. Then:

$$G_1((\widehat{G_n^1})^{-1}(\nu)) = G_1(G_1^{-1}(\nu)) + \left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right) G_1'(G_1^{-1}(\nu)) + \frac{\Theta_1}{2}\phi_n \frac{\log(n)^2}{n} \tag{2.21}$$

and $|\phi_n| \leq 1$. Let:

$$D_n(u) = \left[\widehat{G_n^1}(u) - \widehat{G_n^1}(G^{-1}(\nu))\right] - \left[G_1(u) - G_1(G_1^{-1}(\nu))\right], \tag{2.22}$$

$$H_n = \sup\{|D_n(u)|, u \in J_n\}. \tag{2.23}$$

Since $(\widehat{G_n^1})^{-1}(\nu) \in J_n$ we have:

$$D_n((\widehat{G_n^1})^{-1}(\nu)) = \psi_n H_n \tag{2.24}$$

where $|\psi_n| \leq 1$. Notice that $\widehat{G_n^1}((\widehat{G_n^1})^{-1}(\nu)) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Y_i^1 \leq Y_{(k_n)}^1\}} = n^{-1}k_n$ where $k_n = \lceil n\nu \rceil$. It follows that:

$$\begin{aligned} D_n((\widehat{G_n^1})^{-1}(\nu)) &= \psi_n H_n \\ &= \frac{k_n}{n} - \widehat{G_n^1}(G_1^{-1}(\nu)) - \left(G_1((\widehat{G_n^1})^{-1}(\nu)) - G_1(G_1^{-1}(\nu))\right). \end{aligned} \tag{2.25}$$

Combining (2.21), (2.24) and (2.25):

$$\begin{aligned} \frac{k_n}{n} &= \widehat{G_n^1}(G_1^{-1}(\nu)) + G_1((\widehat{G_n^1})^{-1}(\nu)) - G_1(G_1^{-1}(\nu)) + \psi_n H_n \\ &= \widehat{G_n^1}(G_1^{-1}(\nu)) + \left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right) G_1'(G_1^{-1}(\nu)) + \frac{\Theta_1}{2}\phi_n \frac{\log(n)^2}{n} + \psi_n H_n. \end{aligned}$$

Rearranging the terms:

$$(\widehat{G_n^1})^{-1}(\nu) = G_1^{-1}(\nu) + \frac{k_n/n - \widehat{G_n^1}(G_1^{-1}(\nu))}{G_1'(G_1^{-1}(\nu))} + \frac{2^{-1}\Theta_1\phi_n n^{-1}\log(n)^2 + \psi_n H_n}{G_1'(G_1^{-1}(\nu))}. \tag{2.26}$$

Since $k_n/n = \lceil n\nu \rceil/n$, we have $\nu \leq k_n/n \leq \nu + 1/n$, so there exists κ_n, $|\kappa_n| \leq 1$ such that $k_n/n = \nu + \kappa_n/n$. Equation (2.26) becomes:

$$(\widehat{G_n^1})^{-1}(\nu) = G_1^{-1}(\nu) + \frac{\nu - \widehat{G_n^1}(G_1^{-1}(\nu))}{g_1(G_1^{-1}(\nu))} + \frac{2^{-1}\Theta_1\phi_n n^{-1}\log(n)^2 + \psi_n H_n + n^{-1}\kappa_n}{g_1(G_1^{-1}(\nu))}.$$

We obtain the following explicit form for $A_{n,1}(\nu)$ defined in (2.19) when $(\widehat{G_n^1})^{-1}(\nu) \in J_n$ and $n \geq N_1$:

$$A_{n,1}(\nu) = \frac{2^{-1}\Theta_1\phi_n n^{-1}\log(n)^2 + \psi_n H_n + n^{-1}\kappa_n}{g_1(G_1^{-1}(\nu))}. \tag{2.27}$$

We want to uniformly upper bound $n^{3/4}\log(n)^{-3/2}\mathbb{E}\left(|A_n(\nu)|\,|P = \tilde{P}\right)$ in n, ν and $\tilde{P}$. Since $|\psi_n| \leq 1$, $|\phi_n| \leq 1$, $|\kappa_n| \leq 1$ and $1/g_1(G_1^{-1}(\nu)) \leq A^{-1}\|f'\|_{L^\infty(\mathcal{R})}$ conditional on $P_1 = \tilde{P}$, all we

need to prove that $n^{3/4}\log(n)^{-3/2}\mathbb{E}\left(H_n|P_1=\tilde{P}\right)$ is uniformly bounded, where H_n is defined in (2.23).
In the following, let

$$\begin{aligned}
E_n &= \{-n^{1/4}, -n^{1/4}+1, \dots, n^{1/4}-1\},\\
\eta_{r,n} &= G^{-1}(\nu) + \frac{\log(n)}{n^{3/4}} r \quad \text{where } r \in E_n,\\
z_{r,n} &= |G_1(\eta_{r,n}) - G_1(G_1^{-1}(\nu))|,\\
J_{r,n} &= [\eta_{r,n}, \eta_{r+1,n}],\\
\alpha_{r,n} &= G_1(\eta_{r+1,n}) - G_1(\eta_{r,n}).
\end{aligned}$$

Since $\widehat{G_n^1}$ and G_1 are non decreasing, for all $u \in J_{r,n}$:

$$\begin{aligned}
D_n(u) &\le \widehat{G_n^1}(\eta_{r+1,n}) - \widehat{G_n^1}(G_1^{-1}(\nu)) - G_1(\eta_{r,n}) + G_1(G_1^{-1}(\nu))\\
&= D_n(\eta_{r+1,n}) + \alpha_{r,n}.
\end{aligned}$$

Similarly we have for all $u \in J_{r,n}$: $D_n(u) \ge D_n(\eta_{r,n}) - \alpha_{r,n}$. Meaning that:

$$\begin{aligned}
H_n &\le \max_{r\in E_n}\{|D_n(\eta_{r,n})|\} + \max_{r\in E_n}\{\alpha_{r,n}\}\\
&=: K_n + \beta_n \quad \text{say.}
\end{aligned} \tag{2.28}$$

Let us first deal with $\beta_n = \max_{r\in E_n}\{\alpha_{r,n}\}$. Let $r \in E_n$. Recall that $\alpha_{r,n} = G_1(\eta_{r+1,n}) - G_1(\eta_{r,n})$. The mean value theorem ensures that there exists $\gamma_{r,n} \in (\eta_{r,n}, \eta_{r+1,n})$ such that $\alpha_{r,n} = n^{-3/4}\log(n)G_1'(\gamma_{r,n})$. Observe that $\gamma_{r,n} \in f(\mathcal{R})$, hence using (2.14), $G_1'(\gamma_{r,n}) \le B\,\|1/f'\|_{L^\infty(\mathcal{R})}$ for all $r \in E_n$, meaning that $\beta_n \le B\,\|1/f'\|_{L^\infty(\mathcal{R})}\, n^{-3/4}\log(n)$, so that:

$$\frac{n^{3/4}}{\log(n)}\beta_n \le B\,\|1/f'\|_{L^\infty(\mathcal{R})}\,. \tag{2.29}$$

Now we will focus on the term $n^{3/4}\log(n)^{-3/2}\mathbb{E}\left(K_n|P_1=\tilde{P}\right)$. Recall that $K_n = \max_{r\in E_n}\{|D_n(\eta_{r,n})|\}$. First, since $z_{r,n} = |G_1(\eta_{r,n}) - G_1(G_1^{-1}(\nu))|$, the mean value theorem ensures that there exists $\delta_{r,n} \in (\eta_{r,n}, G_1^{-1}(\nu)) \subset f(\mathcal{R})$ such that:

$$\begin{aligned}
z_{r,n} &= g_1(\delta_{r,n})\frac{\log(n)}{n^{3/4}}|r|\\
&\le B\,\|1/f'\|_{L^\infty(\mathcal{R})}\,\frac{\log(n)}{n^{3/4}}|r|\\
&\le B\,\|1/f'\|_{L^\infty(\mathcal{R})}\,\frac{\log(n)}{n^{1/2}}\\
&=: z_n \quad \text{say.}
\end{aligned}$$

Let:

$$\Gamma_n = \frac{n^{3/4}}{\log(n)^{3/2}}\mathbb{E}\left(K_n|P_1=\tilde{P}\right) \tag{2.30}$$

$$= \mathbb{E}\left(\frac{n^{3/4}}{\log(n)^{3/2}} \max_{r\in E_n} |D_n(\eta_{r,n})| \,|P_1 = \tilde{P}\right)$$

where D_n is defined by (2.22). Now it is easily seen by definition of $z_{r,n}$ that $|D_n(\eta_{r,n})| \sim \frac{1}{n}|B(n, z_{r,n}) - nz_{r,n}|$ conditional on P_1 where $B(n,p)$ denotes the Binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$. Letting $a_n = n^{-1/4}\log(n)^{-1/2}$, we have:

$$\begin{aligned}
\Gamma_n &= \mathbb{E}\left(\frac{n^{3/4}}{\log(n)^{3/2}} \max_{r\in E_n} \frac{1}{n} |B(n, z_{r,n}) - nz_{r,n}| \,|P_1 = \tilde{P}\right) \\
&= \frac{1}{\log(n)}\mathbb{E}\left(\max_{r\in E_n} \{a_n |B(n, z_{r,n}) - nz_{r,n}|\} \,|P_1 = \tilde{P}\right) \\
&= \frac{1}{\log(n)}\mathbb{E}(T_n|P_1 = \tilde{P}) \quad \text{say.} \qquad (2.31)
\end{aligned}$$

We now want to upper bound $\mathbb{E}(T_n|P = \tilde{P})$ and show that it can not exceed $\log(n)$. A natural way of proceeding would be to upper bound the expectation of the maximum by the sum of the expectations, and use classic concentration inequalities on the deviation probabilities to control each expectation. The error injected by the concentration inequality adds up for each term of the sum and makes the upper bound diverges at rate precisely equal to the cardinal of the sum $|E_n|$, which is of order $n^{1/4}$. Using Jensen's inequality [40] beforehand directly on T_n allows for exploiting the structure of the moment generating function of the Binomial distribution and bypass this issue. The following two technical results are required. See the Appendix for a proof.

Lemma 3. *Let $t \geq 0$, $0 \leq z \leq 1/2$ and $n \in \mathbb{N}$. If $X \sim B(n,z)$, then:*

$$\mathbb{E}\left(e^{t(X-nz)}\right) \geq \mathbb{E}\left(e^{-t(X-nz)}\right).$$

Lemma 4. *Let $\lambda = 2^{-1/4}\log(2)^{-1/2}$, $n \in \mathbb{N}$. If $X \sim B(n,z)$, then the mapping $\Phi_n : [0,1] \longrightarrow \mathbb{R}$ defined by*

$$\Phi_n(z) = \mathbb{E}\left(\exp\{a_n\left(B(n,z) - nz\right)\}\right)$$

Is increasing on $[0, (1-e^{\lambda})^{-1} + \lambda^{-1}] \approx [0, 0.42]$ for all $n \geq 2$.

Note that there exists a fixed $N_2 \in \mathbb{N}$ such that $z_{r,n} \leq z_n \leq (1-e^{\lambda})^{-1} + \lambda^{-1} \approx 0.42$ for all $n \geq N_2$. Choose $n \geq N_1 \vee N_2$. Now, using Jensen's inequality we have:

$$\begin{aligned}
e^{\mathbb{E}(T_n|P=\tilde{P})} &\leq \mathbb{E}\left(e^{T_n}|P_1 = \tilde{P}\right) \\
&= \mathbb{E}\left(\max_{r\in E_n} \exp\{a_n |B(n, z_{r,n}) - nz_{r,n}|\}|P_1 = \tilde{P}\right) \\
&\leq \sum_{r\in E_n} \mathbb{E}\left(\exp\{a_n |B(n, z_{r,n}) - nz_{r,n}|\}|P_1 = \tilde{P}\right) \\
&\leq 2\sum_{r\in E_n} \mathbb{E}\left(\exp\{a_n(B(n, z_{r,n}) - nz_{r,n})\}|P_1 = \tilde{P}\right) \\
&\leq 4n^{1/4}\mathbb{E}\left(\exp\left\{a_n(B(n, z_n) - nz_n)\right\}\right), \qquad (2.32)
\end{aligned}$$

Where the last two inequalities are a direct consequence of Lemma 3 combined with the fact that $e^{|x|} \leq e^x + e^{-x}$ and Lemma 4. The conditioning on P_1 was dropped since z_n is independent of P_1. Now we can compute the limit in (2.32). Using that the moment generating function of the Binomial distribution with parameters n and p is $t \mapsto (1 + p(e^t - 1))^n$, and a Taylor expansions we obtain:

$$\begin{aligned}
\mathbb{E}\left(e^{a_n(B(n,z_n)-nz_n)}\right) &= e^{-na_nz_n}\mathbb{E}\left(e^{a_nB(n,z_n)}\right) \\
&= e^{-na_nz_n}(1 + z_n(e^{a_n} - 1))^n \\
&= e^{\frac{nz_na_n^2}{2}+o(nz_na_n^2)} \\
&\underset{n\to\infty}{\longrightarrow} e^{\frac{B}{2}\|1/f'\|_{L^\infty(\mathcal{R})}} < \infty
\end{aligned}$$

hence $\mathbb{E}\left(e^{a_n(B(n,z_n)-nz_n)}\right)$ is bounded by a constant Θ_2 say, for all $n \in \mathbb{N}$. From (2.32)

$$e^{\mathbb{E}(T_n|P_1=\tilde{P})} \leq 4\Theta_2 n^{1/4},$$

hence:

$$\mathbb{E}(T_n|P_1 = \tilde{P}) \leq \log(n^{1/4}) + \log(4\Theta_2) \tag{2.33}$$

Combining (2.31) and (2.33):

$$\begin{aligned}
\Gamma_n &= \frac{1}{\log(n)}\mathbb{E}(T_n|P_1 = \tilde{P}) \\
&\leq \frac{\log(n^{1/4})}{\log(n)} + \frac{\log(4\Theta_2)}{\log(n)} \\
&\leq \frac{1}{4} + 2\log(4\Theta_2),
\end{aligned} \tag{2.34}$$

since $\log(n)^{-1} \leq 2$. Using (2.30) and (2.34), this shows that:

$$\frac{n^{3/4}}{\log(n)^{3/2}}\mathbb{E}(K_n|P_1 = \tilde{P}) \leq \frac{1}{4} + 2\log(4\Theta_2). \tag{2.35}$$

Combining (2.28), (2.29) and (2.35), we have:

$$\frac{n^{3/4}}{\log(n)^{3/2}}\mathbb{E}(H_n|P_1 = \tilde{P}) \leq \frac{1}{4} + B\|1/f'\|_{L^\infty(\mathcal{R})} + 2\log(4\Theta_2). \tag{2.36}$$

Using the form of $A_{n,1}(\nu)$ in (2.27) and the result in (2.36) we finally have:

$$\frac{n^{3/4}}{\log(n)^{3/2}}\mathbb{E}(|A_{n,1}(\nu)|\mathbf{1}_{\{(\widehat{G_n^1})^{-1}(\nu)\in J_n\}}|P_1 = \tilde{P}) \leq \frac{\|f'\|_{L^\infty(\mathcal{R})}}{A}\left(\frac{9}{4} + \frac{\Theta_1}{2} + B\|1/f'\|_{L^\infty(\mathcal{R})} + 2\log(4\Theta_2)\right) \tag{2.37}$$

for all $n \geq N_1 \vee N_2$.

Case 2: $(\widehat{G_n^1})^{-1}(\nu) \notin J_n$

Using the Cauchy-Schwarz inequality we have:

$$\mathbb{E}\left(|A_{n,1}(\nu)|\mathbf{1}_{\{(\widehat{G_n^1})^{-1}(\nu)\notin J_n\}}|P_1 = \tilde{P}\right) \leq \sqrt{\mathbb{E}\left(A_{n,1}(\nu)^2|P_1 = \tilde{P}\right)\mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) \notin J_n|P_1 = \tilde{P}\right)} \tag{2.38}$$

We derive upper bounds for both terms in the product on the RHS of (2.38). Using (2.19) we have:

$$A_{n,1}(\nu)^2 \leq 2\left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right)^2 + 2\left(\frac{\nu - \widehat{G_n^1}(G_1^{-1}(\nu))}{g_1(G_1^{-1}(\nu))}\right)^2 \tag{2.39}$$

But $\nu, \widehat{G_n^1}(G_1^{-1}(\nu)) \in [0,1]$ and we know that since $\nu \in F_1(\mathcal{R}_{\delta/2})$, $1/g_1(G_1^{-1}(\nu)) \leq A^{-1}\left\|f'\right\|_{L^\infty(\mathcal{R})}$ conditional on P_1. So:

$$\mathbb{E}\left[\left(\frac{\nu - \widehat{G_n^1}(G_1^{-1}(\nu))}{g_1(G_1^{-1}(\nu))}\right)^2 \middle| P_1 = \tilde{P}\right] \leq A^{-2}\left\|f'\right\|^2_{L^\infty(\mathcal{R})} \tag{2.40}$$

and we are done with this term. Then:

$$\begin{aligned}
\mathbb{E}\left(\left((\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)\right)^2 | P_1 = \tilde{P}\right) &= \int_0^\infty t\mathbb{P}\left(|(\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)| > t | P_1 = \tilde{P}\right) dt \\
&= \int_0^\infty t\mathbb{P}\left(|Y^1_{(k_n)} - G_1^{-1}(\nu)| > t | P_1 = \tilde{P}\right) dt \\
&\leq n\int_0^\infty t\mathbb{P}\left(|Y_1^1 - G_1^{-1}(\nu)| > t | P_1 = \tilde{P}\right) dt \\
&= n\mathbb{E}\left((Y_1^1 - G_1^{-1}(\nu))^2 | P_1 = \tilde{P}\right) \\
&\leq 2n\left(\mathbb{E}((Y_1^1)^2 | P_1 = \tilde{P}) + \mathbb{E}(G_1^{-1}(\nu)^2 | P_1 = \tilde{P})\right) \\
&\leq 2n\left(\mathbb{E}((Y_1^1)^2 | P_1 = \tilde{P}) + f(b)^2\right)
\end{aligned} \tag{2.41}$$

because $\nu \in F(\mathcal{R}_{\delta/2})$ implies $G_1^{-1}(\nu) \in f(\mathcal{R}) = (f(a), f(b))$, meaning that $G_1^{-1}(\nu) < f(b)$ $\mathbb{M}$-a.s. Also, note that $\mathbb{E}((Y_1^1)^2 | P_1 = \tilde{P}) < \infty$ since $\mathbb{E}((Y_1^1)^2) < \infty$ by hypothesis. So, combining (2.39), (2.40) and (2.41):

$$\mathbb{E}\left(A_{n,1}(\nu)^2 | P_1 = \tilde{P}\right) \leq 2n\left(\mathbb{E}((Y_1^1)^2 | P_1 = \tilde{P}) + f(b)^2\right) + 2A^{-2}\left\|f'\right\|^2_{L^\infty(\mathcal{R})} \tag{2.42}$$

Now let us focus on the term $\mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) \notin J_n | P_1 = \tilde{P}\right)$. Using the definition of J_n in (2.18):

$$\begin{aligned}
\mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) \notin J_n | P_1 = \tilde{P}\right) &= \mathbb{P}\left(|(\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)| > \frac{\log(n)}{\sqrt{n}} | P_1 = \tilde{P}\right) \\
&=: U_1 + U_2,
\end{aligned}$$

where:

$$\begin{aligned}
U_1 &= \mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) > \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu) | P_1 = \tilde{P}\right), \\
U_2 &= \mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) < -\frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu) | P_1 = \tilde{P}\right).
\end{aligned}$$

We will derive an upper bound for U_1 only, the case for U_2 being identical. First, recall that $(\widehat{G_n^1})^{-1}(\nu) = Y^1_{(k_n)}$ where $k_n = \lceil n\nu \rceil$. Since:

$$\left\{ Y^1_{(k_n)} \geq \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu) \right\} = \left\{ \sum_{i=1}^n \mathbf{1}_{Y_i^1 \geq \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu)} \geq n - k_n + 1 \right\},$$

we have:

$$\mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) > \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu) | P_1 = \tilde{P} \right) = \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}_{Y_i^1 \geq \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu)} \geq n - k_n + 1 | P_1 = \tilde{P} \right).$$

Noting $V_i = \mathbf{1}_{\{Y_i^1 \geq \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu)\}}$, $S_n = \sum_{i=1}^n V_i$ and $t_n = n - k_n + 1 - \mathbb{E}S_n$, it follows:

$$\mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) > \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu) | P_1 = \tilde{P} \right) = \mathbb{P}\left(S_n - \mathbb{E}S_n \geq t_n | P_1 = \tilde{P} \right). \tag{2.43}$$

Since $S_n | P_1 \sim \mathrm{B}(n, p_n)$ where $p_n = 1 - G_1(G_1^{-1}(\nu) + \log(n)/\sqrt{n})$:

$$\begin{aligned} t_n &= n - k_n + 1 - \mathbb{E}S_n \\ &= n - k_n + 1 - n\left[1 - G_1\left(G_1^{-1}(\nu) + \frac{\log(n)}{\sqrt{n}} \right) \right] \\ &= 1 + n\left[G_1\left(G_1^{-1}(\nu) + \frac{\log(n)}{\sqrt{n}} \right) - \frac{\lceil n\nu \rceil}{n} \right] \end{aligned}$$

Notice that there exists ϕ_n, $0 \leq \phi_n \leq 1/n$, such that $\lceil n\nu \rceil / n = \nu + \phi_n$, hence:

$$\begin{aligned} t_n &= 1 + n\left[G_1\left(G_1^{-1}(\nu) + \frac{\log(n)}{\sqrt{n}} \right) - \frac{\lceil n\nu \rceil}{n} \right] \\ &= 1 + n\left[G_1\left(G_1^{-1}(\nu) + \frac{\log(n)}{\sqrt{n}} \right) - \nu - \phi_n \right] \\ &= 1 - n\phi_n + n\left[G_1\left(G_1^{-1}(\nu) + \frac{\log(n)}{\sqrt{n}} \right) - \nu \right]. \end{aligned}$$

Because G_1 is strictly increasing and $n\phi_n \in [0, 1]$, we have $t_n > 0$. Moreover, by a Taylor expansion we have:

$$\begin{aligned} t_n &= 1 - n\phi_n + n\left(g_1(G_1^{-1}(\nu)) \frac{\log(n)}{\sqrt{n}} + g_1'(\theta_n) \frac{\log(n)^2}{2n} \right) \\ &= \sqrt{n}\log(n) g_1(G_1^{-1}(\nu)) + \left(1 - n\phi_n + g_1'(\theta_n)\log(n)^2/2 \right) \end{aligned} \tag{2.44}$$

where $\theta_n \in \left(G_1^{-1}(\nu), G_1^{-1}(\nu) + \log(n)/\sqrt{n} \right)$. In particular we have $|\theta_n - G_1^{-1}(\nu)| < \log(n)/\sqrt{n}$, meaning that for all $n \geq N_1$, $\theta_n \in f(\mathcal{R})$ because $\nu \in F_1(\mathcal{R}_{\delta/2})$. Applying Hoeffding's inequality on (2.43):

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t_n | P_1 = \tilde{P}) \leq e^{-2t_n^2/n} \tag{2.45}$$

Recall (2.44):

$$\begin{aligned} t_n &= \sqrt{n}\log(n) g_1(G_1^{-1}(\nu)) + (1 - n\phi_n + g'(\theta_n)\log(n)^2/2) \\ &= \sqrt{n}\log(n) g_1(G_1^{-1}(\nu)) + \alpha_n \quad \text{say.} \end{aligned} \tag{2.46}$$

Combining (2.44) and (2.46):

$$\frac{t_n^2}{n} = \log(n)^2 g_1(G_1^{-1}(\nu))^2 + \frac{\alpha_n^2}{n} + 2\alpha_n . \frac{\log(n) g_1(G_1^{-1}(\nu))}{\sqrt{n}}$$

Using (2.20) and (2.20), we know that $|g_1'(u)| \leq \Theta_1$, $g_1(u) \leq B \left\| 1/f' \right\|_{L^\infty(\mathcal{R})}$ and $g_1(u) \geq A \left\| f' \right\|_{L^\infty(\mathcal{R})}^{-1}$ for all $u \in f(\mathcal{R})$. Moreover since $\alpha_n = 1 - n\phi_n + g_1'(\theta_n)\log(n)^2/2$, we derive the following upper bound:

$$e^{-2t_n^2/n} \leq e^{\Theta_1 B \left\| 1/f' \right\|_{L^\infty(\mathcal{R})} \log(n)^3/\sqrt{n}} e^{-2\log(n)^2 A^2 \left\| f' \right\|_{L^\infty(\mathcal{R})}^{-2}}. \tag{2.47}$$

Using that $\log(n)^3/\sqrt{n} \leq 11$ along with (2.43), (2.45) and (2.47) we have:

$$\mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) > \frac{\log(n)}{\sqrt{n}} + G_1^{-1}(\nu) | P_1 = \tilde{P} \right) \leq \Theta_3 e^{-\Theta_4 \log(n)^2}$$

where $\Theta_3 = e^{22\Theta_1 B \left\| 1/f' \right\|_{L^\infty(\mathcal{R})}}$, $\Theta_4 = A^2 \left\| f' \right\|_{L^\infty(\mathcal{R})}^{-2}$ are constants depending only on the parameters of the model. The same bound can be derived for $\mathbb{P}((\widehat{G_n^1})^{-1}(\nu) < -\log(n)/\sqrt{n} + G_1^{-1}(\nu) | P_1 = \tilde{P})$, finally showing that:

$$\begin{aligned} \mathbb{P}\left(|(\widehat{G_n^1})^{-1}(\nu) - G_1^{-1}(\nu)| > \frac{\log(n)}{\sqrt{n}} | P_1 = \tilde{P} \right) &= \mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) \notin J_n | P_1 = \tilde{P} \right) \\ &\leq 2\Theta_3 e^{-\Theta_4 \log(n)^2} \end{aligned} \tag{2.48}$$

for all $n \geq N_1$. Recall (2.38):

$$\mathbb{E}\left(|A_{n,1}(\nu)| \mathbf{1}_{\{(\widehat{G_n^1})^{-1}(\nu) \notin J_n\}} | P_1 = \tilde{P} \right) \leq \sqrt{\mathbb{E}\left(A_{n,1}(\nu)^2 | P_1 = \tilde{P} \right) \mathbb{P}\left((\widehat{G_n^1})^{-1}(\nu) \notin J_n | P_1 = \tilde{P} \right)},$$

hence combining (2.41) and (2.48) we have, for all $n \geq N_1$:

$$\begin{aligned} &\mathbb{E}(|A_{n,1}(\nu)| \mathbf{1}_{\{(\widehat{G_n^1})^{-1}(\nu) \notin J_n\}} | P_1 = \tilde{P}) \\ &\qquad \leq \sqrt{\left(2n[\mathbb{E}((Y_1^1)^2 | P_1 = \tilde{P}) + f(b)^2] + A^{-2} \left\| f' \right\|_{L^\infty(\mathcal{R})}^2 \right) 2\Theta_3 e^{-\Theta_4 \log(n)^2}}. \end{aligned} \tag{2.49}$$

The RHS of (2.49) is in the form of $C_1 n^{1/2} e^{-C_2 \log(n)^2}$ where C_1, C_2 are positive constants. Multiplying both sides by $n^{3/4}\log(n)^{-3/2}$ will make the RHS of the form $C_1 n^{5/4}\log(n)^{-3/2} e^{-C_2 \log(n)^2}$, which goes to zero hence is bounded. Combining (2.37), (2.49) and the remark above, we obtain that there exist $\mu_1, \mu_2, \mu_3 > 0$ that depend only on parameters of the model such that, for all $n \geq N_1 \vee N_2$:

$$\frac{n^{3/4}}{\log(n)^{3/2}} \mathbb{E}\left(|A_{n,1}(\nu)| \, \mathbf{1}_{\{\nu \in F_1(\mathcal{R}_{\delta/2})\}} | P_1 = \tilde{P} \right) \leq \mu_1 + \sqrt{\mu_2 \mathbb{E}\left((Y_1^1)^2 | P_1 = \tilde{P} \right) + \mu_3}$$

meaning that we have for any $\nu \in [0,1]$, $\tilde{P} \in \mathcal{Z}$ and $n \geq N_1 \vee N_2$:

$$\begin{aligned} h_n(\nu, \tilde{P}) &= \mathbb{E}\left(|A_{n,1}(\nu)|\mathbf{1}_{\{\nu \in F_1(\mathcal{R}_{\delta/2})\}}|\widehat{F_n^1}(x) = \nu, P_1 = \tilde{P}\right) \\ &= \mathbb{E}\left(|A_{n,1}(\nu)|\mathbf{1}_{\{\nu \in F_1(\mathcal{R}_{\delta/2})\}}|P_1 = \tilde{P}\right) \\ &\leq \frac{\log(n)^{3/2}}{n^{3/4}}\left(\mu_1 + \sqrt{\mu_2 \mathbb{E}((Y_1^1)^2|P_1 = \tilde{P}) + \mu_3}\right). \end{aligned}$$

Recall that:

$$\begin{aligned} \mathbb{E}\left(|A_{n,1}(\widehat{F_n^1}(x))|\mathbf{1}_{\{\widehat{F_n^1}(x) \in F_1(\mathcal{R}_{\delta/2})\}}\right) &= \mathbb{E}\left[\mathbb{E}\left(|A_{n,1}(\widehat{F_n^1}(x))|\mathbf{1}_{\{\widehat{F_n^1}(x) \in F_1(\mathcal{R}_{\delta/2})\}}|\widehat{F_n^1}(x), P_1\right)\right] \\ &= \mathbb{E}\left(h_n(\widehat{F_n^1}(x), P_1)\right), \end{aligned}$$

hence for all $n \geq N_1 \vee N_2$, using (2.10) and (2.17), integrating over $\mathbb{M}$ the distribution of P_1 and using the fact that $\mathbb{E}((Y_1^1)^2) < \infty$ we obtain:

$$\begin{aligned} I_2 &= \mathbb{P}\left(\left|\frac{1}{M}\sum_{i=1}^{M} A_{n,i}(\widehat{F_n^i}(x))\right| > K/4, \xi_{n,M}\right) \\ &\leq \frac{4}{K}\mathbb{E}\left(|A_{n,1}(\widehat{F_n^1}(x))|\mathbf{1}_{\{\widehat{F_n^1}(x) \in F_1(\mathcal{R}_{\delta/2})\}}\right) \\ &\leq \frac{4}{K}\frac{\log(n)^{3/2}}{n^{3/4}}\left(\mu_1 + \sqrt{\mu_2 \mathbb{E}((Y_1^1)^2) + \mu_3}\right) \end{aligned}$$

finally proving that there exist a constant Θ_5 that depends only on parameters of the model such that

$$I_2 \leq \frac{4\Theta_5}{K}\frac{\log(n)^{3/2}}{n^{3/4}}. \tag{2.50}$$

- The term II.

By a Taylor expansion we obtain:

$$G_i^{-1} \circ \widehat{F_n^i}(x) - G_i^{-1} \circ F_i(x) = (G_i^{-1})'(F_i(x)) \cdot (\widehat{F_n^i}(x) - F_i(x)) + \frac{(G_i^{-1})''(\Delta_n^i)}{2} \cdot (\widehat{F_n^i}(x) - F_i(x))^2$$

where $\Delta_n^i \in \left(F_i(x), \widehat{F_n^i}(x)\right)$. Letting:

$$\begin{aligned} II_1 &= \mathbb{P}_{\xi_{n,M}}\left(\left|\frac{1}{M}\sum_{i=1}^{M}(G_i^{-1})'(F_i(x)) \cdot (\widehat{F_n^i}(x) - F_i(x))\right| > K/4\right) \\ II_2 &= \mathbb{P}_{\xi_{n,M}}\left(\left|\frac{1}{M}\sum_{i=1}^{M}\frac{(G_i^{-1})''(\Delta_n^i)}{2} \cdot (\widehat{F_n^i}(x) - F_i(x))^2\right| > K/4\right) \end{aligned}$$

we have:

$$II \leq II_1 + II_2. \tag{2.51}$$

We will first derive an upper bound for II_1. Using (2.12), the inverse of G is:

$$G^{-1}(u) = f \circ F^{-1}(u) \quad \forall\ u \in F(\mathcal{R}). \tag{2.52}$$

Differentiating (2.52) we obtain:

$$\begin{aligned} G^{-1}(u)' &= \frac{f' \circ F^{-1}(u)}{F' \circ F^{-1}(u)} \\ &\leq \frac{1}{A} \left\| f' \right\|_{L^\infty(\mathcal{R})} \end{aligned} \tag{2.53}$$

hence:

$$(G_i^{-1})'(F_i(x)) \leq \frac{1}{A} \left\| f' \right\|_{L^\infty(\mathcal{R})}.$$

Using Markov's inequality:

$$\begin{aligned} II_1 &\leq \mathbb{P}\left(\frac{1}{A} \left\| f' \right\|_{L^\infty(\mathcal{R})} \left| \frac{1}{M} \sum_{i=1}^{M} (\widehat{F_n^i}(x) - F_i(x)) \right| > K/4, \xi_{n,M} \right) \\ &\leq \frac{1}{K^2} \frac{16}{A^2} \left\| f' \right\|^2_{L^\infty(\mathcal{R})} \mathbb{E}\Big(\Big| \frac{1}{M} \sum_{i=1}^{M} \widehat{F_n^i}(x) - F_i(x) \Big|^2 \Big) \\ &\leq \frac{1}{MK^2} \frac{16}{A^2} \left\| f' \right\|^2_{L^\infty(\mathcal{R})} \mathbb{E}\left(\left[\widehat{F_n^1}(x) - F_1(x) \right]^2 \right) \\ &\leq \frac{1}{MK^2} \frac{16}{A^2} \left\| f' \right\|^2_{L^\infty(\mathcal{R})} \mathbb{E}\left(\mathbb{E}\left(\left[\widehat{F_n^1}(x) - F_1(x) \right]^2 | P_1 \right) \right) \\ &\leq \frac{1}{MK^2} \frac{16}{A^2} \left\| f' \right\|^2_{L^\infty(\mathcal{R})} \mathbb{E}\left(\frac{F_1(x)(1 - F_1(x))}{n} \right) \\ &\leq \frac{1}{MnK^2} \frac{4}{A^2} \left\| f' \right\|^2_{L^\infty(\mathcal{R})} \end{aligned} \tag{2.54}$$

Since $F_1(x)(1 - F_1(x)) \leq 1/4$ $\mathbb{M}-$a.s.

For II_2, recall that we are on the event $\xi_{n,M} = \bigcap_{i=1}^{M} \{\widehat{F_n^i}(x) \in F_i(\mathcal{R}_{\delta/2})\}$. Since $\Delta_n^i \in \Big(F_i(x), \widehat{F_n^i}(x)\Big)$, then $\Delta_n^i \in F_i(\mathcal{R})$. Using (2.12) the second derivative of the inverse of G is given by:

$$G^{-1}(u)'' = -(G^{-1}(u)')^3 \cdot G'' \circ G^{-1}(u),$$

hence for $u \in F(\mathcal{R})$, combining (2.53), (2.20):

$$|G^{-1}(u)''| \leq \frac{\Theta_1}{A^3} \left\| f' \right\|^3_{L^\infty(\mathcal{R})} =: \Theta_6. \tag{2.55}$$

It follows that:

$$|(G_i^{-1})''(\Delta_n^i)| \leq \Theta_6.$$

So we have:

$$II_2 = \mathbb{P}_{\xi_{n,M}}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}\frac{(G_i^{-1})''(\Delta_n^i)}{2}\cdot(\widehat{F_n^i}(x)-F_i(x))^2\Big| > K/4\Big)$$
$$\leq \frac{4\Theta_6^2}{K^2}\mathbb{E}\Big(\Big|\frac{1}{M}\sum_{i=1}^{M}(\widehat{F_n^i}(x)-F_i(x))^2\Big|^2\Big). \tag{2.56}$$

Letting $W_i = \widehat{F_n^i}(x) - F_i(x)$, and using the Cauchy-Schwarz inequality we have:

$$\mathbb{E}\left(\Big|\frac{1}{M}\sum_{i=1}^{M}W_i^2\Big|^2\right) \leq \mathbb{E}(W_1^4)$$

But $\mathcal{L}(W_1|P_1) = \frac{1}{n}(B(n,F_1(x)) - nF_1(x))$. The 4-th central moment of a Binomial distribution with parameters (n,p) is

$$\mathbb{E}\left[(B(n,p)-np)^4\right] = n[p(1-p)^4 + p^4(1-p)] + 3n(n-1)p^2(1-p)^2$$
$$\leq n^2.$$

Hence:

$$\mathbb{E}(W_1^4) = \mathbb{E}(\mathbb{E}(W_1^4|P_1))$$
$$\leq \frac{1}{n^2}$$

meaninig that (2.56) becomes:

$$II_2 \leq \frac{4\Theta_6^2}{n^2K^2}. \tag{2.57}$$

Finally, combining (2.54) and (2.57):

$$II \leq II_1 + II_2$$
$$\leq \frac{4}{A^2}\|f'\|^2_{L^\infty(\mathcal{R})}\frac{1}{MnK^2} + 4\Theta_6^2\frac{1}{n^2K^2}. \tag{2.58}$$

Combining (2.9),(2.10) and (2.51) we have:

$$\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K, \xi_{n,M}) \leq I_1 + I_2 + II_1 + II_2,$$

and combining (2.58), (2.16) and (2.50) it follows:

$$\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K, \xi_{n,M}) \leq \frac{4}{A^2}\|f'\|^2_{L^\infty(\mathcal{R})}\frac{1}{MnK^2} + \frac{4\Theta_5}{K}\frac{\log(n)^{3/2}}{n^{3/4}} \tag{2.59}$$
$$+ \frac{4}{A^2}\|f'\|^2_{L^\infty(\mathcal{R})}\frac{1}{MnK^2} + 4\Theta_6^2\frac{1}{n^2K^2} \tag{2.60}$$

Recall that we wanted to control

$$\mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K) \leq \mathbb{P}(|\widehat{f}_{n,M}(x) - f(x)| > K, \xi_{n,M}) + \mathbb{P}(\xi^c_{n,M})$$

hence combining (2.8) and (2.60) achieves the proof □

2.5.2 Proof of Theorem 5

We will first derive the result conditional on $P \in \mathcal{Z}$ and then conclude with the dominated convergence theorem. Let $P \sim \mathbb{M}$, denote by F the cumulative function of P and let $Q = f_{\#}P$ and denote by G the cumulative function of Q and g its density. Recall that in this case we have $G = F \circ f^{-1}$. Let $(Y_1, \dots, Y_n)$ be independent random variables with common cumulative distribution function G. Let G_n^{-1} be the generalized inverse of the empirical cumulative function $G_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq x\}}$. For $\alpha \in F(\mathcal{R})$ let

$$Z_n(\alpha) = \sqrt{n}(G_n^{-1}(\alpha) - G^{-1}(\alpha)).$$

By definition of $\mathcal{Z}$, we know that F is differentiable on $\mathcal{R}$. Since $G = F \circ f^{-1}$ and $\alpha \in F(\mathcal{R})$, it follows that G is differentiable at $G^{-1}(\alpha)$ and that $g \circ G^{-1}(\alpha) > 0$. Using the well-known result on the convergence of the empirical process (see [23] for instance):

$$Z_n(\alpha) \underset{n \to \infty}{\longrightarrow} \mathcal{N}\left(0, \sigma_\alpha^2\right) \tag{2.61}$$

in Λ_n-distribution $\mathbb{M}-$a.s. where Λ_n is the distribution of the random sample $(Y_1, \dots, Y_n)$ conditional on P defined in (2.3), and $\sigma_\alpha^2 = \frac{\alpha(1-\alpha)}{G' \circ G^{-1}(\alpha)^2}$. Let $0 < \alpha_1 < \alpha_2 < 1$ and $T = [\alpha_1, \alpha_2]$ where $\alpha_1, \alpha_2 \in F(\mathcal{R})$. There exists $\varepsilon > 0$ such that G is continuously differentiable on the interval $[G^{-1}(\alpha_1) - \epsilon, G^{-1}(\alpha_2) + \epsilon]$ with strictly positive derivative. Hence it holds that the process Z_n weakly converges in $l^\infty(T)$ to $\mathbb{B}/g \circ G^{-1}$ where $\mathbb{B}$ is a standard Brownian bridge (see for instance Corollary 21.5 in Van Der Vaart, Asymptotic Statistics [89]) under Λ_n. Providing that $T \subset F(\mathcal{R})$ $\mathbb{M}-$a.s. for $P \in \mathcal{Z}$, the limiting process $Z = \{\mathbb{B}(t)/g \circ G^{-1}(t), t \in T\}$ is tight since in this case $1/g \circ G^{-1}(t) \leq \|f'\|_{L^\infty(\mathcal{R})} A^{-1}$ $\mathbb{M}-$a.s. uniformly for all $t \in T$, $P \in \mathcal{Z}$. It follows that Z is uniformly bounded in L^1 over T, hence tight. To prove the existence of such T, recall that $x \in \mathcal{R}_\delta = [a + \delta, b - \delta]$. So there exists $\delta_{x,b} \in [x, b] \subset \mathcal{R}$, $F(b) - F(x) = (b - x)F'(\delta_{x,b})$. Because $F' \in \mathcal{F}_{\mathcal{R}}(A, B, C)$, we have in addition that $F'(\delta_{x,b}) \geq A$. Then:

$$\begin{aligned} F(x) &= F(x) - F(b) + F(b) \\ &= F(b) - (b - x)F'(\delta_{x,b}) \\ &\leq F(b) - \delta F'(\delta_{x,b}) \\ &\leq 1 - \delta A. \end{aligned}$$

Similarly we prove that:

$$F(x) \geq \delta A.$$

This means that we have the following uniform bounds in $P \in \mathcal{Z}$ and $x \in \mathcal{R}_\delta$:

$$\delta A \leq F(x) \leq 1 - \delta A. \tag{2.62}$$

Since δ is arbitrary, choose it such that $0 < 1 - \delta A < 1$, i.e. $0 < \delta < A^{-1}$. For this choice of δ, letting $\alpha_1 = \delta A$ and $\alpha_2 = 1 - \delta A$ ensures that $T \subset F(\mathcal{R})$ $\mathbb{M}-$a.s. Now, since weak convergence in $l^\infty(T)$ to a tight element implies stochastic equicontinuity of the converging process (see for instance

Theorem 18.14 in Van Der Vaart, Asymptotic Statistics [89]), $\{Z_n(\alpha), \alpha \in T\}$ is asymptotically equicontinuous, meaning that for all $\eta, \epsilon > 0$, $\exists \gamma > 0$:

$$\limsup_{n\to\infty} \Lambda_n \left(\sup_{s,t\in T, |s-t|<\gamma} |Z_n(s) - Z_n(t)| > \eta \right) < \epsilon.$$

Let $\alpha \in T$. Stochastic equicontinuity guarantees that for any sequence (α_n) such that $\alpha_n \underset{n\to\infty}{\longrightarrow} \alpha$ in Λ_n-probability $\mathbb{M}-$a.s. we have $Z_n(\alpha_n) \underset{n\to\infty}{\longrightarrow} Z(\alpha)$ in Λ_n-probability $\mathbb{M}-$a.s. Indeed, let $\eta, \epsilon > 0$. Choose $\gamma > 0$ such that:

$$\limsup_{n\to\infty} \Lambda_n \left(\sup_{s\in T, |s-\alpha|<\gamma} |Z_n(s) - Z_n(\alpha)| > \eta \right) < \epsilon$$

using equicontinuity. Then there exists $n_1 \in \mathbb{N}$ such that:

$$\Lambda_n \left(\sup_{s\in T, |s-\alpha|<\gamma} |Z_n(s) - Z_n(\alpha)| > \eta \right) < 2\epsilon$$

for all $n \geq n_1$. Since $\alpha_n \underset{n\to\infty}{\longrightarrow} \alpha$ in Λ_n-probability $\mathbb{M}-$a.s., there exists $n_2 \in \mathbb{N}$ such that:

$$\lambda_n \left(|\alpha_n - \alpha| \geq \gamma \right) < \epsilon$$

for all $n \geq n_2$. Then notice that:

$$\begin{aligned}
\Lambda_n \left(|Z_n(\alpha_n) - Z_n(\alpha)| > \eta \right) &\leq \Lambda_n \left(|Z_n(\alpha_n) - Z_n(\alpha)| > \eta, |\alpha_n - \alpha| < \gamma \right) + \Lambda_n \left(|\alpha_n - \alpha| \geq \gamma \right) \\
&\leq \Lambda_n \left(\sup_{s\in T, |s-\alpha|<\gamma} |Z_n(s) - Z_n(\alpha)| > \eta \right) + \Lambda_n \left(|\alpha_n - \alpha| \geq \gamma \right) \\
&\leq 3\epsilon
\end{aligned}$$

for all $n \geq n_1 \vee n_2$, meaning that $Z_n(\alpha_n) - Z_n(\alpha) \underset{n\to\infty}{\longrightarrow} 0$ in Λ_n-probability $\mathbb{M}-$a.s. Finally, writing $Z_n(\alpha_n) = Z_n(\alpha_n) - Z_n(\alpha) + Z_n(\alpha)$ allows us to conclude that:

$$Z_n(\alpha_n) \underset{n\to\infty}{\longrightarrow} Z(\alpha)$$

in Λ_n-distribution $\mathbb{M}-$a.s. Since $F_n(x) \underset{n\to\infty}{\longrightarrow} F(x)$ in Λ_n-probability $\mathbb{M}-$a.s., it holds that:

$$Z_n(F_n(x)) \underset{n\to\infty}{\longrightarrow} Z(F(x))$$

in Λ_n-distribution $\mathbb{M}-$a.s. Now, write:

$$\begin{aligned}
\sqrt{n}(G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)) &= Z_n(F_n(x)) + \sqrt{n}(G^{-1} \circ F_n(x) - G^{-1} \circ F(x)) \\
&= Z_n(F_n(x)) \pm Z_n(F(x)) + \sqrt{n}(G^{-1} \circ F_n(x) - G^{-1} \circ F(x)).
\end{aligned}$$

Stochastic equicontinuity of Z_n implies that $Z_n(F_n(x)) - Z_n(F(x)) \underset{n\to\infty}{\longrightarrow} 0$ in Λ_n-probability $\mathbb{M}-$a.s. The central limit theorem and the delta method guarantee that

$$\sqrt{n}(G^{-1} \circ F_n(x) - G^{-1} \circ F(x)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \tau^2\right)$$

in Λ_n-distribution $\mathbb{M}-$a.s. where

$$\tau^2 = \frac{F(x)(1-F(x))}{\left(g \circ G^{-1} \circ F(x)\right)^2},$$

and using (2.61) it follows that

$$Z_n(F(x)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \tau^2\right)$$

in Λ_n-distribution $\mathbb{M}-$a.s. Since $Z_n(F(x))$ and $\sqrt{n}(G^{-1} \circ F_n(x) - G^{-1} \circ F(x))$ are independent conditional on P we conclude that:

$$\sqrt{n}(G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \sigma_F^2\right)$$

in Λ_n-distribution $\mathbb{M}-$a.s. where $\sigma_F^2 = 2\tau^2$, or equivalently

$$\frac{\sqrt{n}}{\sigma_F}(G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, 1\right)$$

in Λ_n-distribution $\mathbb{M}-$a.s. for any $P \in \mathcal{Z}$. We finally conclude that

$$\frac{\sqrt{n}}{\sigma_F}(G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, 1\right)$$

in $\mathbb{P}-$distribution by proving convergence of the characteristic functions using the dominated convergence theorem. □

2.5.3 Proof of Proposition 4

The strategy is to establish stochastic equicontinuity of the kernel density estimator conditional on the event $P = \tilde{P}$, and conclude by the dominated convergence theorem. For any $\eta, \varepsilon > 0$, we first prove that there exists $\gamma > 0$ such that:

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\substack{s,t\in f(\mathcal{R}) \\ |s-t|<\gamma}} |g_n(s) - g_n(t)| > \eta | P = \tilde{P} \right) < \epsilon. \tag{2.63}$$

We have:

$$\begin{aligned} \mathbb{P}\left(\sup_{\substack{s,t\in f(\mathcal{R}) \\ |s-t|<\gamma}} |g_n(s) - g_n(t)| > \eta | P = \tilde{P} \right) \leq & \mathbb{P}\left(\sup_{\substack{s,t\in f(\mathcal{R}) \\ |s-t|<\gamma}} |g(s) - g(t)| > \eta/2 | P = \tilde{P} \right) \\ & + 2\mathbb{P}\left(\sup_{\substack{s,t\in f(\mathcal{R}) \\ |s-t|<\gamma}} |g_n(t) - g(t)| > \eta/4 | P = \tilde{P} \right) \end{aligned} \tag{2.64}$$

The first term of the right hand side is easily handled because g is continuous on the compact set $f(\mathcal{R})$ implying uniform continuity, hence there exists $\gamma > 0$ such that this term is equal to 0.

Conditioning on $P = \tilde{P}$ makes g a constant with respect to the conditional distribution, so we will apply Dominik Wied and Rafael Weißbach results in [96].

Theorem 7 (Almost sure uniform convergence of the kernel density estimator)**.** *Let $(X_1, \ldots, X_n)$ be independent random variables with common cumulative function F absolutely continuous with respect to the Lebesgue measure, and denote by f their density. The kernel density estimator of f is given by*

$$f_n(t) = \frac{1}{nh_n} \sum_{j=1}^{n} K\left(\frac{X_j - t}{h_n}\right)$$

where $h_n > 0$, $h_n \to 0$ and K is a kernel function. Suppose that the following hold

i) K is right continuous,

ii) K has bounded variations,

iii) $\lim_{|x|\to\infty} K(x) = 0$,

iv) $\sum_{n\leq 1} e^{-\gamma n h_n^2} < \infty$ for all $\gamma > 0$,

v) f is uniformly continuous on $\mathbb{R}$,

then the kernel density estimator uniformly converges almost surely on $\mathbb{R}$.

For our purposes, the choice of bandwidth and Kernel is not important so hypothesis *i*), *ii*), *iii*) and *iv*) of Theorem 7 are satisfied in Proposition 4. We do not have uniform continuity of the true density, but this is easily overcome because we do not need uniform convergence on $\mathbb{R}$: uniform convergence on the compact set $f(\mathcal{R})$ is enough. By the triangle inequality we have:

$$\|g_n - g\|_{L^\infty(f(\mathcal{R}))} \leq \|g_n - \mathbb{E}g_n\|_{L^\infty(f(\mathcal{R}))} + \|\mathbb{E}g_n - g\|_{L^\infty(f(\mathcal{R}))}. \tag{2.65}$$

We refer the reader to [96] for a proof that the first term on the RHS of equation (2.65) goes to 0. The hypothesis of uniform continuity is not necessary to establish convergence of that term, hence the proof in [96] can be followed step by step to obtain

$$\sup_{t\in f(\mathcal{R})} |g_n(t) - \mathbb{E}g_n(t)| \underset{n\to\infty}{\longrightarrow} 0.$$

To prove that the second term goes to 0, since our Kernel is compactly supported we have for any $t \in f(\mathcal{R})$:

$$\begin{aligned}
|\mathbb{E}g_n(t) - g(t)| &= \left|\mathbb{E}\left(\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{Y_i - t}{h_n}\right)\right) - g(t)\right| \\
&= \left|\int_{-\infty}^{+\infty} K(u)g(t + uh_n)du - g(t)\right| \\
&= \left|\int_{-\infty}^{+\infty} K(u)\left[g(t + uh_n) - g(t)\right]du\right|
\end{aligned}$$

but $h_n \to 0$, g is continuous and both $f(\mathcal{R})$ and the support $\mathrm{Supp}(K)$ of K are compact, hence:

$$\sup_{t' \in f(\mathcal{R}), u' \in \mathrm{Supp}(K)} |g(t' + u'h_n) - g(t')| \underset{n\to\infty}{\longrightarrow} 0$$

meaning that

$$\sup_{t \in f(\mathcal{R})} |\mathbb{E}g_n(t) - g(t)| \underset{n\to\infty}{\longrightarrow} 0$$

finally leading

$$\sup_{t \in f(\mathcal{R})} |g_n(t) - g(t)| \underset{n\to\infty}{\longrightarrow} 0.$$

So we have uniform convergence of g_n on $f(\mathcal{R})$, hence the two term in (2.64) go to 0 as n goes to infinity, proving (2.63).

From Theorem 5, we know that $G_n^{-1} \circ F_n(x) \underset{n\to\infty}{\longrightarrow} G^{-1} \circ F(x)$ in Λ_n-probability $\mathbb{M}-$a.s., hence $g_n \circ G_n^{-1} \circ F_n(x) \underset{n\to\infty}{\longrightarrow} g \circ G^{-1} \circ F(x)$ in Λ_n-probability $\mathbb{M}-$a.s. The reason for this is a direct consequence of (2.63) and the splitting of the deviation probability on the events $\{|G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)| > \gamma\}$ and $\{|G_n^{-1} \circ F_n(x) - G^{-1} \circ F(x)| \leq \gamma\}$. It is clear that $F_n(x)(1 - F_n(x))$ converges to $F(x)(1 - F(x))$ conditional on $P = \tilde{P}$. Hence one has:

$$\mathbb{P}(|\sigma_n(F)^2 - \sigma(F)^2| > \varepsilon | P = \tilde{P}) \longrightarrow 0.$$

We conclude by the dominated convergence theorem. □

2.5.4 Proof of Proposition 5

The strategy here is the same as before: we prove that the convergence in distribution conditional on P_1 happens $\mathbb{M}-$a.s., and conclude by the dominated convergence theorem. Using Theorem 5, we have:

$$\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta}_{n,x} - \boldsymbol{\theta}_x) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, I_M)$$

in Λ_n-distribution $\mathbb{M}-$a.s. Using the representation theorem for weakly convergent sequences by Skorokod [9], we derive that there exists a probability space $(\Omega', \mathcal{A}', \mathbb{P}')$ and random variables $Z, (\xi_n)_{n\geq 1}$ defined on $(\Omega', \mathcal{A}', \mathbb{P}')$ satisfying:

$$\begin{aligned} &Z \sim \mathcal{N}(0, I_M), \\ &\xi_n \underset{n\to\infty}{\overset{\mathbb{P}'}{\longrightarrow}} 0, \\ &\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta}_{n,x} - \boldsymbol{\theta}_x) \overset{\mathcal{L}}{=} Z + \xi_n. \end{aligned} \tag{2.66}$$

In (2.66), note that the equality in distribution is to be understood with respect to the conditional distribution λ_n for the term on the left, and with respect to $\mathbb{P}'$ for the term on the right. For ease

of writing, define $\boldsymbol{W}_n = \sqrt{n}(\boldsymbol{\theta}_{n,x} - \boldsymbol{\theta}_x)$. Because the distributions $P_1, \dots, P_M$ are constants with respect to Λ_n, so is $\boldsymbol{\Sigma}$. Multiplying on both sides (2.66) by $\boldsymbol{R}\boldsymbol{\Sigma}^{1/2}$, we have:

$$\boldsymbol{R}\boldsymbol{W}_n \overset{\mathcal{L}}{=} \boldsymbol{R}\boldsymbol{\Sigma}^{1/2} Z + \boldsymbol{R}\boldsymbol{\Sigma}^{1/2}\xi_n.$$

Observing that $\boldsymbol{R}\boldsymbol{\Sigma}^{1/2} Z \sim \mathcal{N}(0, \boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)$, it follows:

$$(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)^{-1/2}\boldsymbol{R}\boldsymbol{W}_n \overset{\mathcal{L}}{=} \mathcal{N}(0, I_{M-1}) + (\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{1/2}\xi_n$$

hence

$$(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)^{-1/2}\boldsymbol{R}\boldsymbol{W}_n \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, I_{M-1})$$

in Λ_n-distribution $\mathbb{M}-$a.s. hence:

$$\sqrt{n}(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)^{-1/2}\boldsymbol{R}(\boldsymbol{\theta}_{n,x} - \boldsymbol{\theta}_x) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, I_{M-1}).$$

in Λ_n-distribution $\mathbb{M}-$a.s. Under H_0^x, it follows that:

$$\sqrt{n}(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)^{-1/2}\boldsymbol{R}\boldsymbol{\theta}_{n,x} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, I_{M-1}). \tag{2.67}$$

in Λ_n-distribution $\mathbb{M}-$a.s. Now observe that the matrix $\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t$ is symmetric positive definite. The symmetry and non-negativity are obvious. Let $x \in \mathbb{R}^{M-1}$. Then

$$x^t\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t x = \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^t x\right\|^2 \geq 0.$$

Moreover

$$\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}^t x\right\|^2 = 0 \iff \boldsymbol{R}^t x = 0$$

because $\boldsymbol{\Sigma}^{1/2}$ is invertible. Now recall that

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 & & \\ \vdots & & \ddots & \\ 1 & & & -1 \end{bmatrix}$$

hence $\boldsymbol{R}^t x = 0$ if and only if $x = 0$ hence $\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t$ is definite. The inverse and square root mappings are continuous over the set of symmetric positive definite matrices. Hence by the continuous mapping theorem and Proposition 4 we have:

$$(\boldsymbol{R}\boldsymbol{\Sigma}_{\boldsymbol{n}}\boldsymbol{R}^t)^{-1/2} \underset{n\to\infty}{\longrightarrow} (\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^t)^{-1/2}$$

in Λ_n-probability $\mathbb{M}-$a.s. proving that

$$\sqrt{n}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}^t)^{-1/2}\boldsymbol{R}\boldsymbol{\theta}_{n,x} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, I_{M-1}). \tag{2.68}$$

in Λ_n-probability $\mathbb{M}-$a.s. under H_0^x. We conclude by the dominated convergence theorem.

□

2.6 The non increasing case

Until now we assumed f increasing and $f \in \mathcal{C}^2(\mathbb{R})$, see Assumption 1, and proved that in this case it can be estimated. Our estimator of f is an increasing function, so clearly if f is non increasing, meaning that there exists $u < v, f(u) \geq f(v)$, it cannot estimate f correctly everywhere. That being said, one might wonder if it is possible that for non increasing f, there exists an increasing function h satisfying:

$$f_{\#}P = h_{\#}P \quad \forall P \in \mathcal{Z}.$$

If such h exists, then our estimator will converge to it at rates as in Theorem 4. From a practical perspective, f and h are completely indistinguishable: no experiment can be designed to determine whether the data was generated with f or h because the samples are independent and not paired, and both functions transport the probability distributions $P \in \mathcal{Z}$ in the same way. For instance, for a random variable X with symmetric distribution, X and $-X$ will have the same distribution and are perfectly indistinguishable. This suggests the introduction of an equivalence relation on the set $\mathcal{C}^2(\mathbb{R})$ such that

$$f \sim g \text{ if and only if } f_{\#}P = h_{\#}P \text{ for all } P \in \mathcal{Z}.$$

Denote $[f] = \{h \in \mathcal{C}^2(\mathbb{R}) | h \sim f\}$ the equivalence class of f. If $f \in \mathcal{C}^2(\mathbb{R})$ is non increasing and is the true transport that generated the data, we guarantee that our estimator converges to $h \in [f]$ at rates in Theorem 4 if and only if h is increasing. The problem then boils down to the question of the existence of an increasing $h \in [f]$ when one only assumes $f \in \mathcal{C}^2(\mathbb{R})$.

Proposition 6. *(Sufficient condition for the existence of an increasing function $h \in [f]$) Suppose $f \in \mathcal{C}^2(\mathbb{R})$ and let $P \in \mathcal{Z}$, $Q = f_{\#}P$. Denote by F, G the cumulative functions of P and Q respectively. Suppose that the function $h = G^{-1} \circ F$ is independent of the choice of P. Then*

$$f_{\#}P = h_{\#}P \text{ for all } P \in \mathcal{Z}.$$

Proof. Let $P \in \mathcal{Z}$, $Q = f_{\#}P$ and let X, Y be random variables with distributions P, Q respectively. Then

$$Y \overset{\mathcal{L}}{=} f(X)$$

since $\mathcal{L}(Y) = Q = f_{\#}P$. Moreover

$$Y \overset{\mathcal{L}}{=} G^{-1} \circ F(X)$$

by the inversion theorem. By assumption $h = G^{-1} \circ F$ is independent of $P \in \mathcal{Z}$, hence we conclude

$$f_{\#}P = h_{\#}P \text{ for all } P \in \mathcal{Z}$$

□

Proposition 7. *Suppose $f \in \mathcal{C}^2(\mathbb{R})$ and let $P \in \mathcal{Z}$. Denote by F, G the cumulative functions of P and $Q = f_{\#}P$ respectively. If there exists an increasing function $h \in [f]$, we have:*

$$P(]-\infty, x]) = P(f^{-1}(]-\infty, h(x)])). \tag{2.69}$$

Proof. Let $P \in \mathcal{Z}, x \in \mathbb{R}$, and $h \in [f]$ such that h is increasing. Let X be a random variable with distribution P. We have:

$$\begin{aligned} P(]-\infty, x]) &= \mathbb{P}(X \leq x) && (2.70)\\ &= \mathbb{P}(h(X) \leq h(x))) \quad \text{because } h \text{ is increasing on } \mathbb{R} && (2.71)\\ &= \mathbb{P}(f(X) \leq h(x))) \quad \text{because } h(X) \overset{\mathcal{L}}{=} f(X) && (2.72)\\ &= \mathbb{P}(X \in f^{-1}(]-\infty, h(x)])) && (2.73)\\ &= P(f^{-1}(]-\infty, h(x)])). && (2.74) \end{aligned}$$

□

Proposition 7 tells us that if f is non increasing, the existence of an increasing $h \in [f]$ severely constrains the family of distributions $\mathcal{Z}$, making the existence of such h very unlikely in general.

Example 1. *Let $f \in \mathcal{C}^2(\mathbb{R})$ be a decreasing function, and assume that there exists an increasing $h \in [f]$. There are two cases:*

i) $\forall x \in \mathbb{R}, f(x) \neq h(x)$

Since f, h are smooth functions, $\forall x \in \mathbb{R}, f(x) < h(x)$ or $\forall x \in \mathbb{R}, f(x) > h(x)$. Because f is decreasing and h increasing, it also holds that $\forall x, y \in \mathbb{R}, f(x) > h(y)$ or $\forall x, y \in \mathbb{R}, f(x) < h(y)$. Let $P \in \mathcal{Z}$ and X a random variable with distribution P. Let $x \in \mathbb{R}$:

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}(h(X) \leq h(x)) \\ &= \mathbb{P}(f(X) \leq h(x)) \end{aligned}$$

meaning that $\forall x \in \mathbb{R}$, $\mathbb{P}(X \leq x) = 0$ or $\forall x \in \mathbb{R}$, $\mathbb{P}(X \leq x) = 1$, a contradiction.

ii) $\exists a \in \mathbb{R}, f(a) = h(a)$

In this case:

$$\begin{aligned} \mathbb{P}(X \leq a) &= \mathbb{P}(f(X) \geq f(a)) \\ &= \mathbb{P}(h(X) \geq f(a)) \\ &= \mathbb{P}(h(X) \geq h(a)) \\ &= \mathbb{P}(X \geq a) \\ &= 1 - \mathbb{P}(X \leq a) \end{aligned}$$

hence $\mathbb{P}(X \leq a) = 1/2$. Meaning that if such h exists, the median of all distributions in $\mathcal{Z}$ must be a. For instance if all distributions $P \in \mathcal{Z}$ are symmetric with center of symmetry a, then $h = f(2a - \cdot)$ is increasing and $h \in [f]$. One checks that indeed $P(]-\infty, a]) = 1/2$ for all $P \in \mathcal{Z}$ and $f(a) = h(a)$ in this case.

Example 2. *Consider $f : x \mapsto |x|$ and suppose there exists an increasing $h \in [f]$. Suppose $\mathcal{Z}$ contains a symmetric distribution P with strictly increasing c.d.f. on $\mathbb{R}$ and let $X \sim P$. For $y \geq 0$ we have:*

$$\mathbb{P}(h(X) \leq y) = \mathbb{P}(f(X) \leq y) \quad \textit{because } h \in [f]$$

$$
\begin{aligned}
&= \mathbb{P}(|X| \leq y) \\
&= \mathbb{P}(-y \leq X \leq y) \\
&= 2(\mathbb{P}(X \leq y) - F(0)) \quad \textit{because } P \textit{ is symmetric} \\
&= 2\mathbb{P}(X \leq y) - 1.
\end{aligned}
$$

Since h is increasing, $\mathbb{P}(h(X) \leq y) = \mathbb{P}(X \leq h^{-1}(y))$. Noting F the c.d.f. of X, we have:

$$F \circ h^{-1}(y) = 2F(y) - 1, \quad \forall\ y \geq 0.$$

The function $y \mapsto 2F(y) - 1$ is strictly increasing on $\mathbb{R}_+$ with image $[0, 1]$. Its inverse is the mapping $\alpha \mapsto F^{-1}(\frac{\alpha+1}{2})$ for $\alpha \in [0, 1]$. Since all $\alpha \in [0, 1]$ are of the form $\alpha = F(x)$ for some $x \in \mathbb{R}$ we have:

$$h(x) = F^{-1}\left(\frac{F(x) + 1}{2}\right) \quad \forall\ x \in \mathbb{R}.$$

Since $h \in [f]$, and h is increasing, it must hold that $h = G^{-1} \circ F$ for all $P \in \mathcal{Z}$ where F is the c.d.f. of P and G the c.d.f. of $f(X)$. But here, h clearly depends on the choice of P. For instance $\lim_{x \to -\infty} h(x) = F^{-1}(1/2)$ and $h(0) = F^{-1}(3/4)$. If $\mathcal{Z}$ contains another symmetric distribution for instance, then the expression for h will be different, which is impossible since $h \in [f]$, hence the contradiction.

The existence of an increasing $h \in [f]$ for non increasing f has very strong implications on the nature of $\mathcal{Z}$. Even for simplistic non increasing f, we saw that it is unreasonable to consider the existence of such h. So our problem of distribution matching can not systematically be reframed in terms of increasing f.

Nevertheless one might still be wondering if there are specific points at which f can still be estimated when f in non increasing and there is no increasing $h \in [f]$. For a point $x_0 \in \mathbb{R}$, the only thing absolutely needed in order to estimate $f(x_0)$ at rates in Theorem 1 is that for any $P \in \mathcal{Z}$ we have:

$$P(]-\infty, x_0]) = P(f^{-1}(]-\infty, f(x_0)]))$$

which is equivalent to saying that $G^{-1} \circ F(x_0) = f(x_0)$ for all $P \in \mathcal{Z}$ where F, G are the c.d.f. of X and $f(X)$ respectively.

Proposition 8. *Let $x_0 \in \mathbb{R}$. If f is strictly increasing and invertible at x_0, then:*

$$P(]-\infty, x_0]) = P(f^{-1}(]-\infty, f(x_0)])) \quad \forall\ P \in \mathcal{Z}$$

or equivalently:

$$f(x_0) = G^{-1} \circ F(x_0)$$

for all $P \in \mathcal{Z}$ where F, G are the c.d.f. of P and $f_{\#}P$ respectively.

Proof. Let $P \in \mathcal{Z}$. Since f is strictly increasing and invertible at x_0, the sets $\{\omega, X(\omega) \leq x_0\} = \{\omega, f(X(\omega)) \leq f(x_0)\}$ are equal. This implies $\mathbb{P}(X \leq x_0) = \mathbb{P}(f(X) \leq f(x_0))$, meaning that $P(]-\infty, x_0]) = P(f^{-1}(]-\infty, f(x_0)]))$, or, equivalently, $f(x_0) = G^{-1} \circ F(x_0)$ where F, G are the c.d.f. of P and $f_{\#}P$ respectively. □

Proposition 8 gives a geometric sufficient condition on the graph of f about the points where it can be estimated. Denote $\mathcal{G}(f) = \{(x, f(x))|x \in \mathbb{R}\}$ the graph of f, and let $NW(t) = \{(x, y) \in \mathbb{R}^2|x < t, y > f(t)\}$, $SE(t) = \{(x, y) \in \mathbb{R}^2|x > t, y < f(t)\}$. $NW(t)$ and $SE(t)$ are the north-west and south-east quadrants of the plane at $(t, f(t))$. The sufficient condition in Proposition 8 is equivalent to saying that $\mathcal{G}(f) \cap [NW(x_0) \cup SE(x_0)] = \emptyset$. The converse of Proposition 8 is false in general, but a partial converse can be stated as follows:

Proposition 9. *Suppose there exists $x_0 \in \mathbb{R}$ such that:*

$$f(x_0) = G^{-1} \circ F(x_0)$$

for all $P \in \mathcal{Z}$ where F, G are the c.d.f. of P and $f_{\#}P$ respectively. Then either one of the following holds:

i) $\mathcal{G}(f) \cap NW(x_0) = \emptyset$ *and* $\mathcal{G}(f) \cap SE(x_0) = \emptyset$

ii) $\mathcal{G}(f) \cap NW(x_0) \neq \emptyset$ *and* $\mathcal{G}(f) \cap SE(x_0) \neq \emptyset$

Proof. By contradiction. Let $P \in \mathcal{Z}$, $X \sim P$ and F the c.d.f. of X. Suppose exactly one of $\mathcal{G}(f) \cap NW(x_0)$ and $\mathcal{G}(f) \cap SE(x_0)$ is non empty. For instance, say that $\mathcal{G}(f) \cap NW(x_0) \neq \emptyset$ and $\mathcal{G}(f) \cap SE(x_0) = \emptyset$. Since $\mathcal{G}(f) \cap NW(x_0) \neq \emptyset$ and f is smooth, there exists a non empty interval $I \subset \{x \in \mathbb{R} \ | x \leq x_0\}$ such that $\forall\ x \in I$, $f(x) > f(x_0)$. Moreover, since $\mathcal{G}(f) \cap SE(x_0) = \emptyset$, $\{f(X) \leq f(x_0)\} \subset \{X \leq x_0\}$, meaning that $\mathbb{P}(f(X) \leq f(x_0)) = \mathbb{P}(f(X) \leq f(x_0), X \leq x_0)$. But:

$$\begin{aligned}\mathbb{P}(X \leq x_0) &= \mathbb{P}(X \leq x_0, f(X) \leq f(x_0)) + \mathbb{P}(X \leq x_0, f(X) > f(x_0)) \\ &\geq \mathbb{P}(f(X) \leq f(x_0)) + \mathbb{P}(X \in I)\end{aligned}$$

Since I is a non empty interval and F strictly increasing on $\mathbb{R}$, $\mathbb{P}(X \in I) > 0$, i.e. $\mathbb{P}(X \leq x_0) > \mathbb{P}(f(X) \leq f(x_0))$, which contradicts the assumption. The other case is proven in a similar way. □

2.7 Appendix

Proof of Lemma 3. Let $n \geq 2$ and $t_n = n^{1/4}\log(n)^{-1/2}$. Recall that the moment generating function of a binomial distribution $B(n, z)$ is $\mathbb{E}(e^{tB(n,z)}) = (1 - z + ze^t)^n$. Then:

$$\begin{aligned}\Phi_n(z) &= \mathbb{E}\left(\exp\{\frac{n^{1/4}}{\sqrt{\log(n)}}\frac{1}{\sqrt{n}}(B(n, z) - nz)\}\right) \\ &= \mathbb{E}\left(\exp\{\frac{t_n}{\sqrt{n}}(B(n, z) - nz)\}\right) \\ &= e^{-t_n\sqrt{n}z}\mathbb{E}\left(\exp\{\frac{t_n}{\sqrt{n}}B(n, z)\}\right) \\ &= e^{-t_n\sqrt{n}z}(1 - z + ze^{t_n/\sqrt{n}})^n \\ &= e^{zu_n}(1 - zv_n)^n\end{aligned}$$

where $u_n = -t_n\sqrt{n} < 0$ and $v_n = 1 - e^{t_n/\sqrt{n}} < 0$. Φ_n is differentiable and:

$$\Phi_n'(z) = u_n e^{u_n z}(1 - v_n z)^n - nv_n e^{u_n z}(1 - v_n z)^{n-1}$$

$$= (1 - v_n z)^{n-1} e^{u_n z}(u_n(1 - v_n z) - n v_n).$$

Since $(1 - v_n z)^{n-1} e^{u_n z} \geq 0$ the sign of Φ_n' is the sign of $(u_n(1 - v_n z) - n v_n)$, solving in z it follows:

$$\Phi_n'(z) \geq 0 \quad \text{if and only if} \quad z \leq \frac{1}{1 - e^{t_n/\sqrt{n}}} + \frac{\sqrt{n}}{t_n}.$$

Let $h : u \mapsto (1 - e^u)^{-1} + u^{-1}$. h is differentiable for all $u > 0$ and one easily checks

$$h'(u) \leq 0 \quad \text{if and only if} \quad u^2 + 2 \leq 2\cosh(u).$$

Expanding the hyperbolic cosine in power series:

$$\cosh(u) = \sum_{i=0}^{\infty} \frac{u^{2i}}{(2i)!} = 1 + \frac{u^2}{2} + \sum_{i=2}^{\infty} \frac{u^{2i}}{(2i)!}.$$

Since $\sum_{i=2}^{\infty} u^{2i}/(2i)! \geq 0$, we have that $u^2 + 2 \leq 2\cosh(u)$ is true for all $u > 0$, so $h'(u) \leq 0$ for all $u > 0$ and h is decreasing on $(0, \infty)$. Moreover an elementary calculations shows that $0 \leq h(u) \leq 1/2$ for all $u > 0$, meaning that

$$\Phi_n'(z) \geq 0 \quad \text{if and only if} \quad z \leq h(t_n/\sqrt{n}) = h(n^{-1/4}\log(n)^{-1/2}),$$

with h positive, bounded by $1/2$ and decreasing. In particular, it implies that Φ_n is increasing on $[0, h(2^{-1/4}(\log 2)^{-1/2})] = [0, (1 - e^{\lambda})^{-1} + \lambda^{-1}]$ for all $n \geq 2$. □

Proof of Lemma 4. Since $X \sim B(n, z)$, $\mathbb{E}(e^{t(X-nz)}) = e^{-tnz}(1 - z + ze^t)^n$ and $\mathbb{E}(e^{-t(X-nz)}) = e^{tnz}(1 - z + ze^{-t})^n$. So, solving in t we have:

$$\mathbb{E}e^{t(X-nz)} \geq \mathbb{E}e^{-t(X-nz)} \quad \text{if and only if} \quad 0 \leq 1 - z + ze^t - e^{2tz}(1 - z + ze^{-t}) =: f_z(t)$$

f_z is differentiable and

$$f_z(t)' = ze^t - 2z(1 - z)e^{2zt} - z(2z - 1)e^{t(2z-1)}$$

so that:

$$\frac{f_z(t)'}{ze^{2zt}} = e^{t(1-2z)} - 2(1 - z) + (1 - 2z)e^{-t}.$$

It follows:

$$\frac{f_z(t)'}{ze^{2zt}} \geq 0 \quad \text{if and only if} \quad e^{t(1-2z)} - 1 \geq (1 - 2z)(1 - e^{-t}).$$

Using twice the fact that $\forall x \in \mathbb{R}, e^x - 1 \geq x$ and that $\forall 0 \leq z \leq 1/2, 1 - 2z \geq 0$, we conclude that the last inequality is true, hence $f_z(t)' \geq 0$ and f_z is non decreasing. Noting that $f_z(0) = 0$ achieves the proof. □

Chapter 3

Internet Latency measurements modelling using Fourier decomposition and ARMA Seasonal-GARCH models

Abstract

Internet content of large audience websites is typically copied and stored on numerous points of presence around the world by *CDN*, in order to spread the audience. *CDN* are companies that rent Websites a system of distributed servers that deliver webpages to their Internet users depending, for instance, on their geographic locations. A multi-*CDN* architecture refers to a Website that uses more than one *CDN*. Latency measures the time in millisecond taken by a request to reach the destination server, and for the response to get back to the host server. Determining the *CDN* with minimal latency to route users efficiently in the Network has become a priority in the industry. We propose a simple time series approach to model and forecast the object of interest in the industry: the so-called median-process of Internet latency, obtained by aggregating measurements over regular partitions of the interval $[0, T]$ with varying mesh $\Delta > 0$. The modeling reveals strong mean and variance dynamics using Fourier series and ARMA Seasonal GARCH, namely an ARMA-GARCH process with additional deterministic seasonal components in the volatility. The parameters of the conditional mean model are estimated by ordinary least squares and the conditional variance model are estimated with a one stage quasi maximum likelihood. The importance of choice of Δ is discussed and we show that the median latency process can not be predicted for $\Delta < 120$s. The forecasting performance of the model is compared against natural baselines used in the industry and a notion of predictability of time series is discussed. A new test for residual information based on a certain entropic criterion is introduced.

3.1 Introduction

3.1.1 Setting and motivation

The task of spreading the Internet users across the network in order to minimize page load time has increasingly become a priority in the digital market. The impact of slow page load times on attendance, online sales and conversion rates is gaining more and more attention. It has been long known that page load time plays a significant role in e-commerce [39] and it is now a global concern: the overload of a server can cause significant increase in page load time. Large audience websites can overcome this issue by duplicating the content of their servers in order to multiply the number of access points. To avoid building expensive and time consuming private infrastructures, more and more websites go through third party companies that manage the duplication, storage and delivery of content using their own infrastructure. Each of these companies, called *CDN*, deliver a certain performance measured by latency. Latency measures the time in millisecond taken by a request to reach the destination server, and for the response to get back to the host server. The lower the latency the faster the communication between the servers, and conversely. The market of Internet content delivery has become very competitive as websites are now increasingly concerned by page load time. Choosing the right *CDN* at the right time is an effective way to reduce page load time.

As seen in Chapter 2, large audience websites that duplicate their content near two or more *CDN* must choose which one is going to deliver the content each time a user connects to their website. This process is called *load-balancing*, and is operated by *load-balancers* (see Chapter 1 Section 1.2 for a presentation of the notion). The time a user will wait in page loadings depends on the *CDN*'s latency, but when two or more *CDN* are available, the process of selecting the *CDN* that will deliver the content also takes time. Hence *load-balancers* must face the following tradeoff: the more time spent selecting the right *CDN*, the less latency and vice and versa. In order to address the *CDN* selection problem when a new user needs to be routed, there are two main possibilities:

- Individual approach
- Community approach

Before describing those two possibilities, we place ourselves in the case where a user I connects to a Website whose content can be delivered by K *CDN* $C_1, \ldots, C_K$. Those *CDN* are said in competition.

Individual approach consists in making the new user I performs latency measurements on all competing *CDN* and select the one with minimal latency. This procedure gives strong guarantees on the performance of the selected *CDN*, but this *CDN* selection process is time intensive: of the order of a couple seconds.

In the community approach, the *load-balancer* geographically partitions the surface of the globe, and produces predictions of future latency for the *CDN* $C_1, \ldots, C_K$ for each element of the geographic partition and Internet service provider, using latency measurements from the *CDN*

$C_1, \ldots, C_K$ generated by other Network users. When the new user I tries to connect to the Website, the *load-balancer* identifies his Internet service provider as well as its geographic location and requests in a database the latency predictions for each *CDN* $C_1, \ldots, C_K$. The *load-balancer* then orders the latency predictions of the competing *CDN* in ascending order and selects the *CDN* having the lower predicted value. Since the predictions are made upstream, the time required to route a user in the community approach is limited to this database request, which is of the order of a few milliseconds.

The individual approach may seem more attractive than the collective approach because it guarantees that the fastest *CDN* is always selected. However, the three orders of magnitude of difference in the time necessary to choose a *CDN* between the two methods makes it totally ineffective in practice. Indeed, even in the event of a prediction error in the collective approach, it is very unlikely that the user would have loaded the Web page faster in the individual approach. The collective approach is therefore widely preferred.

Each time an Internet user is connected to a *CDN* and has finished loading the content, multiple other *CDN* monitored by Citrix are chosen randomly and latency measurement are performed on each one of them until the user tries to load new content. The latency measurement consists in downloading a small object contained in a single TCP packet through an HTTP request and is computed as the elapsed time between the sending of the first bit of the request and the receiving of the first bit of the response. Citrix collects those measurements and use them to make online latency predictions for each *CDN* in the Network. When a new Internet user needs to access a customer website, the *CDN* with the lowest predicted latency calculated with measurements triggered by users close in space and time is selected. The samples used for prediction are typically composed of latency measurements performed by thousands of different individuals at a high frequency, located in different areas with different internet offers and speed. This results in particularly irregular and noisy time series. A way around this problem is the aggregation of the latency time series by computing rolling medians over regular time intervals of length Δ, resulting in a regular time series with increased signal to noise ratio. In this context, we address the problem of modeling and predicting this so-called median latency process. As Δ decreases, the variance explodes rapidly and, a phenomenon reminiscent of microstructure noise [44] [5] [75] [76]. Conversely, as Δ increases, a clear, almost noise free, seasonal signal that captures diurnal and nocturnal cycles of Internet activities is revealed. This suggests that for large Δ the median process can be accurately predicted whereas for small Δ the signal to noise ratio is too weak and predictions are intractable. For *load-balancing* purposes it is crucial to make accurate predictions with the smallest Δ possible. At a given time t when a user needs to be directed to *CDN* A or B, knowing that A will perform better than B in 2 hours has no value. But if Δ is chosen too small, then the ability to make accurate predictions is questionable.

3.1.2 Main results

Let $T > 0$ and $[0, T]$ be a time interval. Timestamps of measurements collected by Citrix are rounded up to the second hence a natural structure for the data generating process is that of a discrete-time stochastic process with the time index expressed in seconds. Formally, we observe

the process $Z = \{Z_t | t \in \{0, \dots, T\}\}$, where Z_t is the empirical measure defined by:

$$Z_t = \frac{1}{N_t} \sum_{k=1}^{N_t} \delta_{Y_k^t}$$

where δ_x is the Dirac measure at $x \in \mathbb{R}$, N_t is the (random) number of latency measurements received at time t and $(Y_k^t)_{K \in \{1,\dots,N_t\}}$ are the latency measurements received at time t. The object of interest for *load-balancing* purposes is not Z, but a certain functional of Z. For $\Delta > 0$, $n \in \mathbb{N}$, let $t_n = n\Delta$ and define the series:

$$X_{t_n}^{\Delta} = \text{Median}\left(Y_k^t\right)_{\substack{t \in]t_{n-1}, t_n] \\ k \in \{1,\dots,N_t\}}}$$

$X_{t_n}^{\Delta}$ is the median of all measurements with timestamps falling in the interval $]t_{n-1}, t_n]$, and the process $(X_{t_n}^{\Delta})_{0 \leq n\Delta \leq T}$ is called the median-process at frequency Δ, see Figure 3.1. Modeling the median process instead of the underlying true data generating process is based on industry standards that rely on quantiles, especially the median, of the latency measurements. The reason is that latency measurements are generated by thousands of different users at a very high frequency, making a distributional approach that relies on the median more manageable and interpretable since the load balancer can not use the latency measurements of a single user that needs to be routed, but instead only sees the whole distribution of latency measurements across all users at once. Another reason is that latency measurements have heavy-tailed distributions. The robustness of the median to outliers makes it appealing.

The choice of the sampling frequency Δ has important consequences on the aggregated series. Because of the very slow evolution of Network performance throughout the day, the measurements exhibit local stationarity, typically over periods of time of the order of 1 hour. In fact, we will see that over short periods of time, measurements can always be approximated as an i.i.d. sample. Hence for two sampling frequencies $\Delta_1 < \Delta_2 <$3600 seconds or 1 hour, the median estimates at time t will have the same expectation but Δ_1 induces estimates computed over larger periods of time including more measurements, resulting in estimates with fewer variance. As Δ increases, the median process exhibits strong structure both in conditional mean and variance that were mostly overshadowed by noise for smaller Δ. Users will typically stay on a webpage for a short amount of time: at most 5 minutes in the vast majority of cases [59], hence predicting the state of each *CDN* in the short term is essential for *load-balancing* purposes. This implies that large Δ are most likely irrelevant. But as Δ shrinks, the median process looses its structure and accurate predictions become intractable.

In this chapter, we provide insight into both mean and variance dynamics into the formation of the median latency process. The spectral density of the aggregated series reveals strong seasonal periods at 8,12 and 24h across all Δ. A Fourier decomposition of the series with $K = 3$ regressors is performed and captures all seasonal components in the mean dynamic. The number of Fourier regressors is selected through AIC minimization, and the estimated parameters of the mean model are equal across $30 \leq \Delta \leq 10800$s. The innovations from the Fourier decomposition present serial correlation and clustered volatility that is accurately described by an ARMA seasonal GARCH model, i.e. an ARMA-GARCH process with additional seasonal components in the GARCH part, fitted in one stage by quasi maximum likelihood estimation. We show that common baselines to

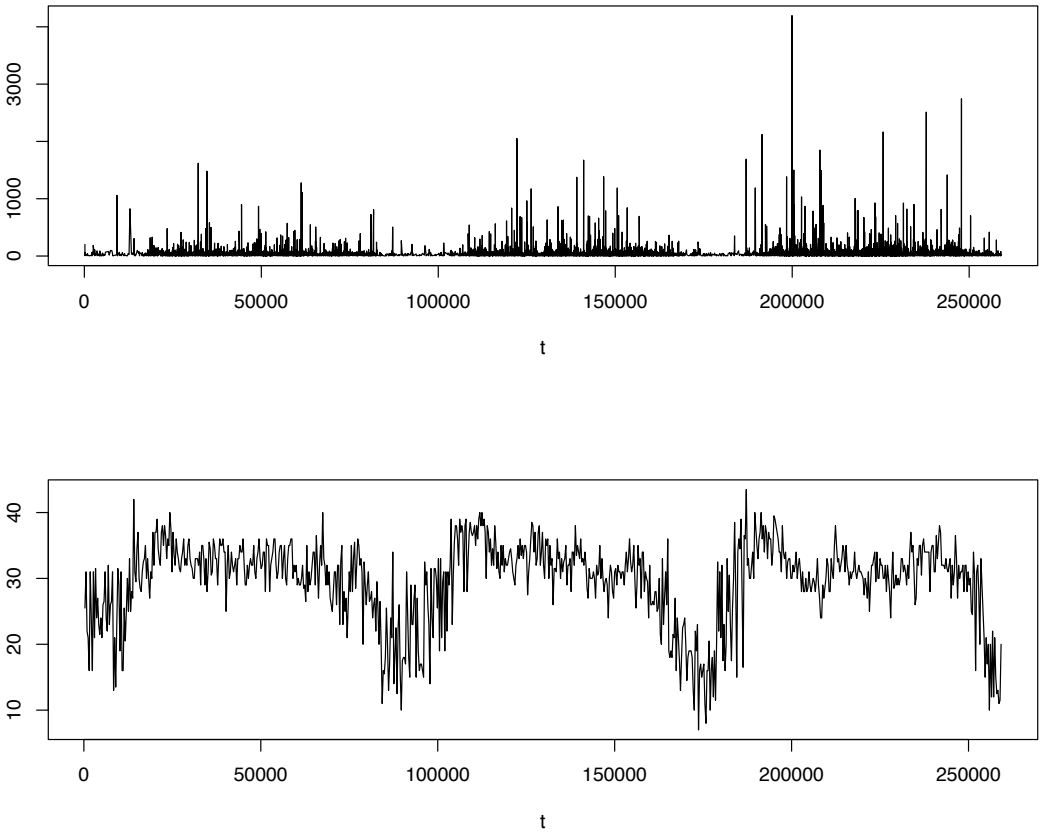

Figure 3.1: *3 days of raw measurements (top), and corresponding aggregating series* X_n^{Δ}*, for* $\Delta = 300$ *(down).*

latency prediction like the NAIVE and AVG predictions, defined in (3.10) and (3.11), that produce forecasts that are equal to the last observed value and average of the the last k observed values respectively are outperformed by our simple model uniformly over all tested Δ. Finally, through the use of certain entropic criterion, we introduce a measure of time series predictability that tests simultaneously for independence and identical distribution that we apply to the residuals of our model as a way to test for remaining information not captured by the model. Existing diagnostic tests and our new test align, and our model accurately describes latency median processes.

3.1.3 Organisation of the chapter

In Section 3.2, we describe the data set and structure. We build the mean and variance models for the aggregated series in Sections 3.3 and 3.4, and discuss point forecasts in Section 3.5. Finally, we propose a new test for independence and identical distribution in Section 3.6 based on a certain

entropic criterion.

3.2 Data analysis and modeling

3.2.1 Latency measurements.

The data presented in this chapter are from Google *CDN*, Paris, ISP Orange, from October 13 to 26, 2018. During that period 1.145.077 data points were collected.

Overall data.

Figure 3.2 and Table 3.1 present summary statistics of the data set.

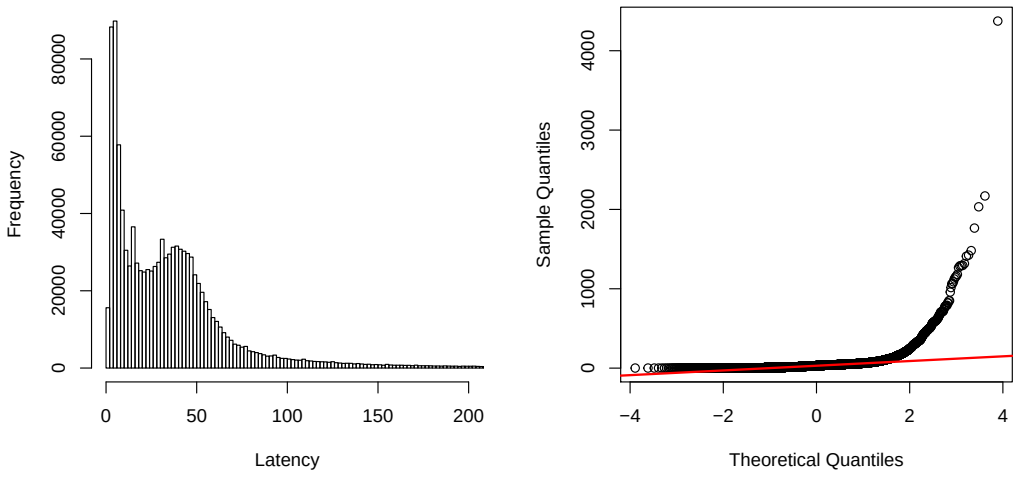

Figure 3.2: *Distribution of latency measurements over the 2 weeks period (left), normal Q-Q plot (right).*

Min	1st Qu.	Median	3rd Qu.	Max	Mean	SD	Kurtosis	Skweness
1	10	31	50	12236	47.4	103	538	15

Table 3.1: *Summary statistics of latency measurements.*

Log transforming the data reduces the variance and help stabilize it, hence we will work on log-latency measurements hereinafter, see [62] and [12]. Moreover, the log-transform reduces drastically the very high skewness of the original data and will make outliers more manageable for inference purposes, see Figure 3.3 and Table 3.2.

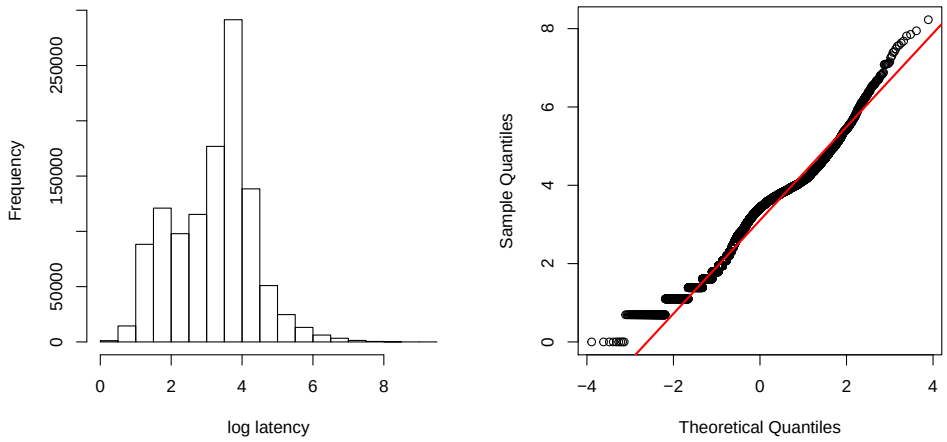

Figure 3.3: *Distribution of log-latency measurements over the 2 weeks period (left), normal Q-Q plot (right).*

Min	1st Qu.	Median	3rd Qu.	Max	Mean	SD	Kurtosis	Skweness
0	2.30	3.43	3.91	9.41	3.19	1.14	2.99	-0.0514

Table 3.2: *Summary statistics of log-latency measurements.*

Daily and hourly latency measurements.

Because of strong persistent daily seasonality in the behaviors of Internet users, the daily descriptive statistics are consistent from day to day. In particular, perhaps surprisingly, no critical differences are to be noted between weekdays and weekends. All summary statistics are consistent, and a day of the week effect, if present, is negligible, see Table 3.3.

Rather than the day of week, the most critical effect is the time of the day as seen in Table 3.4.

The dissection of the data at the hour level reveals the daily patterns. The latency attains its minimum at around 01:00 CET then continuously increases until 06:00 CET. The day time latency is fairly stable with a very slight downward trend up until 22:00 CET, then drops to attain its minimum at 01:00 CET the next day. The kurtosis and skewness evolution also have distinct patterns following the same timestamps. The kurtosis being strongly positively correlated with the median latency, while skewness is strongly negatively correlated with median latency. The evolution of median latency also correlates with the number of person online. This is related to the fact that the available bandwidth is limited, hence the bandwidth must be shared between users. The more people online, the lesser the fraction of bandwidth each user will get. Hence the minimum latency, i.e. best performance, coincide with the moment where less people are connected to the

Day	Min	1st Qu.	Median	3rd Qu.	Max	Mean	SD	Kurtosis	Skweness
13/10/18	0.000	2.197	3.466	3.912	8.198	3.171	1.150	2.905	-0.080
14/10/18	0.000	2.197	3.466	3.932	9.003	3.167	1.172	2.832	-0.037
15/10/18	0.000	2.398	3.434	3.932	8.342	3.214	1.135	3.012	-0.059
16/10/18	0.000	2.398	3.466	3.932	9.099	3.240	1.123	3.084	-0.061
17/10/18	0.000	2.303	3.434	3.912	8.337	3.204	1.143	3.020	-0.041
18/10/18	0.000	2.398	3.434	3.912	8.595	3.203	1.137	3.035	-0.055
19/10/18	0.000	2.398	3.434	3.932	8.284	3.214	1.133	3.011	-0.079
20/10/18	0.000	2.197	3.434	3.892	8.567	3.152	1.161	2.856	-0.016
21/10/18	0.000	2.079	3.401	3.912	8.383	3.123	1.176	2.778	-0.007
22/10/18	0.000	2.398	3.401	3.912	9.412	3.206	1.128	3.086	-0.027
23/10/18	0.000	2.485	3.466	3.932	8.706	3.246	1.091	3.161	-0.082
24/10/18	0.000	2.398	3.401	3.892	8.481	3.189	1.117	3.050	-0.056
25/10/18	0.000	2.303	3.367	3.892	8.256	3.150	1.134	2.977	-0.019
26/10/18	0.000	2.303	3.401	3.892	8.319	3.174	1.129	2.999	-0.062

Table 3.3: *Summary statistics of log-latency measurements aggregated per day.*

network, and vice and versa, explaining in part the day/night cycle of Internet performances[1].

3.2.2 Number of measurements.

The timestamps are irregularly spaced, and the number of measurements N_t received at time t is integer valued. The unconditional distribution of N_t is geometric with parameter 0.51 (χ^2 p-value = 0.23): 44% of timestamps contained no measurements, 30% contained 1 measurement, 26% more

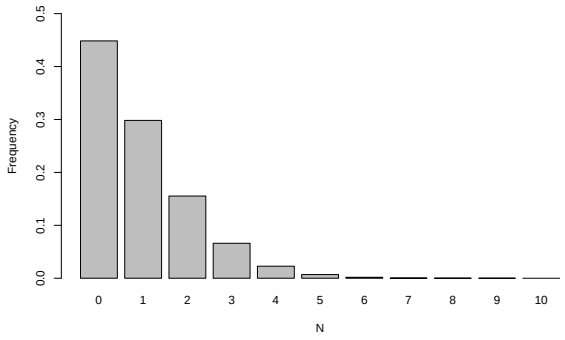

Figure 3.4: *Distribution of N_t*

[1]Some highly performing Networks with very high bandwidth are able to deliver constant bandwidth per user at all times. Those Networks are typically not *CDN*, but correspond to private infrastructures operated by companies like Google for instance. In this case, no seasonal effect is present and performance is stationary. We will focus on those Networks in the next Chapter.

Hour	Min	1st Qu.	Median	3rd Qu.	Max	Mean	SD	Kurtosis	Skewness
01:00 CET	0	1.792	2.944	3.714	7.987	2.839	1.185	2.720	0.247
02:00 CET	0	1.792	3.045	3.714	8.275	2.871	1.139	2.668	0.092
03:00 CET	0	1.946	3.091	3.767	7.858	2.912	1.148	2.738	0.066
04:00 CET	0	1.946	3.332	3.850	8.554	3.042	1.170	2.845	0.015
05:00 CET	0	2.303	3.497	3.912	8.076	3.198	1.146	3.008	-0.176
06:00 CET	0	2.565	3.555	3.951	9.099	3.286	1.117	3.227	-0.230
07:00 CET	0	2.708	3.526	3.951	8.420	3.322	1.093	3.394	-0.174
08:00 CET	0	2.565	3.466	3.932	8.567	3.259	1.114	3.212	-0.087
09:00 CET	0	2.485	3.401	3.912	8.107	3.226	1.101	3.114	-0.048
10:00 CET	0	2.485	3.434	3.912	8.556	3.231	1.111	3.133	-0.028
11:00 CET	0	2.485	3.466	3.932	8.263	3.256	1.106	3.183	-0.080
12:00 CET	0	2.565	3.497	3.932	8.284	3.274	1.101	3.204	-0.117
13:00 CET	0	2.398	3.401	3.912	8.481	3.202	1.112	3.058	-0.062
14:00 CET	0	2.303	3.401	3.912	8.595	3.188	1.130	2.977	-0.048
15:00 CET	0	2.398	3.401	3.912	8.383	3.197	1.122	3.026	-0.055
16:00 CET	0	2.303	3.401	3.892	9.412	3.161	1.124	2.913	-0.042
17:00 CET	0	2.303	3.434	3.932	8.319	3.192	1.149	2.925	-0.029
18:00 CET	0	2.197	3.401	3.912	8.225	3.150	1.176	2.831	-0.014
19:00 CET	0	2.197	3.466	3.932	8.131	3.202	1.170	2.872	-0.070
20:00 CET	0	2.197	3.434	3.932	8.195	3.182	1.171	2.875	-0.005
21:00 CET	0	2.079	3.401	3.912	8.706	3.128	1.178	2.781	0.028
22:00 CET	0	2.079	3.367	3.892	9.003	3.110	1.185	2.819	0.055
23:00 CET	0	1.946	3.178	3.850	8.331	2.985	1.204	2.765	0.148
00:00 CET	0	1.792	3.045	3.761	7.864	2.897	1.175	2.716	0.204

Table 3.4: *Summary statistics of log-latency measurements aggregated per hour.*

than 1 measurements, with an average value of 0.99, i.e. ≈ 1 measurement is received per second on average. Of course, due to nocturnal and diurnal cycles, N_t strongly depends on t. Table 3.5 presents the summary statistics of the distribution of the elapsed time in seconds between two consecutive latency measurements, and Figure 3.5 is a plot of the hourly number of measurements.

Minimum	25^{th} percentile	Median	Mean	75^{th} percentile	Maximum
1	1	1	1.8	2	91

Table 3.5: *Summary statistics of the distribution of the elapsed time in seconds between two consecutive timestamps in the sample.*

3.2.3 Underlying generating process and aggregated time series.

Before introducing the aggregated process, we propose a model for the data generating process. As

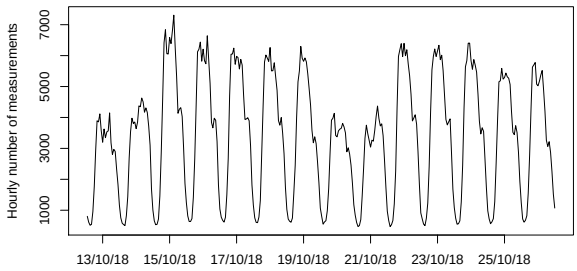

Figure 3.5: *Hourly number of measurements.*

seen above in Section 3.1.2, the timestamps of the measurements are precise down to the second, meaning that latency can not be sampled at arbitrary frequencies. If we let $\eta = 1$ second be the accuracy on the timestamps, it means that we only observe latency measurements at times t of the form $n\eta$ for some $n \in \mathbb{N}$. η is the highest frequency available. Moreover, multiple measurements can be received on the same timestamp. Figure 3.6 is a snapshot of an 8 seconds windows of measurements.

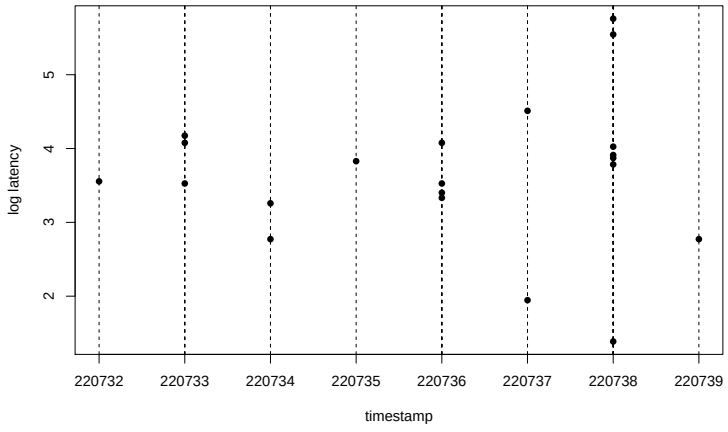

Figure 3.6: *8 seconds window of latency measurements.*

Those data can be seen as the empirical realization of a latent discrete time process taking values in a set of random probability distributions. Formally let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space and let $\mathcal{P}_+$ the set of probability distributions on $\mathbb{R}_+$. We equip $\mathcal{P}_+$ with the Wasserstein metric of order p where $p \geq 1$, denoted W_p [90]. Let $\eta > 0$, $K \in \mathbb{N}$, $T = K\eta$, and let $t_0 \leq t_1 \leq \ldots \leq t_K$ be the regular partition of the time interval $[0, T]$ such that for $i \in [\![0, K]\!] = \{0, \ldots, K\}$, $t_i = i\eta$.

Consider the following process:

$$P : \Omega \times [\![0, K]\!] \longrightarrow \mathcal{P}_+.$$

P is a stochastic discrete process taking values in $\mathcal{P}_+$. In particular, for any $i \in [\![0, K]\!]$, $P(\cdot, t_i)$ is a Markov kernel i.e. for all $\omega \in \Omega$, $P(\omega, t_i)(\cdot)$ is a probability measure on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ and for all $B \in \mathcal{B}(\mathbb{R}_+)$, $P(\cdot, t)(B)$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$. $P_{t_i} := P(\cdot, t_i)$ is the (random) probability distribution of latency at time t_i. One typically never have access to P, the underlying generating process of our data. Instead, at a given time t_i, one only observes an empirical measure associated with P_{t_i} with random number of points. Define:

$$N : \Omega \times [\![0, K]\!] \longrightarrow \mathbb{N}$$

the process that counts the number of measurements received at each timestamp t_i. What we observe is the empirical measure:

$$Z(t_i) = \frac{1}{N(t_i)} \sum_{k=1}^{N(t_i)} \delta_{X_k(t_i)}$$

where δ_x is the Dirac measure at $x \in \mathbb{R}$ and zero elsewhere, and for $(\omega, t_i) \in \Omega \times [\![0, K]\!]$, $(X_k(\omega, t_i))_{k \in \{1, \cdots, N(\omega, t_i)\}}$ are i.i.d. $P(\omega, t_i)$ random variables, i.e. :

$$X_1(\omega, t_i) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$$

such that for all $B \in \mathcal{B}(\mathbb{R}_+)$, $\mathbb{P}(X_1(\omega, t_i) \in B) = \mathbb{P}(\{\omega' \in \Omega | X_1(\omega, t_i)(\omega') \in B\}) = P(\omega, t_i)(B)$. For ease of reading, we will drop the ω dependency and simply write:

$$Z_i := \frac{1}{N_i} \sum_{k=1}^{N_i} \delta_{X_k(t_i)}$$

where $N_i := N(\cdot, t_i)$. Z_i (resp. N_i) is the empirical measure received at time t_i (resp. number of points in the sample) for $i \in \{0, \ldots, K\}$. Because it is possible that at time t_i no measurement is received, the following convention is used:

$$Z_i = 0 \quad \text{when } N_i = 0.$$

Because of the intrinsic structure of the data, and the high noise contamination, it is unreasonable to try to predict Z_t. The operational engineers reduce the problem by aggregating the measurements on non overlapping time intervals of length Δ and then compute the median. The resulting object is much more manageable in the sense that it is a regularly spaced time series on which classical forecast routines may be applied. The tradeoff is the following: empirically it has been observed that as Δ gets larger, the resulting time series exhibit a clear sine wave like pattern with decreasing variance in the noise. But large Δ are not appealing because at a given time t, a user needs ideally to be load balanced based on the real time state of the competing *CDN*, and not their state in 2h, which is in general irrelevant. We would like to choose the smallest Δ possible. Unfortunately, as Δ decreases towards η, the aggregated series gets heavily contaminated by noise, in the sense that the series stops exhibiting any distinct signal and starts behaving like white noise, hence prediction is irrelevant either. We are interested in determining the value Δ

that characterizes this dichotomy.

Let λ be some random probability measure on $\mathbb{R}_+$, i.e. $\lambda : \Omega \times \mathcal{B}(\mathbb{R}_+) \longrightarrow [0,1]$ such that $\lambda(\cdot, B)$ is a random variable and $\lambda(\omega, \cdot)$ is a probability measure. For $\omega \in \Omega$ define by $q^{\lambda}_{1/2}(\omega)$ the median of the probability distribution $\lambda(\omega, \cdot)$:

$$\lambda\left(\omega, (-\infty; q^{\lambda}_{1/2}(\omega)]\right) \geq \frac{1}{2},$$
$$\lambda\left(\omega, [q^{\lambda}_{1/2}(\omega); +\infty)\right) \geq \frac{1}{2}.$$

Now we can define recursively the median of a set of empirical probability measures. Let $(X_i)_{i\geq 1}$ be a collection of independent random variables with distribution μ, and define:

$$M_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

the associated empirical measure. Then $q^{M_n}_{1/2}(\omega)$ is the median of the discrete uniform distribution over the set $\{X_1(\omega), \ldots, X_n(\omega)\}$, and is usually referred to as the empirical median of the sample $(X_1(\omega), \ldots, X_n(\omega))$. Now consider a sequence $(P^i_{n_i})_{i=1,\ldots,N}$ of empirical measures, i.e. for all i, $P^i_{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{i,j}}$, where $(X_{i,j})_{j=1,\ldots,n_i}$ are i.i.d. $P^i_{n_i}$. Define the normalized sum of empirical measures as:

$$\bigvee_{i=1}^{N} P^i_{n_i} := \frac{\sum_{i=1}^{N} n_i P^i_{n_i}}{\sum_{i=1}^{N} n_i}.$$

For $N = 2$ we will simply write:

$$P^1_{n_1} \vee P^2_{n_2} := \bigvee_{i=1}^{2} P^i_{n_i} = \frac{\sum_{i=1}^{2} n_i P^i_{n_i}}{\sum_{i=1}^{2} n_i} = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1} \delta_{X_{1,i}} + \sum_{i=1}^{n_2} \delta_{X_{2,i}}\right).$$

$\bigvee_{i=1}^{N} P^i_{n_i}$ is the random probability measure of the sample $(X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{N,n_N})$ and then $q^{\bigvee_{i=1}^{N} P^i_{n_i}}_{1/2}$ is the corresponding median.

Let $\Delta(k) = k\eta \geq \eta$, $k \in \{1, \ldots, K\}$ be the sampling frequency of the aggregated series. For $k \in \{1, \ldots, K\}$ and $n \in \mathbb{N}$ such that there exists $i \in \{1, \ldots, k\}$, $Z_{(n-1)k+i} \neq 0$, define:

$$X^{\Delta(k)}_n = q^{\mathbb{G}^{\Delta}_n(k)}_{1/2}$$

where

$$\mathbb{G}^{\Delta}_n(k) = \bigvee_{i=1}^{k} Z_{(n-1)k+i}$$

$\mathbb{G}^{\Delta}_n(k)$ is the random probability measure associated with all measurements received at times t_i, $i \in E^{\Delta(k)}_n = \{i \in \mathbb{N} | (n-1)\Delta < i\eta \leq n\Delta\} = \{(n-1)k+1, \ldots, nk\}$. $E^{\Delta(k)}_n$ is the n-th integers run of length k, and $X^{\Delta(k)}_n$ is the then the median of all measurements falling in the interval $](n-1)\Delta, n\Delta]$. Observe that for $k = 1$, $\mathbb{G}^{\Delta(k)}_n = \mathbb{G}^{\eta}_n = Z_n$. Hence the aggregated series at the highest frequency

$X_n^{\Delta(1)} = X_n^{\eta}$ is obtained by computing the median of measurements with identical timestamps, with no time aggregation. In the case where $\mathbb{G}_n^{\Delta}(k) = 0$, $q_{1/2}^{\mathbb{G}_n^{\Delta}(k)}$ is not defined. In this case, let $\underline{n} = \inf\{m \in \mathbb{N} | \mathbb{G}_n^{\Delta}(k) \neq 0\}$, $\overline{n} = \sup\{m \in \mathbb{N} | \mathbb{G}_n^{\Delta}(k) \neq 0\}$, and let $X_n^{\Delta(k)}$ be the linear interpolation between $X_{\underline{n}}^{\Delta(k)}$ and $X_{\overline{n}}^{\Delta(k)}$, see Table 3.6.

$\Delta = 1 (= \eta)$	$\Delta = 30$	$\Delta = 60$	$\Delta = 120$
45%	0.5%	0.004%	0%

Table 3.6: *Percentage of missing values in the aggregated series for different* Δ.

3.2.4 Local stationarity

Visual inspection of the aggregated series reveal a clear seasonal pattern with decreasing variance as Δ increases, as seen in Figure 3.7.

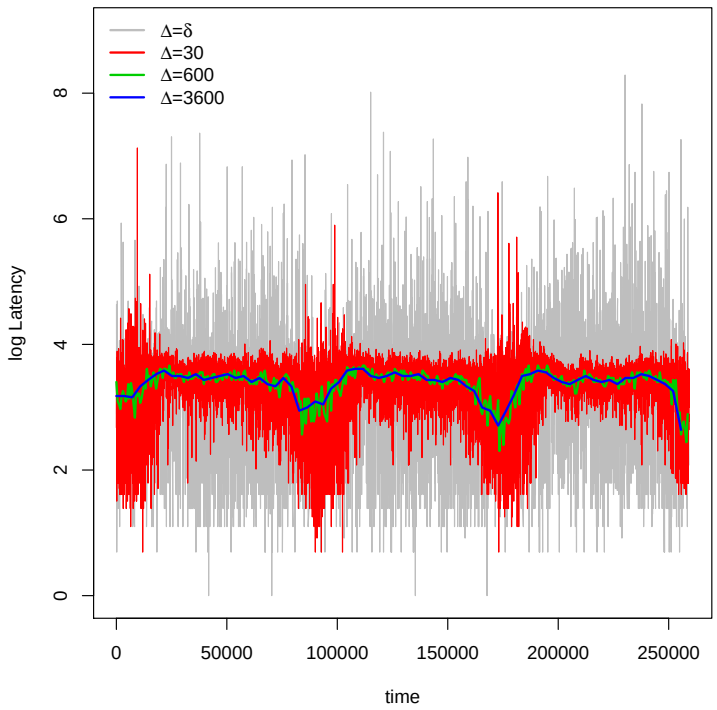

Figure 3.7: *3 days of measurements aggregated at different frequencies.*

For small Δ the data are heavily contaminated by noise and seem to loose structure. The

notion of local stationarity was first introduced by R. Dahlhaus in 1996 [25], see also [77] and [91]. Heuristically speaking, a stochastic process is called locally stationary if it behaves approximately has a stationary process locally in time. Classic routines to assess stationarity like the Augmented Dickey-Fuller (ADF) test, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and Ljung-Box test (see [30], [38], [55], [60]) performed on aggregated series with small Δ over periods of time of the order of 1h typically fail to reject stationarity. We briefly recall those tests before presenting the results.

Augmented Dickey-Fuller Test

The augmented Dickey-Fuller (ADF) is a unit root test for stationarity. The null hypothesis is the presence of a unit root in the series, the alternative is stationarity of the series. Given a series $(x_t)_t$ the test fits the following model:

$$\Delta x_t = \alpha + \beta t + \gamma x_{t-1} + \sum_{i=1}^{k} \delta_i \Delta x_{t-i} + \varepsilon_t$$

where k is the lag order to be used and $\alpha, \beta, \gamma, (\delta_i)_{1 \leq i \leq k}$ are the parameters of the model, and the ε_t's are error terms assumed to be independently and identically distributed with zero mean. The lag is typically chosen using the Akaike Information Criteria or AIC first introduced by Akaike in 1973 (see [2] and [79]) by minimizing:

$$AIC = -2 \log L + 2k$$

where L is the likelihood of the model. The test is performed under the null hypothesis that $\gamma = 0$ against $\gamma \neq 0$. The parameters are fitted by ordinary least squares and the test statistics $\hat{\gamma}/\mathrm{sd}(\hat{\gamma})$ is compared against the Dickey Fuller distribution. The heuristics is that if the series is integrated, adding a term x_{t-1} gives no additional information in predicting Δx_t.

KPSS test

This test assumes the following model for the data:

$$x_t = r_t + \varepsilon_t$$

with ε_t a stationary non integrated process, meaning that ε_t as an MA(∞) representation with square summable moving average coefficients, and $r_t = r_{t-1} + u_t$ where u_t is a Gaussian white noise with mean 0 and variance σ^2. The null hypothesis is stationarity and is reduced to testing $\sigma^2 = 0$, corresponding in this case to:

$$x_t = r_0 + \varepsilon_t.$$

A linear regression of x_t is then performed on a constant. The sum of squared residuals renoramalized by the Newey-West estimator of σ^2 [68] forms the test statistic and is compared to its limiting distribution.

Ljung-Box test

This test is used to test the null hypothesis that the data points are independently distributed by testing the significance of the autocorrelations. The test statistic is:

$$Q = n(n+2) \sum_{k=1}^{p} \frac{\hat{\rho}_k^2}{n-k}$$

where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k and p is the number of lag to be tested. Under the null hypothesis H_0 that no autocorrelation is present in the series up to lag p we have $Q \sim \chi^2(p)$. Simulation studies [88] suggest to take $p \approx \log(n)$.

Choosing $\Delta = 30$s as the tradeoff between high enough frequency and low percentage of missing values in the aggregated series, see Table 3.6, we compute the percentage of times ADF, KPSS, Ljung-Box reject the hypothesis of stationarity using Monte-Carlo methods by randomly selecting 1000 timestamps $(T_1, \dots, T_{1000})$ and performing the 3 tests on the sub-series $(X^{\Delta}_{t_n})_{t_n \in E_i}$ where E_i is the time interval $[T_i, T_i + L]$ where L is the length of the period tested. We choose four values for L, $L = 30$ minutes, $L = 1$ hour, $L = 2$ hours and $L = 4$ hours. The significance level is $\alpha = 0.05$. The results are summed up in Table 3.7.

	ADF	KPSS	Ljung-Box
$L = 30$ min	0.096	0.062	0.032
$L = 1$ hour	0.13	0.15	0.081
$L = 2$ hour	0.49	0.45	0.19
$L = 4$ hour	0.81	0.70	0.53

Table 3.7: *Percentage of times stationarity of the aggregated series $X^{\Delta}_{t_n}$ was rejected for $\Delta = 30$ over randomly selected time intervals of length 30 minutes, 1 hour, 2 hours and 4 hours.*

Locally in time the aggregated series at high frequencies behave like noise. Even over 2 hours periods, the tests failed to reject stationarity of the aggregated series for $\Delta = 30$ more than 50% of the time. Over periods of 30 minutes, the tests almost never reject stationarity. The ACF and PACF [13] functions can be used to assess the degree of serial correlations within a time series $(x_t)_t$. For a lag $h \in \mathbb{N}$, the ACF $\alpha(h)$ measures the correlation between x_t and x_{t+h}, i.e.:

$$\alpha(h) = \text{corr}(x_t, x_{t+h})$$

while the PACF $\rho(h)$ measures the *direct* correlation between x_t and x_{t+h}, i.e.:

$$\rho(h) = \widehat{\beta_h}$$

where $\widehat{\beta_h}$ is the ordinary least squares estimation of β_h in the model:

$$x_t = \beta_0 + \sum_{i=1}^{h} \beta_i x_{t+i} + e_t.$$

where e_t is a zero-mean white noise. In other words, ACF measures the correlation between x_t and x_{t+h} while PACF measures the correlation between x_t and x_{t+h} after removing the linear dependency between intermediate observations. Results show evidence that for small Δ the process $(X^{\Delta}_n)_n$ is locally stationary, see Figure 3.8.

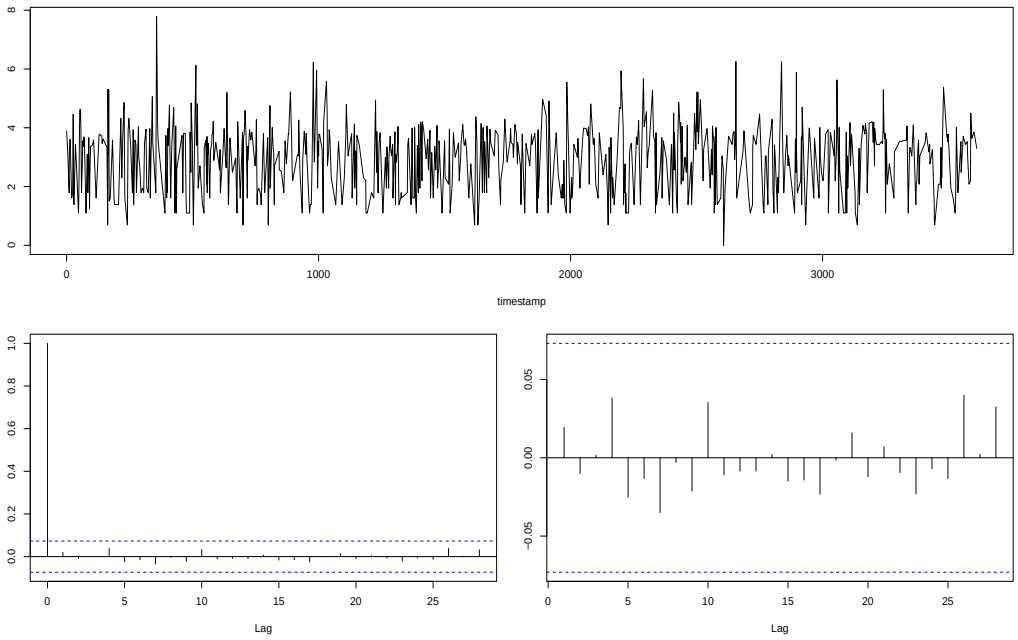

Figure 3.8: *Top: the process $X^{\Delta}_{t_n}$ over a 1 hour period, $\Delta = 5s$. Bottom left (resp right): ACF (resp PACF).*

3.3 Conditional mean model

3.3.1 Methodology

The aggregated series across Δ exhibit a seasonal pattern with a fundamental period of 24h confirmed by a spectral analysis. The periodogram [61] [80], or squared modulus of the Fourier transform, of the aggregated series reveals strong periodicity at multiple frequencies, namely 24, 12 and 8 hours, see Figure 3.9 and Table 3.8.
A natural way to model those seasonal time series is to perform a Fourier decomposition of the signal [54], i.e. we choose a model of the form:

$$X^{\Delta}_n = \mu^{\Delta} + \sum_{k=1}^{K} a^{\Delta}_k \sin\left(\frac{2k\pi t^{\Delta}_n}{\phi}\right) + \sum_{k=1}^{K} b^{\Delta}_k \cos\left(\frac{2k\pi t^{\Delta}_n}{\phi}\right) + \varepsilon^{\Delta}_n, \tag{3.1}$$

where:

- X^{Δ}_n is the n-th observation.
- $t^{\Delta}_n = (n - 1/2)\Delta$ is the corresponding timestamp.
- K is the number of Fourier regressors.
- ϕ is the fundamental period of the series, i.e. $\phi = 24\text{h} = 86400\text{s}$.

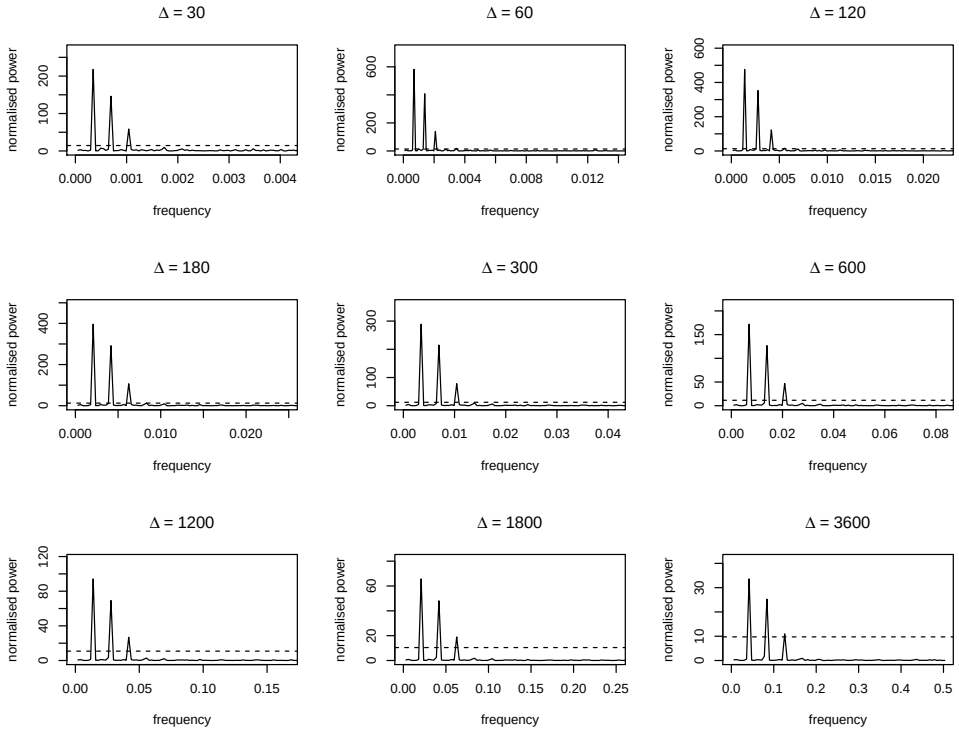

Figure 3.9: *Periodogramm of the aggregated series $X^{\Delta}_{t_n}$ for various Δ. All series exhibit three major peaks in the power spectrum corresponding respectively to 24h, 12h and 8h periods.*

- μ, $(\alpha_i)_{1\leq i\leq K}$ and $(\beta_i)_{1\leq i\leq K}$ are the coefficients of the model and are estimated by linear regression.
- ε is an innovation.

We shall use the notation

$$f^{\Delta}_n = \mu^{\Delta} + \sum_{k=1}^{K} a^{\Delta}_k \sin\left(\frac{2k\pi t^{\Delta}_n}{\phi}\right) + \sum_{k=1}^{K} b^{\Delta}_k \cos\left(\frac{2k\pi t^{\Delta}_n}{\phi}\right) \tag{3.2}$$

to denote the mean dynamic of our model. The number of Fourier regressors K is chosen through AIC minimization, see Table 3.9. $K = 4$ is the minimizer for all series with $\Delta \geq 120$ and $K = 7$ when $\Delta < 120$, but in all cases the AIC significantly drops from $K = 1$ to $K = 2$ and $K = 2$ to $K = 3$, and very little improvement is noticed after. This is in good alignment with the spectral analysis in Figure 3.9 and Table 3.8. The aggregated series $X^{\Delta}_{t_n}$ have a fundamental period of 24h, and three peaks in the power spectrum are present at periods 24h, 12h and 8h, suggesting indeed that the signal has three modes of vibration at frequencies $2k\pi/\phi$, $k = 1, 2, 3$. We choose $K = 3$ for all series.

	Significant periods (highest to lowest power)
$\Delta = 30$	23.98 - 11.99 - 8.00
$\Delta = 60$	23.97 - 11.99 - 7.99 - 16.79 - 4.81
$\Delta = 120$	24.00 - 12.00 - 7.98 - 5.99
$\Delta = 180$	23.99 - 11.99 - 7.99
$\Delta = 300$	23.98 - 12.01 - 7.99
$\Delta = 600$	23.97 - 11.98 - 7.99
$\Delta = 1200$	23.95 - 11.97 - 7.91
$\Delta = 1800$	23.92 - 11.96 - 7.97
$\Delta = 3600$	23.85 - 11.92 - 7.95

Table 3.8: *Significant periods in the spectrums of the aggregated series $X^{\Delta}_{t_n}$. All exhibit very strong power at the 24,12 and 8h periods.*

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$
$\Delta = 30$	26527.29	26441.66	26373.94	26370.37	26365.13	26367.01	26363.40	26366.96
$\Delta = 60$	6898.97	6821.88	6756.72	6747.88	6745.22	6748.13	6741.64	6744.28
$\Delta = 120$	367.72	294.84	229.13	214.98	215.04	218.43	213.56	216.93
$\Delta = 180$	-1164.37	-1230.15	-1290.72	-1301.78	-1300.91	-1297.47	-1300.12	-1296.70
$\Delta = 300$	-1624.17	-1688.21	-1747.09	-1759.76	-1757.72	-1754.29	-1755.89	-1752.45
$\Delta = 600$	-1374.07	-1435.27	-1491.46	-1502.15	-1499.86	-1496.33	-1496.86	-1493.36
$\Delta = 1200$	-803.58	-865.42	-926.26	-937.69	-936.01	-932.70	-934.82	-931.68
$\Delta = 1800$	-544.30	-605.35	-664.03	-673.79	-671.89	-668.22	-669.02	-667.00
$\Delta = 3600$	-264.11	-320.62	-379.26	-393.13	-389.98	-386.77	-388.91	-388.71

Table 3.9: *AIC of the Fourier model for varying Δ and K.*

The Fourier decomposition allows us to accurately model the evolution of day/night cycles and is in fact strongly related to the number of measurements received, or, equivalently, the number of people connected on the Network. Across all sampling frequencies, the average correlation between the number of measurements per time interval and the corresponding value of the Fourier decomposition is 0.8. This seasonal component encapsulates the information about attendance.

3.3.2 Results

Table 3.10 presents the important reduction of the autocorrelation in the aggregated after applying the Fourier decomposition.
Figure 3.10 confirms that the Fourier decomposition correctly removes all seasonal periods in the aggregated series. Despite that the Fourier decomposition captures the seasonality and autocorrelation in the aggregated series, the residuals' volatility is clustered as suggested by the autocorrelation in $|\varepsilon^{\Delta}|$, suggesting heteroscedasticity, see Table 3.10. The coefficients $(a_k^{\Delta})_k$, $(b_k^{\Delta})_k$ of equation (3.1) are computed via ordinary least squares. Since the residuals violate the homoscedasticity assumption of the classical linear model, we cannot compute standard confidence intervals. Using the fact that the residuals of the Fourier decomposition are uncorrelated, we may assume that they

	$r(1)$	$r(\phi/\Delta)$	$r(\phi/2\Delta)$	$r(\phi/3\Delta)$	Significance level
$\Delta = 30$					
X^Δ	0.090	0.160	-0.054	-0.058	
ε^Δ	0.038	0.009	0.006	0.004	(0.013)
$\lvert\varepsilon^\Delta\rvert$	0.298	0.261	-0.144	-0.109	
$\Delta = 60$					
X^Δ	0.315	0.256	-0.083	-0.085	
ε^Δ	0.051	0.020	-0.001	0.001	(0.020)
$\lvert\varepsilon^\Delta\rvert$	0.356	0.283	-0.156	-0.127	
$\Delta = 120$					
X^Δ	0.451	0.364	0.093	-0.121	
ε^Δ	0.106	0.029	0.018	-0.008	(0.026)
$\lvert\varepsilon^\Delta\rvert$	0.364	0.296	-0.156	-0.138	
$\Delta = 180$					
X^Δ	0.547	0.439	-0.114	-0.133	
ε^Δ	0.144	0.051	0.031	0.001	(0.032)
$\lvert\varepsilon^\Delta\rvert$	0.351	0.289	-0.158	-0.118	
$\Delta = 300$					
X^Δ	0.652	0.520	-0.126	-0.159	
ε^Δ	0.205	0.051	0.047	-0.005	(0.041)
$\lvert\varepsilon^\Delta\rvert$	0.358	0.289	-0.144	-0.133	
$\Delta = 600$					
X^Δ	0.772	0.604	-0.154	-0.188	
ε^Δ	0.323	0.080	0.055	-0.002	(0.058)
$\lvert\varepsilon^\Delta\rvert$	0.392	0.310	-0.157	-0.135	
$\Delta = 1200$					
X^Δ	0.816	0.691	-0.173	-0.212	
ε^Δ	0.332	0.128	0.070	0.003	(0.082)
$\lvert\varepsilon^\Delta\rvert$	0.463	0.285	-0.112	-0.114	
$\Delta = 1800$					
X^Δ	0.817	0.712	-0.175	-0.233	
ε^Δ	0.314	0.086	0.095	-0.010	(0.10)
$\lvert\varepsilon^\Delta\rvert$	0.480	0.252	-0.136	-0.112	
$\Delta = 3600$					
X^Δ	0.782	0.662	-0.184	-0.256	
ε^Δ	0.208	0.031	0.051	0.058	(0.141)
$\lvert\varepsilon^\Delta\rvert$	0.365	0.248	-0.121	-0.054	

Table 3.10: *Autocorrelations of the aggregated series and residuals after Fourier decomposition. $r(k)$ is the autocorrelation at lag k. Because $\eta = 1s$, $r(\phi/\Delta)$, $r(\phi/2\Delta)$ and $r(\phi/3\Delta)$ correspond to the autocorrelation at 24h, 12h and 8h, i.e. the one corresponding to significant peaks in the power spectrum.*

have diagonal variance Σ^Δ. Following on White's heteroscedesticity-consistent estimators [94], we start by rewriting our model for X_n^Δ in matrix form. We have:

$$\boldsymbol{X^\Delta} = \boldsymbol{F^\Delta}\theta^\Delta + \boldsymbol{\varepsilon^\Delta}$$

where $\boldsymbol{F^\Delta}$ is the Fourier design matrix, i.e. the matrix containing all Fourier regressors:

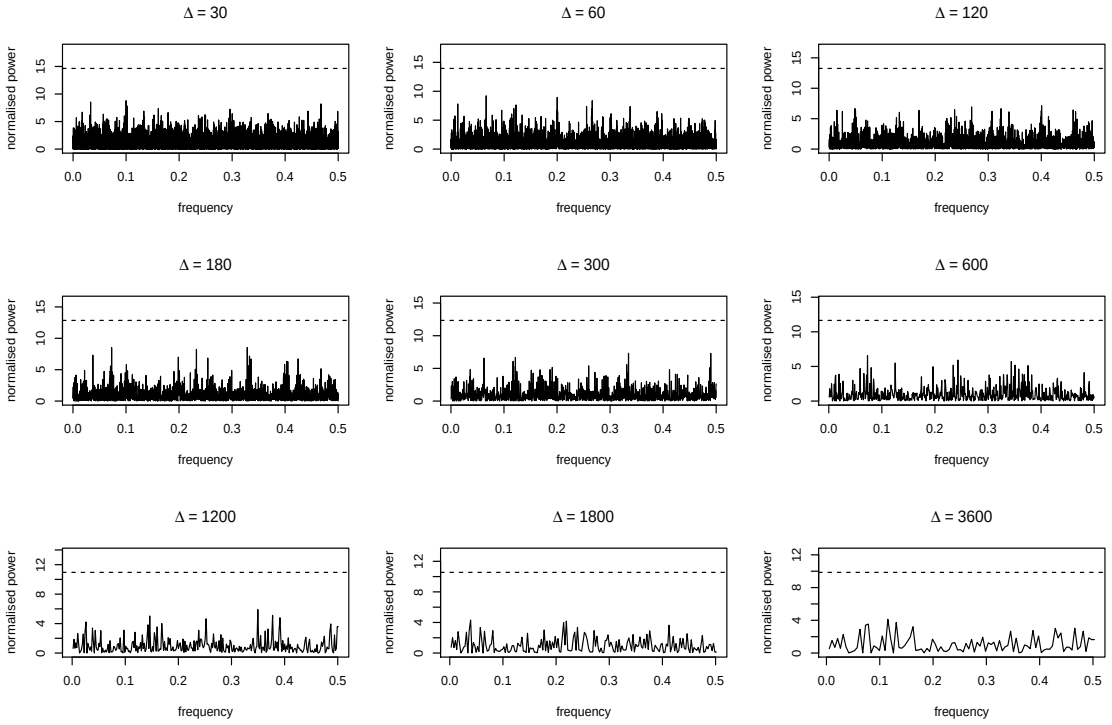

Figure 3.10: *Periodogramm of the residuals of the Fourier decomposition for various* Δ*. All significant periods have been removed.*

$$\boldsymbol{F^{\Delta}}_{i,j} = \begin{cases} \sin(\pi j t_i^{\Delta}/\phi) \text{ if } j \text{ is even} \\ \cos(\pi(j-1)t_i^{\Delta}/\phi) \text{ if } j \text{ is odd} \end{cases}$$

i.e:

$$\boldsymbol{F^{\Delta}} = \begin{bmatrix} 1 & \sin\left(\frac{2\pi t_1^{\Delta}}{\phi}\right) & \cos\left(\frac{2\pi t_1^{\Delta}}{\phi}\right) & \sin\left(\frac{4\pi t_1^{\Delta}}{\phi}\right) & \cos\left(\frac{4\pi t_1^{\Delta}}{\phi}\right) & \dots & \sin\left(\frac{2K\pi t_1^{\Delta}}{\phi}\right) & \cos\left(\frac{2K\pi t_1^{\Delta}}{\phi}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \sin\left(\frac{2\pi t_N^{\Delta}}{\phi}\right) & \cos\left(\frac{2\pi t_N^{\Delta}}{\phi}\right) & \sin\left(\frac{4\pi t_N^{\Delta}}{\phi}\right) & \cos\left(\frac{4\pi t_N^{\Delta}}{\phi}\right) & \dots & \sin\left(\frac{2K\pi t_N^{\Delta}}{\phi}\right) & \cos\left(\frac{2K\pi t_N^{\Delta}}{\phi}\right) \end{bmatrix},$$

$$\theta^{\Delta} = (\mu^{\Delta}, a_1^{\Delta}, b_1^{\Delta}, \dots, a_K^{\Delta}, b_K^{\Delta}),$$

and

$$\boldsymbol{\varepsilon^{\Delta}} = (\varepsilon_1^{\Delta}, \dots, \varepsilon_N^{\Delta}).$$

$\boldsymbol{F^{\Delta}}$ is an $N \times p$ matrix, where N is the number of rows and $p = 2K + 1$ is the number of columns. The analysis of the residuals suggests that $\mathbb{V}(\boldsymbol{\varepsilon^{\Delta}}) = \Sigma^{\Delta}$ where Σ^{Δ} is diagonal. The ordinary least

squares estimate for θ^{Δ} gives:

$$\hat{\theta}^{\Delta} = (\boldsymbol{F}^{\boldsymbol{\Delta}\prime}\boldsymbol{F}^{\boldsymbol{\Delta}})^{-1}\boldsymbol{F}^{\boldsymbol{\Delta}\prime}\boldsymbol{X}^{\boldsymbol{\Delta}}.$$

Assuming normality, we have a limiting distribution:

$$\sqrt{N}\Omega_N^{-1/2}(\hat{\theta}^{\Delta} - \theta^{\Delta}) \longrightarrow \mathcal{N}(0, I_{2K+1})$$

where

$$\Omega_N = (\boldsymbol{F}^{\boldsymbol{\Delta}\prime}\boldsymbol{F}^{\boldsymbol{\Delta}})^{-1}\boldsymbol{F}^{\boldsymbol{\Delta}\prime}\Sigma^{\Delta}\boldsymbol{F}^{\boldsymbol{\Delta}}(\boldsymbol{F}^{\boldsymbol{\Delta}\prime}\boldsymbol{F}^{\boldsymbol{\Delta}})^{-1}.$$

Σ^{Δ} is unknown but can be estimated from the residuals as:

$$\widehat{\Sigma^{\Delta}_{(i,i)}} = (X_i^{\Delta} - \boldsymbol{F}^{\boldsymbol{\Delta}\prime}{}_{(i,\cdot)}\hat{\theta}^{\Delta})^2.$$

Confidence bands for the parameters $\hat{\theta}^{\Delta}$ are then easily derived. We present below in Tables 3.11 and 3.12 and Figure 3.11 the summary of the fit for f^{Δ} defined in equation (3.2).

	μ^{Δ}	a_1^{Δ}	b_1^{Δ}	a_2^{Δ}
$\Delta = 30$	3.284 (3.274,3.294)	-0.002 (-0.013,0.008)	-0.175 (-0.187,-0.163)	-0.06 (-0.072,-0.049)
$\Delta = 60$	3.301 (3.289,3.313)	-0.006 (-0.019,0.007)	-0.178 (-0.193,-0.163)	-0.067 (-0.082,-0.053)
$\Delta = 120$	3.317 (3.302 ,3.331)	-0.004 (-0.021,0.012)	-0.165 (-0.184,-0.146)	-0.064 (-0.081,-0.046)
$\Delta = 180$	3.324 (3.307,3.340)	-0.004 (-0.022,0.014)	-0.158 (-0.179,-0.137)	-0.061 (-0.081,-0.041)
$\Delta = 300$	3.328 (3.309,3.348)	-0.001 (-0.023,0.02)	-0.155 (-0.18,-0.13)	-0.06 (-0.083,-0.036)
$\Delta = 600$	3.335 (3.311,3.359)	0.003 (-0.023,0.029)	-0.146 (-0.177,-0.116)	-0.052 (-0.081,-0.024)
$\Delta = 1200$	3.339 (3.308,3.369)	0.004 (-0.029,0.038)	-0.144 (-0.183,-0.105)	-0.051 (-0.087,-0.014)
$\Delta = 1800$	3.343 (3.308,3.377)	0.006 (-0.032,0.044)	-0.136 (-0.18,-0.093)	-0.046 (-0.087,-0.006)
$\Delta = 3600$	3.345 (3.299,3.391)	0.009 (-0.042,0.059)	-0.14 (-0.198 ,-0.081)	-0.043 (-0.097,0.012)

Table 3.11: *μ^{Δ}, a_1^{Δ} and b_1^{Δ} and a_2^{Δ} as a function of Δ and corresponding confidence intervals with level $\alpha = 0.05$.*

Except for b_1^{Δ}, and μ^{Δ}, $\Delta = 30, 60$, there is a strong stability in the estimates of the coefficients, suggesting that f^{Δ} is independent of Δ as seen in Figure 3.12.

3.3.3 Discussion

Perhaps surprisingly, for Δ ranging from 30 to 10800, the estimators are within the same margin of error, suggesting that the mean dynamic is independent of the choice of such Δ. The fact that all

	b_2^Δ	a_3^Δ	b_3^Δ	$\mathrm{sd}(\varepsilon^\Delta)$
$\Delta = 30$	-0.132 (-0.143,-0.12)	-0.044 (-0.056,-0.033)	-0.064 (-0.075,-0.053)	0.417
$\Delta = 60$	-0.141 (-0.155,-0.127)	-0.052 (-0.066,-0.038)	-0.072 (-0.086,-0.058)	0.318
$\Delta = 120$	-0.139 (-0.156,-0.122)	-0.048 (-0.065,-0.03)	-0.072 (-0.09,-0.055)	0.244
$\Delta = 180$	-0.137 (-0.156,-0.117)	-0.046 (-0.065,-0.026)	-0.075 (-0.095,-0.055)	0.207
$\Delta = 300$	-0.135 (-0.159,-0.112)	-0.047 (-0.07,-0.024)	-0.075 (-0.099,-0.052)	0.173
$\Delta = 600$	-0.133 (-0.162,-0.104)	-0.042 (-0.071,-0.013)	-0.075 (-0.104,-0.047)	0.130
$\Delta = 1200$	-0.133 (-0.17,-0.097)	-0.041 (-0.078,-0.005)	-0.076 (-0.113,-0.04)	0.107
$\Delta = 1800$	-0.129 (-0.171,-0.088)	-0.035 (-0.076,0.006)	-0.074 (-0.115,-0.033)	0.090
$\Delta = 3600$	-0.132 (-0.187,-0.077)	-0.036 (-0.09,0.019)	-0.078 (-0.132,-0.023)	0.080

Table 3.12: *b_2^Δ and a_3^Δ and b_3^Δ as a function of Δ, standard deviation of ε^Δ and corresponding confidence intervals with level $\alpha = 0.05$.*

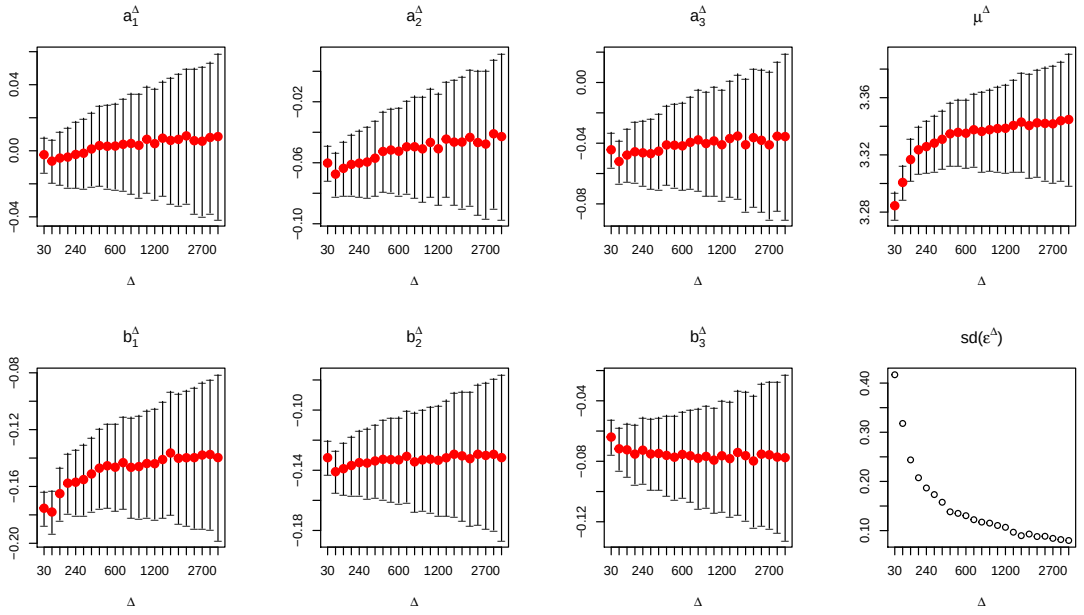

Figure 3.11: *Graphs of the evolution of μ^Δ, $(a_i^\Delta)_i$ and $(b_i^\Delta)_i$ as a function of Δ and corresponding confidence intervals with level $\alpha = 0.05$.*

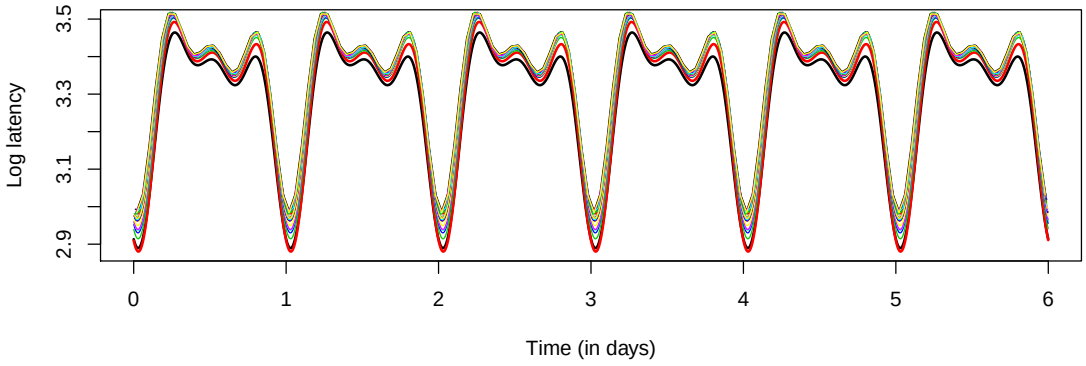

Figure 3.12: *Conditional mean model estimate for Δ=30 to 3600. The red and black thic lines are* $\Delta = 30, 60$.

mean dynamics across the tested Δ are essentially equal is a consequence of the very slow evolution of performance during the day. Because the measurements locally behave like noise, computing the median over entangled sub intervals leads similar estimates.

Of course this has a limit, for $\Delta = 86400$, i.e. 24h, the aggregated series would simply average measurements over whole days and would display a series with constant mean dynamic. The underlying dynamic does not need be constant for the different aggregated series to display the same pattern, it is enough to have monotonous trend. In this case again, computing median on entangled intervals would lead similar results, with higher variance as the interval shrinks. The only case where one would expect different behaviors between aggregated series would be when the mean dynamic attains a local extremum. In this case, provided Δ is large enough, aggregating the measurements should be expected to underestimate or overestimate the true value depending on whether the extremum is a maximum or minimum respectively.

As seen by the estimated mean dynamics in Figure 3.12, there are 6 equally distributed local extrema per day. Hence the measurements median changes direction every 3 hours on average. Because the median is particularly robust with a breakdown point of 0.5 (see [78]), even in those regimes where direction changes, results are expected to be consistent through scales provided Δ is small enough. It can be seen in figure 3.12 that indeed it is around local extremums that the main differences are displayed. As Δ increases, the mean dynamic increases in crest and decreases in peaks. The effect is marginal though.

3.3.4 Optimization of parameters computation

Assuming standard algorithms are used, the computational complexity of multiplying two matrices of size $n \times m$ and $m \times p$ is $\mathcal{O}(nmp)$, and the computational complexity of inverting a $n \times n$ matrix is $\mathcal{O}(n^3)$. Faster algorithms exist, but we will not discuss them here, see for instance [10]. Since

$\boldsymbol{F}^{\boldsymbol{\Delta}} \in \mathcal{M}_{n\times p}(\mathbb{R})$, the complexity of computing $\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}}$ is $\mathcal{O}(np^2)$ and inverting it is $\mathcal{O}(p^3)$. Since $n > p$, the overall complexity for computing $(\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}})^{-1}$ is $\mathcal{O}(np^2)$. This time can be cut down to $\mathcal{O}(p)$ by exploiting the specific structure of the Fourier design matrix.

Proposition 10. *Let $\Delta > \eta$ such that $\phi/\Delta \in \mathbb{N}^*$, N be the length of X^Δ such that $N = J\phi/\Delta$ where $J \in \mathbb{N}^*$ and $\boldsymbol{F}^{\boldsymbol{\Delta}}$ be the Fourier design matrix, i.e.:*

$$\boldsymbol{F}^{\boldsymbol{\Delta}} = \begin{bmatrix} 1 & \sin\left(\frac{2\pi t_1^\Delta}{\phi}\right) & \cos\left(\frac{2\pi t_1^\Delta}{\phi}\right) & \sin\left(\frac{4\pi t_1^\Delta}{\phi}\right) & \cos\left(\frac{4\pi t_1^\Delta}{\phi}\right) & \dots & \sin\left(\frac{2K\pi t_1^\Delta}{\phi}\right) & \cos\left(\frac{2K\pi t_1^\Delta}{\phi}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \sin\left(\frac{2\pi t_N^\Delta}{\phi}\right) & \cos\left(\frac{2\pi t_N^\Delta}{\phi}\right) & \sin\left(\frac{4\pi t_N^\Delta}{\phi}\right) & \cos\left(\frac{4\pi t_N^\Delta}{\phi}\right) & \dots & \sin\left(\frac{2K\pi t_N^\Delta}{\phi}\right) & \cos\left(\frac{2K\pi t_N^\Delta}{\phi}\right) \end{bmatrix}$$

where $t_n^\Delta = \frac{2n-1}{2}\Delta$ is the timestamp of the n-th element of X_n^Δ. Then:

$$\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}} = N \begin{bmatrix} 1 & & & \\ & 1/2 & & \\ & & \ddots & \\ & & & 1/2 \end{bmatrix}$$

Proof. We recall the following technical result on sums of sine and cosine. Let $x, y \in R$. Then:

$$S_1 := \sum_{n=1}^{N} \cos(x + ny) = \cos\left(x + (N+1)y/2\right) \frac{\sin\left(Ny/2\right)}{\sin\left(y/2\right)} \tag{3.3}$$

and:

$$S_2 := \sum_{n=1}^{N} \sin(x + ny) = \sin\left(x + (N+1)y/2\right) \frac{\sin\left(Ny/2\right)}{\sin\left(y/2\right)}. \tag{3.4}$$

Now for $1 \le i \le p$, let $\boldsymbol{F}^{\boldsymbol{\Delta}\,i}$ be the i-th column of $\boldsymbol{F}^{\boldsymbol{\Delta}}$. Then the (i,j) entry of $\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}}$ is:

$$(\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}})_{(i,j)} = \langle\, \boldsymbol{F}^{\boldsymbol{\Delta}\,i}, \boldsymbol{F}^{\boldsymbol{\Delta}\,j} \rangle$$

It is obvious from the definition of $\boldsymbol{F}^{\boldsymbol{\Delta}}$ that $\boldsymbol{F}^{\boldsymbol{\Delta}}{}_{(1,1)} = N$. Now let $k = 2q$, $q \in \mathbb{N}^*$.

$$\begin{aligned} (\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}})_{(k,k)} &= \left\| \boldsymbol{F}^{\boldsymbol{\Delta}\,k} \right\|^2 \\ &= \sum_{n=1}^{N} \sin\left(\frac{k\pi t_n^\Delta}{\phi}\right)^2 \\ &= \sum_{n=1}^{N} \sin\left(\frac{\Delta}{\phi} q\pi(2n-1)\right)^2 \end{aligned}$$

Let $\tau = \phi/\Delta$ then $x \mapsto \sin\left(\frac{\Delta}{\phi} q\pi(2x-1)\right)^2$ is τ-periodic. By assumption, τ is an integer, hence:

$$\sum_{n=1}^{N} \sin\left(\frac{q\pi}{\tau}(2n-1)\right)^2 = J \sum_{n=1}^{\tau} \sin\left(\frac{q\pi}{\tau}(2n-1)\right)^2$$

$$= \frac{J}{2}\left(\tau - \sum_{n=1}^{\tau} \cos\left(\frac{2q\pi}{\tau}(2n-1)\right)\right)$$
$$= \frac{J}{2}\left(\tau - \sum_{n=1}^{\tau} \cos\left(x + ny\right)\right)$$

where $x = -2q\pi/\tau$, $y = 4q\pi/\tau$. Using (3.3) it follows:

$$\sum_{n=1}^{N} \sin\left(\frac{q\pi}{\tau}(2n-1)\right)^2 = \frac{J}{2}\left(\tau - \cos\left(x + \frac{(\tau+1)y}{2}\right)\frac{\sin(\tau y/2)}{\sin(\tau/2)}\right).$$

But $\tau y/2 = 2q\pi$, and since $q \in \mathbb{N}$, we finally have:

$$(\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}})_{(k,k)} = \frac{J\tau}{2} = \frac{N}{2}$$

The proof for odd k is identical since $\cos(x)^2$, like $\sin(x)^2$, can also be linearized in terms of $\cos(2x)$. Now for the off diagonal elements, observe that if either $i = 1$ or $j = 1$ the result follows directly from the (3.3) and (3.4). Now, if $i, j > 1$ and $i \neq j$, then:

$$(\boldsymbol{F}^{\boldsymbol{\Delta}'}\boldsymbol{F}^{\boldsymbol{\Delta}})_{(i,j)} = \langle\, \boldsymbol{F}^{\boldsymbol{\Delta}^i}, \boldsymbol{F}^{\boldsymbol{\Delta}^j}\rangle$$

will involve sums with general term of the forms:

- $\sin\left(\frac{k\pi}{\tau}(2n-1)\right)\cos\left(\frac{l\pi}{\tau}(2n-1)\right)$
- $\sin\left(\frac{k\pi}{\tau}(2n-1)\right)\sin\left(\frac{l\pi}{\tau}(2n-1)\right)$
- $\cos\left(\frac{k\pi}{\tau}(2n-1)\right)\cos\left(\frac{l\pi}{\tau}(2n-1)\right)$

where $k, l \in \mathbb{N}$ and $k \neq l$, and are potentially different from line to line. Using the product-to-sum trigonometric identities and (3.3) and (3.4) concludes the proof. □

3.4 Innovation

In analogy with financial data, supported by the reminiscent microstructure effects evoked in Section 3.1.1, we shall use the term *volatility* to denote the innovations' variability. The reason to model the volatility of the innovations is mainly to decrease computational complexity of the prediction algorithm without impacting prediction accuracy. As seen in sections 3.1.2 and 3.3, the innovations of the Fourier regression display seasonal clustered volatility. Being able to predict volatility of the measurements may help anticipate periods of low predictability of the series.

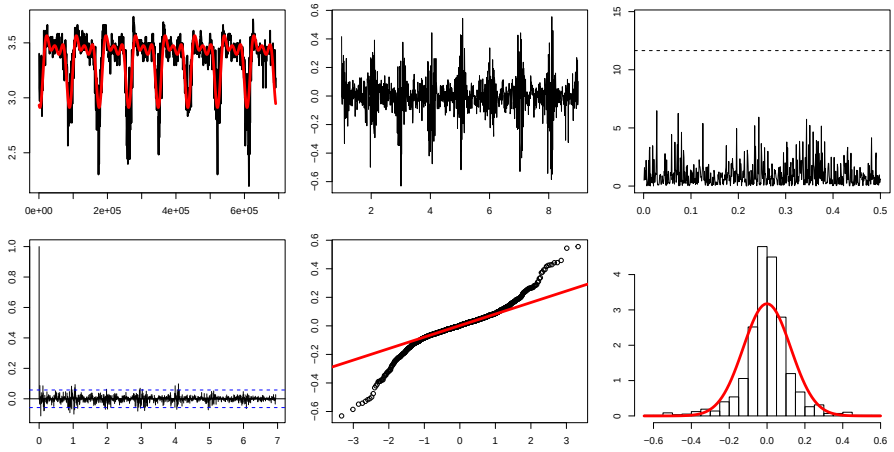

Figure 3.13: *Residuals analysis of the model fitted on 8 days, $\Delta = 600$. Left to right, top to bottom: aggregated series and fitted values; residuals; power spectrum; acf of the residuals; normal Q-Q plot of the residuals; histogram of the residuals.*

3.4.1 Naive GARCH

The residuals ε^{Δ} are uncorrelated and the Fourier decomposition removes seasonality of the series, but exhibit strong autocorrelation in the second moment because of clustered volatility.

Figure 3.13 shows an example of fit of the model for $\Delta = 600$s. The seasonality is corrected by the Fourier decomposition and the autocorrelation is contained into the confidence bands. But clearly, the variance is clustered. The distribution of the residuals is locally Gaussian around the mean, but with heavier tails. In order to account for those effects, we propose to first fit a GARCH model on the residuals. GARCH models were introduced by Bollerslev [11], following on the work of Engle [32], to describe the variable volatility of certain time series. Let $p, q \in \mathbb{N}$, we say that a time series ε_t is GARCH(p, q) if there exists positive coefficients $\omega, (\alpha_i)_{1\leq i\leq q}, (\beta_i)_{1\leq i\leq p}$ such that

$$\begin{aligned}\varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2\end{aligned}$$

where the random variables z_t are independent, identically distributed, with mean 0 and variance 1. Those models have been intensively used to describe real world time series, like financials series, see [49] for instance.

We fit a GARCH model on the residuals of the Fourrier decomposition with skewed Student's t distribution for the error terms. The skewed Student's t distribution [35] is a family of distributions that aim at accommodating the skewness and high kurtosis often present in real world data [35]. The skewed Student's t distribution has probability density function:

$$f(x; m, \tau, \nu, \xi) = \frac{2s\sigma}{\tau(\xi + \xi^{-1})} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\pi\nu}} \left(1 + \frac{s^2}{\nu} \frac{(\frac{x-m}{\tau}\sigma + \mu)^2}{\xi^{2\text{sign}(\frac{x-m}{\tau}\sigma+\mu)}}\right)^{-\frac{\nu+1}{2}},$$

with

$$
\begin{aligned}
m_1 &= \frac{2\sqrt{\nu-2}}{(\nu-1)B(\frac{1}{2},\frac{\nu}{2})},\\
\mu &= m_1(\xi-\xi^{-1}),\\
\sigma &= \sqrt{(1-m_1^2)(\xi^2+\xi^{-2})+2m_1^2-1},\\
s &= \sqrt{\frac{\nu}{\nu-2}},
\end{aligned}
$$

where

$$
B(a,b)=\int_0^1 x^{a-1}(1-x)^{b-1}dx, \quad a,b>0
$$

is the Beta function and

$$
\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt, \quad x>0
$$

is the Gamma function. The parameters $m \in \mathbb{R}$ and $\tau > 0$ are the mean and standard deviation of the distribution, while the parameters $\nu > 2$ and $\xi > 0$ are the shape and skewness parameters respectively. The shape parameter ν controls the heaviness of the tails, whereas the skew parameter controls if the distribution is left or right skewed: for $\xi < 1$ the distribution is left skewed, for $\xi > 1$ the distribution is right skewed and for $\xi = 1$ the distribution is symmetric. The skewed Student's t distribution embeds several well known distributions like the Student's t, Normal or Laplace distributions, among many others. For instance, the Gaussian distribution is obtained by setting $\xi = 1$ and letting $\nu \to \infty$. Indeed, it is easily verified that:

$$
\lim_{\nu\to\infty} f(x;m,\tau,\nu,1)=\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{1}{2}\frac{(x-m)^2}{\tau^2}\right).
$$

The high flexibility of the distribution allows for excellent fit of various real world data, especially financial data.

The diurnal and nocturnal cycles of Internet activities are responsible for the sine-cosine like patterns in the aggregated series: people connect more on the Internet during day time than during night time. The more people on the Network, the higher the latency because of finite bandwith, hence we observe this daily cycle where latency increases rapidily at dawn, slowly decreases during the day and rapidily decreases around midnight. In addition to this pattern, we also observe cycles in the volatility of the measurements. There are potentially two competing effects here. Possibly, as less people access the Network the latency decreases and the volatility increases because there are fewer measurements. But this only marginally explains the phenomenon as those daily cycles in volatility are still present when we under-sample every time interval so that each point in the aggregated series is built with the same number of measurements. Which is precisely the reason why the aggregation is performed using the median, because of its robustness. Indeed, the sample median has the highest breakdown value possible of 0.5, when the sample mean as a breakdown value of 0 [78]. Intuitively, the breakdown value is the proportion of outliers in a sample the estimator can handle before behaving abnormally.

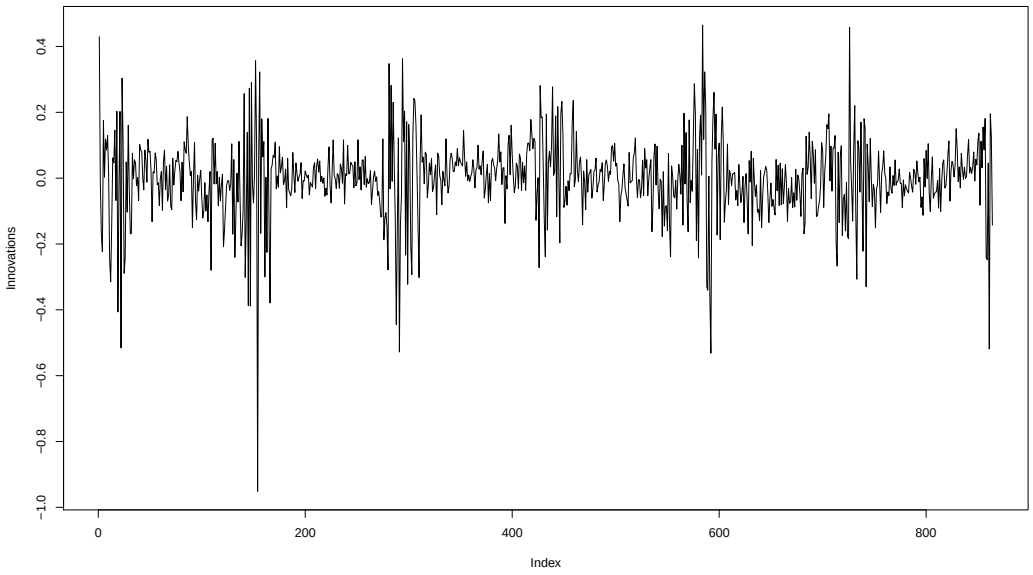

Figure 3.14: *Innovations of the Fourier decomposition (errors) on a test set of 6 days.* $\Delta = 600$.

The volatility of the measurements appear to be clustered. One way to model clustered volatility is to fit an ARMA-GARCH model on the innovations of our Fourier decomposition.
Figure 3.14 clearly reveals the clustered volatility. Understanding how this volatility can be modeled will give us insight on the relevant Δ to choose. Using AIC and BIC for model selection, ARMA(1,1)-GARCH(1,1) was selected for all Δ, see examples of selection in Figures 3.15 and 3.16.
The GARCH model appears to be appropriate. The weighted Ljung-Box Test on standardized and squared standardized residuals, used to evaluate the dependence of the first and second moments in the residuals with a time lag are not significant (p >5%). Sign bias tests [33], used to verify whether previous positive and negative shocks have a different impact on heteroscedasticity, are not significant either (p >5%), suggesting good specification of the model and no asymmetric effects. Nyblom tests [70] are not significant for all parameters (p >5%), suggesting that the parameters of the model are constant across time. Finally, the Chi-squared goodness of fit test is also not significant, suggesting that the residuals follow the target distribution. The model accurately captures the structure of the data. Our model for X_n^Δ now has the form:

$$
\begin{aligned}
X_n^\Delta &= \mu^\Delta + \sum_{k=1}^{K} \alpha_k \sin\left(\frac{2k\pi t_n^\Delta}{\phi}\right) + \sum_{k=1}^{K} \beta_k \cos\left(\frac{2k\pi t_n^\Delta}{\phi}\right) + \varepsilon_n^\Delta \\
\varepsilon_n^\Delta &= \nu^\Delta + \kappa^\Delta \varepsilon_{n-1} + u_n^\Delta + \rho^\Delta u_{n-1} \\
u_n^\Delta &= \sigma_n^\Delta z_n \\
{\sigma_n^\Delta}^2 &= \omega^\Delta + \lambda^\Delta {u_{n-1}^\Delta}^2 + \gamma^\Delta {\sigma_{n-1}^\Delta}^2
\end{aligned}
\tag{3.5}
$$

where ε_n^Δ is the innovation of the Fourier decomposition defined in Section 3.3, ν^Δ is the mean level

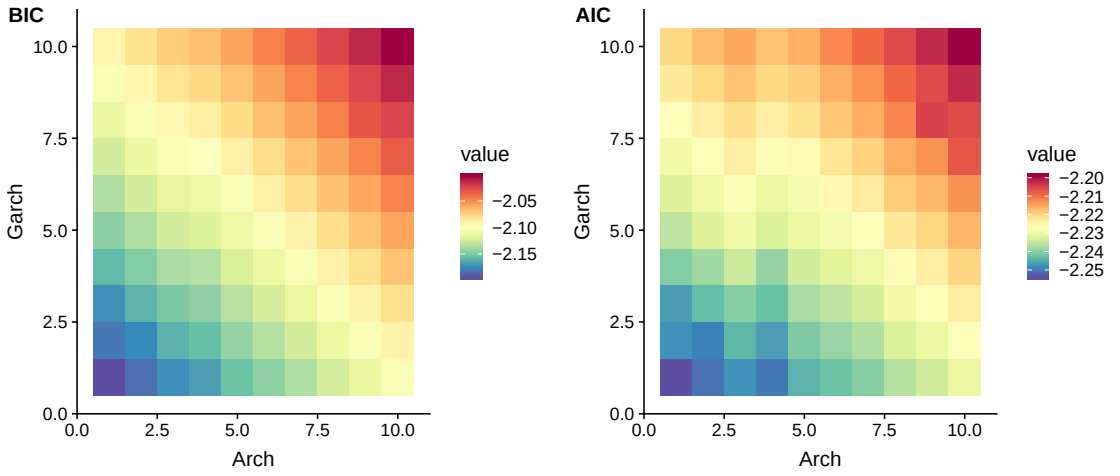

Figure 3.15: *Heat map of the model selection using AIC and BIC for* $\Delta = 600$*. Ranging from 1 to 10, x-axis is the number of ARCH terms, y-axis is the number of GARCH terms.*

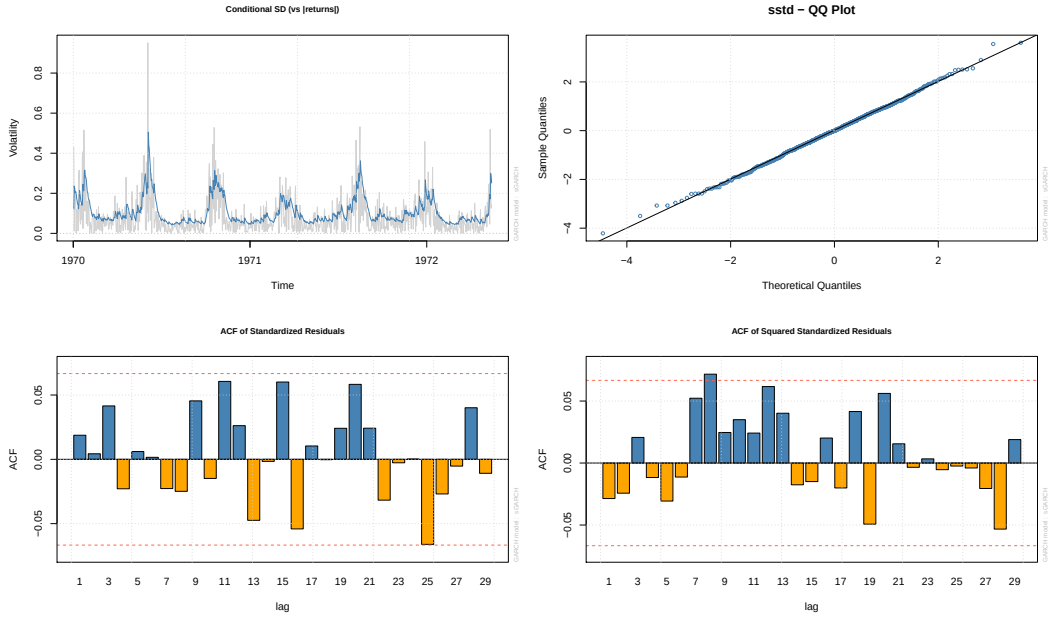

Figure 3.16: *Garch model of the innovations of the errors in the Fourier decomposition,* $\Delta = 600$*.*

of the innovations, κ^Δ is the autoregressive term, ρ^Δ is the moving average term, the parameters of the GARCH part satisfy $\omega^\Delta > 0$, $\lambda^\Delta > 0$ and $\gamma^\Delta > 0$ and the random variables z_n have zero mean and variance 1.

It is worse noting that as $\Delta \to \infty$ the conditional SD loses its structure and amplitude, and flattens, see Figure 3.17. It can be linked to the number of measurements per windows. The

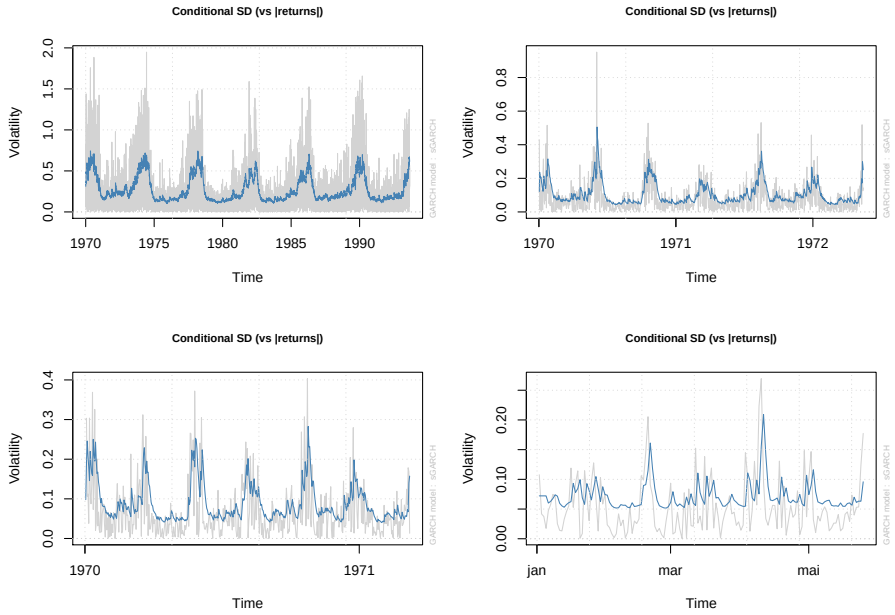

Figure 3.17: *Example of 4 conditional SD. Left to right, top to bottom:* $\Delta = 1$ *min,* $\Delta = 10$ *min,* $\Delta = 20$ *min,* $\Delta = 60$ *min.*

number of measurement does play a role in the clustered variance, especially in high frequencies, but this is still persistent with equal subsampling, even in large frequencies.

Diagnostic tests

The diagnostic tests rely on Ljjung Box tests for serial correlation in the residuals and squard residuals, and Anderson Darling (see [4]), Cramer Von Mises (see [92], [22]) and Kolmogorov Smirnov (see [65]) tests for goodness of fit. The target distribution is skewed Student's t and depends on the two unknown parameters of shape and skewness. We propose a methodology to test for relevant values. We fit our model in (3.5) on a training set of past history from the same *CDN* and extract the estimated shape and skewness parameters for varying Δ. We can observe a strong regularity in the evolution of those parameters as a function of Δ, see Figure 3.18. A logistic function is fitted to the skewness and a linear function is fitted to the shape, suggesting a model for the shape and skewness parameters, denoted respectively by $\mathrm{Sh}(\Delta)$ and $\mathrm{Sk}(\Delta)$:

$$\mathrm{Sk}(\Delta) = \frac{1}{1 + e^{-0.07\sqrt{\Delta}}} \tag{3.6}$$

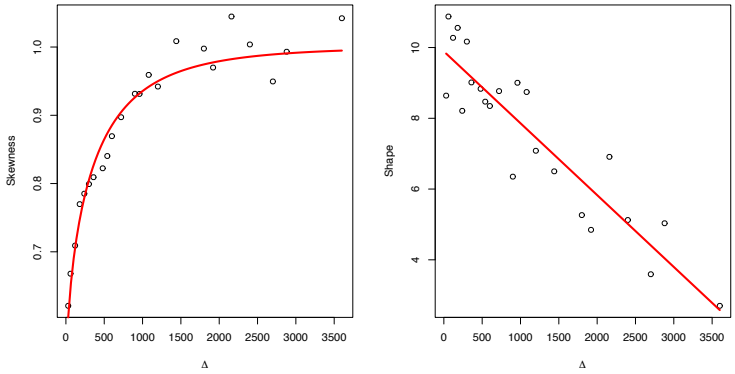

Figure 3.18: *Evolution of the skewness and shape parameters on a training set for* $\Delta = 30$ *to* 3600 *and fit using a logistic (right) and linear (left) function.*

and

$$\mathrm{Sh}(\Delta) = 9.9 - 0.002\Delta \tag{3.7}$$

The goodness of fit for this model on the training set as measured with the coefficient of determination leads $R^2 = 0.89$ for the skewness and 0.84 for the shape. The Skewed student distribution with mean 0, standard deviation 1, Skewness $\mathrm{Sk}(\Delta)$ and Shape $\mathrm{Sh}(\Delta)$ as defined in equations (3.6) and (3.7) will be used as the target distribution in the null hypothesis for the goodness of fit of our model at scale Δ. Diagnostics are presented in Tables 3.13, 3.14 and 3.15.

The skewness of the standardized residuals converges to 1 as Δ increases, indicating that the distribution of the residuals gets more symmetric. The shape on the other hand has a slight downwards trend, indicating that the tails of the innovations get lighter as Δ increases. The goodness of fit tests indicates very good fit to the theoretical skewed student distribution, with an average of 4.35% rejection rate at level $\alpha = 5\%$. The Ljung box tests also reveal an average of 4.35% rejection rate at lag 1 and $\log(n)$ for the standardized residuals, but jumps to 20% rejection rate for the squared residuals, suggesting that the Naive Garch failed to captured the patterns in the second moment of the residuals. The spectral analysis revealed periodicity in the squared volatility, but GARCH processes can not produce seasonal patterns, suggesting that GARCH model might not be optimal to fit those data. A way around this problem is to force deterministic seasonal lags in the volatility.

3.4.2 Periodic GARCH

It is clear from the the previous section that the volatility of the innovations is intimately related to the seasonality of the series, which means that the volatility is periodic, which was not taken into account into the previous models. The estimated conditional volatility of the innovations are clearly periodic, but GARCH(1,1) processes are not periodic. The conditional volatility is estimated through the optimization of a quasi-likelihood that embeds the squared residuals of the model. Because those are seasonal, they contaminate the estimators and produce this seasonal pattern in the conditional volatility as seen in Figure 3.17 for instance. Indeed, the estimated

	Lag = 1	Lag = log(n)
$\Delta = 30$	0.138	0.245
$\Delta = 60$	0.531	0.151
$\Delta = 120$	0.149	0.084
$\Delta = 180$	0.836	0.289
$\Delta = 240$	0.544	0.237
$\Delta = 300$	0.866	0.252
$\Delta = 360$	0.419	0.427
$\Delta = 480$	0.492	0.354
$\Delta = 540$	0.609	0.812
$\Delta = 600$	0.335	0.651
$\Delta = 720$	0.47	0.862
$\Delta = 900$	0.659	0.552
$\Delta = 960$	0.481	0.416
$\Delta = 1080$	0.738	0.546
$\Delta = 1200$	0.529	0.85
$\Delta = 1440$	0.223	0.536
$\Delta = 1800$	0.456	0.593
$\Delta = 1920$	0.561	0.448
$\Delta = 2160$	0.62	0.058
$\Delta = 2400$	0.5	0.196
$\Delta = 2700$	0.715	0.072
$\Delta = 2880$	0.479	0.165
$\Delta = 3600$	0.02	0.04

Table 3.13: *Standardized residuals Ljung-Box tests p-values.*

	Lag = 1	Lag = log(n)
$\Delta = 30$	0.104	0.175
$\Delta = 60$	0.022	0.036
$\Delta = 120$	0.116	0.286
$\Delta = 180$	0.03	0.25
$\Delta = 240$	0.077	0.09
$\Delta = 300$	0.06	0.199
$\Delta = 360$	0.033	0.455
$\Delta = 480$	0.676	0.612
$\Delta = 540$	0.639	0.727
$\Delta = 600$	0.276	0.659
$\Delta = 720$	0.247	0.441
$\Delta = 900$	0.699	0.912
$\Delta = 960$	0.96	0.823
$\Delta = 1080$	0.434	0.234
$\Delta = 1200$	0.956	0.986
$\Delta = 1440$	0.681	0.757
$\Delta = 1800$	0.92	0.737
$\Delta = 1920$	0.958	0.556
$\Delta = 2160$	0.409	0.912
$\Delta = 2400$	0.41	0.734
$\Delta = 2700$	0.588	0.918
$\Delta = 2880$	0	0
$\Delta = 3600$	0	0

Table 3.14: *Squared standardized residuals Ljung-Box tests p-values.*

conditional volatility $\hat{\sigma}^2$ is fitted according to the equation:

$$\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha} u_{t-1}^2 + \hat{\beta} \hat{\sigma}_{t-1}^2$$

where $\hat{\omega}$, $\hat{\alpha}$ and $\hat{\beta}$ are the parameters fitted by quasi likelihood estimation. In order to initialize the recurrence, $\hat{\sigma}_0$ is set to the standard deviation of the innovations. It is clear that if u_t^2 has a seasonal component, so will $\hat{\sigma}_t^2$. In this case, we need to take that seasonality into account in the modeling. In order to do that, we can simply add seasonal terms in the GARCH part, i.e. fit a model of volatility for the innovations of $X_{t_n}^{\Delta}$ of the form:

$$\sigma_t^2 = \omega + \sum_{k=1}^{P} \alpha_k u_{t-k}^2 + \sum_{k=1}^{Q} \beta_k \sigma_{t-k}^2 + \sum_{k=1}^{R} \lambda_{1,k} \left| \cos\left(\gamma_{1,k} + \frac{k\pi t}{\phi}\right) \right|^a + \lambda_{2,k} \left| \sin\left(\gamma_{2,k} + \frac{k\pi t}{\phi}\right) \right|^a$$

where the parameters $\omega, (\alpha_k)_{1 \leq k \leq P}, (\beta_k)_{1 \leq k \leq Q}, (\lambda_{1,k})_{1 \leq k \leq R}$ and $(\lambda_{2,k})_{1 \leq k \leq R}$ are positive and $(\gamma_{1,k})_{1 \leq k \leq R}, (\gamma_{2,k})_{1 \leq k \leq R}$ are real numbers. For ease of reading, we dropped the Δ subscript. This model naturally incorporates the seasonality. The GARCH structure has been adjusted to integrate deterministic terms at the seasonal periods, on top of which we added a base of absolute

	Anderson Darling	Cramer-Von Mises	Kolmogorov Smirnov
$\Delta = 30$	0.005	0.013	0.004
$\Delta = 60$	0.125	0.127	0.105
$\Delta = 120$	0.543	0.502	0.616
$\Delta = 180$	0.614	0.583	0.608
$\Delta = 240$	0.86	0.828	0.74
$\Delta = 300$	0.949	0.958	0.95
$\Delta = 360$	0.938	0.857	0.675
$\Delta = 480$	0.824	0.658	0.449
$\Delta = 540$	0.956	0.897	0.809
$\Delta = 600$	0.959	0.892	0.676
$\Delta = 720$	0.856	0.785	0.719
$\Delta = 900$	0.994	0.979	0.92
$\Delta = 960$	0.903	0.866	0.86
$\Delta = 1080$	0.968	0.949	0.991
$\Delta = 1200$	0.991	0.983	0.944
$\Delta = 1440$	1	0.999	0.997
$\Delta = 1800$	0.993	0.985	0.966
$\Delta = 1920$	0.803	0.736	0.495
$\Delta = 2160$	0.993	0.974	0.94
$\Delta = 2400$	0.81	0.834	0.731
$\Delta = 2700$	0.999	0.994	0.93
$\Delta = 2880$	0.975	0.978	0.883
$\Delta = 3600$	0.942	0.942	0.976

Table 3.15: *Goodness of fit tests p-values.*

sine and cosine functions to reflect the seasonal oscillations of the volatility of the innovations. The power term a acts as a way to sharpen the spikes, see Figure 3.19.

Across Δ we observe three significant spikes in the spectrum of u^2 at the scale Δ at periods ϕ/Δ, $\phi/2\Delta$ and $\phi/3\Delta$, corresponding to 24,12 and 8 hours periods, suggesting $R = 3$. In what follows we set $P = Q = 1$. The other parameters will be estimated via quasi likelihood optimization. The classic quasi likelihood used in GARCH estimation consists in replacing the unobserved sequence $(\sigma_t)_t$ with a sequence $(\tilde{\sigma}_t)_t$ that follows the same equation of evolution as σ_t, with the difference that we postulate a value for $\tilde{\sigma}_0$, typically the unconditional standard deviation of all the innovations. Since one naturally suspects that non seasonal effects are present in the dynamic, we account for those in the modeling by specifying an autoregressive and moving average model on the innovations. Hence we propose the following model for the innovations:

$$\begin{aligned}
\varepsilon_t &= \mu + \psi(\varepsilon_{t-1} - \mu) + \theta u_{t-1} + u_t \\
u_t &= \sigma_t z_t \\
\sigma_t^2 &= \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \sum_{k=1}^{R} \lambda_{1,k} \left| \cos\left(\gamma_{1,k} + \frac{k\pi t}{\phi}\right) \right|^a + \lambda_{2,k} \left| \sin\left(\gamma_{2,k} + \frac{k\pi t}{\phi}\right) \right|^a
\end{aligned}$$

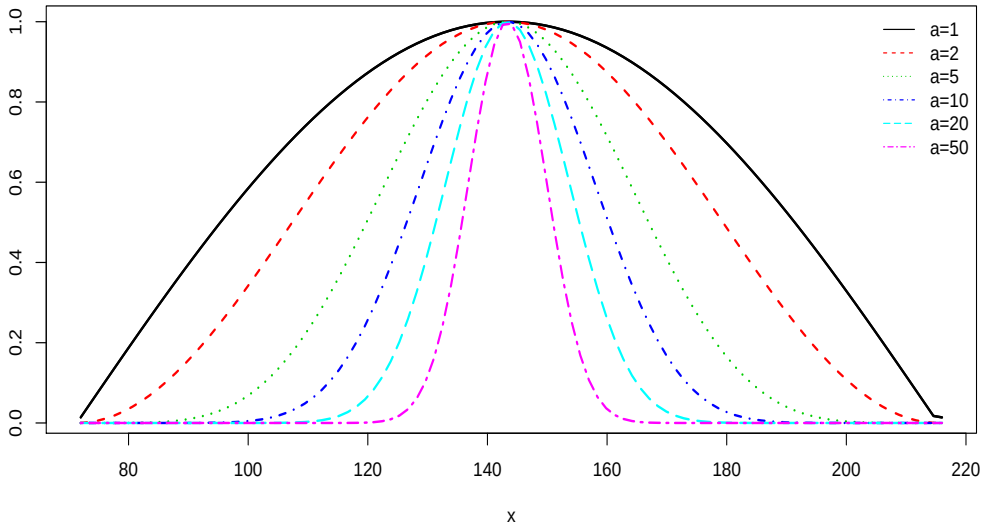

Figure 3.19: *Graph of* $x \mapsto \left|\cos\left(0.014 + \frac{\pi x}{144}\right)\right|^a$ *for* $a = 1, 2, 5, 10, 20, 50$.

where ε_n^Δ is the innovation of the Fourier decomposition defined in Section 3.3, μ is the mean level of the innovations, ψ is the autoregressive term, θ is the moving average term, the parameters of the GARCH part satisfy $\omega > 0$, $\alpha > 0$ and $\beta > 0$, $(\lambda_{1,k})_{1\leq k\leq R}, (\lambda_{2,k})_{1\leq k\leq R}$ are positive, $(\gamma_{1,k})_{1\leq k\leq R}, (\gamma_{2,k})_{1\leq k\leq R}$ are real numbers, $a > 0$ and the random variables z_t have zero mean and variance.

We shall now discuss how this model is estimated. Denote by F the distribution of z_1 and $f = F'$ its density. F will be chosen to be the skewed Student's t distribution with shape ν and skewness ξ. Denote:

$$\Theta = (\mu, \psi, \theta, \omega, \alpha, \beta, (\lambda_{1,k})_{1\leq k\leq 2}, (\lambda_{2,k})_{1\leq k\leq 2}, (\gamma_{1,k})_{1\leq k\leq 2}, (\gamma_{2,k})_{1\leq k\leq 2}, \nu, \xi)$$

the vector of unknown parameters. We choose the quasi likelihood approach. Let $\mathcal{F}_t = \sigma(\varepsilon_s | s \leq t)$ be the sigma algebra generated by the observations up to t. Then conditional on $\mathcal{F}_t$, ε_t has a density with respect to the Lebesgue measure, denoted by f_t given by:

$$f_t(x) = \frac{1}{\sigma_t} f\left(\frac{x - \mu - \psi(\varepsilon_{t-1} - \mu) - \theta u_{t-1}}{\sigma_t}\right)$$

Now, for any random variables $Y_1, \ldots, Y_n$ the joint probability distribution can be expressed as the product of the conditional distributions:

$$p(Y_1, \ldots, Y_n) = \prod_{k=1}^{n} p(Y_k | Y_1, \ldots, Y_{k-1})$$

meaning that the log-likelihood of the observations is given by:

$$\begin{aligned} l(X|\Theta) &= \sum_{t=3}^{T} \log f_t(\varepsilon_t) \\ &= \sum_{t=3}^{T} \left(-\log(\sigma_t) + \log f \left(\frac{\varepsilon_t - \mu - \psi(\varepsilon_{t-1} - \mu) - \theta u_{t-1}}{\sigma_t} \right) \right) \end{aligned} \tag{3.8}$$

We drop ε_1 and ε_2 because those are defined recursively using the unobserved variables ε_{-1} and ε_0. All parameters need to be estimated in one stage as estimating first the ARMA parameters on ε_t using the equation:

$$\varepsilon_t = \mu + \psi(\varepsilon_{t-1} - \mu) + \theta u_{t-1} + u_t$$

then estimating the GARCH parameters on the residuals using the equation:

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \sum_{k=1}^{R} \lambda_{1,k} \left| \cos\left(\gamma_{1,k} + \frac{k\pi t}{\phi} \right) \right|^a + \lambda_{2,k} \left| \sin\left(\gamma_{2,k} + \frac{k\pi t}{\phi} \right) \right|^a$$

leads inconsistent estimators because ARMA models assume conditional homoscedasticity while GARCH models specifically assumes conditional heteroscedasticity. Hence, we need to solve in Θ:

$$\arg\max_{\Theta} \sum_{t=3}^{T} -\log(\sigma_t) - \log f \left(\frac{\varepsilon_t - \mu - \psi(\varepsilon_{t-1} - \mu) - \theta u_{t-1}}{\sigma_t} \right). \tag{3.9}$$

Since we do not observe the sequences u_t and σ_t, the likelihood in (3.8) can not be optimized directly. We chose a quasi maximum likelihood approach. The dynamics satisfied by the sequences u_t and σ_t are deterministic in the sense that if one observes the initial values u_1 and σ_1, then conditional on $\sigma(\varepsilon_1, \ldots, \varepsilon_T)$, the sequences σ_t and u_t are completely determined. Instead of solving (3.9), we solve a similar optimization problem where we replace the sequences σ_t and u_t by $\tilde{\sigma}_t$ and $\tilde{u}_t$ where the latter satisfy the same dynamics as the former and such that $\tilde{\sigma}_1$ and $\tilde{u}_1$ are given. The choice of the initial values is not important, and consistency and asymptotic normality of the quasi likelihood estimator can be proved (see [49]). In practice, we initialize $\tilde{\sigma}_1$ to the unconditional standard deviation of the observations and $\tilde{u}_1$ to 0. Then we apply the Nelder-Mead simplex search method (see [67]) to optimize the function that aims at minimizing an n dimensional function by comparing the values at the vertices of a general $(n+1)$-simplex, denoted $x_1, \ldots, x_{n+1}$, such that $f(x_1) \leq \ldots \leq f(x_{n+1})$. At each step, the reflected point to x_{n+1} with respect to the center of mass of $x_1, \ldots, x_n$, denoted x_r, is computed and the simplex will either expand if $f(x_r) < f(x_1)$ or contract if $f(x_r) > f(x_n)$. Otherwise, x_{n+1} is replaced by x_r and a new reflection point is computed.

The model captures most features of the innovations. The correlation and seasonality of the original series and squared series are well explained by the model, see Figures 3.20 and 3.21 for an example of fit for $\Delta = 600$.

Diagnostic tests

Similar to the Naive ARMA-GARCH in section 3.4.1, we present the diagnostic tests on the same data for the ARMA Seasonal GARCH model. Just as for the Naive GARCH, skewness of

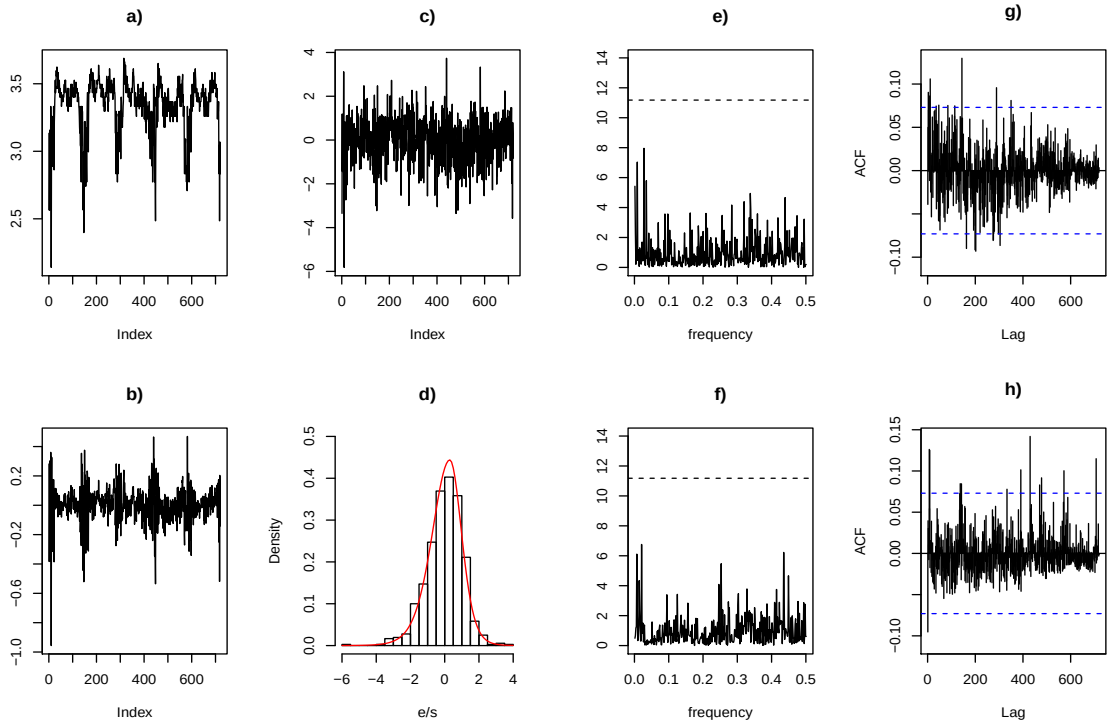

Figure 3.20: *a) Original series. b) Innovations of the Fourier decomposition. c) Residuals of the seasonal GARCH. d) Residuals histogram of the seasonal GARCH with theoretical fit. e) Periodogram of the residuals. f) Periodogram of the squared residuals. g) ACF of the residuals. h) ACF of the squared residuals.*

the standardized residuals converges to 1 as Δ increases, indicating that the distribution of the residuals gets more symmetric and the shape also has a downwards trend indicating that the tails get lighter as Δ increases. The goodness of fit tests indicates very good fit to the theoretical skewed student distribution with parameters $\text{Sh}(\Delta)$ and $\text{Sk}(\Delta)$, with an average of 6.5% rejection rate at level $\alpha = 5\%$, see Table 3.18. The Ljung box tests reveal an average of 6% rejection rate at lag 1 and $\log(n)$ for the standardized residuals, and only 2% rejection rate for the squared residuals, resulting in a great improvement over the Naive GARCH. The model accurately captures the serial correlation in both moments, and the mean seasonal effects, again in both moments, see Tables 3.16 3.17.

3.5 Point Forecast

In this section we assess the accuracy of our model for the conditional mean using the root mean squared error (RMSE) as a measure of performance, i.e. for a time series $(x_t)_{t=1,\dots,T}$ and corre-

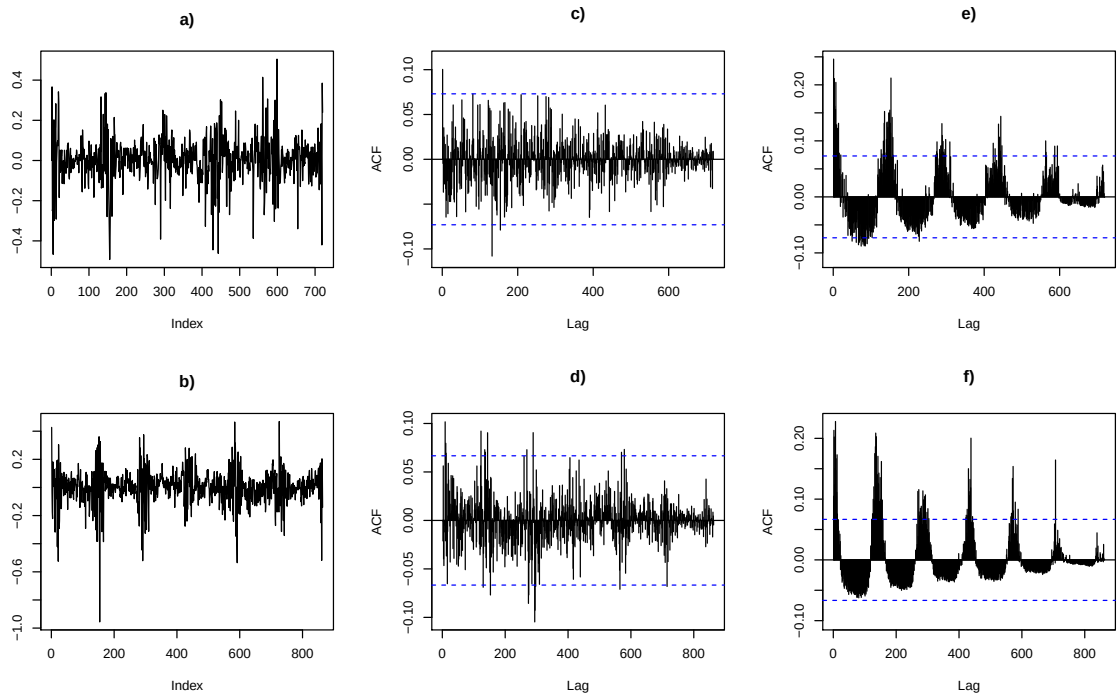

Figure 3.21: *a) Original Fourier innovations. b) Simulated path of the seasonal GARCH. c) ACF of the innovations. d) ACF of the simulated path. e) ACF of the squared innovations. f) ACF of the squared simulated path.*

sponding predictors $(\hat{x}_t)_{t=1,\dots,T}$, RMSE [51] is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{x}_t - x_t)^2}{T}}$$

We evaluate the forecast relative to two common benchmarks: the NAIVE and the AVG models. Given a time series $(y_t)_{t\in\mathbb{N}}$, the NAIVE prediction produces forecasts that are equal to the last observed value, i.e.:

$$\hat{y}_{t+1} = y_t. \tag{3.10}$$

The AVG prediction produces forecasts that are equal to the average of the last k observed values where $k \leq t$, i.e.:

$$\hat{y}_{t+1} = \frac{1}{k}\sum_{i=1}^{k} y_{t+1-i}. \tag{3.11}$$

AVG prediction can be used when mean reversion effect is suspected. We trained our model for f^{Δ} defined in (3.2) on 8 days of measurements and tested it on the following 6.

	Lag = 1	Lag = log(n)
$\Delta = 30$	0.138	0.245
$\Delta = 60$	0.531	0.151
$\Delta = 120$	0.149	0.084
$\Delta = 180$	0.255	0.174
$\Delta = 240$	0.606	0.542
$\Delta = 300$	0.726	0.561
$\Delta = 360$	0.275	0.054
$\Delta = 480$	0.589	0.523
$\Delta = 540$	0.067	0.002
$\Delta = 600$	0.649	0.262
$\Delta = 900$	0.238	0.709
$\Delta = 1080$	0.958	0.152
$\Delta = 1200$	0.336	0.888
$\Delta = 1800$	0.547	0.261
$\Delta = 1920$	0.283	0.417
$\Delta = 2400$	0.498	0.166
$\Delta = 2700$	0.453	0.835
$\Delta = 3600$	0.967	0.012

Table 3.16: *Residuals LB tests p-values.*

	Lag = 1	Lag = log(n)
$\Delta = 30$	0.66	0.561
$\Delta = 60$	0.215	0.222
$\Delta = 120$	0.977	0.384
$\Delta = 180$	0.276	0.844
$\Delta = 240$	0.522	0.27
$\Delta = 300$	0.334	0.16
$\Delta = 360$	0.094	0.093
$\Delta = 480$	0.318	0.307
$\Delta = 540$	0.358	0.07
$\Delta = 600$	0.078	0.945
$\Delta = 900$	0.512	0.167
$\Delta = 1080$	0.346	0.529
$\Delta = 1200$	0.153	0.582
$\Delta = 1800$	0.622	0.321
$\Delta = 1920$	0.086	0.225
$\Delta = 2400$	0.891	0.602
$\Delta = 2700$	0.943	0
$\Delta = 3600$	0.796	0.63

Table 3.17: *Squared residuals LB tests p-values.*

Our model improves uniformly the two benchmarks, but we shall discuss some nuances. First, we shall represent more visually those results by plotting the log-log plots of the RMSE for the 3 models and the ratio of our predictive model relative to each benchmark.

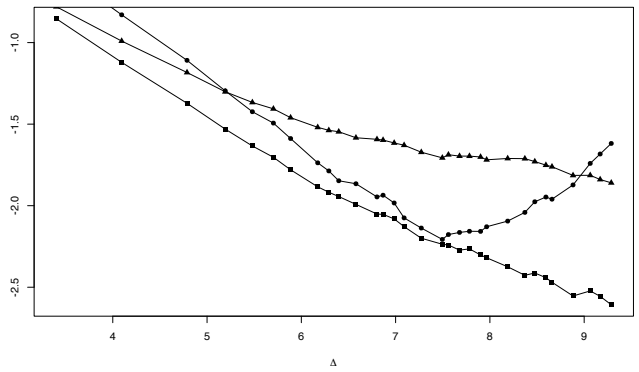

Figure 3.22: *Fourier (squares), naive (circles) and mean (triangles) prediction RMSE in log-log scale.*

Relative to the AVG prediction, the Fourier decomposition improves as Δ increases and stabilizes for $\Delta = 3600$ and onwards, whereas relative to the NAIVE prediction, the ratio peaks to 0.96

	Anderson Darling	Cramer-Von Mises	Kolmogorov Smirnov
$\Delta = 30$	0.005	0.013	0.004
$\Delta = 60$	0.125	0.127	0.105
$\Delta = 120$	0.543	0.502	0.616
$\Delta = 180$	0	0.023	0.007
$\Delta = 240$	0.825	0.88	0.797
$\Delta = 300$	0.069	0.246	0.293
$\Delta = 360$	0.375	0.345	0.489
$\Delta = 480$	0.313	0.329	0.227
$\Delta = 540$	0.375	0.286	0.305
$\Delta = 600$	0.342	0.279	0.332
$\Delta = 900$	0.987	0.972	0.907
$\Delta = 1080$	0.906	0.825	0.8
$\Delta = 1200$	0.804	0.801	0.805
$\Delta = 1800$	0.869	0.876	0.872
$\Delta = 1920$	0.104	0.1	0.081
$\Delta = 2400$	0.786	0.697	0.68
$\Delta = 2700$	0.956	0.915	0.923
$\Delta = 3600$	0.574	0.54	0.198

Table 3.18: *Goodness of fit tests of the model across* Δ.

	$\hat{f}^{\Delta}$	Naive prediction	Mean prediction
$\Delta = 30$	0.42	0.57	0.43
$\Delta = 60$	0.32	0.43	0.34
$\Delta = 120$	0.25	0.32	0.27
$\Delta = 180$	0.21	0.27	0.23
$\Delta = 240$	0.19	0.24	0.21
$\Delta = 300$	0.18	0.22	0.22
$\Delta = 360$	0.16	0.2	0.23
$\Delta = 480$	0.14	0.17	0.21
$\Delta = 540$	0.14	0.16	0.21
$\Delta = 600$	0.13	0.15	0.2
$\Delta = 720$	0.13	0.15	0.2
$\Delta = 900$	0.12	0.14	0.2
$\Delta = 960$	0.12	0.13	0.19
$\Delta = 1080$	0.12	0.13	0.19
$\Delta = 1200$	0.11	0.12	0.19

Table 3.19: RMSE for the 3 models as a function of Δ.

	$\hat{f}^{\Delta}$	Naive prediction	Mean prediction
$\Delta = 1440$	0.1	0.11	0.18
$\Delta = 1800$	0.1	0.1	0.17
$\Delta = 1920$	0.1	0.11	0.18
$\Delta = 2160$	0.1	0.11	0.18
$\Delta = 2400$	0.1	0.11	0.18
$\Delta = 2700$	0.09	0.11	0.17
$\Delta = 2880$	0.09	0.11	0.17
$\Delta = 3600$	0.09	0.12	0.17
$\Delta = 4320$	0.08	0.12	0.17
$\Delta = 4800$	0.08	0.14	0.17
$\Delta = 5400$	0.08	0.14	0.16
$\Delta = 5760$	0.07	0.14	0.16
$\Delta = 7200$	0.07	0.15	0.16
$\Delta = 8640$	0.07	0.17	0.15
$\Delta = 9600$	0.08	0.18	0.15

Table 3.20: RMSE for the 3 models as a function of Δ.

at $\Delta = 1800$ and continuously decreases onwards. The NAIVE prediction's RMSE is U-shaped and attains its minimum at $\Delta = 1800$, whereas the Fourier's RMSE is decreasing. For $\Delta < 1800$, the NAIVE prediction's RMSE decrease is steeper and both RMSE are approximately equal at

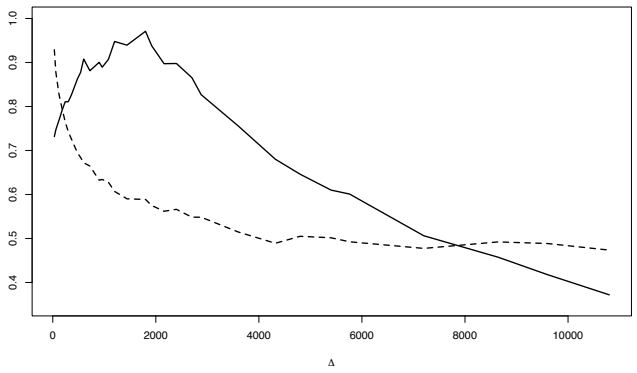

Figure 3.23: *RMSE ratio of Fourier model over NAIVE prediction (solid) and AVG prediction (dashed).*

$\Delta = 1800$. The standard deviation of the differenced series, i.e. the change between consecutive observations, and RMSE of the NAIVE prediction are essentially equal. This is because the aggregated series can be assumed to be bounded in probability. Under this hypothesis the RMSE for the NAIVE prediction is indeed asymptotically equal to the standard deviation of the differenced series. For any time series $(x_t)_t$ bounded in probability, the NAIVE prediction at time t is defined as $\hat{x_t} = x_{t-1}$ and the differenced series as $z_t = x_t - x_{t-1}$. Noting $\overline{z_N} = n^{-1} \sum_{t=1}^{N} z_t$, the variance of the differenced series up to time N is:

$$\begin{aligned}
\hat{\sigma}_N^2 &= \frac{1}{N} \sum_{t=1}^{N} z_t^2 - \overline{z_N}^2 \\
&= \frac{1}{N} \sum_{t=1}^{N} (x_t - x_{t-1})^2 - \overline{z_N}^2 \\
&= \frac{1}{N} \sum_{t=1}^{N} (x_t - \hat{x_t})^2 - \overline{z_N}^2 \\
&= \mathrm{RMSE}^2 - \overline{z_N}^2
\end{aligned}$$

but $\overline{z_N} = n^{-1}(z_N - z_0) \to 0$ in probability because the series is bounded in probability. As seen earlier, the mean model for the aggregated series is independent of the choice of Δ, and Δ is simply the elapsed time between consecutive observations. This has an important consequence: there are two competing effect that explain the U-shaped form of the NAIVE prediction's RMSE. As seen above, the NAIVE prediction's RMSE is the standard deviation of the differenced series. For small Δ, the difference between consecutive images of the mean model will contribute marginally to the the difference between consecutive observations. But as Δ increases, the contribution becomes more significant. Because the standard deviation of the differenced innovations keep decreasing, the increase of the volatility in the differenced series after $\Delta > 1800$ is a consequence of the contribution of the differenced mean model, i.e. the seasonal cycle. Hence the value $\Delta = 1800$ for the NAIVE prediction's RMSE corresponds to a tradeoff between variance in the differenced

innovations and variance in the differenced mean model. For $\Delta < 1800$, the volatility between consecutive observations is explained by the innovations' volatility, whereas for $\Delta > 1800$, the volatility between consecutive observations is explained by the seasonal cycles that overwhelm the innovations's volatility, resulting in this U-shaped form. The Fourier prediction's RMSE decrease for $\Delta < 1800$ is less steep than that of the NAIVE prediction: a slope of -0.32 in the power law signature for the Fourier's RMSE versus -0.41 for the NAIVE prediction's RMSE ($R^2 = 0.98$, p-value $< 2.2 \times 10^{-16}$ and $R^2 = 0.98$, p-value $< 2.2 \times 10^{-16}$ and $R^2 = 0.99$, p-value $< 2.2 \times 10^{-16}$ respectively), resulting in the observed increasing ratio up to this point. The reason though is unknown.

Just like the Fourier's RMSE, the AVG prediction's RMSE displays a power law, with a significantly smaller power index of -0.17 ($R^2 = 0.92$, p-value $< 2.2 \times 10^{-16}$), the quotient resulting in a power law with power index of -0.14 ($R^2 = 0.98$, p-value $< 2.2 \times 10^{-16}$). For the smallest scale $\Delta = 30$, the ratio attains a maximum of 0.93, and continuously decreases to 0.48 for the $\Delta = 10800$. Because of the strong seasonality, it is not surprising that the AVG prediction's RMSE is close to the standard deviation of the aggregated series. For small Δ, the volatility of the innovations is way larger in amplitude than the range of our mean model (up to a factor 10), hence for such Δ, it is expected that the Fourier model will perform approximately as the AVG prediction.

3.5.1 Optimal sampling frequency

As briefly explained in Section 3.4, one reason to use GARCH model for volatility is to decrease computational complexity of the prediction algorithm without impacting prediction accuracy in case the practitioner wants to use a very small value for Δ. As seen in the previous Section 3.5, the Fourier decomposition improves the RMSE over the two baselines uniformly for $\Delta = 30$ to $\Delta = 7200$ but the benefit from using the Fourier decomposition becomes increasingly smaller. In high frequencies the edge is marginal. It is clear that and our model outperforms the baselines partly because there is transition phase in the latency measurements twice a day: a sudden increase at dawn and a sudden decrease at mid-night. The baselines will perform very poorly because of that, whereas the Fourier decomposition actually captures that pattern. So the difference in high frequencies is mainly due to a few minutes of transition twice a day. Can we identify those transition phases, and what happens RMSE-wise when we look at them individually? In high frequencies one possible way to compare the RMSE is to compute it separately on periods of high/low variability. One way to identify those periods is to use the conditional standard deviation modeled with GARCH. The conditional standard deviation of the aggregated series bursts everyday, so using our model for the volatility we can precisely predict those bursts. We can then compare the performance of the Fourier decomposition and baselines separately during periods of high and low variability. We propose to identify those periods by setting a threshold on the conditional standard deviation. The threshold can simply be set to a quantile of the conditional standard deviation distribution. In the following numerical experiment, the threshold we use is 70%, see Figure 3.24.

In Table 3.21 we present the RMSE ratio between the NAIVE prediction and the Fourier decomposition, and the same ratio conditional on the periods of high and low variability. Figure 3.25 is a plot of the RMSE for the different models as a function of Δ. The difference between Naive and Fourier is less than 1% for $\Delta \leq 120s$ in periods of low variability, which represents with this threshold 70% of the measurements, meaning that in the majority of cases, for $\Delta \leq 120s$, the Fourier decomposition performs equally well as the Naive prediction, suggesting that the simpler

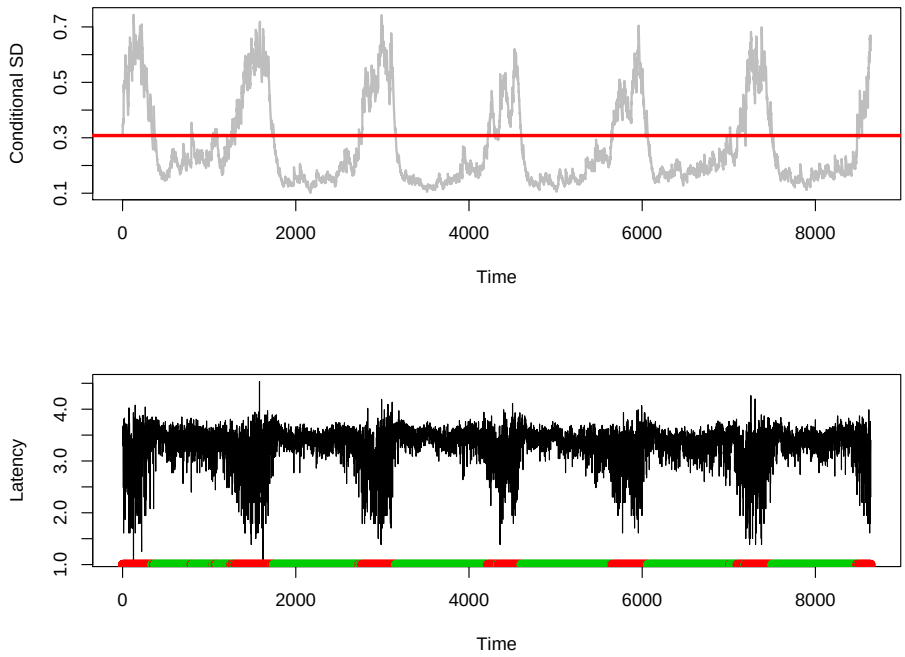

Figure 3.24: *Threshold for the conditional SD (top) and latency measurements with corresponding alternating intervals (bottom). Green intervals for low volatility, red intervals for high volatility.* $\Delta = 60$.

model can be used instead of the Fourier decomposition in order to gain time when producing a prediction. $\Delta < 120s$ can be identified as the values for Δ such that the aggregated series at higher frequencies are not predictable. Adjusting the threshold to be the 70% quantile of the conditional SD partition the horizon interval into alternating sub intervals of high and low volatility.

	Global error ratio	High variability error Ratio	Low variability error ratio
$\Delta = 30$s	1.020	1.032	0.99
$\Delta = 60$s	1.039	1.057	1.007
$\Delta = 120$s	1.067	1.092	1.015
$\Delta = 180$s	1.089	1.115	1.024
$\Delta = 240$s	1.109	1.159	1.056

Table 3.21: *Error ratio between Naive prediction and Fourier decomposition.*

We still observe similar patterns as when we did not split using the volatility, but split that way, it is clear that the edge of the Fourier prediction is absent for $\Delta < 120s$ in low variability periods, the slight edge is only present in high variability periods, as expected. For $\Delta < 120s$, we recommend to use simpler models in periods of high variability to decrease computational

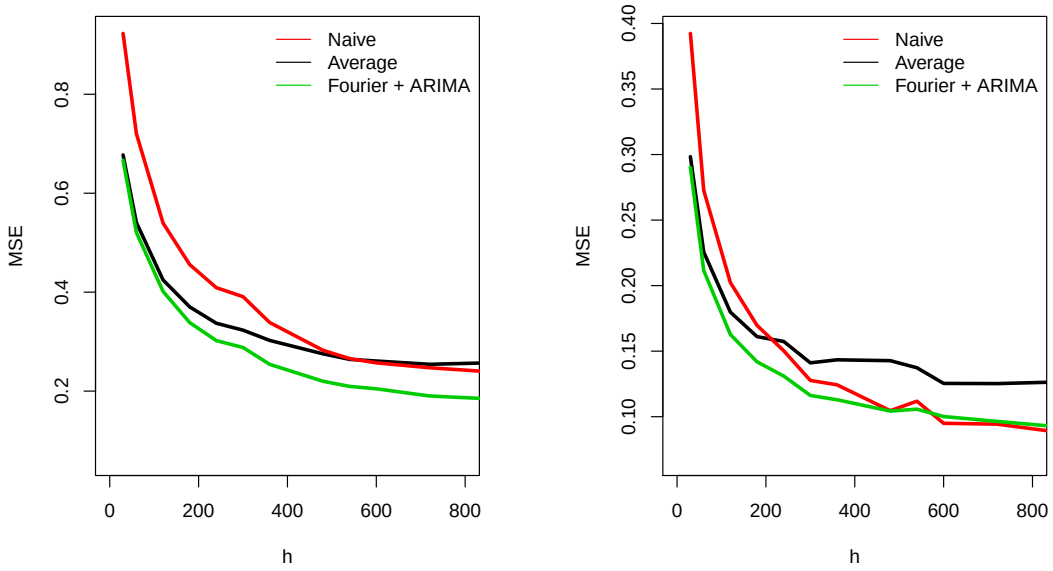

Figure 3.25: *RMSE as a function of* Δ *in low (left) and high (right) volatility regimes for the three models.*

complexity of the predictions.

3.6 Sample entropy as a predictability measure

Richman and Moormanis introduced a tool called sample entropy to study the predictability of time series [74]. Sample entropy (SE) is a measure of unpredictability of a time series. For a given embedding dimension m, tolerance r and number of data points N, sample entropy is the negative logarithm of the probability that if two sets of simultaneous data points of length m have distance $< r$ then two sets of simultaneous data points of length $m+1$ also have distance $< r$.

Definition 6. *Let* $m \in \mathbb{N}$, $r > 0$ *and* $X = (x_1, \ldots, x_N)$ *be a regularly spaced time series. Define* $X_m(i) = (x_i, \ldots, x_{i+m-1})$ *and let* $d = \|\cdot\|_\infty$ *be the sup norm. Then the sample entropy of* X *is defined as:*

$$SE_n^X = -\log \frac{A}{B} \tag{3.12}$$

where: $A = \#\{i \neq j, d(X_{m+1}(i), X_{m+1}(j)) < r\}$, $B = \#\{i \neq j, d(X_m(i), X_m(j)) < r\}$.

Since $A \leq B$, SE is always a positive number. SE is an approximation of the conditional probability that two vectors remain within a distance r at the next sampled point given that they were already at distance r for the first m points. High SE indicates that this probability is low, hence suggesting unpredictability of the series.

In the case where $X = \{x_1, \dots, x_N\}$ is the realization of N i.i.d. random variables $(X_i)_{1 \leq i \leq N}$, then $-\log(A/B)$ is the empirical counterpart of:

$$-\log \frac{\mathbb{P}\left(d((Z_1, \dots, Z_{m+1}), (Y_1, \dots, Y_{m+1})) < r\right)}{\mathbb{P}\left(d((Z_1, \dots, Z_m), (Y_1, \dots, Y_m)) < r\right)} = -\log \frac{\mathbb{P}\left(\max\limits_{1 \leq i \leq m+1} |Y_i - Z_i| < r\right)}{\mathbb{P}\left(\max\limits_{1 \leq i \leq m} |Y_i - Z_i| < r\right)}$$
$$= -\log\left(\mathbb{P}(|Z_1 - Y_1| < r)\right)$$

where $(Z_i)_i$, $(Y_i)_i$ are independent random variables with common distribution that of X_1. In other words, SE for i.i.d. data drawn from F estimates the concentration of F and measures the mean volume of a random ball of radius r. For instance, the theoretical sample entropy of an i.i.d. sample is $\log(2\pi^{-1} \int_0^{r/2\sigma} e^{-t^2} dt)$ in the case $\mathcal{N}(0, \sigma^2)$, or $-\log(2r - r^2)$ in the case $\mathcal{U}(0, 1)$. The

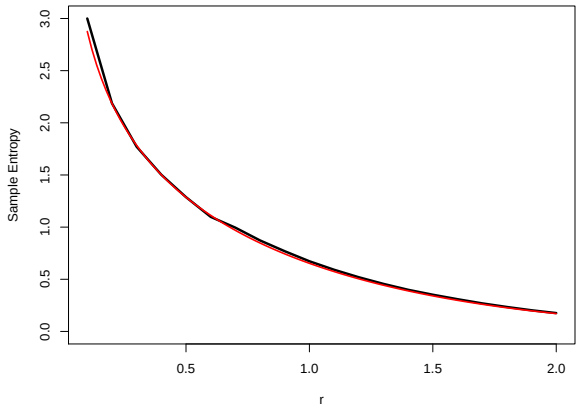

Figure 3.26: *Theoretical (red) and sample entropy (black) as a function of r in the normal $\mathcal{N}(0, 1)$ case.*

choice of r as a function of the standard deviation is to be scale free, while the coefficient 0.2 was motivated by empirical studies (see [74]). We suspect that among all distributions supported over the real line, the maximum entropy should be attained for some distribution, potentially under moment constraints. Indeed, in all generality if Z, Y are independent with common distribution F and density F', then the entropy at point r can be expressed as a fonctionnal of F:

$$-\log\left(\mathbb{P}(|Z - Y| < r)\right) = \int_{\mathbb{R}} \left(F(x + r) - F(x - r)\right) F'(x) dx$$
$$=: J_r(F)$$

3.6.1 Theoretical analysis

Sample entropy was first introduced as a measure of predictability or self similarity of a time series

and has been widely applied since, especially on physiological data, see for instance [17] [84] [56]. We propose to use SE in the context of time series prediction of the median-process of Internet. After fitting a model, one is typically interested in assessing to what extent the model captured the information contained in the data. We propose a way to use SE to achieve this goal, i.e. to perform a single test to assess whether the residuals of a statistical model are independent and identically distributed according to some target distribution.

Sample Entropy or SE is closely related to the correlation integral first introduced by Grassberger and Procaccia [41] [42] to study attractors of some dynamical systems. In what follows, we will first recall some facts about the correlation integrals and the convergence of $U-$statistics under specific conditions to set up the theoretical background for the study of SE. We will conclude this section by proving convergence properties of SE.

Let F be some distribution function over some set S. For a metric d and $r > 0$ define the correlation integral as

$$C(d,r) = \int_S \int_S \mathbf{1}_{\{d(x,y)<r\}} F(dx)F(dy).$$

In many cases, orbits of dynamical systems tend to accumulate in a region of the space called the attractor. Dynamical systems are deterministic, but can be studied under the scope of probability using ergodic theory. For many dynamical systems the existence of an invariant measure has an important consequence: a typical trajectory can be regarded as a outcome of a stationary sequence $(W_n)_{n\in\mathbb{N}}$ equipped with its invariant measure μ with c.d.f. F. Grassberger and Procaccia showed that in many cases $C(d,r) = \text{Cste} \cdot r^{\nu}$ as $r \to 0$ where $\nu > 0$ is called the correlation exponent and is intimately related to the fractal dimension of the underlying attractor. $C(d,r)$ is a measure of the concentration of F and represents the mean volume of a ball of radius r. Indeed if we let μ denote law of X where $X \sim F$, we have:

$$\begin{aligned} C(d,r) &= \int_S \int_S \mathbf{1}_{\{d(x,y)<r\}} F(dx)F(dy) \\ &= \int_S \mu(B(x,r))F(dx) \\ &= \mathbb{E}\mu(B(X,r)) \end{aligned}$$

A natural estimator of $C(d,r)$ is:

$$C_n(d,r) = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \mathbf{1}_{\{d(W_i,W_j)<r\}}.$$

Note that $C_n(d,r)$ is a U-statistic. It is clear that the sample entropy is the log-ratio of two correlation integrals. Before introducing rigorously the sample entropy in terms correlation integrals, we shall state the fundamental theorem concerning the convergence of U-statistics for dependent data. The basic results concerning U-statistics built on an i.i.d. sequence is due to Hoeffding (1948) (see [47] for instance) and the introduction of the so-called Hoeffding-decomposition, but will not apply here as we will focus our attention on U-statistics based on dependent observations. Under the assumption that the process W is a Lipschitz functional of an absolutely regular stationary process with good mixing properties, Denker and Keller [28] along with Cutler [24] extended the

work of Grassberger and Procaccia and proved asymptotic normality of U-statistics for this class of processes. Let us first recall the definition of U-statistics. Let $\boldsymbol{W} = (W_n)_{n\geq 1}$ a collection of $\mathbb{R}^d$ valued random variables and $h : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ be a kernel, i.e. a symmetric and measurable function. The associated U-statistic is:

$$U_n = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} h(W_i, W_j).$$

The main result of Denker and Keller is asymptotic normality of U_n for a large class of process $\boldsymbol{W}$. They first introduced a class of kernels that satisfy a specific variation condition, denoted $\mathcal{C}_P$. For $h : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$, $x \in (\mathbb{R}^p)^2$ and $\varepsilon > 0$, define the oscillation function of h at x:

$$osc(h, \varepsilon, x) = \sup\{|h(y) - h(y')|\colon \|y - x\|_\infty < \varepsilon, \|y' - x\|_\infty < \varepsilon, y, y' \in (\mathbb{R}^p)^2\}$$

where $\|\cdot\|_\infty$ denotes the sup norm. $osc(h, \varepsilon, x)$ is the amplitude of h in a neighborhood of x. Now with respect to a probability measure P on $\mathbb{R}^p$ define the mean oscillation of h:

$$osc(h, \varepsilon) = \int osc(h, \varepsilon, x) dP^2(x)$$

Now define the class $\mathcal{C}_P$ that satisfy the variation condition with respect to the probability measure P:

$$\mathcal{C}_P = \{h : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R} : h \text{ is a kernel and } \exists s > 0,\ M = \sup_{\varepsilon>0} \varepsilon^{-s} osc(h, \varepsilon) < \infty\}$$

Now let $(W_n)_{n\geq 1}$ be a stationary process and $(U_n)_{n\geq 1}$ the associated U-statistic, i.e.:

$$U_n = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} h(W_i, W_j).$$

Assume that for all $n \geq 1$, $W_n = f(Y_n, Y_{n+1}, \dots)$ where $(Y_n)_{n\geq 1}$ is an absolutely regular stationary process and f is measurable. We shall now state the result of Denker and Keller [28] on the asymptotic normality of $U-$statistics. Suppose the following hypothesis holds:

- (H1): $(Y_n)_{n\geq 1}$ has β-mixing coefficients satisfying:

  $$\beta(n)^{\delta/(2+\delta)} = \mathcal{O}(n^{-2-\varepsilon})$$

  for some $\delta, \varepsilon > 0$.

- (H2): The function f is Lipchitz continuous in the sense that that there exists $0 \leq \alpha < 1$:

  $$|f(y_1, y_2, \dots) - f(y_1,' y_2', \dots)| \lesssim \alpha^n$$

  if $y_1 = y_1'$, $\dots$, $y_n = y_n'$.

- (H3): the kernel h is in $\mathcal{C}_P$, i.e. satisfy the variation condition with respect to P.

Before stating the theorem, we need one last notation. Denote by dF the marginal distribution of $(W_n)_{n\geq 1}$, i.e. $\mathbb{P}(W_1 \in A \subset \mathbb{R}^p) = F(A)$. For $x \in \mathbb{R}^p$ let:

$$h_1(x) = \int h(x, y) dF(y)$$

Theorem 8 (Denker and Keller)**.** *Assuming (H1), (H2) and (H3) hold:*

$$U_n \xrightarrow{P} \theta$$

and

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where

$$\theta = \mathbb{E}_F(h_1(W_1)) = \int \int h(x,y) dF(x) dF(y)$$

$$\sigma^2 = 4\left(\mathbb{E}_F(h_1(W_1)^2) - \theta^2 + \sum_{n \geq 2} \mathbb{E}_F(h_1(W_1) - \theta)(h_1(W_n) - \theta) \right).$$

Hereinafter let $(X_n)_{n \geq 1}$ be independent random variables with common distribution F that has a second moment. For $m \in \mathbb{N}$, introduce the process:

$$Z_i = (X_i, \ldots, X_{i+m})$$

for some $m \in \mathbb{N}$. Let $k \in \{m, m+1\}$, and define the functions $d_k : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \longrightarrow \mathbb{R}$ and $h_k : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \longrightarrow \mathbb{R}$ as

$$d_k(x,y) = \max_{1 \leq i \leq k} |x_i - y_i|$$

and

$$h_k(x,y) = \mathbf{1}_{\{d_k(x,y) < r\}}.$$

Then the sample entropy of the random sample $(X_1, \ldots, X_n)$ is:

$$SE_n^X = -\log\left(\frac{C_n(d_m, r)}{C_n(d_{m+1}, r)} \right)$$

where

$$\begin{aligned} C_n(d_k, r) &= \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbf{1}_{\{d_k(Z_i, Z_j) < r\}} \\ &= \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h_k(Z_i, Z_j). \end{aligned}$$

Observe that h_k is a kernel, hence SE_n^X is the ratio of two U-statistics. We will use Theorem 8 to prove that the those two U-statistics have Gaussian asymptotic distribution. Joint asymptotic normality will be a consequence of the Cramer-Wold [21] theorem, and the final result will be a derived using the Delta method.

Proposition 11. *Let $m \in \mathbb{N}$ and $(X_n)_{n \geq 1}$ be independent random variables with common distribution F such that $\mathbb{E}X_1^2 < \infty$. Then the process $(Z_n)_{n \geq 1}$ defined as:*

$$Z_n = (X_n, \ldots, X_{n+m})$$

is stationary and satisfy hypothesis (H1) and (H2).

Proof. Z clearly is stationary. The fact that $(Z_n)_{n\geq 1}$ satisfy (H1) and (H2) is immediate since $(X_n)_{n\geq 1}$ is an i.i.d. sequence. □

Proposition 12. *Let $k \in \{m, m+1\}$. Define $h_k : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \longrightarrow \mathbb{R}$ as:*

$$h_k(x, y) = \mathbf{1}_{\{d_k(x,y)<r\}}$$

where $d_k(x, y) = \max_{1\leq i\leq k}|x_i - y_i|$. Then h_k is a kernel satisfying (H3), i.e. the variation condition, with respect to any measure P of the form $P = \otimes_{i=1}^{m+1} Q$ where Q is an absolutely continuous probability measure on $\mathbb{R}$.

Proof. Throughout the proof we will use the following notation: for any $x \in (\mathbb{R}^{m+1})^2$, let $x^1, x^2 \in \mathbb{R}^{m+1}$ such that $x = (x^1, x^2)$. In particular, $h_k(x) = h_k(x^1, x^2)$.

Let $k \in \{m, m+1\}$. Observe that for $x \in (\mathbb{R}^{m+1})^2$, $osc(h_k, \varepsilon, x) \in \{0, 1\}$ since $h_k(y) \in \{0, 1\}$ for all $y \in (\mathbb{R}^{m+1})^2$. We first observe that if $osc(h, \varepsilon, x) > 0$ then:

$$r - 2\varepsilon \leq d_k(x^1, x^2) \leq r + 2\varepsilon$$

If it were not the case, say if $d_k(x^1, x^2) > r + 2\varepsilon$, then for any $y \in (\mathbb{R}^{m+1})^2$ such that $\|y - x\| < \varepsilon$ we would have:

$$\begin{cases} |x_i^1 - y_i^1| < \varepsilon & \text{for all } i \in \{1, \dots, m+1\} \\ |x_i^2 - y_i^2| < \varepsilon & \text{for all } i \in \{1, \dots, m+1\} \end{cases}$$

hence combining the two inequalities and using the triangle inequality it follows:

$$|y_i^1 - y_i^2| > |x_i^1 - x_i^2| - 2\varepsilon > r$$

for all $1 \leq i \leq m+1$. Then it follows:

$$d_k(y^1, y^2) > d_k(x^1, x^2) - 2\varepsilon > r$$

since $d_k(x^1, x^2) > r + 2\varepsilon$ by assumption. This implies that for all $y \in (\mathbb{R}^{m+1})^2$ such that $\|y - x\| < \varepsilon$, $h_k(y) = 0$ hence $osc(h_k, \varepsilon, x) = 0$, which is a contradiction hence $d_k(x^1, x^2) \leq r + 2\varepsilon$. The proof that $d_k(x^1, x^2) \geq r - 2\varepsilon$ is identical. Hence the following hold true:

$$osc(h_k, \varepsilon, x) > 0 \Rightarrow r - 2\varepsilon \leq d_k(x^1, x^2) \leq r + 2\varepsilon$$

Now, let P a measure on $\mathbb{R}^{m+1}$ of the form $P = \otimes_{i=1}^{m+1} Q$ where Q is an absolutely continuous probability measure on $\mathbb{R}$. By definition, the mean oscillation of h with respect to P is:

$$\begin{aligned} osc(h, \varepsilon) &= \int osc(h, \varepsilon, x) dP^2(x) \\ &= \mathbb{E}_P osc(h, \varepsilon, X) \end{aligned}$$

where $X = (X^1, X^2) \sim P \otimes P$. We know that $osc(h, \varepsilon, X) \in \{0, 1\}$ and $osc(h, \varepsilon, X) > 0 \Rightarrow r - 2\varepsilon \leq d_k(x^1, x^2) \leq r + 2\varepsilon$, hence:

$$\begin{aligned} osc(h, \varepsilon) &= \mathbb{P}_P(osc(h, \varepsilon, X) = 1) \\ &\leq \mathbb{P}_P(r - 2\varepsilon \leq d_k(X^1, X^2) \leq r + 2\varepsilon) \end{aligned}$$

Let G_k be the c.d.f. of $d_k(X^1, X^2)$. Then:

$$osc(h, \varepsilon) \leq G_k(r + 2\varepsilon) - G_k(r - 2\varepsilon).$$

The joint distribution of $X = (X^1, X^2)$ is $\otimes_{i=1}^{2(m+1)} Q$, hence:

$$\begin{aligned} G_k(u) &= \mathbb{P}(d_k(X^1, X^2) \leq u) \\ &= \mathbb{P}(\max_{1 \leq i \leq k} |X_i^1 - X_i^2| \leq u) \\ &= \mathbb{P}(|U - V| \leq u)^k \end{aligned}$$

where U, V are independant with common distribution Q. Since Q is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, $|U - V|$ is also absolutely continuous with respect to the Lebesgue measure, hence G_k is differentiable almost everywhere. Finally write:

$$0 < \varepsilon^{-1} osc(h, \varepsilon) \leq \frac{G_k(r + 2\varepsilon) - G_k(r - 2\varepsilon)}{\varepsilon}.$$

The right hand side converges as $\varepsilon \to 0$ since G_k is differentiable. So we can conclude:

$$M = \sup_{\varepsilon > 0} \varepsilon^{-1} osc(h, \varepsilon) < \infty$$

i.e. $h_k \in \mathcal{C}_P$, hence satisfies (H3). □

Proposition 13. *Let $k \in \{m, m+1\}$. Let:*

$$C_n(d_k, r) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbf{1}_{\{d_k(Z_i, Z_j) < r\}}$$

Then there exists $\sigma_k^2 > 0$ such that:

$$\sqrt{n}(C_n(d_k, r) - \theta_k) \xrightarrow{d} \mathcal{N}(0, \sigma_k^2)$$

where $\theta_k = \mathbb{P}(|X_1 - X_2| < r)^k$.

Proof. Direct application of Theorem 8 and Propositions 11 and 12. □

Proposition 14. *$C_n(d_m, r)$ and $C_n(d_{m+1}, r)$ have joint asymptotic Gaussian distribution, i.e. there exists a symmetric positive definite matrix Σ of size 2×2 such that:*

$$\sqrt{n} \left[\begin{pmatrix} C_n(d_m, r) \\ C_n(d_{m+1}, r) \end{pmatrix} - \begin{pmatrix} \theta_m \\ \theta_{m+1} \end{pmatrix} \right] \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\theta_k = \mathbb{P}(|X_1 - X_2| < r)^k$, $k \in \{m, m+1\}$.

Before proving this proposition, we will need the following lemma:

Lemma 5. *Let $l \in \mathbb{N}$ and P be a probability measure on $\mathbb{R}^l$. Suppose that $h_1 : \mathbb{R}^l \times \mathbb{R}^l \longrightarrow \mathbb{R}$, $h_2 : \mathbb{R}^l \times \mathbb{R}^l \longrightarrow \mathbb{R}$ are two kernels that satisfy (H3) with respect to P. Then any linear combination of h_1 and h_2 satisfy (H3).*

Proof. Let $\lambda_1, \lambda_2 \in \mathbb{R}$, $h = \lambda_1 h_1 + \lambda_2 h_2$ and $\varepsilon > 0$. Recall that

$$osc(h, \varepsilon) = \int osc(h, \varepsilon, x) dP^2(x).$$

Let $y, y' \in (\mathbb{R}^l)^2$. By the triangle inequality:

$$\begin{aligned} |h(y) - h(y')| &= |\lambda_1 h_1(y) + \lambda_2 h_2(y) - \lambda_1 h_1(y') - \lambda_2 h_2(y')| \\ &\leq |\lambda_1||h_1(y) - h_1(y')| + |\lambda_2||h_2(y) - h_2(y')| \end{aligned}$$

But $osc(h, \varepsilon, x) = \sup\{|h(y) - h(y')| : \|y - x\| < \varepsilon, \|y' - x\| < \varepsilon, y, y' \in (\mathbb{R}^l)^2\}$, hence:

$$osc(h, \varepsilon, x) \leq |\lambda_1| osc(h_1, \varepsilon, x) + |\lambda_2| osc(h_2, \varepsilon, x).$$

It follows:

$$osc(h, \varepsilon) \leq |\lambda_1| osc(h_1, \varepsilon) + |\lambda_2| osc(h_2, \varepsilon)$$

and since h_1, h_2 satisfy (H3), clearly h does too. □

Proof of Proposition 14. Let $\lambda = (\lambda_m, \lambda_{m+1}) \in \mathbb{R}^2$, $\lambda \neq 0$. Then:

$$\begin{aligned} \lambda_m C_n(d_m, r) + \lambda_{m+1} C_n(d_{m+1}, r) &= \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \lambda_m h_m(Z_i, Z_j) + \lambda_{m+1} h_{m+1}(Z_i, Z_j) \\ &=: \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h^\lambda(Z_i, Z_j) \quad \text{say.} \end{aligned}$$

Then clearly h^λ is a kernel satisfying (H3) according to Lemma 5 and Proposition 12. Hence applying Theorem 8 with kernel function h^λ and letting $h_1^\lambda(x) = \int h^\lambda(x, y) dF(y)$, we have that there exists $\sigma_\lambda^2 > 0$:

$$\sqrt{n} \left(\lambda_m C_n(d_m, r) + \lambda_{m+1} C_n(d_{m+1}, r) - \mu_\lambda \right) \xrightarrow{d} \mathcal{N}(0, \sigma_\lambda^2)$$

where:

$$\begin{aligned} \mu_\lambda &= \int \int h^\lambda(x, y) dF(x) dF(y) \\ &= \lambda_m \int \int h_m(x, y) dF(x) dF(y) + \lambda_{m+1} \int \int h_{m+1}(x, y) dF(x) dF(y) \\ &= \lambda_m \theta_m + \lambda_{m+1} \theta_{m+1} \end{aligned}$$

and:

$$\sigma_\lambda^2 = 4 \left(\mathbb{V}(h_1^\lambda(Z_1)) + \sum_{t=1}^{m+1} \operatorname{cov}(h_1^\lambda(Z_1), h_1^\lambda(Z_{1+t})) \right).$$

We have established that for all $\lambda = (\lambda_m, \lambda_{m+1}) \in \mathbb{R}^2$, $\lambda \neq 0$, if

$$Y_n = \sqrt{n}\begin{pmatrix} C_n(d_m, r) - \theta_m \\ C_n(d_{m+1}, r) - \theta_{m+1} \end{pmatrix},$$

we have

$$\langle \lambda, Y_n \rangle \xrightarrow{d} \mathcal{N}(0, \sigma_\lambda^2).$$

We quickly recall the Cramer-Wold theorem. If $(A_n)_{n\geq 1}$ is a sequence of $\mathbb{R}^p$ valued random variables such that for all $\xi \in \mathbb{R}^p$ there exists a real valued random variable B_ξ such that $\langle \xi, A_n \rangle \xrightarrow{d} B_\xi$, then A_n weakly converges to a limit $A_\infty \in \mathbb{R}^p$ and $B_\xi = \langle \xi, A_\infty \rangle$. Moreover, by definition, a vector $A \in \mathbb{R}^p$ is a Gaussian vector if for all $\xi \in \mathbb{R}^p$, $\langle \xi, A \rangle$ is a real valued Gaussian variable. Hence there exists Y_∞ such that:

$$\langle \lambda, Y_n \rangle \xrightarrow{d} \langle \lambda, Y_\infty \rangle$$

and Y_∞ has Gaussian $\mathcal{N}(0, \Sigma)$ distribution. Because Σ is a 2×2 covariance matrix, there exists $a, b > 0$ and $c \in \mathbb{R}$ such that:

$$\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

We shall now derive the exact expression for Σ using the limiting distribution of $\langle \lambda, Y_n \rangle$ for specific choices of λ.

For $\lambda = (1, 0)$ we have:

$$\langle \lambda, Y_n \rangle = \sqrt{n}\,(C_n(d_m, r) - \theta_m) \xrightarrow{d} \mathcal{N}(0, \sigma_{(1,0)}^2).$$

For $\lambda = (0, 1)$ we have:

$$\langle \lambda, Y_n \rangle = \sqrt{n}\,(C_n(d_{m+1}, r) - \theta_{m+1}) \xrightarrow{d} \mathcal{N}(0, \sigma_{(0,1)}^2).$$

Finally, for $\lambda = (1, 1)$ we have:

$$\langle \lambda, Y_n \rangle = \sqrt{n}\,(C_n(d_m, r) + C_n(d_{m+1}, r) - \theta_m - \theta_{m+1}) \xrightarrow{d} \mathcal{N}(0, \sigma_{(1,1)}^2).$$

By identification we have

$$\begin{cases} \sigma_{(1,0)}^2 = a \\ \sigma_{(0,1)}^2 = b \\ \sigma_{(1,1)}^2 = a + b + 2c \end{cases}$$

hence the asymptotic covariance term c satisfies

$$c = \frac{1}{2}(\sigma_{(1,1)}^2 - \sigma_{(1,0)}^2 - \sigma_{(0,1)}^2)$$

where

$$\sigma_\lambda^2 = 4\sum_{t=0}^{m} \operatorname{cov}(h_1^\lambda(Z_1), h_1^\lambda(Z_{1+t}))$$

for any $\lambda \in \mathbb{R}^2$.

□

3.6.2 Asymptotic variance estimation

In order to estimate σ_λ^2, one needs to estimate $\kappa^\lambda(t) = \text{cov}(h_1^\lambda(Z_1), h_1^\lambda(Z_{1+t}))$ for $t = 0, 1, \ldots, m+1$. First let:

$$p^\lambda(j) = \frac{1}{n-1} \sum_{i \neq j} h^\lambda(Z_i, Z_j)$$

$$C_n^\lambda = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h^\lambda(Z_i, Z_j)$$

then the estimator for $\kappa^\lambda(t)$ is:

$$\hat{\kappa}^\lambda(t) = \frac{1}{n-t} \sum_{i=1}^{n-t} p^\lambda(j) p^\lambda(j+t) - {C_n^\lambda}^2,$$

and the estimator $\hat{\Sigma}$ for Σ easily follows. Finally we derive asymptotic normality of the sample entropy.

Proposition 15. *Let $\theta = -\log(\mathbb{P}(|X_1 - X_2| < r))$. Then the following hold:*

$$\begin{cases} SE_n^X \xrightarrow{P} \theta \\ \sqrt{n}(SE_n^X - \theta) \xrightarrow{d} \mathcal{N}(0, \nabla g(\theta) \Sigma \nabla g(\theta)^t) \end{cases}$$

where $g : (x, y) \mapsto \log(x) - \log(y)$.

Proof. Direct consequence of the asymptotic normality of $(C_n(d_m, r), C_n(d_{m+1}, r)$ and the Delta method. □

The reason why sample entropy is important is because under the null hypothesis that the data are i.i.d. F, then its exact behavior under H_0 is known, and will significantly shift away from H_0 under $H_1 = \neg H_0$. The typical use case where one would care to apply this test procedure would be on the residuals of some statistical model to assess to what extent the model captured the information contained in the series. We performed the test at level $\alpha = 0.05$ on the residuals of the Fourier seasonal-GARCH. The distribution of the residuals under H_0 was specified as the Skewed Student distribution with parameters Sk(Δ) and Sh(Δ) defined in (3.6) and (3.7). The results are presented in Figure 3.27, and suggest that our model successfully captured the relevant information in the aggregated series. Instead of using the test on the residuals, it can be applied directly to the aggregated series, leading a different interpretation of the results. When applied on the residuals, the test will measure how good the model captured the information within the series, but when applied directly to the original series, it measures the quantity of information that can be used to predict it.

3.7 Conclusion

In this chapter we presented an empirical analysis of Internet traffic latency using a Fourier ARMA Seasonal-GARCH model to explain both conditional mean and variance dynamics in the formation

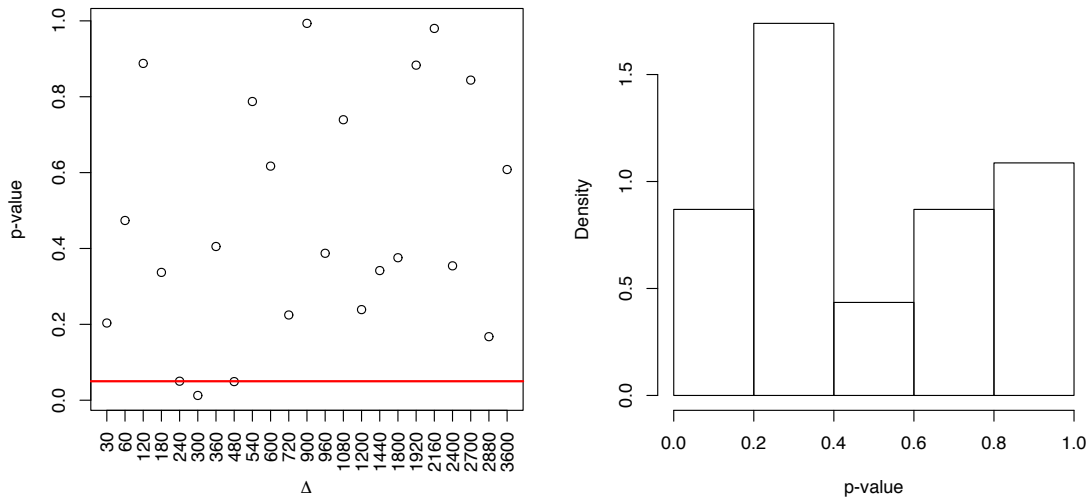

Figure 3.27: *P-values of all test for* $\Delta \leq 3600$ *with horizontal line* $y = 0.05$ *in red (left). Histogram of the p-values.*

of the median process of log-latency measurements obtained after aggregating measurements over consecutive and non overlapping time intervals of length $\Delta > 0$. Modeling this median process instead of the underlying true data generating process is based on operational standards than rely on quantiles, and especially the median, of the latency measurements. The reason is that measurements are generated by thousands of different users at a very high frequency, making a distributional approach that relies on the quantiles more manageable and interpretable since the load balancer can not use the latency performance of a single user that needs to be routed, but instead only sees the whole distribution of latency across all users at once.

The mean dynamic displays strong persistent daily cycles, captured by a Fourier regression of the aggregated series fitted by ordinary least squares with 3 periods at 24h, 12h and 8h. A particularly important feature is the invariance of the estimates for all $\Delta < 10800$, i.e. 3 hours, suggesting very slow rate of conditional mean evolution. The Fourier decomposition captures the serial autocorrelation of the series and all seasonal cycles.

The squared residuals of this conditional mean model exhibit strong autocorrelation and seasonal volatility clustering. An ARMA seasonal GARCH model, i.e. an ARMA-GARCH process with additional deterministic terms in the GARCH innovations, fitted in one stage by quasi maximum likelihood on the residuals on the mean model captures all volatility features.

Residuals diagnostic show excellent goodness of fit of both mean and variance model. The model allows for time series prediction and is proven to outperform common baselines in the industry like NAIVE prediction and AVG prediction.

Finally, a key problem for *load-balancing* purposes is that prediction should be made at the shortest time horizon possible, i.e. with the smallest Δ possible. But as Δ goes to 0, the volatility explodes and prediction become intractable. We proposed a way to select the optimal sampling frequency by exploiting alternating seasonal periods of low and high volatility. A new way of measuring the predictability of a time series based on sample entropy is proposed, and asymptotic properties are derived. The test is applied on the residuals of our model and shows evidence that the model accurately captures the structure of the median latency process.

Chapter 4

Detecting changes and training set selection in stable networks

Abstract

In computer engineering, latency measures the time in millisecond necessary for a request to reach the destination server, and for the response to get back to the host server. For *load-balancers* (see Chapter 1 Section 1.2 for a presentation of the notion) latency plays a central role: low latency implies low page load time. The other main concern for *load-balancers* is *availability*, which quantifies the degree to which a server functions normally. In rare cases, a server may experience an outage. Informally, an outage occurs when some servers in the network are no longer in good functioning condition and sending traffic to those server will significantly deteriorate the so-called user experience. During an outage, both mean and variance of the distribution of latency measurements increase. Prediction accuracy of Internet latency and the ability to detect outages are the most important matters of concern for *load-balancer*s because optimal performance is necessarily a consequence of the two combined. In this chapter, we address the problem of optimizing the training data selection in a class of networks with a certain stability property that we call $\varepsilon-$stable networks, or $\varepsilon - SN$, and we adapt the result of Daniel Kifer et al. [52] on the detection of changes in the distribution of a data stream to detect outages. We demonstrate how to effectively tune the predictive algorithm that Citrix uses in those special networks to decrease the computational time without impacting accuracy. In a second step we show that the distribution of latency measurements exhibit a power law behavior, meaning that the survival function of latency measurements decreases as $x^{-\alpha}$ for some $\alpha > 0$, and show how our change detection algorithm can be adapted to limit the impact of catastrophic events. We demonstrate a reduction in the false positive detection rate by using weighted empirical distributions of latency measurements that penalize outliers.

4.1 Introduction

Attendance on the Internet is cyclical: more people connect during daytime than during night time. Any server is bandwidth limited, hence very busy Internet infrastructures often produce sea-

sonal performance reflecting attendance oscillations. For this reason, latency measurements often exhibit strong mean and variance varying patterns as described in Chapter 3. But some networks may produce latency measurements with no such statistical properties. Busy infrastructures are more likely to display those varying patterns than less busy ones because high attendance will typically limit the bandwidth available per user hence reducing latency in periods of peak attendance. Distance to the server is also a decisive factor. For instance, a server located in Tokyo may display strong statistical properties when accessed from Taipei, but will lose most of them when accessed from Paris. Finally, the quality of the infrastructure is also an important factor. Powerful companies like Google or Facebook can build highly performing infrastructures that produce extremely stable latency that does not exhibit varying patterns through time, resulting in far less structured latency measurements. In those cases, sophisticated prediction algorithms are less important for two reasons. First, the gain in accuracy compared to simpler models is negligible. Second, building sophisticated models often requires large training sets and are computationally more intensive, meaning more data processing, hence cost more money. After defining those stable networks, we describe a way to assess simultaneously the reliability of the prediction model and the stability of the measurements.

4.1.1 Problem formulation

In this chapter we shall focus our attention on stable networks. A stable network will be precisely defined in section 4.2.2. Heuristically, stable networks are characterized by stationary median latency $(X^{\Delta}_{t_n})_{n\in\mathbb{N}}$ processes with low predictability. Recall from Chapter 3 that $X^{\Delta}_{t_n}$ is the median of all measurements with timestamps falling in the interval $]t_{n-1}, t_n]$, where $t_n = n\Delta$ and $\Delta > 0$. The process $(X^{\Delta}_{t_n})_{0\leq n\Delta\leq T}$ is called the median-process at frequency Δ.

More specifically, a stable network, hereinafter $\varepsilon-$stable network or $\varepsilon - SN$, will be defined as a network producing a purely non deterministic median process such that the ratio between the AVG 1-step ahead Root Mean Squared Error (RMSE) and the best ARMA(p, q) 1-step ahead RMSE is less than $1+\varepsilon$ where $\varepsilon > 0$ is intended to be small and where AVG refers to a prediction model that produces forecasts that are equal to the average of the last k observed values. The AVG prediction is defined formally in Definition 11, the performance criterion RMSE is defined in Definition 10 and the Best ARMA(p, q) prediction in Definition 12. Informally we have:

$$\frac{\text{AVG 1-step ahead RMSE}}{\text{Best ARMA}(p,q)\text{ 1-step ahead RMSE}} \leq (1+\varepsilon).$$

For such median latency processes, sophisticated predictive models are irrelevant: instead of training a predictive algorithm with a large history of data points, basic models trained with few data points perform marginally better for small ε. The reason to favor models that can be trained with little data is economic: building a prediction with a large training set requires more data processing than building a prediction with a small training set, hence costs more money. The reason is the following: *load-balancer*s typically rent computing power to companies like Amazon or Google and those companies charge for each Megabyte of data processed. The natural question hence reduces to what "few" data points is enough and in what sense?

In parallel, we want the predictive algorithm to quickly detect possible outages. In stable networks, the median latency process is expected to be stationary. In case of an outage, the distribution suddenly changes: the typical situation is latency going up and variance exploding. When

outages happen, sending users on that network will deteriorate the so called user experience. We adapt the ideas of Daniel Kifer et al. [52] on the detection of changes in the distribution of a data stream to our latency measurements to build a detecting change algorithm based on a comparison between two certain sliding windows of equal length N called the Ref and Shift windows, using the Wasserstein distance. The Ref window is a fixed snapshot of past performance when no outage occurred, reflecting expected normal behavior of the network, while the Shift window reflects present performance by ingesting new observations of the median process $(X^{\Delta}_{t_n})_n$ online. Each time a new observation is received the algorithm tests if an outage occurred by computing the Wasserstein distance between the two windows Ref and Shift. If it exceeds a certain threshold Q, an outage is reported. The impact of Q along with the number of data points N in the sliding windows are crucial: low values of Q or N will make the algorithm more sensitive to changes hence will potentially result in a large number of false positives. On the contrary, large values of Q or N will under estimate the severity of real outages and delay the detection, if not miss it completely, resulting in a false negative. A false positive in this context is defined has a detection of change at a timestamp when no outage occurred, and a false negative is defined has an absence of detection of change at a timestamp when an outage occurred.

The change detection algorithm naturally defines time periods of confidence and suspicion in the underlying model. When no alert is triggered, we should keep confidence in the fact that the network behaves in a normal fashion and the predictive model should be considered reliable. On the other hand, when an outage is detected, the users should stop being routed to that specific network immediately until normal conditions are met again.

The mathematical problem can be stated as follows: in stable networks, how should we select the minimum training sets in order to minimize the prediction error over the periods of confidence while having as few as possible false positive and negative outages detections? In other words, we want to build a predictive/change detection algorithm with the following properties:

1. Identify the periods of time where the network behaves abnormally.
2. Select the training sets with as few data points as possible.
3. Minimize the prediction error over the periods of confidence.
4. No false negative. Little tolerance for false positive.

4.1.2 Main results and organisation of the chapter

In section 4.2 we present in details the notion of $\varepsilon-$stable networks and the requirements of the algorithm for change detection and optimal training set selection. Namely we want an algorithm that selects the minimum number of training samples without impacting the performance beyond a tolerance $1+\tau$ of the best model, and we want to detect the outages when they happen with no false negative and as few as possible false positives. Note that τ and ε are different parameters. ε will measure the stability of the network, while τ will measure the loss in prediction accuracy resulting from training the prediction algorithm with fewer data points. In section 4.3 we describe the algorithm itself and how it should be used on latency measurement, with a special treatment for various aspects concerning the heaviness of the survival function and the properties of the distance used to detect changes along with how it impacts inference and estimation. Numerical results on

real data are presented in section 4.4. The algorithm is tested on three different stable networks and is proven to perform extremely well: all outages are quickly detected and no false positive are triggered. We show that the predictive algorithm used by Citrix may be fed with much fewer data points with very little accuracy loss.

4.2 Prediction in stable networks

4.2.1 The median process

Let $T > 0$ and $[0, T]$ be a time interval. Timestamps of measurements collected by Citrix are rounded up to the second hence a natural structure for the data generating process is that of a discrete-time stochastic process with the time index expressed in seconds. Formally, we observe the process $Z = \{Z_t | t \in \{0, \dots, T\}\}$, where Z_t is the empirical measure defined by:

$$Z_t = \frac{1}{N_t} \sum_{k=1}^{N_t} \delta_{Y_k^t}$$

where δ_x is the Dirac measure at $x \in \mathbb{R}$, N_t is the (random) number of data points collected at time t and $(Y_k^t)_{K \in \{1,\dots,N_t\}}$ are the observed latency measurements at time t. The specificity of the process Z is that a random number of measurements N_t is received at time t. Z is a process that outputs an empirical distribution at each time, instead of a single observation. The object of interest in the industry is the median process, not the raw measurements directly. We briefly recall the definition of the median process.

Definition 7 (Median latency process)**.** *Let $\Delta > 0$, $n \in \mathbb{N}$, $t_n = n\Delta$ and define the series:*

$$X_n^\Delta = X_{t_n}^\Delta = \text{Median}\left(Y_k^t\right)_{\substack{t \in]t_{n-1}, t_n] \\ k \in \{1,\dots,N_t\}}},$$

i.e. X_n^Δ is the median of all measurements with timestamps in the interval $]t_{n-1}, t_n]$. the process $(X_n^\Delta)_{0 \leq n \leq T/\Delta}$ is called the median process at frequency Δ.

The median process $(X_n^\Delta)_{0 \leq n \leq T/\Delta}$ was formally introduced in Section 3.2.3 of Chapter 3 and is illustrated in Figure 3.7 for different values of the parameter Δ. In this Chapter, we will focus on less structured processes. Predicting the median process instead of the underlying true data generating process is based on operational standards than rely on quantiles, especially the median, of the latency measurements. The reason is that latency measurements are generated by thousands of different users at a very high frequency, making a distributional approach that relies on the median more manageable and interpretable since the *load-balancer* can not use the latency measurements of a single user that needs to be routed, but instead only sees the whole distribution of latency measurements across all users at once, and the median is robust to outliers. We will tackle the problem of outliers in great details in section 4.3.3.

A major source of concern for *load-balancer*s is the ability to predict the median process in order to efficiently route the end users to the fastest network available. Any predictive model for the median process has a certain complexity which roughly corresponds to the number of computational operations needed to output a prediction, meaning that the underlying algorithm takes some time to output the said prediction. We are interested in this Chapter in networks

that have the particularity to produce median processes that have little statistical structure. For such networks, elementary models using as few training samples as possible shall perform only marginally less in terms of prediction accuracy to more sophisticated models. In the following sections, we will define mathematically the stable networks and how to optimize the training set selection.

4.2.2 Stable networks

Before introducing the notion of stable networks, we recall some facts about time series, see for instance [43].

Definition 8 (Deterministic process)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a second order process. For $t \in \mathbb{Z}$ let:*

$$H_{t-1} = \overline{Vect\{X_{t-1}, X_{t-2}, \ldots\}}$$

the closure in L^2 of the vector space $Vect\{X_{t-1}, X_{t-2}, \ldots\}$, i.e. all linear combinations of the form $\sum_{k=0}^{\infty} \lambda_k X_{t-k}$ that converge in L^2. Then we say that $(X_t)_{t\in\mathbb{Z}}$ is deterministic if and only if

$$X_t \in H_{t-1},$$

or in other words, $(X_t)_{t\in\mathbb{Z}}$ is deterministic if and only if

$$X_t = \text{proj}(X_t, H_{t-1})$$

where

$$\text{proj}(X_t, H_{t-1}) = \arg\min_{Y\in H_{t-1}} \|X_t - Y\|_2$$

is the orthogonal projection in L^2 of X_t onto the sub-vector space H_{t-1}.

Example 3 (Deterministic process)**.**
Let A, B two independent standard Gaussian random variables and $\theta \in]-\pi, \pi[$. Consider the harmonic process:

$$X_t = A\cos(\theta t) + B\sin(\theta t) \quad \text{for all } t \in \mathbb{Z}.$$

Then $(X_t)_{t\in\mathbb{Z}}$ is a deterministic process. Indeed, it can be seen that

$$X_t = 2\cos(\theta)X_{t-1} - X_{t-2} \in H_{t-1},$$

i.e. X_t is a linear combination of past observations.

Remark 5. *For a deterministic process $(X_t)_{t\in\mathbb{Z}}$, the prediction error, measured with the Root Mean Squared Error (RMSE), of the 1-step ahead forecast $\widehat{X_t} = \mathbb{E}(X_t|H_{t-1})$, is always 0.*

The following well-known theorem, due to Wold [3], gives a decomposition of weakly stationary processes.

Theorem 9 (Wold's Theorem)**.**
Let $(X_t)_{t\in\mathbb{N}}$ be a zero mean weakly stationary process. Then there exists two random processes $(\varepsilon_t)_{t\in\mathbb{N}}$ and $(d_t)_{t\in\mathbb{N}}$ and real numbers $(\psi_t)_{t\in\mathbb{N}}$ such that:

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} + d_t \quad \textit{for every } t \in \mathbb{Z},$$

where:

i) $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

ii) $(\varepsilon_t)_{t\in\mathbb{N}}$ *is a white noise process, i.e.* $\mathbb{E}(\varepsilon_t) = 0$ *and* $\mathbb{E}(\varepsilon_t\varepsilon_s) = \sigma^2 \mathbf{1}_{\{s=t\}}$.

iii) $(d_t)_{t\in\mathbb{N}}$ *is a deterministic process.*

iv) $\forall s, t$, $\mathbb{E}(d_s\varepsilon_t) = 0$.

Moreover, this decomposition is unique.

Definition 9 (Purely non deterministic process)**.**
Assuming $(X_t)_{t\in\mathbb{N}}$ is a zero mean weakly stationary process, then $(X_t)_{t\in\mathbb{N}}$ is said purely non deterministic if and only if $d_t = 0$ for all t, where $(d_t)_{t\in\mathbb{N}}$ is the deterministic process in the Wold's decomposition of $(X_t)_{t\in\mathbb{N}}$.

Example 4 (Purely non deterministic process)**.**
Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a collection of i.i.d. $\mathcal{N}(0, \sigma^2)$, and let ϕ such that $|\phi| < 1$. Consider the AR(1) process:

$$X_t = \phi X_{t-1} + \varepsilon_t.$$

It is well known that X_t has a MA(∞) representation:

$$X_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}.$$

Because Wold's decomposition is unique, it follows that X_t is purely non deterministic.

It immediately follows from the Wold's decomposition that any purely non deterministic process can be arbitrarily approached with an ARMA(p, q) process. Indeed, let $(X_t)_{t\in\mathbb{N}}$ be a purely non deterministic process, then by Wold's theorem we have:

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} = \psi(L)\varepsilon_t$$

where L is the lag operator such that $LX_t = X_{t-1}$. Then it is easily verified that there exists two polynomials θ and ϕ with finite degree such that

$$\psi(L) \approx \frac{\theta(L)}{\phi(L)}$$

with arbitrary precision. It immediately follows that the process $(Y_t)_{t\in\mathbb{N}}$ solution of:

$$\phi(L)Y_t = \theta(L)\varepsilon_t$$

approaches $(X_t)_{t\in\mathbb{N}}$ with arbitrary precision. Conversely, if an ARMA(p,q) is causal and invertible, then it has an MA(∞) representation [16].

Now, if $(X_t)_{t\in\mathbb{Z}}$ is a stationary process with auto-covariance function γ satisfying:

$$\sum_{h\in\mathbb{Z}}|\gamma(h)|<\infty$$

then $(X_t)_{t\in\mathbb{N}}$ has a spectral density given by:

$$f(\nu)=\sum_{h\in\mathbb{Z}}\gamma(h)e^{-2\pi ih\nu}.$$

A necessary and sufficient condition for $(X_t)_{t\in\mathbb{N}}$ to be purely non deterministic is given in the following theorem du to Kolmogorov [53]:

Theorem 10 (Kolmogorov)**.** *Let $(X_t)_{t\in\mathbb{N}}$ be a second order process with auto-covariance function γ. The process $(X_t)_{t\in\mathbb{N}}$ is purely non deterministic if and only if the following conditions hold:*

i) *F_X is absolutely continuous with respect to the Lebesgue measure,*

ii) *f_X is positive almost everywhere,*

iii) *$\log f_X$ is integrable,*

where F_X and f_X are the spectral distribution and density respectively of $X=(X_t)_{t\in\mathbb{N}}$, meaning that F_X is the cumulative distribution function of the measure whose Fourier coefficients are $\gamma(h)$, $h\in\mathbb{Z}$, i.e. the spectral measure of $(X_t)_{t\in\mathbb{N}}$:

$$\gamma(h)=\int_{-1/2}^{1/2}e^{2\pi i\nu h}F_X(d\nu)$$

and

$$f_X(\nu)=\sum_{h\in\mathbb{Z}}\gamma(h)e^{-2\pi i\nu h}.$$

As we will see below in Proposition 16, $f_X>0$ everywhere implies $\log(f_X)$ integrable. Richard Bradley [15] derived a necessary and sufficient condition for a spectral density of a weakly stationary process to be positive that involves linear dependence coefficients. The condition is technical and up to our knowledge, no statistical test is proposed in the literature for positivity of the spectral density. By visual inspection of the estimated spectral density obtained by computing the Lomb-Scargle Periodogram [61], [80] we derive good intuition on the positivity of the underlying spectral density.

If the spectral density has a zero at some frequency ω, the estimated power spectrum is be expected to show a significant drop at that frequency, see example 5 and Figure 4.2 for instance. The flatness of the estimated power of the process median-process in Figure 4.1 supports the assumption that the underlying spectral density is positive everywhere. This motivates the fact that the median latency processes can be assumed to be be purely non deterministic and accurately descried as ARMA(p,q).

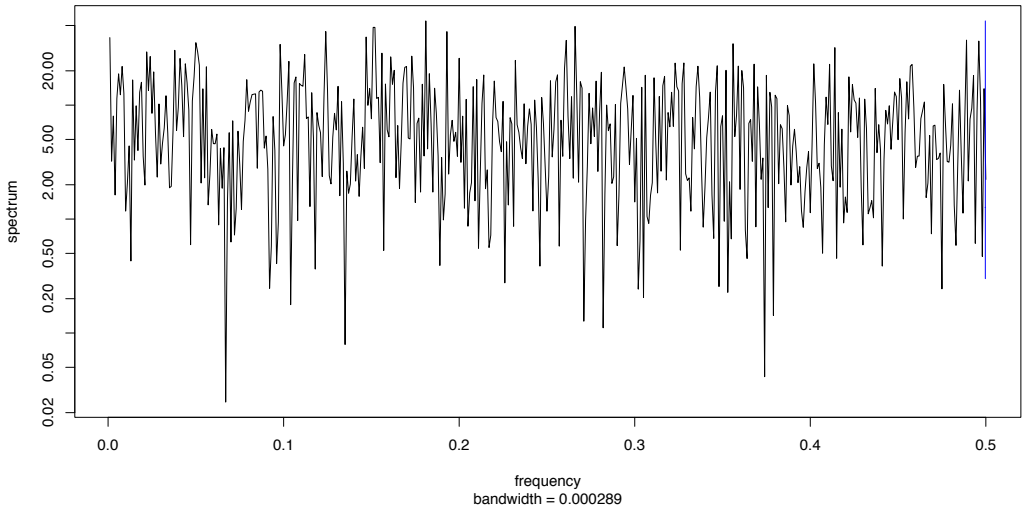

Figure 4.1: *Estimated spectral density for the median process* $(X_n^{\Delta})_{n\in\mathbb{N}}$. *The frequency domains ranges from 0 to 1/2, and the power estimates are in* log*-scale.*

Example 5. *Suppose* $X = (X_t)_{t\in\mathbb{Z}}$ *is a moving average process that satisfies:*

$$X_t = \frac{\varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t}{3}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is a Gaussian white noise with mean 0 and variance* σ^2. *It is easily seen that the auto-covariance function of* X *satisfies:*

$$\gamma_X(h) = \frac{\sigma^2}{9}(3 - |h|) \quad \textit{for } |h| \leq 2, \textit{ and 0 otherwise.}$$

The auto-covariance function is summable, hence it follows that the spectral density f_X *exists and satisfies:*

$$f_X(\nu) = \frac{\sigma^2}{9}(3 + 4\cos(2\pi\nu) + 2\cos(4\pi\nu)).$$

On the interval $[0, 1/2]$, f_X *has exactly 1 zero, at* $\nu = 1/3$. *See Figure 4.2.*

We will make the following assumptions regarding $(X_t^{\Delta})_{t\in\mathbb{Z}}$.

Assumption 3. *The median process* $(X_t^{\Delta})_{t\in\mathbb{Z}}$ *is stationary.*

Assumption 4. *The auto-covariance function of* $(X_t^{\Delta})_{t\in\mathbb{Z}}$, *denoted* γ_X, *is summable, i.e.:*

$$\sum_{h\in\mathbb{Z}} |\gamma_X(h)| < \infty$$

Assumption 5. $f_X^{\Delta}(\nu) > 0$ *for all* $\nu \in [-1/2, 1/2]$

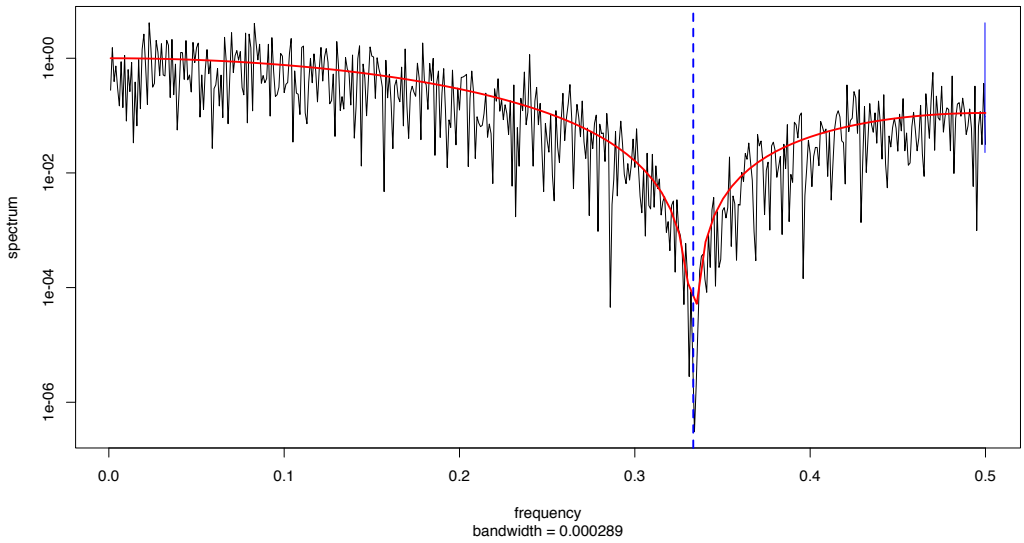

Figure 4.2: *Estimated spectral density of* $X = (X_t)_{t\in\mathbb{Z}}$ *where* $X_t = (\varepsilon_{t-2}+\varepsilon_{t-1}+\varepsilon_t)/3$ *(solid black) along with the true spectral density (solid red) in* log *scale. The frequency domains ranges from 0 to 1/2. Vertical dashed blue is the line* $x = 1/3$*. To be compared to 4.1 in order to support the assumption the our median process has positive spectral density everywhere.*

Assumption 4 implies that the spectral density of $(X_t^\Delta)_{t\in\mathbb{Z}}$ exists, and satisfies:

$$f_X^\Delta(\nu) = \sum_{h\in\mathbb{Z}} \gamma_X(h) e^{-2i\pi h\nu}.$$

Proposition 16. *Under Assumption 3,4,5,* $(X_t^\Delta)_{t\in\mathbb{Z}}$ *is purely non deterministic.*

Proof. We know that $(X_t^\Delta)_{t\in\mathbb{Z}}$ has a positive spectral density everywhere. In particular, the spectral density is continuous, hence there exists $m > 0$, $f_X^\Delta(\nu) > m$ for all $\nu \in [-1/2, 1/2]$. In addition, since γ_X is summable, there exists $M > 0$ such that $f_X^\Delta(\nu) < M$ for all $\nu \in [-1/2, 1/2]$. It immediately follows that $\log(f_X)$ is integrable, hence by Kolmogorov's theorem, $(X_t^\Delta)_{t\in\mathbb{Z}}$ is purely non deterministic. □

Definition 10 (RMSE)**.** *Let* $(X_t)_{t\in\mathbb{Z}}$ *be a second order time series. Let* $\widehat{X_{t+1}}$ *be any measurable function with respect to* $\sigma(X_s | s \le t)$*. The prediction error associated with* $\widehat{X_{t+1}}$ *is defined as:*

$$RMSE(\widehat{X_{t+1}}) = \sqrt{\mathbb{E}\left[(\widehat{X_{t+1}} - X_{t+1})^2\right]}.$$

Definition 11 (AVG prediction)**.**
Let $(X_t)_{t\in\mathbb{Z}}$ *be a second order time series and* $k \in \mathbb{N}$*. For all* $t \in \mathbb{Z}$ *let:*

$$\widehat{X_{t+1}}^{AVG} = \frac{1}{k}\sum_{i=1}^{k} X_{t-i+1}$$

The AVG prediction is the simple average of the last k measurements.

Let $(X_t)_{t\in\mathbb{Z}}$ be a second order time series. For varying $p, q \in \mathbb{N}$, fitting an ARMA(p, q) for $(X_t)_{t\in\mathbb{Z}}$ means estimating the following model:

$$X_t = u_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j u_{t-j}$$

where $(u_t)_t$ is a Gaussian white noise.

Definition 12 (optimal ARMA prediction)**.**
Let $(X_t)_{t\in\mathbb{Z}}$ be a second order time series. For varying $p, q \in \mathbb{N}$, the following model is estimated through maximum likelihood:

$$X_t = u_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j u_{t-j}$$

where $(u_t)_t$ is a Gaussian white noise. The best model is chosen through AIC minimization [2]. The optimal ARMA prediction is defined as the 1-step ahead prediction from that best model:

$$\widehat{X_{t+1}}^{ARMA} = \sum_{i=1}^{p} \hat{\phi}_i X_{t-i} + \sum_{j=1}^{q} \hat{\theta}_j \hat{u}_{t-j}$$

where the $\widehat{u}_t$ are the estimated residuals and the parameters ϕ_i, θ_j are estimated by minimizing the residuals some of squares.

Remark 6. *$\widehat{X_{t+1}}^{ARMA}$ is the natural estimator of X_{t+1} conditional on $\mathcal{F}_t = \sigma(X_s, u_s | s \leq t)$. Indeed, it is easily seen that the optimal forecast, in the sense of minimizing the mean squared error, is:*

$$\mathbb{E}(X_{t+1}|\mathcal{F}_t) = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j u_{t-j}.$$

In other words, $\widehat{X_{t+1}}^{ARMA}$ is an estimation of the optimal predictor given all past information (conditional expectation).

We are now ready to define a notion of $\varepsilon-$stable network at the scale $\Delta > 0$.

Definition 13 ($\varepsilon-$stable network)**.**
Let N be a Network. Let $(X_t^{\Delta})_{t\in\mathbb{Z}}$ be the median process at scale $\Delta > 0$ produced by N. We say that N is a $\varepsilon-$stable network, or $\varepsilon - SN$, if $(X_t^{\Delta})_{t\in\mathbb{Z}}$ is a purely non deterministic process and

$$RMSE\left(\widehat{X_{t+1}}^{AVG}\right) \leq (1+\varepsilon) RMSE\left(\widehat{X_{t+1}}^{ARMA}\right)$$

In other words, an $\varepsilon-$stable network generates purely non deterministic median processes with a relative error between the optimal ARMA prediction and the AVG prediction bounded by $1+\varepsilon$. Using remark 6 above, one sees in particular that an $\varepsilon - SN$ is such that the optimal forecast using all the history of the process leads only marginal improvement over the optimal forecast using only immediate history.

Example 6. *The existence of a $\varepsilon - SN$ is not immediate. In particular for very small choices of $\varepsilon > 0$. For instance, consider an AR(1) process:*

$$X_t = \phi X_{t-1} + u_t$$

where $|\phi| < 1$ and u_t are i.i.d. $\mathcal{N}(0, \sigma^2)$. This process is purely non deterministic stationary. The optimal 1-step ahead forecast is $\widehat{X_{t+1}}^{ARMA} = \mathbb{E}(X_{t+1}|\mathcal{F}_t) = \phi X_t$. The optimal AVG prediction is $\mathbb{E}\left(k^{-1}\sum_{i=1}^{k} X_{t-i}\right) = 0$. As a consequence we have:

$$\begin{aligned} RMSE\left(\widehat{X_{t+1}}^{AVG}\right)^2 &= \mathbb{E}\left[(X_{t+1} - 0)^2\right] \\ &= \mathbb{V}(X_{t+1}) \\ &= \frac{\sigma^2}{1-\phi^2}, \end{aligned}$$

and:

$$\begin{aligned} RMSE\left(\widehat{X_{t+1}}^{ARMA}\right)^2 &= \mathbb{E}\left[(X_{t+1} - \phi X_t)^2\right] \\ &= \mathbb{E}\left[u_{t+1}^2\right] \\ &= \sigma^2 \end{aligned}$$

So $(X_t)_t$ is $\varepsilon-$stable if and only if

$$\frac{\phi^2}{1-\phi^2} \leq \varepsilon.$$

In other words, the larger the autoregressive term, the larger ε and vice and versa. The autocorrelation structure of the process constrains the parameter ε.

4.2.3 Training set in $\varepsilon - SN$

Suppose we observe a median process $(X_t^{\Delta})_{t\in\mathbb{Z}}$ from an $\varepsilon - SN$. By definition, predicting $(X_t^{\Delta})_{t\in\mathbb{Z}}$ with sophisticated models only improves by at most ε the accuracy over the AVG prediction. The reason why *load-balancers* are willing to give up marginal gain in prediction accuracy is because sophisticated models need larger training sets, hence require more processing and computing, hence are more expensive. If one wants to fit an ARMA process, following on Box and Tiao rule of thumb [14], at least 100 data points are needed. If one is interested in the median process $X_{t_n}^{\Delta}$ with $\Delta = 5$min, by construction of the median-process, 100×5 minutes $= 8$ hours and 20 minutes of history must be considered at least. For such models, updates are needed at every new periods, in particular for calculating the estimated residuals. For networks producing measurements at a very high frequency, the problem can become computationally very intensive. The point of developing a notion of $\varepsilon - SN$ is to identify networks in which accuracy can be traded for computational efficiency without significant loss, resulting in potential savings in computing power.

For reasons of confidentiality, the predictive algorithm for $\varepsilon - SN$ that we developed at Citrix will be treated as a black box. At time t, denote $\widehat{X_t^{\Delta}}^{Citrix}$ the associated prediction. We will only treat the determination of the optimal training set. This predictor for the median latency uses at most the last N more recent latency measurements provided they were received within the last M minutes, even if that means ending up with less than N measurements. We shall write

$$\widehat{X_t^{\Delta}}^{Citrix} = \widehat{X_{t,N,M}^{\Delta}}^{Citrix}.$$

We will discuss the problem of tuning (N, M) that meet the requirements of Section 4.4.3 such that we choose the smallest possible training set possible without impacting the accuracy beyond a certain threshold.

The value for Δ will be fixed in this chapter. We choose $\Delta = \lambda$, where λ is the prediction Time To Live or TTL. A TTL is a generic term to quantify the lifespan of any data that is stored for a finite amount of time in a network before being updated. It is used to prevent the predictive algorithms to update the predictions on real time, which would be computationally too intensive. This means that if a prediction is updated at time t, every new users coming between times t and $t + \lambda$ will use the same prediction. Only at time $t + \lambda$ will the prediction be updated. Hence a prediction at time t with lifespan λ must predict the median value for latency measurements over the interval $[t, t + \lambda]$. $\lambda = 60$s is the default value set up by the engineers.

4.3 Detecting outages

A big source of concern for *load-balancer*s are outages. The reasons are numerous: failure, maintenance, system updates etc. and can not be anticipated. For *load-balancing* purposes it is a major concern to be able to react quickly: for as long as an outage is not detected, users can potentially be sent to a down server, failing to providing the content. The data generating process in stable networks produce measurements whose underlying distribution changes abruptly at random times when an outage occurs Outages are characterized by a sudden increase in the mean and variance of the median process $(X_n^{\Delta})_{n\in\mathbb{N}}$, see Figure 4.3. The goal is to be able to detect the moments where such changes occur. In this section we build an algorithm to detect changes in the distribution of the median latency process $(X_n^{\Delta})_{n\in\mathbb{N}}$.

Difficulties arise from the fact that the distribution of latency measurements exhibit a power law behavior with small tail index. Such distributions are known for generating catastrophic events with high probability. The task of detecting outages in this context is challenging because outliers may wrongly suggest a change in the distribution of latency measurements. Following up on the methodology developed by Kifer [52], we build an algorithm based on the Wasserstein distance between two certain sliding windows, and propose to modify the W_p distance by applying specific weights to reduce the number of false positives. We will intensively discuss the outliers issue and how it impacts the detection of changes when using the W_p distance.

Since detection of changes in the median latency process is based on the comparison between two sliding windows, it reduces to the problem of testing whether or not two sample distributions are "close". The two sliding windows will be called Ref and Shift windows hereinafter. The Ref

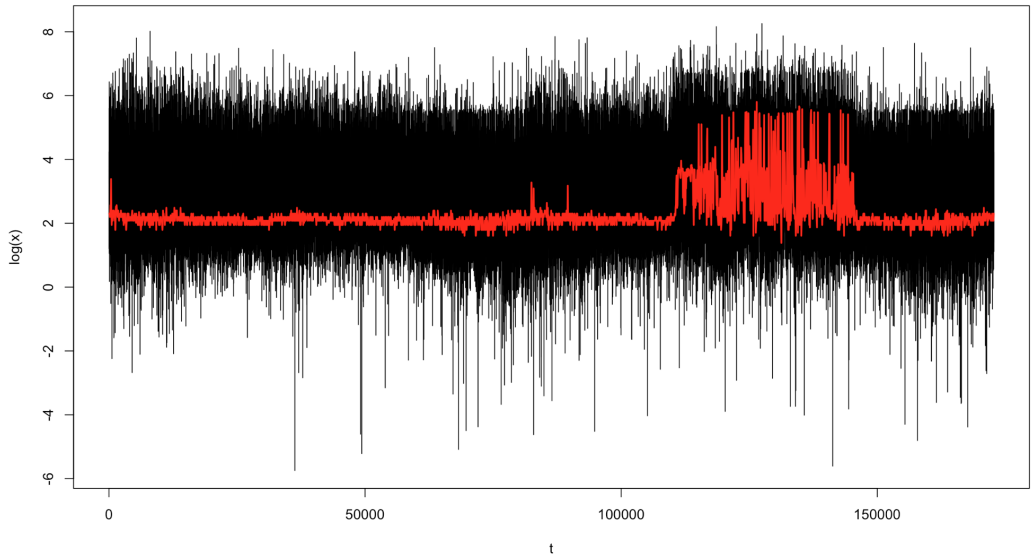

Figure 4.3: *Example of an outage. Raw measurements in black and median-process in red. The consequence of the outage is a sudden increase in both mean and variance of the measurements. The* $y-$*axis is in* log *scale*

window consists of measurements that came in in recent past and reflect what was the last recorded normal situation. The Shift window consists of new measurements being ingested online. Each time the Shift window gets a new data point, a certain distance between the two samples will be measured. As soon as this distance between Ref and Shift exceeds a certain threshold, we declare that the distribution has changed, see Figure 4.4.

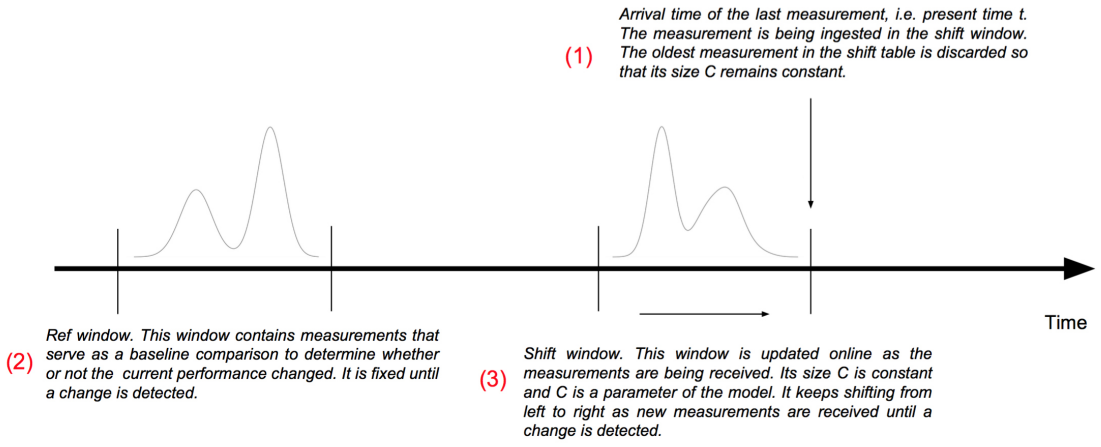

Figure 4.4: *Illustration of the Ref and Shift windows.*

4.3.1 Algorithm description

Formally, let $(Z_t)_{t\in\mathbb{N}}$ be a sequence of observations, typically Z_t will represent the median of latency measurements at time t. For each $t \in \mathbb{N}$, suppose $Z_t \sim P_t$ where P_t is some probability distribution. In a stable network, all distributions P_t are expected to be close, in the sense that there for any $s, t \in \mathbb{N}$, $W_p(P_t, P_s)$ must be bounded by a positive constant M. An outage will be said to have happened at time t if the distance $W_p(P_{t-1}, P_t)$ exceeds M, where W_p is the Wasserstein distance of order p, see equation 4.1. In order to estimate instants when the distribution changes in the data stream, we make online comparison between the first C measurements received after the last known outage (the Ref window), and the last C measurements received in the data stream (the Shift) window. In practice, the true distribution of measurements is unobserved, the W_p distance will be computed between the empirical distributions, and will be called the empirical Wasserstein distance, denoted $W_{p,n}$, see equation 4.2.

The algorithm takes as input two parameters: C the number of data points in each window, and $\beta > 0$ a scaling parameter that will quantify the sensitivity of the detection. The algorithm is divided in two independent stages. First it estimates the maximum W_p distance between two windows of length C via Monte-Carlo simulations by randomly selecting such windows on a training set where no outage happened. Denote by Q the resulting maximum value. The value for Q is typically computed once a day. The second stage is the online procedure for change detection. At current time t_0, the algorithms initializes the Ref and Shift windows, and starts sliding the Shift window by 1 point. If $W_p(\text{Ref}, \text{Shift}) > \beta Q$, the current time is stored and a change is said to have been detected. After a change is detected, both windows are cleared and need to be re-initialized. During this process the algorithm declares the measurements unstable and the underlying predictive model $\widehat{X_t^{\Delta}}^{Citrix}$ not trustworthy, until a new Ref and Shift windows are built and verify $W_p(\text{Ref}, \text{Shift}) \leq \beta Q$. The union of all instants such that the model should not be trusted will be referred to the Red zone. If no outage is detected, confidence in the model at current time is kept until the next comparison. The union of all instants such that the model should be trusted will be referred to the Green zone.

Initialization;
Ref = First C points of stream;
Shift = Next C points of stream ;
while *not at end of stream* **do**
  Slide Shift by 1 point;
  if $W_{p,n}(Ref, Shift) > \beta Q$ **then**
    $t_0 \leftarrow$ current time ;
    Report change at t_0;
    Ref = First C points starting at t_0;
    Shift = First C points starting at $t_0 + C$;
  else
    $t_0 \leftarrow$ current time ;
    Report confidence in model at time t_0;
  end
end

Algorithm 2: Detecting change in a data stream.

A natural choice is $\beta \approx 1$. In this case, changes are declared when the distance between Ref and Shift exceeds the maximum value of the W_p distance between two windows of length C recorded during a period without outages. It is easily seen that small values for β will make the algorithm likely to trigger false positives, artificially increasing the time spent in the Red zone. Indeed, setting $\beta = 0$ will lead to $W_{p,n}(\text{Ref}, \text{Shift}) > \beta Q$ for all windows Ref and Shift, declaring a change at all times. On the contrary, large values of β will make it less likely to detect a change, increasing the number of false positives and the time spent in the Green zone. The parameter C is also decisive: small values will make the algorithm quicker to react to a change, but at the cost of W_p distance estimates with more fluctuation, hence potential false positives. Before presenting the results, we shall discuss the choice of the distance and recall some basic facts, see for instance [90].

4.3.2 The Wasserstein distance

The classical Wasserstein distance

Let $p \geq 1$. The W_p distance between two distribution P, Q on $\mathbb{R}$ is defined as:

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int |x-y|^p \pi(dx,dy) \right)^{1/p} \tag{4.1}$$

where $\Pi(P,Q)$ is the set of probability measures with marginal P and Q respectively and $\|\cdot\|$ is the Euclidean norm. When P, Q are probability distributions over the real line, W_p has the closed form:

$$\begin{aligned} W_p(P,Q) &= \left(\int_{\mathbb{R}} |F(x) - G(x)|^p dx \right)^{1/p} \\ &= \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p} \end{aligned}$$

where F, G are the c.d.f. of P and Q respectively, and F^{-1}, G^{-1} their generalized inverse. If one observes 2 samples, $(X_1, \ldots, X_n)$ independent with common distribution F, and $(Y_1, \ldots, Y_n)$ independent with common distribution G, the empirical Wasserstein distance between the two samples is defined defined as the Wasserstein distance between the two empirical distributions:

$$W_{p,n}(F,G) := W_p(F_n, G_n) = \left(\int_0^1 |F_n^{-1}(u) - G_n^{-1}(u)|^p du \right)^{1/p} \tag{4.2}$$

where $F_n(u) = n^{-1} \sum_{i=1}^n 1_{\{X_i \leq u\}}$ and $G_m(u) = n^{-1} \sum_{j=1}^m 1_{\{Y_j \leq u\}}$ are the empirical cumulative distributions and F_n^{-1} and G_n^{-1} denote their general inverse, or empirical quantile function. The empirical Wasserstein distance has the closed form:

$$W_{p,n}(F,G) = \left(\frac{1}{n} \sum_{i=1}^n |X_{(i)} - Y_{(i)}|^p \right)^{1/p}$$

where $(X_{(i)})_{1 \leq i \leq n}$ and $(Y_{(i)})_{1 \leq i \leq n}$ are the order statistics of the two samples, ie $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$.

The Wasserstein distance is a natural candidate for change detection because compared to other classical metrics over probability distributions, this distance preserves in some sense the underlying geometry of the space. For instance, consider the collection $\{F_a | F_a \sim \mathcal{U}(a, a+1), a \in \mathbb{R}\}$ of uniform distributions over line segments of length 1. Then for any $a, b \in \mathbb{R}$ and $p \geq 1$ we have:

$$\begin{aligned} W_1(F_a, F_b) &= \int_0^1 |F_a^{-1}(u) - F_b^{-1}(u)| du \\ &= |a - b|. \end{aligned}$$

It is easily seen that other classical metrics like TV, L_2, Hellinger or χ^2 return a constant value for any choice of distributions F_a, F_b such that $|a-b| > 1$ while we would intuitively say that F_0 and F_2 are "closer" than F_0 and F_3. The W_p distance solves an optimal transport problem, meaning that this distance represents the effort needed to map one distribution into the other and is well adapted to measure distances from distribution to another. One drawback of the Wasserstein distance is that it is not a robust statistic [78] in the sense that its breakdown point is 0.

Definition 14. *Let $\boldsymbol{X}^n = (X_1 \ldots, X_n)$ be a collection of random variables. Denote by $P_n = n^{-1} \sum_{i=1}^n X_i$ its empirical distribution. Let $Q_{n,k}$ be the empirical distribution of a sample obtained from $\boldsymbol{X}^n$ after replacing at most k of the X_i's with arbitrary values. For T a functional of P_n, the breakdown point of the statistic $T(P_n)$ is defined as:*

$$BP(\boldsymbol{X}^n, T) = \frac{1}{n} \min \left\{ 1 \leq k \leq n, \sup_{Q_{n,k}} |T(P_n) - T(Q_{n,k})| = \infty \right\}.$$

$BP(\boldsymbol{X}^n, T)$ corresponds to the minimum proportion of outliers the statistic T can handle before taking arbitrarily large values.

Example 7 (Empirical Mean)**.** *Let $\boldsymbol{X}^n = (X_1 \ldots, X_n)$ an i.i.d sample and let $P_n = n^{-1} \sum_{i=1}^n X_i$ its empirical distribution. The sample mean*

$$T(P_n) = \frac{1}{n} \sum_{i=0}^n X_i$$

has a breakdown point of 1/n. Indeed, let $\boldsymbol{Y}^n = (y, X_2, \ldots, X_n)$ a sample obtained from $\boldsymbol{X}^n$ after replacing X_1 with an arbitrary real number y. Denote $Q_{n,1}$ its empirical distribution. Then:

$$|T(P_n) - T(Q_{n,1})| = \frac{1}{n} |y - X_1| \longrightarrow \infty$$

as $y \to \infty$. Meaning that $BP(\boldsymbol{X}^n, T) = 1/n$.

Example 8. *Let $\boldsymbol{X}^n = (X_1 \ldots, X_n)$ an i.i.d sample and let $P_n = n^{-1} \sum_{i=1}^n X_i$ its empirical distribution. The sample median*

$$T(P_n) = X_{(\lceil n/2 \rceil)}$$

has a breakdown point of $\lceil n/2 \rceil / n$. Suppose n is odd, i.e. $\exists q \in \mathbb{N}$, $n = 2q+1$. Then $\lceil n/2 \rceil = q+1$. Let $\boldsymbol{Y}^n = (y_1, y_2, \ldots, y_k, X_{k+1}, \ldots, X_n)$ a sample obtained from $\boldsymbol{X}^n$ after replacing the first k

values with a arbitrary real number $y_1, \dots, y_k$. Suppose $k = q + 1$ and that $y_1 < y_2 < \cdots < y_{q+1}$. For y_1 large enough, we have:

$$|T(P_n) - T(Q_{n,1})| = |y_1 - X_{(\lceil n/2 \rceil)}| \longrightarrow \infty$$

as $y_1 \to \infty$. Meaning that $BP(\boldsymbol{X}^n, T) \geq (q+1)/n = \lceil n/2 \rceil / n$. Now if $k = q$, it is easily seen that for any choice of replacement of q values in $\boldsymbol{X}^n$, there exists $i < j$ such that $T(Q_{n,q})$ is bounded between $X_{(i)}$ and $X_{(j)}$, hence $|T(P_n) - T(Q_{n,1})|$ is almost surely finite. Hence $BP(\boldsymbol{X}^n, T) = \lceil n/2 \rceil / n$.

Remark 7. *The breakdown point is often considered in the limit, hence we say that the mean (resp median) has a breakdown point of 0 (resp 1/2). It is easily seen that the Wasserstein distance also has a breakdown point of 0: the statistic is not robust to outliers.*

Internet latency measurements have heavy tailed distributions. For heavy tailed distributions, catastrophic events occur with high probability, and statistics with low breakdown point will behave poorly in this case. We will describe in greater details those stylized facts in the next section. In order to cope with this undesirable property, we can apply weights to the Wasserstein distance.

The weighted Wasserstein distance

In the case where a small number of particularly high latency measurements would be received in a short period of time, the algorithm described above with the classic Wasserstein distance may trigger a false positive. In what follows, we will denote $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ two random samples, that should be thought of as the Ref and Shift windows. One solution to handle outliers is to replace the equal wieghts $\frac{1}{n}$ in the W_p distance definition by adaptive weights to emphasize on the center of the distribution. That is, we propose the weighted W_p distance:

$$\overline{W}_{n,p}(\boldsymbol{X}, \boldsymbol{Y}) = \left(\int_0^1 |F^{-1}_{w^{\boldsymbol{X}}}(u) - G^{-1}_{w^{\boldsymbol{Y}}}(u)|^p du \right)^{1/p}$$

where:

$$F_{w^{\boldsymbol{X}}}(u) = \sum_{i=1}^{n} w_i^{\boldsymbol{X}} 1_{\{X_i \leq u\}} \text{ and } G_{w^{\boldsymbol{Y}}}(u) = \sum_{j=1}^{m} w_j^{\boldsymbol{Y}} 1_{\{Y_j \leq u\}} \tag{4.3}$$

are the weighted empirical cumulative distributions functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The weights $w^{\boldsymbol{X}}$ and $w^{\boldsymbol{Y}}$ satisfy:

$$w_i^{\boldsymbol{X}} \geq 0,\ w_j^{\boldsymbol{Y}} \geq 0,\ \sum_{i=1}^{n} w_i^{\boldsymbol{X}} = 1,\ \sum_{i=1}^{m} w_i^{\boldsymbol{Y}} = 1.$$

A possible choice for the weights is an exponentially decay in the distance to a given sample quantile. Let:

$$\begin{cases} u^{\boldsymbol{X}}_{\pi(i)} = e^{-\lambda \left| X_i - q^{\boldsymbol{X}}_\beta \right|} \\ u^{\boldsymbol{Y}}_{\sigma(j)} = e^{-\lambda \left| Y_j - q^{\boldsymbol{Y}}_\beta \right|} \end{cases}$$

where π and σ are permutations of $\{1, \dots, n\}$ such that $X_{\pi^{-1}(1)} \leq \dots \leq X_{\pi^{-1}(n)}$ and $Y_{\pi^{-1}(1)} \leq \dots \leq Y_{\pi^{-1}(n)}$, $\lambda > 0$ and $q^{\boldsymbol{X}}_\beta, q^{\boldsymbol{Y}}_\beta$ are the empirical quantiles of order $\beta \in (0, 1)$ of X^n and Y^n respectively. And define the weights as :

$$\begin{cases} w_i^{\boldsymbol{X}} = \frac{u_i^{\boldsymbol{X}}}{\sum_j u_j^{\boldsymbol{X}}} \\ w_i^{\boldsymbol{Y}} = \frac{u_i^{\boldsymbol{Y}}}{\sum_j u_j^{\boldsymbol{Y}}} \end{cases}$$

Those weights penalize observations in the samples that are too far from a given sample quantile.

Remark 8. *In the case where $w_i/\min_j w_j \in \mathbb{N}$ for all $1 \leq i \leq n$, by factoring out by $\min_j w_j$ in Eq 4.3 it can be observed that the weighted c.d.f. coincides exactly with the unweighted c.d.f. of a sample of size $1/\min_j w_j$ that is obtained after duplicating the i-th order statistic from the original sample $w_i/\min_j w_j$ times. Indeed, let $(X_1, \ldots, X_n)$ be an i.i.d. sample with cumulative distribution F. The unweighted empirical c.d.f. is:*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}.$$

Now suppose we are given weights $w_i \geq 0$ for all $1 \leq i \leq n$ such that $\sum_i w_i = 1$. the weighted empirical c.d.f. is defined as:

$$F_{n,w}(x) = \sum_{i=1}^{n} w_i \mathbf{1}_{\{X_i \leq x\}}.$$

Now let $\lambda_i = w_i/\min_j w_j$ and suppose $\lambda_i \in \mathbb{N}$ for all $1 \leq i \leq n$. Let $N = \sum_i \lambda_i$. Because the weights sum to 1 we have:

$$N = \sum_{i=1}^{n} \lambda_i \ = \sum_{i=1}^{n} \frac{w_i}{\min_j w_j} = \frac{1}{\min_j w_j}.$$

Now rewrite the empirical weighted c.d.f.:

$$F_{n,w}(x) = \sum_{i=1}^{n} w_i \mathbf{1}_{\{X_i \leq x\}} = \min_j w_j \sum_{i=1}^{n} \frac{w_i}{\min_j w_j} \mathbf{1}_{\{X_i \leq x\}} = \frac{1}{N}\sum_{i=1}^{n} \lambda_i \mathbf{1}_{\{X_i \leq x\}}.$$

Now let $Y^N = (Y_1, \ldots, Y_N)$ be a sample obtained after duplicating each X_i λ_i times. Its empirical c.d.f. G_N satisfies:

$$G_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{Y_i \leq x\}} = \frac{1}{N}\sum_{i=1}^{n} \lambda_i \mathbf{1}_{\{X_i \leq x\}} = F_{n,w}(x)$$

by definition of the sample Y^N. This is the heuristic behind the weighted W_p distance: observations closer to a given quantile will weight more than outliers, hence reducing the influence of the latter.

Let us analyze the effect of weights on a toy example for the estimation of the Wasserstein distance. See Figures 4.5 and 4.6 for an illustration of the weight function.

Using a Monte-Carlo simulation we estimated the distribution of the empirical W_p distance between two samples i.i.d. log-normal $\mathcal{LN}(0,1)$ of size $n = 100$ by computing it $K = 10.000$ times with and without weights, wee Figure 4.7. Recall that if X has a log-normal $\mathcal{LN}(0,1)$ distribution, then $Y = \log(X)$ has the standard normal distribution $\mathcal{N}(0,1)$. Applying the weights on a right skewed distribution to emphasize on the median has the effect to reduce the empirical distance: the right tail is more penalized than the left tail because of the skewness, hence reducing the median. The weights are designed as to give flexibility for operational teams along with robustness because latency measurements are contaminated by outliers. The distribution of latency measurements in stable networks can be modeled with regularly varying tails. For such distributions, using weights in the Wasserstein distance may help reduce the number of false positive when detecting outages.

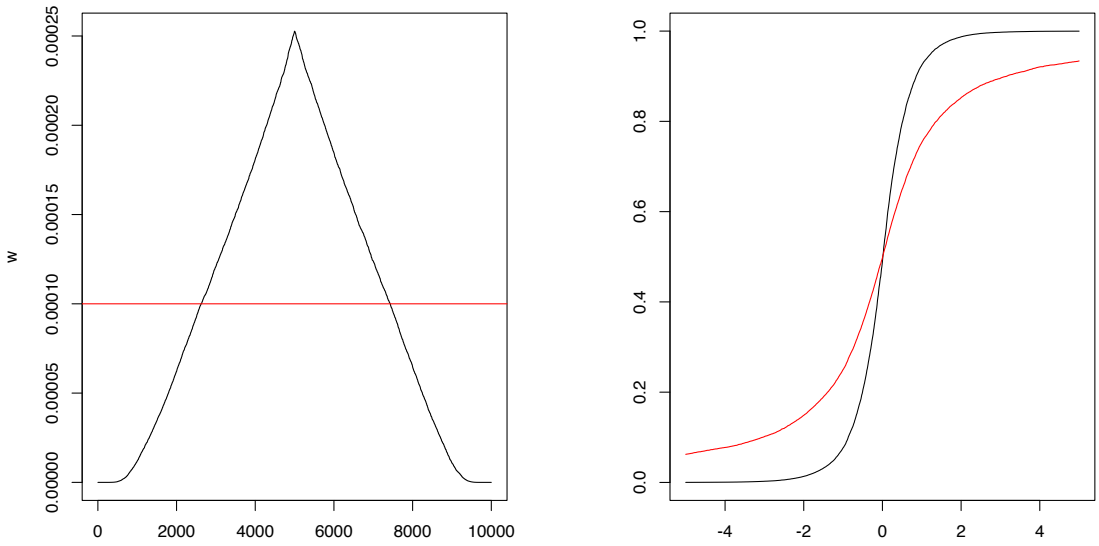

Figure 4.5: *Left: plot of the weights (black) with* $\lambda = 1$, $\beta = 1/2$ *for a sample of* $n = 1e4$ *Cauchy* $\mathcal{C}(0, 1)$ *random variables, along uniform weights* $1/n$ *(red) against the order statistics. Right: Weighted c.d.f. (black) versus unweighted c.d.f. The weights decrease the tails, and put extra mass in the center of the distribution.*

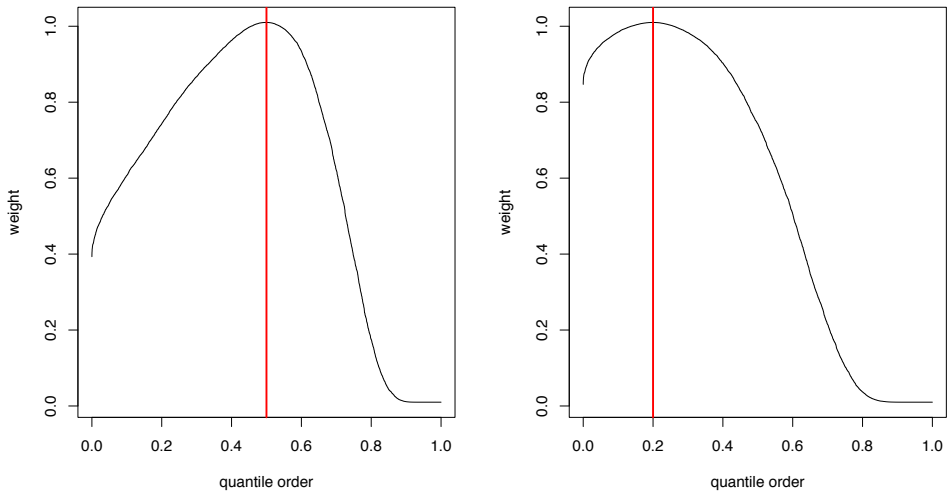

Figure 4.6: *Weight function insisting on the median i.e.* $\beta = 1/2$ *(left) and quantile of order 0.2 i.e.* $\beta = 0.2$ *(right) for log-normal samples.* $\lambda = 1$.

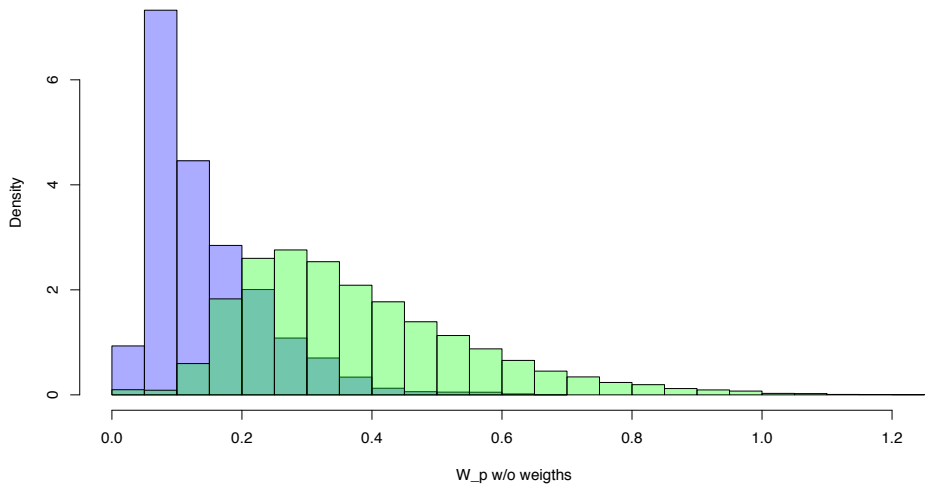

Figure 4.7: *Simulated distribution of the empirical W_p distance for log-normal samples with weigths (blue) and without (green).*

4.3.3 The issue of detecting outages

Heavy-tailed distributions

As seen earlier in section 4.2.1 , the object of importance is the median process $(X_n^{\Delta})_{n\in\mathbb{N}}$, not the raw measurements $(Y_k^t)_{t\in\mathbb{N},1\leq k\leq N_t}$. Individual measurements are irrelevant on their own because the users are routed before they start generating latency measurements. The reason is two-fold: performing latency tests before routing is largely sub-optimal in average because of the time it requires before *load-balancing*, and because what matters most is the state of the network, not the state of the user's connectivity. The distributional approach is more relevant, and simplifies the statistical analysis because the median process is more manageable. At this point there are two choices for the data to feed the change algorithm: either the raw data, or the median process. As we will discuss in greater details in later sections, we choose to analyze changes in the distribution of the medians, not in the distribution of the raw measurements. This choice is motivated for stability reasons and coherence with the object of study, namely the median process. The stability reason is a consequence of the heavy tails of the measurements. Before presenting evidence stylized fact, we recall some facts about heavy tailed distributions. See for instance [37].

Definition 15. *Let F be the c.d.f. of a positive random variable X. F is said to be heavy tailed if:*

$$e^{tx}\mathbb{P}(X > x) = e^{tx}(1 - F(x)) \longrightarrow \infty$$

for all $t > 0$.

An important sub-class of heavy tailed distributions are the regularly varying distributions.

Definition 16. *Let F be the c.d.f. of a positive random variable. F is said to be regularly varying with coefficient $\alpha > 0$ if:*

$$\mathbb{P}(X > x) = 1 - F(x) = L(x)x^{-\alpha}$$

where $L : (0, \infty) \longrightarrow (0, \infty)$ is a slow varying function, i.e.:

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1$$

for all $x > 0$.

It is easily seen that this definition is equivalent to the regular variation of the tail function, hence the name:

Proposition 17. *Let X be a positive random variable with distribution F. Then F is regularly varying with index if and only if there exists $\alpha > 0$*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}$$

for all $x > 0$. α is called the tail index of F.

Proof. Suppose F is regularly varying, then there exists $\alpha > 0$ and L slow varying such that:

$$1 - F(t) = L(t)t^{-\alpha}.$$

Then for any $x > 0$:

$$\begin{aligned}\frac{1 - F(tx)}{1 - F(t)} &= \frac{L(tx)t^{-\alpha}x^{-\alpha}}{L(t)t^{-\alpha}} \\ &= \frac{L(tx)}{L(t)}x^{-\alpha} \\ &\longrightarrow x^{-\alpha}\end{aligned}$$

as $t \to \infty$ because L is slow varying. Now, suppose:

$$\frac{1 - F(tx)}{1 - F(t)} \longrightarrow x^{-\alpha}$$

as $t \to \infty$. Let L such that $L(t) = (1 - F(t))t^{\alpha}$. Then L is slow varying. Indeed:

$$\frac{L(tx)}{L(t)} = \frac{(1 - F(tx))t^{\alpha}x^{\alpha}}{(1 - F(t))t^{\alpha}} = \frac{1 - F(tx)}{1 - F(t)}x^{\alpha} \longrightarrow 1$$

□

It is noteworthy that the class of regularly varying distributions is strictly included in the class of heavy distributions. But in this chapter we will only focus on regularly varying distributions, so the two denominations will be used without distinction.

Example 9. *Let $X \sim \mathcal{E}(1)$ be exponentially distributed. Then $Y = e^X$. Then the cumulative function of Y is regularly varying with coefficient 1. Indeed it immediately follows for the definition of Y that for $x \geq 1$:*

$$\mathbb{P}(Y > x) = x^{-1}$$

Then:

$$\lim_{t\to\infty} \frac{\mathbb{P}(Y > tx)}{\mathbb{P}(Y > t)} = x^{-1}$$

Now we briefly recall the definition of (strict) power-law distributions.

Definition 17. *Let X be of a positive random variable. X is said to have a (strict) power-law distribution if there exists coefficients $x_{\min}, \lambda, \alpha > 0$ such that:*

$$\mathbb{P}(X > x) = \lambda x^{-\alpha}$$

as for all $x \geq x_{x\min}$.

The following lemmas, stated without proof, will be useful in the next section.

Lemma 6. *Let X be of a positive random variable with strict power law distribution, i.e. there exists $x_{\min}, \lambda, \alpha > 0$ such that:*

$$\mathbb{P}(X > x) = \lambda x^{-\alpha}.$$

for all $x \geq x_{min}$. Let Y be a random variable such that $\mathcal{L}(Y) = \mathcal{L}(X|X > x_{\min})$, then Y is Pareto distributed with parameters $(x_{\min}, \alpha)$, i.e.:

$$\mathbb{P}(Y > y) = \left(\frac{x_{\min}}{y}\right)^{\alpha}$$

for all $y > x_{\min}$.

Lemma 7. *Let $X_1, \ldots, X_n$ be i.i.d. Pareto$(x_{\min}, \alpha)$. Then the maximum likelihood estimates for $x_{\min}$ and α are given by:*

$$\begin{cases} \widehat{\alpha}^{ML} = \left(\frac{1}{n}\sum_{i=1}^{n} \log(X_i/\widehat{x}_{\min}^{ML})\right)^{-1} \\ \widehat{x}_{\min}^{ML} = \min_{1\leq i\leq n} X_i \end{cases}$$

Clearly, power-law distributions are heavy tailed. Those distributions are of particular interest because the latency measurements often exhibit heavy tails with power-law behaviors. We now present briefly some descriptive statistics of the data used in this chapter and show evidence of the heavy distributions.

Estimating the tail index

One very important feature of the latency measurements in the $\varepsilon-$stable network is the stationnarity of $n-$samples through time. This feature is key in the estimation of the tail index because it guarantees its stability through time, hence it can be estimated on the whole data set. This

property can be assessed by Monte Carlo simulations. Let $K = 10000$ be the number of simulations. Let $(t_{1,i})_{1\leq i\leq K}$,$(t_{2,i})_{1\leq i\leq K}$ be uniformly distributed timestamps over the time period $[0, T]$, and $(\mathcal{X}^n_{1,i})_{1\leq i\leq K}$, $(\mathcal{X}^n_{2,i})_{1\leq i\leq K}$ be 2 collections of n-samples such that $\mathcal{X}^n_{1,i}$ contains the first n latency measurements received after time $t_{1,i}$ and $\mathcal{X}^n_{2,i}$ contains the first n latency measurements received after time $t_{2,i}$. For all $1 \leq i \leq K$, perform a 2 samples χ^2-test between $\mathcal{X}^n_{1,i}$ and $\mathcal{X}^n_{2,i}$. The distribution of p-values is then compared to the standard Uniform distribution using the Kolmogorov-Smirnov test. We do this test 200 times for varying n ranging from 50 to 1.000. The results of this procedure is 200 p-values, each resulting from a test that aims at determining if the distribution of n-samples is identical trough time. The distribution of those p-values is given in Figure 4.8.

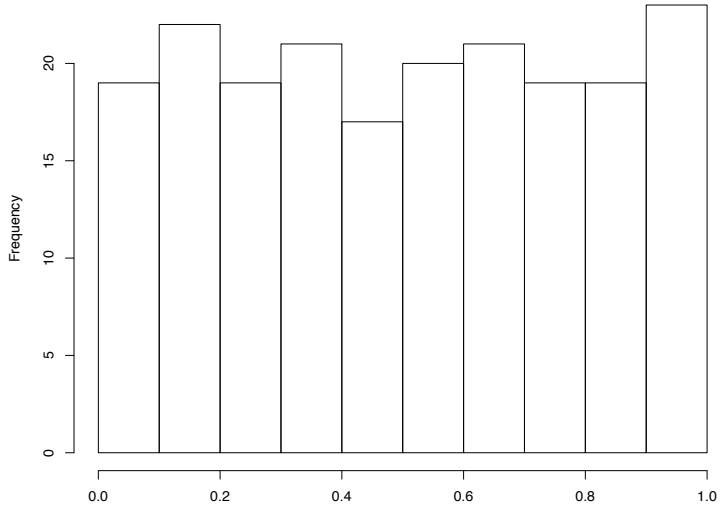

Figure 4.8: *Distribution of the 200 p-values. Each p-value corresponds to a Kolmogorov-Smirnov test of uniformity of the distribution of the K p-values from the χ^2-test between $(\mathcal{X}^n_{1,i})_{1\leq i\leq K}$ and $(\mathcal{X}^n_{2,i})_{1\leq i\leq K}$.*

We obtain again a uniform distribution, which is consistent with what would be obtained if the null hypothesis was true in all cases. We now assumes that the latency measurements in the $\varepsilon-$SN come from the same distribution. Using 4 days of measurements, with sample size $n = 226620$, descriptive statistics are given in Table 4.1.

Min	1st Qu.	Median	3rd Qu.	Max	Mean	SD	Kurtosis	Skweness
2	36	41	49	4208	59.6	79.8	514.3	16.9

Table 4.1: Summary statistics of latency measurements in a stable network.

We shall now characterize the right tail of the latency distribution, see Figure 4.9. Heuristics like

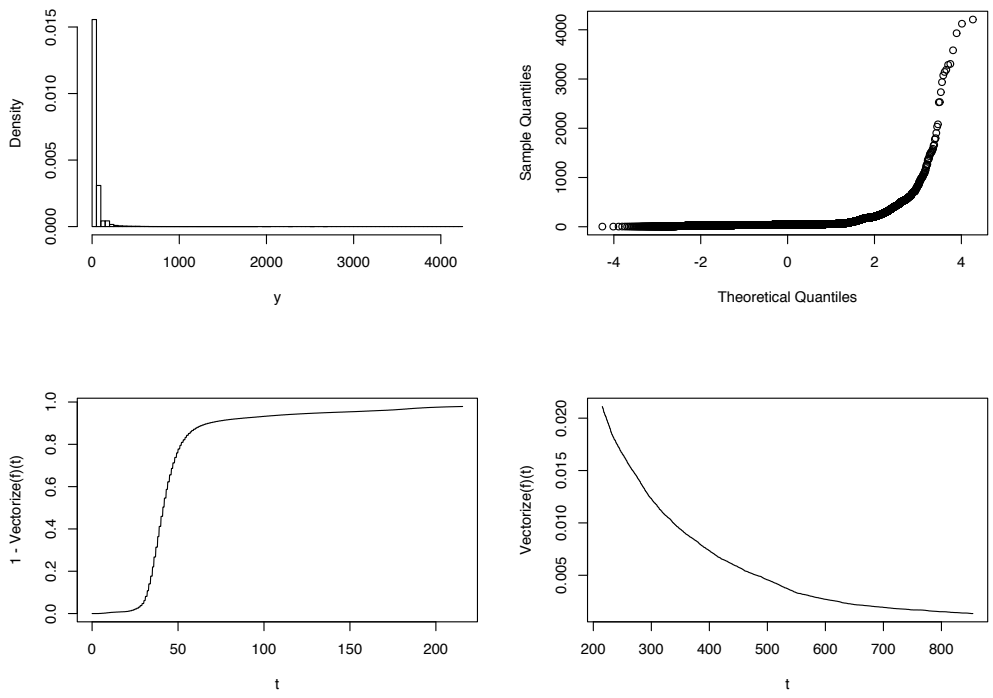

Figure 4.9: *Histogram of the latency measurements (top left), Normal Q-Q plot (top right), empirical cdf between 0 and $\hat{\mu} + 2\hat{\sigma}$ where $\hat{\mu}$, $\hat{\sigma}$ are the empirical mean and standard deviation of the sample (bottom left),empirical survival function between $\hat{\mu} + 2\hat{\sigma}$ and $\hat{\mu} + 5\hat{\sigma}$ (bottom right).*

Pareto-QQ plot [1] and log-log plot of the survival function suggest that the latency measurements exhibit a power-law as seen in Figure 4.10. A simple yet not foolproof way to estimate the tail index consists in performing two linear regressions: one of the survival function in log-log scale and one between the sample quantiles and the theoretical quantiles of the exponential distribution with parameter 1. Those heuristics are a direct consequence of the definition of power law distributions. If X has a strict power law distribution, then for all $x > x_{\min}$:

$$\mathbb{P}(X > x) = \lambda x^{-\alpha}$$

hence taking log on both sides it follows:

$$\log \mathbb{P}(X > x) = \log(\lambda) - \alpha \log(x).$$

Moreover, if we let q_β be the quantile of order β of X, inverting the c.d.f. of X leads to:

$$q_\beta = \left(\frac{\lambda}{1-\beta}\right)^{1/\alpha}$$

[1] A Pareto-QQ plot is a scatterplot created by plotting the empirical quantiles of the log-transformed sample against the theoretical quantiles of the standard exponential distribution.

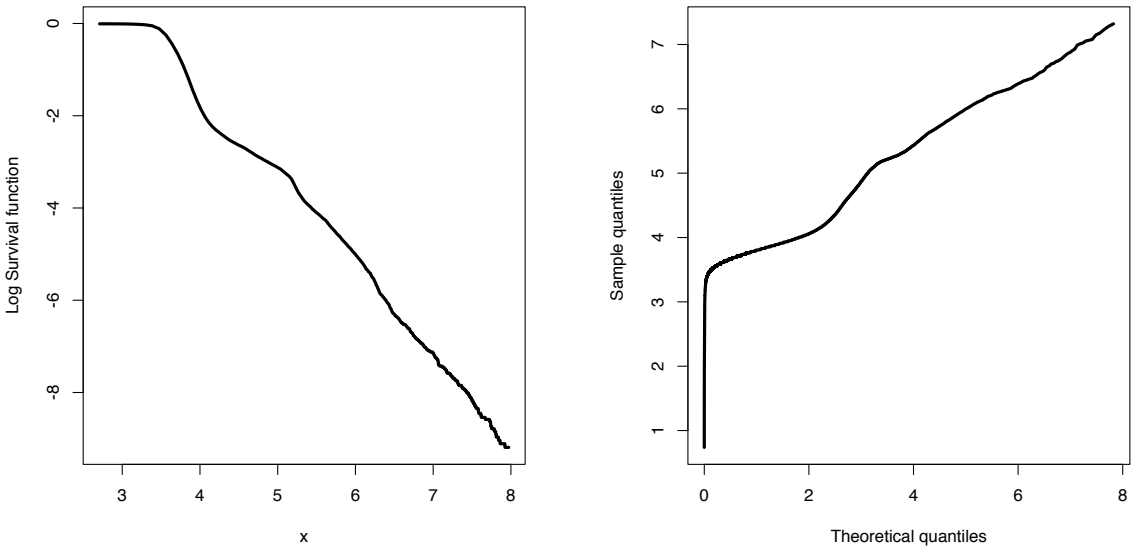

Figure 4.10: $\log - \log$ *plot of the survival function (left) and Pareto-QQ plot (right) of latency measurements. In both cases, a linear relationship seems to be present for large values of the quantiles, suggesting a power law.*

for all $\beta > \mathbb{P}(x \leq x_{\min})$. Taking log on both sides we have:

$$\log(q_\beta) = \frac{\log(\lambda)}{\alpha} - \frac{1}{\alpha}\log(1-\beta) = \frac{\log(\lambda)}{\alpha} + \frac{1}{\alpha} q_\beta^{\mathcal{E}(1)}$$

where $q_\beta^{\mathcal{E}(1)}$ is the quantile of order β of the standard exponential distribution.

In practice one typically does not know where the power law behavior starts, meaning that the value $x_{\min}$ needs to be estimated from the sample. Heuristically, $x_{\min}$ can be estimated visually from the Pareto-QQ plot and log-log plot of the survival function by looking at the point where the the linear relationship starts. These techniques are subjective and known to lead to poor estimates, see [86]. Clauset and al. [19], [20] proposed a more objective way to estimate $x_{\min}$. Their idea is as follows: for each value $x_{\min}$ from a reasonable range, compute the Maximum Likelihood estimate of the tail index $\widehat{\alpha}^{ML}$ using only data points greater than $x_{\min}$, and then compute the Kolmogorov-Smirnov distance between the data being fit and the theoretical Pareto$(x_{\min}, \widehat{\alpha}^{ML})$. Finally, choose the value for $x_{\min}$ that minimizes this distance. Formally, they estimate $x_{\min}$ as:

$$\widehat{x}_{\min} = \min_y \max_x |\hat{F}(x;y) - F(x;\widehat{\alpha}^{ML}, y)| =: \min_y D(y) \tag{4.4}$$

where $\hat{F}(x;y)$ is the empirical distribution function of the data points greater than y, and $F(x;\widehat{\alpha}^{ML}, y)$ is the theoretical distribution of the Pareto$(y, \widehat{\alpha}^{ML})$. Once the parameters $x_{\min}$ and α have been estimated, we propose a significance test following on the work of Lilliefors [58]. The idea is to compute the Kolmogorov-Smirnov test statistic on the latency measurements greater than $x_{\min}$

and test for a Pareto$(x_{\min}, \widehat{\alpha}^{ML})$ distribution. It is well known that performing the Kolmogorov-Smirnov when the parameters of the underlying distribution under the null are estimated from the sample leads far too conservative results [57].

To circumvent this issue, suppose X is a random variable whose distribution depends on some parameters μ, σ, and denote $F_{\mu,\sigma}(\cdot)$ its distribution function. The probability integral transformation of X is defined as the random variable $Y = F_{\mu,\sigma}(X)$ and it is well known that $Y \sim \mathcal{U}(0,1)$. If μ, σ are estimated with μ_n, σ_n from an independent sample $(X_1, \ldots, X_n)$ with common distribution $F_{\mu,\sigma}(\cdot)$, then $Z = F_{\mu_n,\sigma_n}(X)$ no longer follows the uniform distribution. Nevertheless David and al. [27] proved that the distribution of Z is independent of the choice of μ and σ as long as those parameters are location and scale. This is the basis for the so called Lilliefors' test, namely a Kolmogorov-Smirnov where the unknown parameters are estimated from the sample. When the estimated parameters are scale or location, the test statistic

$$T_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_{\mu_n,\sigma_n}(x)|$$

where

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}$$

does not depend on the true values μ, σ. Critical values from the distribution of T_n are typically estimated by Monte-Carlo simulations.

This gives us a way to test the null hypothesis that the distribution of the latency measurements have a power-law behavior. Indeed, under the null, conditional on the event that the measurements are greater than $x_{\min}$, they are Pareto$(x_{\min}, \alpha)$ by Lemma 6. Lemma 8 shows that the logarithm of a Pareto distributed random variable has the translated exponential distribution.

Lemma 8. *Suppose $X \sim$ Pareto$(x_{\min}, \alpha)$, then $Y = \log(X) \sim \mathcal{TE}(x_{\min}, \alpha)$ i.e.:*

$$\mathbb{P}(Y \leq y) = \left(1 - e^{-\alpha(y - x_{\min})}\right) \mathbf{1}_{\{y \geq x_{\min}\}}$$

Since the parameters of a translated exponential are location and scale, the results of David and al. [27] can be used to derive a Lilliefor's like test of the null hypothesis that the distribution of the latency measurements have a power-law behavior by simply computing the distribution of the Kolmogorov-Smirnov statistic between the log of the latency measurements greater than $\hat{x}_{\min}$ against the true translated exponential distribution with parameters $(\log(\widehat{x}_{\min}), \widehat{\alpha}^{ML})$ where $\hat{x}_{\min}$ and $\widehat{\alpha}^{ML}$ are defined in Lemma 7. The distribution of the test statistic is estimated using Monte-Carlo simulations, and a p-value can be derived. The precision of the estimation for α was calculated by Newmann [69] and is $\widehat{\alpha}^{ML}/\sqrt{n}$ where n is the sample size. Results are presented in Figure 4.11.

The results of the estimation are presented in Table 4.2. We conclude with strong confidence that the latency measurements follow a power-law distribution with a tail index $\alpha = 2.1$. The tail index α can also be estimated using the well known Hill's estimator[46].

Definition 18. *Let $(X_n)_{n \geq 1}$ be a sequence of independent and identically distributed positive random variables with c.d.f. F. Suppose F is regularly varying with coefficient α. The Hill*

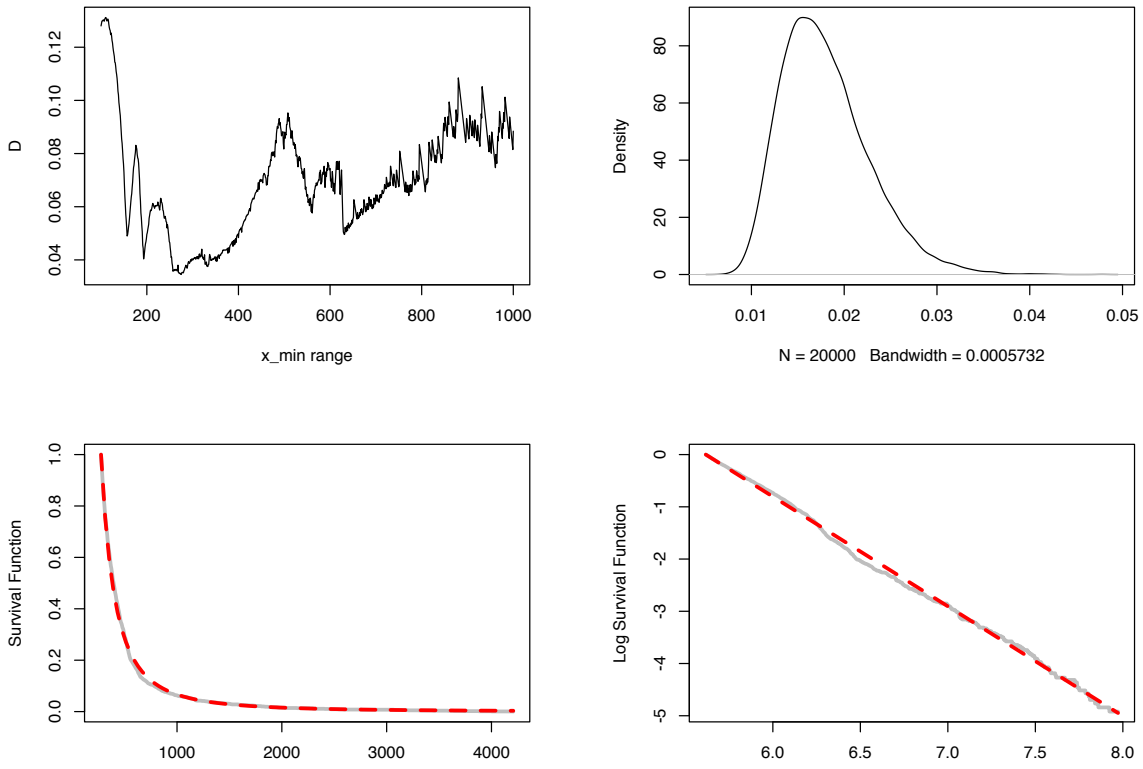

Figure 4.11: *Top left: the function D defined in (4.4) hits a minimum at $\hat{x}_{\min} = 274$. Top right: estimated distribution of the KS-Statistic with estimated parameters $(\widehat{x}_{\min}, \widehat{\alpha}^{ML})$. Bottom left: survival function of the latency measurements greater than $\hat{x}_{\min}$ with Pareto fit. Bottom right:* $\log - \log$ *plot survival function of the latency measurements greater than $\hat{x}_{\min}$ with* $\log -$*Pareto fit.*

$\widehat{x}_{\min}$	$\widehat{\alpha}^{ML}$	Confidence Bands (5%)	Test Statistic	P-value
274	2.10	$[2.05, 2.15]$	0.022	0.19

Table 4.2: *Tail index estimation for the latency measurements.*

estimator for α takes the form:

$$\hat{\alpha}_{k,n} = \left(\frac{1}{k} \sum_{i=0}^{k} \log(X_{n-i:n}) - \log(X_{n-k:n}) \right)^{-1}$$

where $(X_{j:n})_{1 \leq i \leq n}$ is the order statistics, i.e. is the permutation of $(X_i)_{1 \leq i \leq n}$ such that $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$.

Under the hypothesis that k goes to ∞ at a rate such that $k/n \to \infty$, then $\hat{\alpha}_{k,n} \to \alpha$. The

choice of the optimal number of order statistic k was debated in the literature, see for instance [73] [31], but we'll use the infamous "Eye Balling" technique [72] [26] for our purposes. Figure

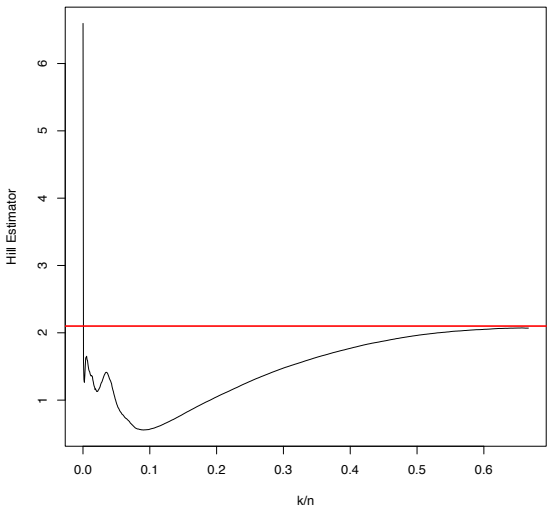

Figure 4.12: *Hill plot. Hill estimators for different values of k (solid black) maximum likelihood estimate* α^{ML}.

4.12 suggests convergence of the Hill plot to the same value of the coefficient calculated using the Lilliefors' like test exposed above.

The catastrophe principle

The issue with power-law distributions is that they tend to produce catastrophic events with high probability. The consequences of the so-called "catastrophe principle" or "principle of single big jumps" [83] [37] on detecting outages in the median process $(X_n^{\Delta})_{n\in\mathbb{N}}$ is addressed in this section.

This "principle" is a property of sub-exponential distributions, see Definition 19. Recall that power law distributions are regularly varying distributions, and regularly varying functions are sub-exponential distributions, see for instance [37] for a proof. This property establishes that when the sum of n independent sub-exponential random variables exceeds some large value x, it is most likely due to the fact that the maximum of those n random variables also exceeds x. As described in section 4.3.3, the distribution of latency measurements exhibit a power law behavior. This is one of the reasons why the median process is considered for running the change detection algorithm instead of the raw measurements: computing the median increases the value of the tail index, hence reducing the probability of catastrophic events. Before describing precisely how this reduction occurs, we recall the definition of sub-exponential distributions and the "catastrophe principle", see [37].

Definition 19. *A positive random variable X is said to have a sub-exponential distribution if:*

$$\frac{\mathbb{P}(X_1 + X_2 > x)}{\mathbb{P}(X_1 > x)} \longrightarrow_{x\to\infty} 2$$

where X_1, X_2 are i.i.d. copies of X. This is equivalent to:

$$\frac{\mathbb{P}(\max(X_1, X_2) > x)}{\mathbb{P}(X_1 + X_2 > x)} \longrightarrow_{x\to\infty} 1$$

The definition of sub-exponential distributions implies the so-called "catastrophe principle" or "principle of single big jump".

Proposition 18. *Let X be a positive random variable with sub-exponential distribution. Then for all $n \geq 1$:*

$$\frac{\mathbb{P}(\max_{1\leq i\leq n} X_i > x)}{\mathbb{P}(\sum_{i=1}^{n} X_i > x)} \longrightarrow_{x\to\infty} 1$$

The catastrophe principle is problematic for our purposes. The distribution of latency measurements exhibit a power-law behavior hence produces catastrophic events. Using the median process will reduce the frequency of such events. We shall now prove that the median of an independent sample with a common F such that F is regularly varying with tail index $\alpha > 0$ also has a regularly varying cumulative function, but with a tail index $\beta > \alpha$. This is an important feature of the median since catastrophic events are likely to wrongly trigger an outage alert.

Proposition 19. *Let $n \in \mathbb{N}$ and $(X_1, \ldots, X_n)$ an i.i.d. sample with distribution F and let $(X_{(1)}, \ldots, X_{(n)})$ the order statistics. For any $k \in 1, \ldots, n$, let $k_n = n - k + 1$. Assume that F is regularly varying, i.e. suppose there exists $\alpha > 0$ and L slow varying such that:*

$$1 - F(x) \sim_{x\to\infty} L(x)x^{-\alpha}.$$

Then the distribution of $X_{(k)}$ is also regularly varying, with tail index $k_n\alpha$, i.e. there exists a slow varying function G such that:

$$\mathbb{P}(X_{(k)} > x) \sim_{x\to\infty} G(x)x^{-k_n\alpha}$$

Proof. First, notice that we have the following equality of events:

$$\{X_{(k)} > x\} = \{Z \geq n - k + 1\}$$

where Z has the binomial $\mathcal{B}(n, p_x)$ distribution and $p_x = \mathbb{P}(X_1 > x)$. Then:

$$\begin{aligned}\mathbb{P}(X_{(k)} > x) &= \mathbb{P}(Z \geq k_n) \\ &= \sum_{i=k_n}^{n} \binom{n}{i} p_x^i (1 - p_x)^{n-i}.\end{aligned}$$

It follows that

$$\frac{\mathbb{P}(X_{(k)} > x)}{\binom{n}{k_n} p_x^{k_n}} = (1 - p_x)^{n-k_n} + \sum_{i=k_n+1}^{n} \binom{n}{i} p_x^i (1 - p_x)^{n-i},$$

but the right hand side clearly converges to 1 as $x \to \infty$. This means:

$$\mathbb{P}(X_{(k)} > x) \sim \binom{n}{k_n} p_x^{k_n} \sim \binom{n}{k_n} L(x)^{k_n} x^{-k_n \alpha}.$$

as $x \to \infty$. Since L is slow varying, clearly $G(x) = \binom{n}{k_n} L(x)^{k_n}$ is also slow varying, and we have that $X_{(k)}$ is regularly varying with tail index $k_n \alpha$:

$$\mathbb{P}(X_{(k)} > x) \sim G(x) x^{-k_n \alpha}.$$

□

In particular, for the median of i.i.d. n-sample $(X_1, \ldots, X_n)$ with regularly varying distribution F with tail index $\alpha > 0$, if we denote the median $Y_n = X_{\lceil n/2 \rceil}$ we have:

$$\mathbb{P}(Y_n > x) \sim G(x) x^{-(n - \lceil n/2 \rceil + 1)\alpha}$$

This quantifies how the probability that the median process will produce catastrophic events decreases with the number of observations. The other advantage with utilizing the median process is because of integrability reasons. We previously showed that the tail index for the latency measurements is greater than 1 but less than 2, meaning that the variance of latency measurements is infinite, limiting a priori the Wasserstein distance of order p to $p = 1$. As soon as $n \geq 4$, the Wasserstein of order 2 may be used. The higher the n, the higher the order possible for the Wasserstein distance. Because at this point no control is guaranteed on the sample size since n is given over every interval $[t, t + \lambda[$, we will stick to $p = 1$.

4.4 Empirical Results

We now present the results of the difference algorithms presented in this chapter on real data.

4.4.1 Data description

Three different networks were selected. Each one experienced an outage over the course of several hours. We selected 4 days of measurements so that the outage appeared in the middle of the 4th day. For the training set selection and tuning of (N, M), only the first 3 days were considered, i.e. the period with no outage. For the change detection algorithm, the first 2 days were used to estimate the value Q for each choice of parameter C, and the next 2 days were the testing set: one day without any issue, the other with issues. For reasons of confidentiality the timestamps were initialized to 0 in all cases, and the providers will be referred to as $P1$, $P2$ and $P3$ and their respective outages as $O1$, $O2$ and $O3$. $O1$, $O2$ and $O3$ happened respectively in the US, Europe and Asia. The outages are summarized in Table 4.3 and Figure 4.13.

4.4.2 Assessing stability of the networks

	Out. Dur.	Out. Var	∅ Out. Var	Out. Mean	∅ Out. Mean
$O1$	4.1h	41.6	1.55	34.5	8.2
$O2$	1.8h	62.6	3.4	77.3	42.4
$O3$	7.3h	53.7	6.3	67.9	54.1

Table 4.3: *Descriptive statistics of the 3 outages. Out. Dur. is the total amount of time the outage lasted, Out. Var. and ∅ Out. Var. are respectively the variance of latency measurements during the outage and outside the outage, Out. Mean and ∅ Out. Mean are respectively the mean of latency measurements during the outage and outside the outage.*

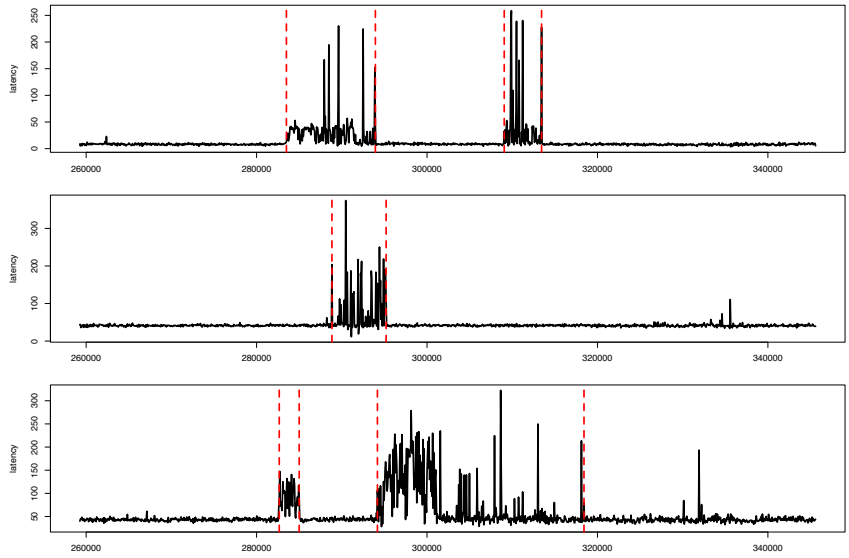

Figure 4.13: *Top to bottom: $P1$, $P2$ and $P3$ media-processes over day 4 where the outages occurred. Red vertical dashed lines frame the outages. The outages O_1 and O_3 occurred in two consecutive episodes.*

Stationarity

The hypothesis of stationarity of the median processes: is assessed with three classical routines: Augmented Dickey-Fuller Test (ADF), Kwiatkowski-Phillips-Schmidt-Shin test (KPSS) and Phillips-Perron test (PP) . Unlike the two others, stationarity is the null hypothesis in the KPSS test. Only the KPSS test in P_1 rejects the hypothesis of stationarity, see Table 4.4.

Ratio between AVG and best ARMA

The largest ratio between AVG and best ARMA was for $P1$, with a value of 1.019. The three networks qualify as $2\% - SN$, see Table 4.5.

	ADF	PP	KPSS
$P1$	< 0.01	< 0.01	< 0.01
$P2$	< 0.01	< 0.01	> 0.1
$P3$	< 0.01	< 0.01	> 0.1

Table 4.4: *p-values for the different tests and networks.*

	$RMSE\left(\widehat{X_{t+1}}^{AVG}\right) \Big/ RMSE\left(\widehat{X_{t+1}}^{ARMA}\right)$
$P1$	1.019
$P2$	1.007
$P3$	1.015

Table 4.5: *Ratio of the RMSE between the AVG prediction and best ARMA prediction for the three networks, computed over the first three days when no outages where reported.*

4.4.3 Training set selection

The goal in this section is determine the minimum training set that does not impact accuracy more than a given tolerance $\tau > 0$. For reasons of business confidentiality, the actual predictive algorithm used by Citrix will be treated as a black box. Only will we treat the determination of the optimal training set. In order to produce a prediction in stable networks, Citrix uses at most the last N more recent measurements provided they were received within the last M minutes, even if that means ending up with less than N measurements. We propose a simple approach by grid searching to tune the optimal parameters (N, M). We first define timestamps $(t_k)_{k=1,\dots,K}$ at which predictions will be made, such that $t_{i+1} - t_i = \lambda$, where $\lambda = 60$s is the prediction TTL. Only the first three days are considered. For each pair (N, M), let $\mathcal{E}(N, M)$ be the prediction error over the timestamps $(t_k)_k$, that is:

$$\mathcal{E}(N,M) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left|\widehat{X^{\lambda}_{t_k,N,M}}^{Citrix} - X^{\lambda}_{t_k}\right|^2}$$

N will vary from 5 to 1000, and M from 1min to 1h. Denote (N^*, M^*) the minimizer of the prediction error, i.e.

$$(N^*, M^*) \in \arg\min_{N,M} \mathcal{E}(N, M)$$

The results are presented in Table 4.6. The value (N^*, M^*) that minimizes the prediction error corresponds to the largest tested values for the parameters N et M: the error associated with the prediction $\widehat{X^{\Delta}_t}^{Citrix}$ decreases as N, M increase. But the error has a particular profile: it rapidly converges to $\mathcal{E}(N^*, M^*)$ as N, M increase, see Figure 4.14. We propose to limit the size of the training set relatively to a tolerance $\tau > 0$: let $\mathcal{C}(\tau)$ be the set of containing all pairs (N, M) satisfying:

$$\mathcal{C}(\tau) = \{(N, M) | \mathcal{E}(N, M) < (1 + \tau)\mathcal{E}(N^*, M^*)\},$$

	N^*	M^*	$\mathcal{E}(N^*, M^*)$
$P1$	940	9min	1.27
$P2$	950	57min	2.34
$P3$	850	45min	4.62

Table 4.6: Results of the (N, M) grid search.

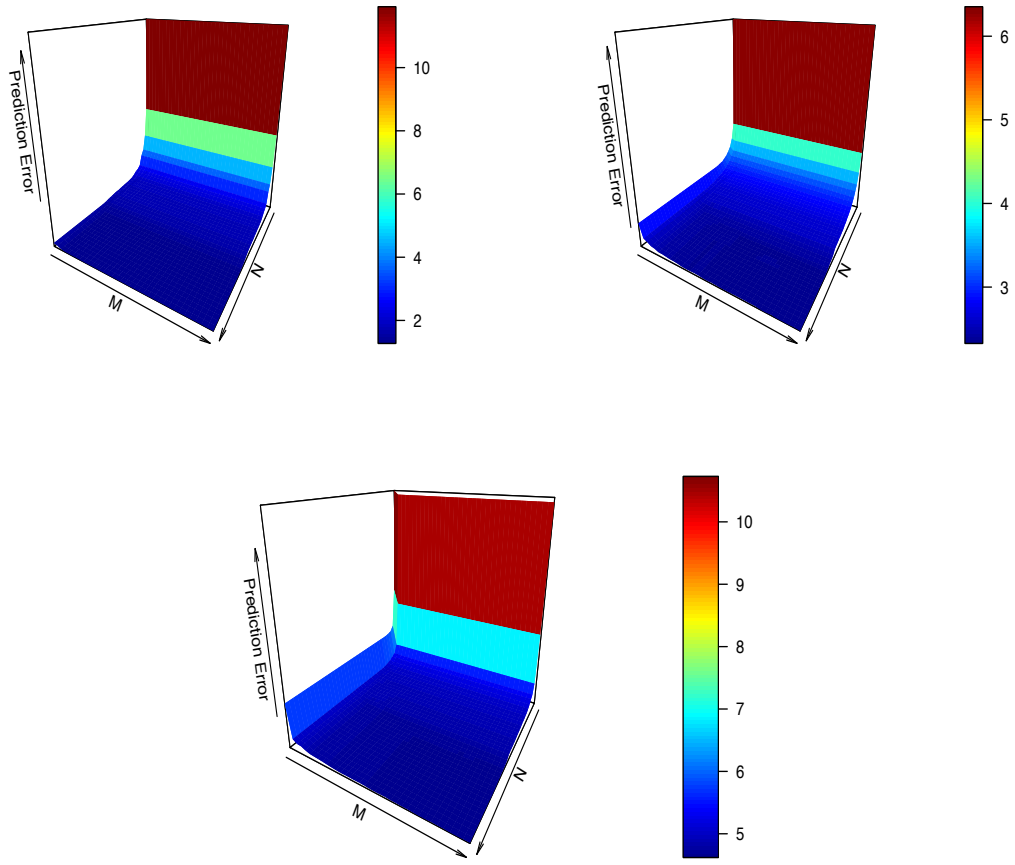

Figure 4.14: *Surface plot of the prediction error associated to* $\widehat{X_t^{\Delta}}^{Citrix}$ *in* $P1$*,* $P2$ *and* $P3$*, as a function of* N, M*. The error decreases very quickly as* N, M *increase, then stabilizes.*

then the pair $(N, M) = (N_\tau, M_\tau)$ that we shall select is given by:

$$M_\tau = \min\{M | (N, M) \in \mathcal{C}(\tau)\}$$
$$N_\tau = \min\{N | (N, M_\tau) \in \mathcal{C}(\tau)\}.$$

In other words, among all pairs (N, M) satisfying $\mathcal{E}(N, M) < (1+\tau)\mathcal{E}(N^*, M^*)$, we select the pair (N_τ, M_τ) that has the smallest M first, and among the remaining candidates, we chose the ones with the minimum N. See Table 4.7 for the results for varying τ.

	τ	N_τ	M_τ	$\mathcal{E}(N_\tau, M_\tau)/\mathcal{E}(N^*, M^*)$
$P1$	10%	300	3min	1.093
	5%	400	4min	1.048
	1%	820	6min	1.009
$P2$	10%	150	3min	1.087
	5%	300	4min	1.049
	1%	600	17min	1.009
$P3$	10%	45	2min	1.099
	5%	100	6min	1.049
	1%	500	20min	1.009

Table 4.7: (N_τ, M_τ) for different values of τ.

4.4.4 Change detection

We present the results of the change detection algorithm described in Algorithm 2. The parameters that need to be tuned are C, the number of data points in each window, and β, the parameter that quantifies the sensitivity to the detection with respect to the maximum value Q taken by the W_p distance between any two windows of size C over periods of time when no outages occurred. C ranges from 5 to 100, and β ranges from 0.5 to 2.

For every value of C, the value of Q is computed on the first two days of measurements by computing the maximum value of the W_p distance between randomly selected windows of size C. Then the algorithm for change detection is run, with parameters (C, β). The prediction algorithm is fed with the parameters (N_τ, M_τ) with $\tau = 10\%$. Recall from Section 4.3.1 that the union of all instants where no outage is detected by algorithm is called the Green Zone denoted by $\mathrm{GZ}(C, \beta)$ and the union of all instants where an outage was detected and trust was lost in the model is called the Red zone, denoted by $\mathrm{RZ}(C, \beta)$. For a pair (C, β) we denote the Green Error and Red Error by $\mathrm{GE}(C, \beta)$ and $\mathrm{RE}(C, \beta)$ respectively the prediction error associated with the prediction $\widehat{X^{\Delta}_{t,N_\tau,M_\tau}}^{Citrix}$ over the Green and Red Zones respectively:

$$\mathrm{GE}(C,\beta) = \sqrt{\frac{1}{\#\mathrm{GZ}(C,\beta)} \sum_{k \in \mathrm{GZ}(C,\beta)} \left| \widehat{X^{\lambda}_{t_k,N_\tau,M_\tau}}^{Citrix} - X^{\lambda}_{t_k} \right|^2}$$

$$\mathrm{RE}(C,\beta) = \sqrt{\frac{1}{\#\mathrm{RZ}(C,\beta)} \sum_{k \in \mathrm{RZ}(C,\beta)} \left| \widehat{X^{\lambda}_{t_k,N_\tau,M_\tau}}^{Citrix} - X^{\lambda}_{t_k} \right|^2}$$

We will also look the time spent in Green zone, denoted by

$$\mathrm{GT}(C,\beta) = \frac{\#\mathrm{GZ}(C,\beta)}{\#\mathrm{GZ}(C,\beta) + \#\mathrm{RZ}(C,\beta)}$$

that is the percentage of timestamps in the Green zone. Finally the ratio Red Error over Green Error $\mathrm{GE}(C,\beta)/\mathrm{RE}(C,\beta)$ will be considered.

Finally if the weights introduced in 4.3.2 were used for the detection of outages, we shall write $RE(C,\beta)_w$, $GE(C,\beta)_w$ and $GT(C,\beta)_w$.

The rule of thumb for the choice of the parameters is $(C,\beta)=(10,1)$. Because we are working with the median process at scale $\Delta=\lambda=60$s where λ is the prediction TTL (see Section 4.4.3) the choice $C=10$ implies that the windows Ref and Shift span intervals of length $C\lambda=10$ minutes. When an outage is reported, by construction the algorithm will declare the underlying network unreliable during the time needed to rebuild the two sliding windows. Each window contains C data points, each data point arrives every 60s, hence during $2C\lambda=20$ minutes after an outage is detected the network is declared unreliable, which is considered reasonable by operational engineers. The choice $\beta=1$ is reasonable because in this case an outage is declared when when the W_p distance between the Ref and Shift windows exceeds the maximum value of the W_p distance between any two windows of length C recorded during a period without outages.

This choice of parameters will be compared to the optimal values (C^*,β^*) obtained by choosing the values that maximizes the ratio between Red Error / Green Error with and without weights:

$$(C^*,\beta^*)\in\arg\max_{(C,\beta)}\frac{\mathrm{RE}(C,\beta)}{\mathrm{GE}(C,\beta)}$$

$$(C^*_w,\beta^*_w)\in\arg\max_{(C,\beta)}\frac{\mathrm{RE}(C,\beta)_w}{\mathrm{GE}(C,\beta)_w}$$

Without weights in the Wasserstein distance

The results of the outage detection algorithm without using the weights are presented in Tables 4.8, 4.9 and Figures 4.15 and 4.16.

With weights in the Wasserstein distance

The results of the outage detection algorithm using the weights are presented in Tables 4.10, 4.11 and Figures 4.17 and 4.18.

Comments

The ratios $\mathrm{GE}(C,\beta)/\mathrm{RE}(C,\beta)$ and $\mathrm{GE}(C,\beta)_w/\mathrm{RE}(C,\beta)_w$ are very high for all networks, indicating that the algorithm accurately identified the outages in all cases. The choice $(C,\beta)=(10,1)$ leads similar results than the choice (C^*,β^*) concerning the ratio Red Error / Green Error for all networks, again with and without using the weights. The value of β^* for $P1$, $P2$ and P_3 is consistent with the intuition in a stable network, ranging from 0.85 to 1, while β^*_w ranges from 0.8 to 1.15.

Without the use of weights, 70% (resp. 73%) of the detections corresponded to the actual outages, on average over P_1, P_2 and P_3 in the case $(C,\beta)=(C^*,\beta^*)$ (resp. $(C,\beta)=(10,1)$). With the use of weights, 85%, resp. 100%, of the detections corresponded to the actual outages,

on average over P_1, P_2 and P_3 in the case $(C, \beta) = (C_w^*, \beta_w^*)$ (resp. $(C, \beta) = (10, 1)$). With and without weights, the algorithm was able to detect precisely the instants of outages: 95% of the durations of the outages were correctly identified without the weights, while 98% of the durations of the outages were correctly identified with the weights.

The best results were obtained with the use of the weights and with parameters $(C, \beta) = (10, 1)$ with perfect detection of the outages with and no false positives.

Because latency measurements have heavy tailed distributions, catastrophic events occur quite often. Since those events are not necessarily the signature of an outage, the false positive rate is greater without weights. In addition, the algorithm was quicker to detect outages with the weights: on average 1.3 (resp. 1.9) observations of the median process X_n^{Δ} were needed to correctly identify the beginning (resp. end) of the outage with the weights whereas on average 2.1 (resp. 2.5) observations of the median process X_n^{Δ} were needed to correctly identify the beginning (resp. end) of the outage without the weights.

	$\mathrm{RE}(C^*, \beta^*)/\mathrm{GE}(C^*, \beta^*)$	C^*	β^*	$\mathrm{GT}(C^*, \beta^*)$
$P1$	23.8	7	0.9	90.1%
$P2$	18.0	8	1	92.0 %
$P3$	8.39	17	0.85	71.2%

Table 4.8: *Optimal values (C^*, β^*) that maximize the ratio Red Error / Green Error. No weights were used in the W_p distance.*

	$\mathrm{RE}(C, \beta)/\mathrm{GE}(C, \beta)$	$\mathrm{GT}(C, \beta)$
$P1$	21.1	89.2%
$P2$	14.2	93.3 %
$P3$	5.9	83.8 %

Table 4.9: *Ratio Red Error / Green Error and Time spent in Green zone for $C = 10$ and $\beta = 1$. No weights were used in the W_p distance.*

	$\mathrm{RE}(C^*, \beta^*)_w/\mathrm{GE}(C^*, \beta^*)_w$	C_w^*	β_w^*	$\mathrm{GT}(C^*, \beta^*)_w$
$P1$	23.9	6	1.15	92.0%
$P2$	17.6	10	0.95	94.6 %
$P3$	6.79	60	0.80	64.8%

Table 4.10: *Optimal values (C^*, β^*) that maximize the ratio Red Error / Green Error. Weights were used in the W_p distance.*

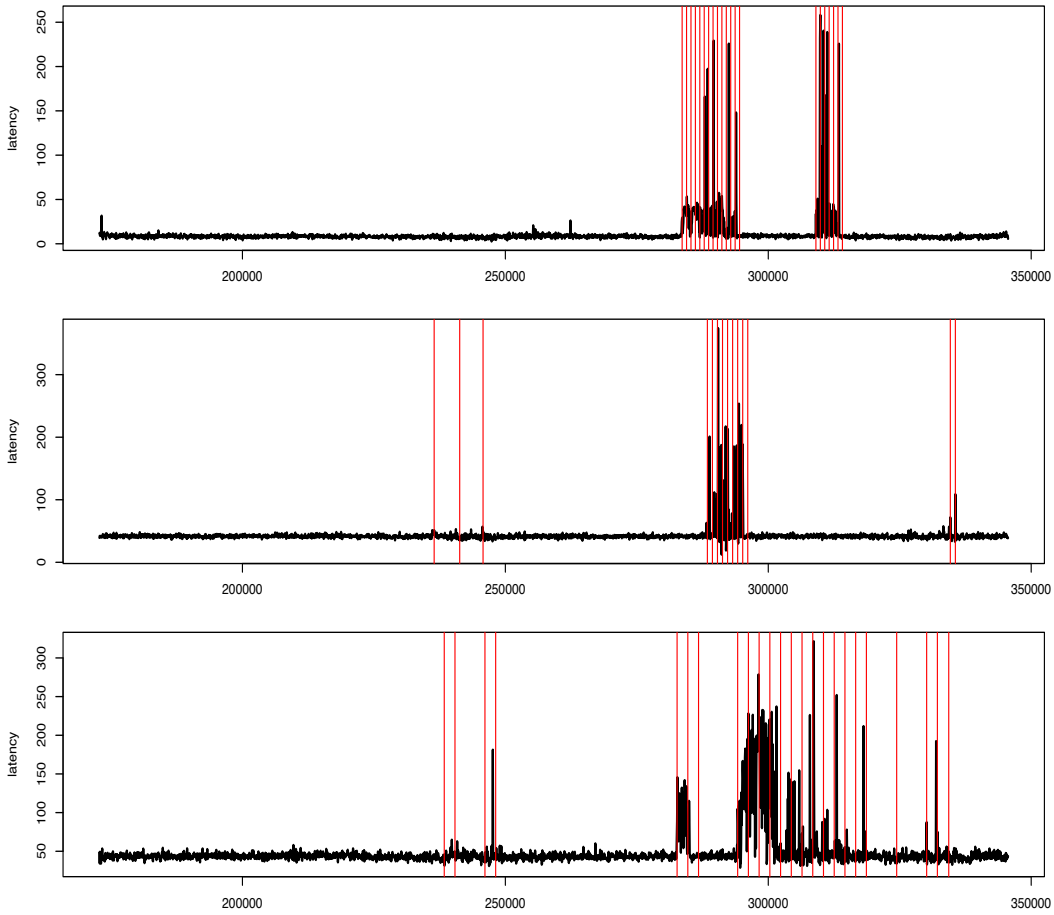

Figure 4.15: *Top to bottom:* $P1$*,* $P2$ *and* $P3$ *change detection output for* C^**,* β^**. No weights were used in the* W_p *distance. Each vertical red line corresponds to an instant where the algorithm triggered an alert.*

	$\text{RE}(C,\beta)_w/\text{GE}(C,\beta)_w$	$\text{GT}(C,\beta)_w$
$P1$	20.3	89.9%
$P2$	15.9	96.3 %
$P3$	5.2	89.2 %

Table 4.11: *Ratio Red Error / Green Error and Time spent in Green zone for* $C = 10$ *and* $\beta = 1$*. Weights were used in the* W_p *distance.*

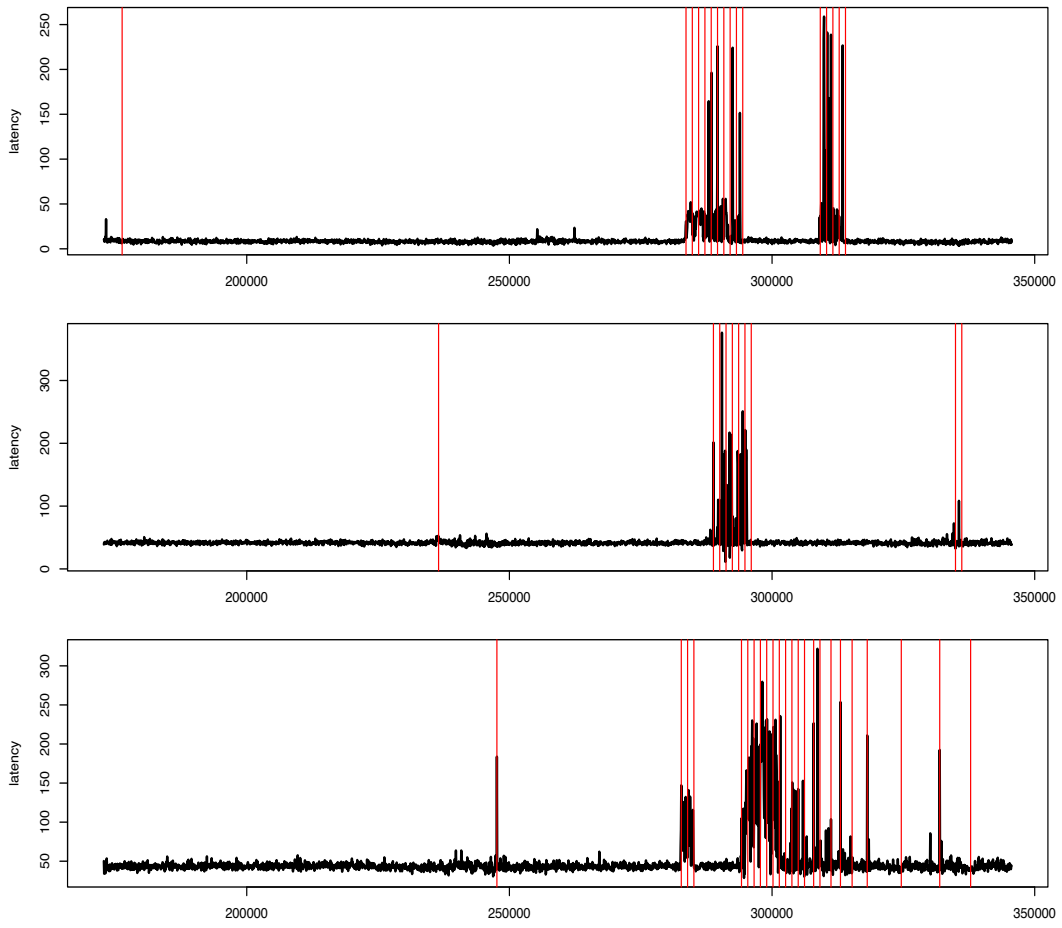

Figure 4.16: *Top to bottom:* $P1$, $P2$ *and* $P3$ *change detection output for* $C = 10$, $\beta = 1$. *No weights were used in the* W_p *distance. Each vertical red line corresponds to an instant where the algorithm triggered an alert.*

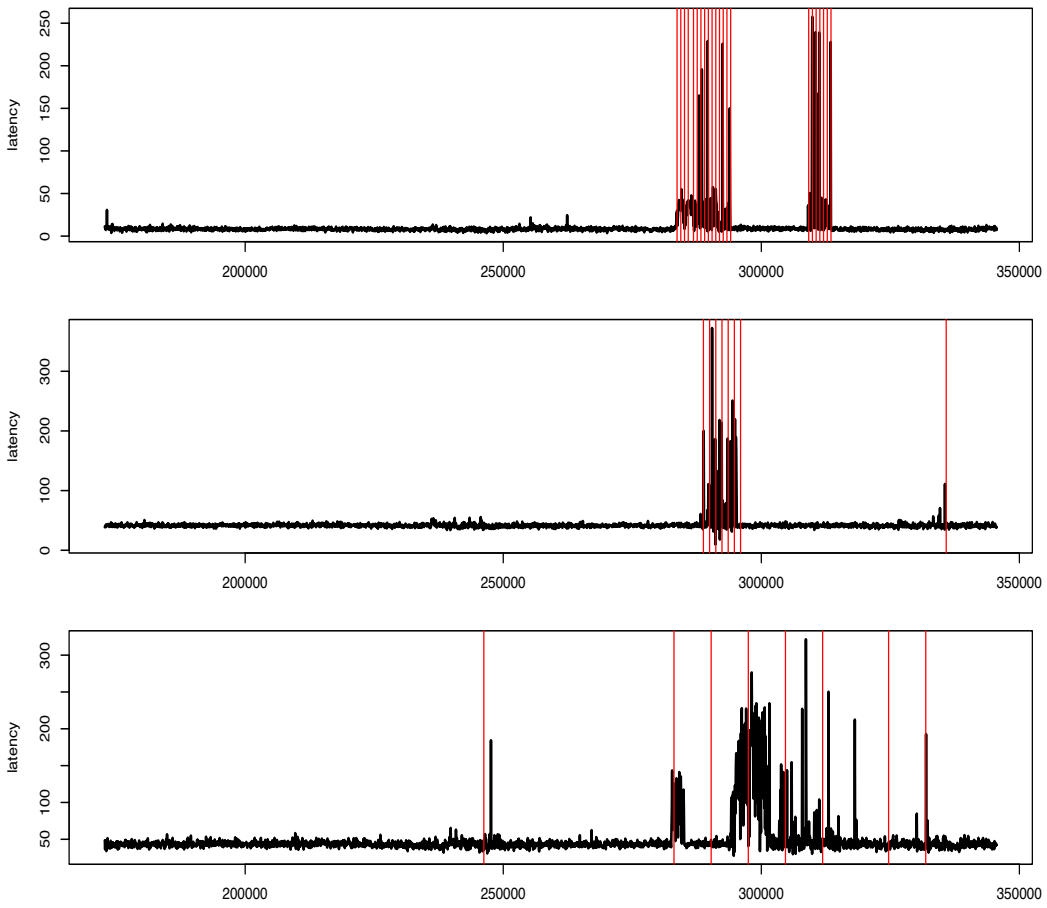

Figure 4.17: *Top to bottom:* $P1$, $P2$ *and* $P3$ *change detection output for* C_w^*, β_w^*. *Weights were used in the* W_p *distance. Each vertical red line corresponds to an instant where the algorithm triggered an alert.*

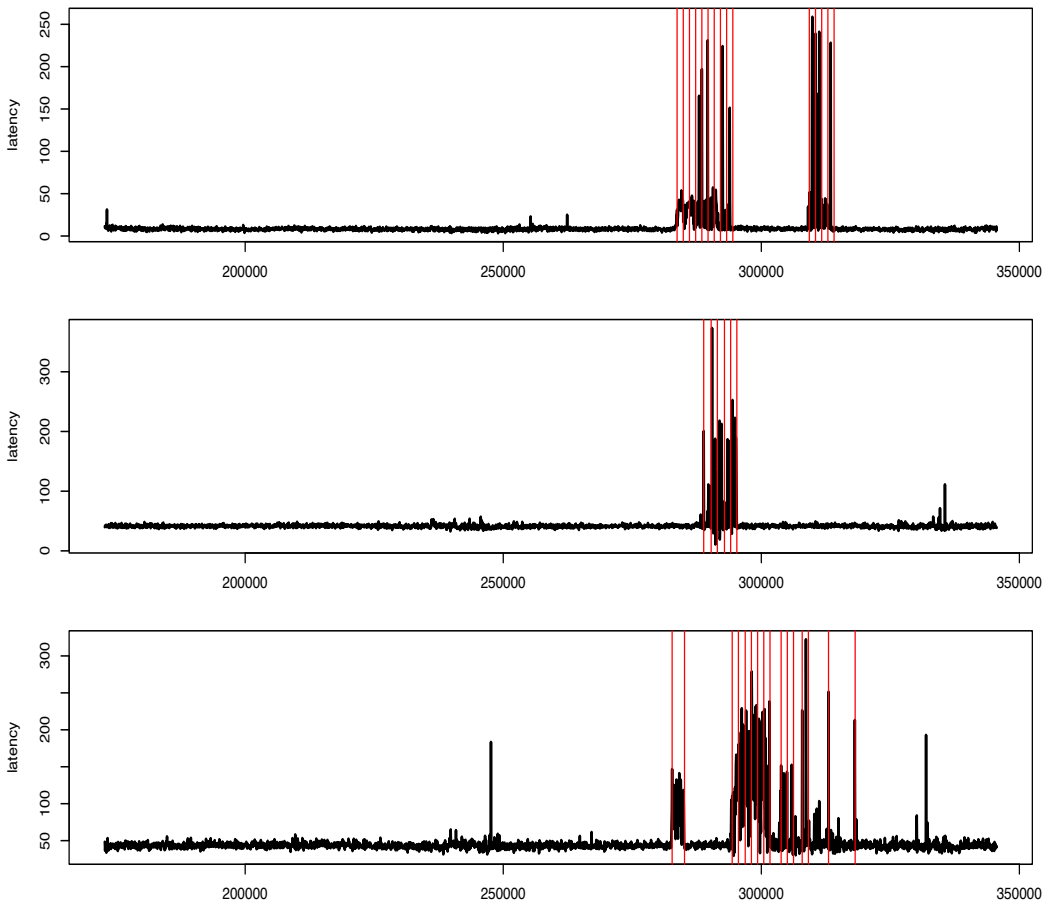

Figure 4.18: *Top to bottom: $P1$, $P2$ and $P3$ change detection output for $C = 10$, $\beta = 1$. Weights were used in the W_p distance. Each vertical red line corresponds to an instant where the algorithm triggered an alert.*

Conclusion

In this Chapter we presented a notion of $\varepsilon-$stable networks that are characterized by stationary median latency processes with low structure, as opposed too highly structured median latency processes described in Chapter 3.

Identifying those networks allows the use of very low computational complexity predictive algorithms, thus reducing overall computational costs. We showed how to select the minimum number of training sets for those predictive algorithms without impacting performance above a certain tolerance.

In a second step, we developed an algorithm that aims at detecting outages. Differentiating outages from outliers in the median latency process is challenging because of the power-law behavior exhibited by the distribution of the latency measurement. We described an algorithm based on a comparison using the Wasserstein distance of two sliding windows. After characterizing the power-law behavior of the distribution of latency measurements, we proposed a way to increase the robustness of the algorithm by weighting the Wasserstein distance in order to limit influence of outliers without impacting the sensitivity of the outage detection. An experiment conducted on real data gave us empirical evidence that the use of the weights allow to react quicker to outages and to decrease the number of false positives, while keeping a very high true positive rate. We suggested to implement the algorithm for detecting changes in the median latency process presented in this Chapter with the values $(C, \beta) = (10, 1)$ and with the weighted W_p distance.

Bibliography

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[3] T. W. Anderson. *The statistical analysis of time series*. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

[4] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statistics*, 23:193–212, 1952.

[5] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quant. Finance*, 13(1):65–77, 2013.

[6] R. R. Bahadur. A note on quantiles in large samples. *Ann. Math. Statist.*, 37:577–580, 1966.

[7] Gaurav Banga and Peter Druschel. Measuring the capacity of a web server under realistic loads. *World Wide Web*, 2(1):69–83, 1999.

[8] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.

[9] Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[10] Dario Bini and Victor Y. Pan. *Polynomial and Matrix Computations (Vol. 1): Fundamental Algorithms*. Birkhauser Verlag, CHE, 1994.

[11] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327, 1986.

[12] George E. P. Box and Gwilym M. Jenkins. *Time series analysis: forecasting and control*. Holden-Day, San Francisco, Calif.-Düsseldorf-Johannesburg, revised edition, 1976. Holden-Day Series in Time Series Analysis.

[13] George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time series analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, fourth edition, 2008. Forecasting and control.

[14] George EP Box and George C Tiao. Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, 70(349):70–79, 1975.

[15] Richard C. Bradley. On positive spectral density functions. *Bernoulli*, 8(2):175–193, 2002.

[16] Peter J. Brockwell and Richard A. Davis. *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2002. With 1 CD-ROM (Windows).

[17] Ying Chen and Tuan D. Pham. Sample entropy and regularity dimension in complexity analysis of cortical surface structure in early alzheimer's disease and aging. *Journal of Neuroscience Methods*, 215(2):210 – 217, 2013.

[18] Marc Claesen and Bart De Moor. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

[19] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009.

[20] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. On the frequency of severe terrorist events. *Journal of Conflict Resolution*, 51(1):58–87, 2007.

[21] H. Cramér and H. Wold. Some Theorems on Distribution Functions. *J. London Math. Soc.*, 11(4):290–294, 1936.

[22] Harald Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):141–180, 1928.

[23] Harald Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.

[24] Colleen D. Cutler. A review of the theory and estimation of fractal dimension. In *Dimension estimation and models*, volume 1 of *Nonlinear Time Ser. Chaos*, pages 1–107. World Sci. Publ., River Edge, NJ, 1993.

[25] R. Dahlhaus. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, 62(1):139–168, 1996.

[26] Jon Danielsson, Lerby Murat Ergun, Laurens de Haan, and Casper G de Vries. Tail index estimation: Quantile driven threshold selection. *Available at SSRN 2717478*, 2016.

[27] F. N. David and N. L. Johnson. The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35:182–190, 1948.

[28] Manfred Denker and Gerhard Keller. Rigorous statistical procedures for data from dynamical systems. *J. Statist. Phys.*, 44(1-2):67–93, 1986.

[29] Luc Devroye. *Nonuniform random variate generation*. Springer-Verlag, New York, 1986.

[30] David A. Dickey and Wayne A. Fuller. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49(4):1057–1072, 1981.

[31] Holger Drees, Laurens de Haan, and Sidney Resnick. How to make a hill plot. *The Annals of Statistics*, 28(1):254–274, 2000.

[32] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[33] Robert F. Engle and Victor K. Ng. Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5):1749–1778, 1993.

[34] Michael J Evans and Jeffrey S Rosenthal. *Probability and statistics: The science of uncertainty*. Macmillan, 2004.

[35] Carmen Fernandez and Mark F. J. Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[36] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings, 2018.

[37] Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An introduction to heavy-tailed and subexponential distributions*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2011.

[38] Wayne A. Fuller. *Introduction to statistical time series*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1996. A Wiley-Interscience Publication.

[39] Dave Gehrke and Efraim Turban. Determinants of successful website design: relative importance and recommendations for effectiveness. *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, Track5:8 pp.–, 1999.

[40] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Elsevier/Academic Press, Amsterdam, eighth edition, 2015. Translated from the Russian, Translation edited and with a preface by Daniel Zwillinger and Victor Moll, Revised from the seventh edition [MR2360010].

[41] Peter Grassberger and Itamar Procaccia. Characterization of strange attractors. *Phys. Rev. Lett.*, 50:346–349, Jan 1983.

[42] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Phys. D*, 9(1-2):189–208, 1983.

[43] James D. Hamilton. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994.

[44] Peter R. Hansen and Asger Lunde. Realized variance and market microstructure noise. *J. Bus. Econom. Statist.*, 24(2):127–218, 2006. With comments and a rejoinder by the authors.

[45] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[46] Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174, 1975.

[47] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948.

[48] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[49] Jianhua Huang. *GARCH models: structure, statistical inference and financial applications* [book review, Wiley, Hoboken, NJ, 2010]. *J. Amer. Statist. Assoc.*, 107(498):847–848, 2012.

[50] Soosung Hwang and Pedro Valls Pereira. Small sample properties of garch estimates and persistence. *European Journal of Finance*, 12:473–494, 02 2006.

[51] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006.

[52] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. pages 180–191, 04 2004.

[53] Andrei N. Kolmogorov. *Selected works. II. Probability theory and mathematical statistics.* Springer Collected Works in Mathematics. Springer, Dordrecht, 2019. Translated from the Russian by G. Lindquist, Edited by Albert N. Shiryaev, Reprint of the 1992 edition [MR1153022].

[54] T. W. Körner. *Fourier analysis.* Cambridge University Press, Cambridge, 1988.

[55] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159 – 178, 1992.

[56] Douglas E Lake, Joshua S Richman, M Pamela Griffin, and J Randall Moorman. Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 283(3):R789–R797, 2002.

[57] Hubert W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

[58] Hubert W. Lilliefors. On the kolmogorov-smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, 64(325):387–389, 1969.

[59] Chao Liu, Ryen W. White, and Susan Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 379–386. Association for Computing Machinery, Inc., October 2010.

[60] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

[61] N. R. Lomb. Least-Squares Frequency Analysis of Unequally Spaced Data. *apss*, 39(2):447–462, Feb 1976.

[62] Helmut Lütkepohl and Fang Xu. The role of the log transformation in forecasting economic variables. *Empirical Economics*, 42(3):619–638, 2012.

[63] Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)*, 383(9921):999–1008, 03 2014.

[64] Spyros Makridakis, Steven Wheelwright, and Rob J Hyndman. *Forecasting: Methods and Applications, 3rd Ed.* John Wiley & Sons, United States of America, 1997.

[65] George Marsaglia, Wai Wan Tsang, and Jingbo Wang. Evaluating kolmogorov's distribution. *Journal of Statistical Software, Articles*, 8(18):1–4, 2003.

[66] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. The fundamentals of heavy-tails: Properties, emergence, and identification. pages 387–388, 06 2013.

[67] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7(4):308–313, 1965.

[68] Whitney K. Newey and Kenneth D. West. Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, 61(4):631–653, 10 1994.

[69] MEJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[70] Jukka Nyblom. Testing for the constancy of parameters over time. *Journal of the American Statistical Association*, 84(405):223–230, 1989.

[71] David Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

[72] Sidney Resnick and Cătălin Stărică. Smoothing the Hill estimator. *Adv. in Appl. Probab.*, 29(1):271–293, 1997.

[73] Sidney Resnick and Cătălin Stărică. Smoothing the hill estimator. *Advances in Applied Probability*, 29(1):271–293, 1997.

[74] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[75] Christian Yann Robert and Mathieu Rosenbaum. Volatility and covariation estimation when microstructure noise and trading times are endogenous. *Math. Finance*, 22(1):133–164, 2012.

[76] Mathieu Rosenbaum. A new microstructure noise index. *Quantitative Finance*, 11(6):883–899, 2011.

[77] François Roueff and Andrés Sánchez-Pérez. Prediction of weakly locally stationary processes by auto-regression. *ALEA Lat. Am. J. Probab. Math. Stat.*, 15(2):1215–1239, 2018.

[78] Peter J. Rousseeuw and Sabine Verboven. Robust estimation in very small samples. *Comput. Stat. Data Anal.*, 40(4):741?758, October 2002.

[79] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike information criterion statistics*, volume 1 of *Mathematics and its Applications (Japanese Series)*. D. Reidel Publishing Co., Dordrecht; SCIPRESS, Tokyo, 1986. With a preface by Tosio Kitagawa, Translated from the Japanese.

[80] J. D. Scargle. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *apj*, 263:835–853, Dec 1982.

[81] Robert H. Shumway and David S. Stoffer. *Time series analysis and its applications*. Springer Texts in Statistics. Springer-Verlag, New York, 2000.

[82] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[83] Stephen Smale. Book Review: Catastrophe theory: Selected papers,. *Bull. Amer. Math. Soc.*, 84(6):1360–1368, 1978.

[84] Yuedong Song, Pietro Liò, et al. A new approach for epileptic seizure detection: sample entropy based feature extraction and extreme learning machine. *Journal of Biomedical Science and Engineering*, 3(06):556, 2010.

[85] J. Michael Steele. *The Cauchy-Schwarz master class*. MAA Problem Books Series. Mathematical Association of America, Washington, DC; Cambridge University Press, Cambridge, 2004. An introduction to the art of mathematical inequalities.

[86] Stilian A. Stoev, George Michailidis, and Murad S. Taqqu. Estimating heavy-tail exponents through max self-similarity. *IEEE Trans. Inform. Theory*, 57(3):1615–1636, 2011.

[87] Panayiotis Theodossiou. Financial data and the skewed generalized t distribution. *Management Science*, 44(12):1650–1661, 1998.

[88] Ruey S. Tsay. *Analysis of financial time series*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, third edition, 2010.

[89] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[90] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[91] Michael Vogt. Nonparametric regression for locally stationary time series. *Ann. Statist.*, 40(5):2601–2633, 2012.

[92] Richard von Mises. *Probability, statistics and truth*. Dover Publications, Inc., New York, english edition, 1981.

[93] M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London, 1995.

[94] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[95] E. T. Whittaker and G. N. Watson. *A course of modern analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1996. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions, Reprint of the fourth (1927) edition.

[96] Dominik Wied and Rafael Weiß bach. Consistency of the kernel density estimator: a survey. *Statist. Papers*, 53(1):1–21, 2012.

RÉSUMÉ

La vitesse des échanges d'information dans le réseau Internet se mesure à l'aide de la latence: une durée mesurant le temps écoulé entre l'envoi du premier bit d'information d'une requête et la reception du premier bit d'information de la réponse. Dans cette thèse réalisée en collaboration avec la société Citrix, nous nous intéressons à l'étude et à la modélisation des données de latence dans un contexte d'optimisation de traffic Internet.
Citrix collecte des données via deux canaux différents, générant des mesures de latence soupçonnées de partager des propriétés communes. Dans un premier temps, nous nous intéressons à un problème d'ajustement distributionnel où les co-variables et les réponses sont des mesures de probabilité images l'une de l'autre par un transport déterministe, et les observables des échantillons indépendants tirés selon ces lois. Nous proposons un estimateur de ce transport et démontrons ses propriétés de convergence. On montre que notre estimateur peut être utilisé pour faire correspondre les distributions des mesures de latence générées par les deux canaux.
Dans un second temps nous proposons une stratégie de modélisation pour prédire le processus obtenu en calculant la médiane mobile des mesures de latence sur des partitions régulières de l'intervalle $[0, T]$ avec un maillage $\Delta > 0$. On montre que la moyenne conditionnelle de ce processus, qui joue un rôle majeur dans l'optimisation du traffic Internet, est correctement décrite par une décomposition en séries de Fourier et que sa variance conditionnelle s'organise en clusters qu'on modélise à l'aide d'un processus ARMA Seasonal-GARCH, c'est à dire un processus ARMA-GARCH avec ajout de termes saisonniers déterministes. Les performances prédictives de ce modèle sont comparées aux modèles de référence utilisés dans l'industrie. Une nouvelle mesure de la quantité d'information résiduelle non captée par le modèle basée sur un certain critère entropique est introduite.
Nous abordons ensuite le problème de la détection de panne dans le réseau Internet. Nous proposons un algorithme de détection de changement dans la distribution d'un stream de données de latence basé sur la comparaison de deux fenêtres glissantes à l'aide d'une certain distance de Wasserstein pondérée.
Enfin, nous décrivons comment sélectionner les données d'entraînement des algorithmes prédictifs de manière à réduire leur taille pour limiter les coûts de calculs sans impacter la précision.

MOTS CLÉS

Ajustement distributionnel, Séries temporelles, Processus ARMA-GARCH, Prévision, Détection de changement, Distance de Wasserstein, Latence Internet

ABSTRACT

Information exchange speed on the Internet is measured with latency: the duration of the elapsed time between the sending of the first bit of a request and the reception of the first bit of the response. In this thesis carried out in collaboration with the company Citrix, we are interested in the analysis and modeling of latency data in a context of Internet traffic optimization.
Citrix collects data through two different channels generating latency measurements suspected to share common properties. First, we study a probability distribution matching problem where the outputs are the transported probability distributions of the inputs under an unknown deterministic transport, and where the observables are independent samples drawn according to these probability distributions. We study an estimator of this transport and prove its convergence properties. We show that our estimator can be used to match the distributions of latency measurements from the two channels.
Then, we propose a modeling strategy to predict the process obtained by calculating the moving median of latency measurements on regular partitions of the interval $[0, T]$ with mesh $\Delta > 0$. We show that the conditional mean of this process, which plays a major role in Internet traffic optimization, is correctly described by a decomposition into Fourier series and that its conditional variance forms clusters which are modeled using an ARMA Seasonal-GARCH process, i.e. an ARMA-GARCH process with additional deterministic seasonal terms. The predictive performance of this model is compared to benchmark models used in the industry. A new measure of the amount of residual information not captured by the model based on a certain entropy criterion is introduced.
We then address the problem of outage detection in the Internet. We propose a change detection algorithm in the distribution of a latency data stream based on the comparison of two sliding windows using a certain weighted Wasserstein distance.
Finally, we describe how to minimize the size of the training data sets used by the predictive algorithms to limit the calculation costs without impacting accuracy.

KEYWORDS

Probability distribution matching, Time series, ARMA-GARCH process, Forecasting, Change detection, Wasserstein distance, Internet latency